Essays on Healthcare Payments and Operations

Shima Nassiri

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2017

Reading Committee:

Hamed Mamani, Chair

Maria Shunko

Yongpin Zhou

Program Authorized to Offer Degree:
Business Administration

University of Washington

Abstract

Essays on Healthcare Payments and Operations

Shima Nassiri

Chair of the Supervisory Committee:
Professor Hamed Mamani
Information Systems and Operations Management

A common objective of my dissertation is to use analytical and empirical models to understand the impact of current healthcare policies on outcomes for stakeholders, cost containment, quality of care, and efficiency of the providers. Consequently, this dissertation focuses on two areas:

1. payment models in healthcare (Chapters 1 and 2), and
2. healthcare behavioral operations (Chapter 3).

Healthcare spending accounts for 17.5% of the GDP in the United States and has been rising year after year, which makes this a significant concern for society. There have been many proposals specifically related to the reimbursement models from insurers to providers to curb the healthcare cost. However, many of these proposals are not studied or thoroughly evaluated before they are implemented. I believe this has created a special opportunity to bridge the gap between practice and academia by developing new models to study the decisions and models that have been proposed. To this extent, the first two chapters of my thesis focus on healthcare payments, where I use mathematical models to investigate the current policies in healthcare payments, propose new payments that can better align

the incentives of the stakeholders to the system optimum outcomes, compare and contrast different payment systems, and support the derivations with empirical evidence.

The current reimbursement system for hospitals is *fee-for-service* (FFS), under which the hospital is reimbursed for every procedure, test, etc. This payment model motivates the hospitals to treat patients more than is needed and to perform unnecessary tests that contribute to excessive healthcare expenditures. There have been many proposals for alternative payment models especially from the Center for Medicare and Medicaid services (CMS). One way of controlling healthcare expenditures is by modifying the reimbursement schemes to share some of the insurers' risk with providers. This possibility, results in the first chapter of this dissertation which is now also a published paper [4]. The second chapter is related to impacting healthcare spending through cost-sharing mechanisms with patients, which affects the providers' market shares and, subsequently their pricing strategies. The questions that I am trying to address are: what is the impact of new payment schemes and are they achieving their premises or are there some unintended consequences?

In the third chapter, I mainly focus on the emerging problems that can affect the behavior of providers in healthcare operations, such as changing the facility layout of the exam rooms so that providers will share some resources. This change can potentially impact provider's efficiency, scheduling, and assignments. I use an empirical approach to address the effect of the layout change benefiting from over a year of patient-level data from an outpatient Clinic. The clinic recently implemented a flexible suite layout for their outpatient department in which a physician can share the bed and some examination equipments with another provider. This is in contrast to the traditional exam rooms where each provider is assigned to a room with all the equipments. We investigate the impact of new intervention, multitasking, competition, and load on provider behavior and efficiency.

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGMENTS

I would like to express my sincere gratitude to my advisor Prof. Hamed Mamani, for the continuous support of my PhD studies over the past five years. I have benefited immensely from his knowledge, patience, and intuition. He has been a source of inspiration for me through some of the most difficult times of my PhD life. His guidance has helped me in my research, teaching, presentation skills, choosing my career path, and focusing on my longterm career goals. He has made my PhD experience very enjoyable and I could not imagine having a better advisor and mentor.

I would also like to thank Elodie Adida, my coauthor on the first two chapters of my dissertation. She is a great researcher and person to work with. I have learned a lot from her knowledge and attention to details. Her help extended beyond a coauthor and she has kindly advised and helped me through out my career search. Although, I only met her a few times in conferences, she has always been available through Skype and phone for advice and feedback. Her responsiveness, patience, and knowledge were greatly valuable to improving my research skills.

In the past year I had the pleasure of working with Masha Shunko who has also accepted to be part of the reading committee for my dissertation. She has been very supportive and she has guided me on a new stream of research that I want to pursue in future. She is a great person to work with and I have considerably benefited from her insight.

I am grateful to Michael Wagner who has guided and advised me during the first two years of my PhD studies. He has helped me build the foundation of conducting academic research. My sincere thanks goes to Yong-Pin Zhou who has advised and helped me throughout the PhD program as PhD coordinator and part of my dissertation committee. His feedback has

been very valuable to me and truly helped me on the academic job market.

I would also like to thank my fellow Foster PhD students for the sleepless nights of working in Mackenzie Hall and for all the fun we have had in the past five years. I am so grateful to Elnaz Jalilipour-Alishah, Behnaz Ghahestani-Bojd, Amir Fazli, Matthew Denes, and Tony Chen for their support and friendship. My special thanks to my old friend Maryam Agahi who has been supporting me along the way.

Last but not least, I am grateful to my family: my encouraging and supportive parents and brother, and my always energetic, enthusiastic, inspiring, and loving sister who has taught me to dream big and has always been my cheerleader. Finally, I would like to thank my husband, Shahab, who has greatly helped my personal growth in the past three years. I thank him for his unconditional love and support, and for being the most understanding partner I could wish for.

DEDICATION

To my beloved parents,

Mansoor Nassiri and Soheila Esfandiari

Chapter 1

BUNDLED PAYMENT VS. FEE-FOR-SERVICE: IMPACT OF PAYMENT SCHEME ON PERFORMANCE

1.1 Introduction

The much debated Affordable Care Act aims to make drastic changes to many aspects of the healthcare system in the US. In particular, a key part of the legislation is designed to control for rising healthcare costs by transforming the way healthcare providers are paid. Under the current fee-for-service (FFS) payment system, medical providers are compensated based on the volume of services performed, such as the number of tests and treatment procedures provided to the patient. Many healthcare experts criticize such a payment system on the basis that it rewards providers for spending more without necessarily increasing the quality of care, instead of focusing on delivering value and improving health outcomes [45].

Providers often have a choice of treatment options to follow for a patient and the option selected is not necessarily the most beneficial from a system cost-benefit standpoint. [92] notes that "Americans (...) are typically prescribed more expensive procedures and tests than people in other countries". Widespread abuses in the current system have received some media attention in the recent past. For example, the same article remarks that colonoscopies "are often prescribed and performed more frequently than medical guidelines recommend" and that "while several cheaper and less invasive tests to screen for colon cancer are recommended as equally effective by the federal government's expert panel on preventive care – and are commonly used in other countries – colonoscopy has become the go-to procedure in the United States." [95] points to a dermatological procedure called Mohs surgery, noting that "while it offers clear advantages in certain cases, it is more expensive than simply cutting or freezing off a lesion" and tends to be overused with an increase of 400% in the last

decade. [2] bring attention to a drug known by the brand name Lucentis that is injected as often as once a month as treatment for a kind of age-related macular degeneration in elderly patients, contributing to a $3.3 billion spending from Medicare to about 3,300 ophthalmologists, when "a cancer drug that is used as an alternative can cost much less". The FFS system gives providers financial incentives to treat more, not better, thus partly contributing to the nation's rising healthcare costs. With Medicare spending approaching $600 billion a year, even a small fraction improvement in average spending may have a significant impact on the bottom line.

To help address this issue, better align incentives, and rein in costs, the Centers for Medicare and Medicaid Services (CMS) has been experimenting a new payment initiative, called Bundled Payment for Care Improvement (BPCI), since 2013. Under the bundled payment (BP) system, the provider is compensated with one lump sum for a whole episode of care, regardless of the exact tests and procedures implemented and regardless of eventual complications[1]. Currently CMS proposes several models depending on whether the episode of care includes hospital stay and/or post-acute care of various time windows. In the FFS system, when a beneficiary needs to undergo a given treatment, the insurer covers the cost of each test, x-ray, specialist consultation, skilled nursing visit, days of hospital stay, etc., including those incurred in case of potential complications and even readmission. In contrast, under BP, the insurer only pays the pre-specified bundled payment value upfront to cover all possible services rendered to the patient within a specified time window around the treatment, including eventual complications. If actual costs are lower than the bundled payment, the provider makes a profit; if total treatment costs (including possible complications and readmission) exceed the lump-sum amount, the provider incurs a loss.

Proponents of BP claim that such a payment system promotes high quality of care while keeping costs under control [18]. Since complications and readmissions after discharge do not lead to further reimbursements from the payer, providing high quality of care from the start

[1]Note that we use the term *provider* for any group of providers serving a patient within a certain episode of care. The lump sum is paid jointly to all the providers serving the patient and is divided amongst them.

of treatment increases the chances of making a profit for the healthcare provider [79]. In addition, BP removes incentives to implement unnecessary procedures and hence is expected to lower costs [79]. Opponents of BP, on the other hand, argue that implementing this scheme could jeopardize quality of care by means of increasing efficiencies and keeping costs low [45]. Furthermore, BP could lead to patient selection as healthcare providers would have financial incentives to turn down those patients with high healthcare needs or potentially high treatment costs [18]. Additionally, some are concerned that bundled payments impose significant financial risks on the provider as they incur a loss whenever treatment costs exceed the set reimbursement amount [79, 77]. A high level of risk for the provider may not only lead to much resistance toward adopting the new payment scheme but also in the long term increases the risk of bankruptcy for healthcare providers and thus may lead to diminished access to care. It may also result in provider decisions that are not optimal from a system's perspective in terms of patient selection and treatment level.

The risk borne by the provider under BP stems from three sources. One is the chance that the patient develops complications following the first-stage (initial) treatment, as the patient will then incur further costs (in a second stage) which may lead to financial losses for the provider. This risk may be lowered by implementing the right treatment level on the patient (e.g., schedule nurse visits after the procedure, follow-up with the patient to ensure they are taking their medications, etc). Another source of risk, for a given patient risk profile, is the variability of the actual second-stage cost. A high variability increases the risk exposure of the provider (potential loss). The third source of risk is the patient type mix within the population. Specifically, a provider serving a potentially costlier population (many patients with a high expected second-stage cost) is likely to incur further losses, and a provider with more variability in the patient types will see more variability in her total utility, again increasing her exposure. The latter effect is related to the population size: in a small patient population, an outlier patient with a very high second-stage cost is less likely to be compensated by low-cost patients.

While the BPCI initiative has just started its pilot program, there have been many

initiatives in the past aiming at testing payment systems that distance themselves from a fee-for-service approach in favor of a pay-for-performance, capitation, bundling, or diagnostic-related-group (DRG) approach [65]. Among these, the largest-scale program was arguably the prospective payment system (PPS) enacted in 1983. This program paid the provider an amount depending on the patient's classification within a certain DRG [74]. The key distinctions of BPCI with respect to PPS are as follows: (i) PPS applies only to inpatient hospital stays, while BPCI could apply also to outpatient procedures, (ii) PPS includes only hospital services, as opposed to physician services, while BPCI bundles all services received by the patient from a variety of providers, (iii) PPS covers a relatively short time period – a single hospital stay – when BPCI includes services provided during the hospital stay *and after*, for a pre-determined duration [63]; as a result, when complications arise and a patient must be readmitted, under BPCI the providers do not receive any new reimbursement, (iv) under PPS, a given provider's reimbursement is set based on the cost at comparable facilities [83], as opposed to BPCI, where the reimbursement is based on the historical cost at this specific provider [77]. PPS and other similar programs have been studied in the literature, often from an empirical standpoint. There are few attempts in the literature at modeling the effect of payment systems within an analytical framework to compare their performance.

Our goal is to compare payment systems vis-a-vis a variety of performance measures, and test whether the claims of proponents and opponents of BP are justified. More specifically, we propose to answer the following research questions: (1) Do the payment schemes under consideration give incentives for patient selection? (2) What is the treatment level selected by the provider, and how does it compare with what would be system-optimal? (3) How does the financial risk borne by the provider compare across different payment schemes? (4) How do the utility of the provider and the total system payoff compare across the different payment schemes? (5) Is there another payment system that could alleviate the shortcomings of schemes currently under consideration? (6) What role does the provider's risk aversion play?

We consider a population consisting of a finite number of beneficiaries (patients) seeking

treatment for a given episode of care, a provider[2], and an insurer. Under the BP system, the provider receives a fixed payment for the episode of care. The provider decides whether to accept the beneficiary and, for beneficiaries receiving care, selects in a first stage the treatment level that maximizes her expected risk-averse utility. In a second stage, after the treatment is provided, the beneficiary may face complications and require further treatment, the likelihood of which depends on the first stage treatment level. In case of complications, the provider incurs additional treatment costs that will not be further reimbursed by the insurer. We use a similar framework to model the FFS system with the exception that the payment to the healthcare provider is proportional to the cost of treatment offered to the beneficiary. Furthermore, in case of complications, the provider receives an additional payment that is proportional to the complication cost. We assume that the cost of treating a beneficiary in the second stage is a random variable whose distribution depends on the beneficiary's "type". The beneficiary type is characterized by the beneficiary's expected complication cost and is observed by the provider. Therefore, the provider chooses to only accept those beneficiaries that are expected to generate a non-negative utility; this allows us to model the issue of patient selection.

In this chapter we introduce a new model of healthcare payment systems that incorporates heterogeneous patients, considers provider risk aversion, allows for patient selection and includes treatment level flexibility. We derive the optimal treatment intensities, patient selection levels, and expected utilities under the BP and FFS payments, as well as at the system optimum (that is, a Pareto-optimal outcome). Qualitatively, our findings are consistent with the observations made in the public health policy literature. We find that FFS can never induce the system-optimal patient selection level. We also show that under FFS, the provider takes advantage of variable payments and generally selects the highest possible treatment level. In contrast, we show that BP may lead to treatment levels that are either

[2]While we frame the discussion around a beneficiary obtaining treatment from one provider, our general model and findings apply to situations where multiple medical providers make treatment decisions, since under the bundled payment system all providers must coordinate care and split the lump sum payment among themselves.

lower or higher than the system optimum depending on the provider's risk aversion and other factors. While we find that BP may yield a higher utility for the provider and a higher system payoff than FFS, in general the performance of the BP mechanism is extremely sensitive to the selection of the bundled payment value as well as the provider's risk aversion. Furthermore, we show that the BP system can induce suboptimal patient selection levels and expose providers to high levels of risk.

Our results indicate that, in general, no BP or FFS payment system can achieve the system optimum. However, minor adjustments can alleviate the shortcomings of both FFS and BP in many scenarios. Specifically, inspired by various risk-sharing mechanisms in the operations and supply chain management literature, we design two practical payment schemes: (1) a hybrid payment system that is a combination of FFS and BP mechanisms, and (2) a stop-loss protection mechanism that is a variation of the BP system. The hybrid payment mechanism improves various performance measures and can achieve the system optimum when the provider is not very risk averse; the stop-loss protection mechanism achieves the same when the provider is highly risk averse. They each accomplish this by offering the provider incentives to exert optimal treatment efforts and implement patient selection according to what is Pareto-optimal for the system. We also show that for a limited range of parameters when the provider is moderately risk averse *and* the treatment success probability is high enough, none of these payment schemes can be coordinating. Finally, we investigate the provider's overall risk burden by studying the relationship between population size and risk exposure.

1.2 Literature Review

Since the currently tested bundled payment system is recent, there is little literature directly addressing it. However, current and past payment systems have been studied in the literature from a variety of perspectives. Our work is related to two main research streams that have been built with contributions from the operations management, health economics and health policy literatures.

First, our work is related to the quantitative assessment of payment system reforms and their effects, which is typically done using empirical methods. As mentioned in Section 1.1, the prospective payment system (PPS), based on patients' diagnosis related group (DRG), presents some similarity with the bundled payment (BP) system by using episode-based payments. Given that PPS was established about 30 years ago, there are a number of empirical studies that have been conducted to evaluate its effectiveness. [75] performs an empirical analysis of PPS vs. FFS reimbursement incentives. They conclude that the PPS payment scheme contributes to patient selection while the FFS system contributes to an increase in the intensity of care for patients with certain conditions. This empirical evidence matches our analytical results for BP and FFS.

While PPS applies only to *hospital inpatient* care, other care settings are subject to the same issues caused by the FFS payment system. [62] focus on *home health agencies*, which have been subject in the past 20 years to payment reforms aiming at shifting reimbursement away from FFS towards episode-based payments. They consider a payment system including a fixed and a marginal reimbursement, somewhat similar to our hybrid system. They study the effect of lowering either or both payment components on the treatment level and on patient selection, as well as hospital readmission and mortality. Using data to develop empirical strategies, they find that lowering only the marginal payment decreases admissions and increases only slightly the use of resources, while lowering both the fixed and variable payment decreases both. In both cases, they find little evidence of patient selection and limited effects on readmissions and mortality. They suggest that reforms such as bundled payments are likely to impact provider behavior, and that the level of payment could influence whether the reduction in the use of resources would benefit the provider or the insurer. This paper contrasts with ours in three main aspects: (i) we use optimization techniques rather than empirical strategies to obtain the provider decisions (ii) we do not focus on a specific policy shift, but obtain the provider decisions for a given reimbursement structure (iii) we consider alternative payment systems, such as the hybrid and stop-loss protection payments, in the context of coordinating decisions to a system optimum. Similarly, [104] consider

inpatient rehabilitation facilities and the effect of initiatives taken by Medicare aiming at reducing the marginal reimbursement and increasing the fixed reimbursement. The authors investigate the extent to which providers respond to these changes in the payment system by adjusting the number of admitted patients, types of patients admitted and intensity of care. Using an empirical approach they show that the treatment intensity decreases as the payment system moves towards a prospective payment system, despite an increase in payments to the facilities. Along the same lines, [71] evaluate reforms in the payment system for dialysis providers for *Medicare's End-Stage Renal Disease program* to shift toward a "pay-for-compliance" system with limited risk adjustments to encourage providers to conform to standardized guidelines of best practice. The authors use an empirical approach to develop an evidence-based optimal procedure for incorporating full risk adjustment and pay-for-compliance into the payment system. They show that the payment scheme proposed by Medicare would not provide the desired incentives, but the design they introduce would improve outcomes with no additional expenditures. [63] provide a meta-analysis of research on bundled payments. They review 58 studies on the topic (excluding research on PPS) as well as 4 review articles on PPS, with the goal of identifying the effect on quality of care and health care spending. Their review includes a total of 20 different bundled payment interventions that aggregate costs over time for a single provider, across providers, and/or involve warranties regarding the occurrence of complications, in the US and abroad, and in a variety of settings (e.g. nursing homes, rehabilitation facilities, etc.). They find weak but consistent evidence that bundled payment programs do succeed in containing costs without significantly affecting quality of care, spending and utilization rate. Most of the studies considered a single provider and were descriptive and observational.

There have also been initiatives involving "pay-for-performance" payment schemes to incentivize providers to administer better quality of care. [96] do a meta-analysis of reports of quality incentive programs from 1998 to 2003. They show that, although these mechanisms, by rewarding good performance rather than *improvements* in performance, increase the quality of care for some providers, low-quality providers often are not motivated to make

the necessary investments to improve their performance, which limits the impact of these programs.

Some empirical studies were developed in anticipation of BPCI with the goal of influencing the design of the pilot program. Using existing Medicare data, [103] investigate which episodes of care are more suitable to be included in the pilot program and what the episode length should be, based on the potential cost savings and the financial risk on the providers. [37] use recent beneficiary level claims data to make recommendations on how the bundled payment system should be designed, including the conditions to include, episode length, pricing of the bundle, risk adjustments and other design considerations.

Our work is also related to the stream of research that uses an analytical approach or economic reasoning for understanding a variety of issues related to payment systems such as patient selection, moral hazard, efficiency incentives, and contracts. Economists have been interested in designing mechanisms that induce better outcomes than the FFS system. In a seminal paper, [99] proposes a "yardstick competition" mechanism in which the payment that a firm receives depends on the average cost at identical firms, as they reflect the attainable cost level. This provides each firm with incentives to reduce its cost below that of others. The author shows that this mechanism yields the system optimum for identical firms. This mechanism is in line with the way reimbursements in PPS are set. [80] examines trade-offs related to risk and selection in a fee-for-service and a prospective payment system from an economic standpoint. He shows that when some of the yardstick competition model assumptions are relaxed, PPS alone no longer yields optimal outcomes. Instead he proposes a mixed payment scheme incorporating features from both systems, which is consistent with the hybrid system that we analyze.

The operations management literature has also contributed to healthcare payment systems research. In particular, some papers have taken a modeling and analytical approach aligned with our work. Focusing on hospice care, [13] introduce a dynamic model to understand how the payment system in place for these facilities may be causing an increasing number of hospice bankruptcies, mainly because of an annual cap. They also analyze how

Medicare's reimbursement policy may give incentives for sometimes selecting short-lived patients and may influence treatment choices. They propose an alternative that alleviates these issues.

One main issue that has been studied is that of moral hazard within a principal-agent framework [84]: the provider, possibly enjoying hidden information, makes treatment decisions that the insurer does not necessarily observe (hidden action), but that directly affect the insurer's payoff[3]. When treatment is not observable, payment terms must be based on patient outcomes. Motivated by Medicare's End-Stage Renal Disease program and the fact that Medicare was considering capitated payments, [48] find an optimal payment system that induces system-optimal treatment choices in a dynamic setting for a risk-averse provider. The optimal payment is outcome adjusted and consists of two components: a prospective payment per patient and a retrospective payment adjustment based on adverse short-term patient outcomes, which is reminiscent of the hybrid system analyzed in our paper (but in our model the treatment intensity is observable and dictates reimbursement in FFS, and the insurer's payoff does not depend on the provider's treatment decision in BP, hence there is no moral hazard). Unlike our model they do not consider the issue of patient selection. In the presence of moral hazard and asymmetric information, coordinating contracts can help align incentives and obtain the system optimum. [121] consider a preventive procedure such as a screening test administered based on a threshold policy selected by the provider. They find that when the number of patients seeking the intervention is verifiable, there exists a coordinating contract, but otherwise the FFS system does not coordinate the channel as the provider selects a too low level of effort. In the context of an online appointment scheduling system which enables the provider to allocate service capacity under access-to-care requirements, [66] study optimal contracts between a purchaser and a provider, where performance is achieved when a waiting-time target is met.

---

[3]In much of the literature, it is assumed that the insurer is able to verify the diagnosis and health outcome of the patient. Later [85] argue that this claim is not always valid, as they find empirically that operational conditions such as the system workload influence physicians' diligence of paperwork execution.

One of the coordinating mechanisms studied in our model (the hybrid payment scheme) is related to the notion of two-part tariffs from the economics literature (e.g., [23, chapt. 9, 10], [116], and [57]). However, unlike the classic two-part tariff mechanism that coordinates one decision, our proposed mechanism can coordinate two decisions: treatment level and patient selection level. Furthermore, often, especially in economics, a menu of two-part tariffs is used to segment the market according to different customer types whereas in our model, the "customer" being offered the contract is the provider (of a single type), and the purpose is to coordinate the provider's *decisions*.

Finally, our work is also related to a stream of literature, outside the healthcare area, that endogenizes future implications of decisions, like avoidable medical complications occurring because of inadequate treatment intensity in our model. In project management, the concept of Design-Build-Operate-Maintain captures how the initial "design" and "build" phases of a project influence the "operate" and "maintain" phases, and how future costs can be reduced by integrating these phases [30, 16]. Similarly, the practice of "servicizing" a product by selling the functionality of the product rather than the product itself and being responsible for maintenance and repairs, gives incentives to improve the quality of the product and extend product life cycles [5, 117, 108]. This also recalls the idea of providers being responsible for treating medical complications at their own cost. In the aerospace and defense industry, performance-based-contracts have emerged as an alternative to time and material contracts, under which the supplier is compensated for the amount of resources consumed [53].

1.3 Model

1.3.1 Modeling framework

In this section we outline the model framework and its assumptions. Table 1.1 summarizes the notations used for the parameters and variables used in our model. We consider a population consisting of a finite number (N) of beneficiaries (patients), a provider and an insurer. Beneficiaries wish to undergo treatment for a certain non-emergency medical

condition (episode of care). The provider may accept or reject beneficiaries. A beneficiary receives payoff V when she is given the treatment. Without loss of generality, we assume that beneficiaries have a reservation utility equal to zero when they are denied treatment. If the beneficiary is accepted, the provider selects the treatment level $t \in [\underline{t}, \overline{t}]$ for this beneficiary so as to maximize her expected utility. We model the provider as risk averse with a constant absolute risk aversion (CARA) utility function. We denote the provider's risk-aversion coefficient by θ. The first-stage treatment cost incurred by the provider, $c_1(t)$, increases with the intensity of treatment. Treatment results in "success" or "failure", where failure means that the beneficiary is subject to complications requiring further treatment (e.g., readmission). The probability of success is denoted by $q(t)$. If the treatment fails, the beneficiary suffers disutility T^B, and the provider receives penalty T^P (e.g. representing the effect on her reputation). In case of failure in the first stage, we assume that the provider does not make any more treatment level decision in the second stage. We denote by c_2 the treatment cost in case of complications. Note that c_2 incorporates the costs of treatment until the complication is resolved. In other words, we assume that all beneficiaries will be eventually treated and discharged; death or life-long treatments are neglected for the episodes of care that we consider (in particular, we do not consider chronic diseases). Figure 1.1 illustrates the sequence of events in our model.

The second-stage cost c_2 is a random variable with support $[\underline{c}, \overline{c}]$, which captures patient heterogeneity. Beneficiaries are characterized by their "type", μ, that is the *expected* second-stage treatment cost. We assume a continuum of beneficiary types $[\underline{\mu}, \overline{\mu}]$ with probability distribution function $f(\cdot)$; the types of distinct beneficiaries are independent of each other. The provider is able to identify the beneficiary type (e.g. using family history, health assessment, prior test results, etc.) before deciding whether or not to accept the beneficiary. For a beneficiary of a given type μ, the second-stage treatment cost follows a conditional probability distribution function $g_\mu(\cdot)$ with conditional mean μ and conditional variance s_μ^2. In particular, the beneficiaries' second-stage costs are not identically distributed.

We make the following assumption on the conditional probability distribution function

t	treatment level for a beneficiary, to be selected by the provider within $[\underline{t}, \overline{t}]$
$q(t)$	probability of "success" of the treatment
$c_1(t)$	first-stage cost of treatment incurred by the provider
c_2	second-stage random cost of treatment in case of first-stage treatment failure, within $[\underline{c}, \overline{c}]$
μ	beneficiary type, defined as the average second-stage cost of treatment, within $[\underline{\mu}, \overline{\mu}]$
$g_\mu(\cdot)$	conditional probability distribution function of c_2 given μ; has mean μ and variance s_μ^2
$f(\cdot)$	probability distribution function of μ; has support $[\underline{\mu}, \overline{\mu}]$
κ	payment factor under FFS: insurer pays provider κ times the treatment cost in both stages
B	lump-sum payment from the insurer to the provider for treating the beneficiary under BP
V	payoff of beneficiary for receiving treatment
T^B	disutility of the beneficiary due to unsuccessful treatment ($T^B > 0$)
T^P	penalty of the provider due to unsuccessful treatment ($T^P > 0$)
θ	provider's risk-aversion parameter
$U^P(w)$	utility of the provider when receiving payoff w
$U^B(w)$	utility of the beneficiary when receiving payoff w
$\pi^j(t)$	total expected utility of agent j for a beneficiary of a given type under treatment level t, where $j = P, I, B$ for the provider, insurer and beneficiary respectively
N	size of beneficiary population seeking treatment
w^j	payoff of agent j, where $j = P, I, B, S$ for the provider, insurer, beneficiary and system respectively
γ	discount rate used to obtain the bundle payment value from the historical spending

Table 1.1: Notations

$g_\mu(.)$.

Assumption 1.1 *The family of conditional probability distribution functions $g_\mu(.)$ has the monotone likelihood ratio property.*

This assumption is not very restrictive and many commonly used distributions have this property (e.g., normal, uniform, exponential, gamma).

Definition 1.1 *Suppose that the distribution of X is in a parametric family of density functions $\{g_\mu(x)\}_{\mu\in[\underline{\mu},\overline{\mu}]}$ indexed by a parameter μ taking values in an interval $[\underline{\mu}, \overline{\mu}]$. The family*

of distribution is said to have a monotone likelihood ratio if for any $\mu_1 \leq \mu_2 \in [\underline{\mu}, \overline{\mu}]$, *the ratio* $g_{\mu_2}(x)/g_{\mu_1}(x)$ *is a non-decreasing function of* x.

Lemma 1.1 *Suppose that* $\{g_\mu(x)\}_{\mu\in[\underline{\mu},\overline{\mu}]}$ *has a monotone likelihood ratio. If* ψ *is a non-decreasing function, then* $\phi(\mu) = E[\psi(X)]$ *is a non-decreasing function of* μ.

Proof Let $\mu_1 \leq \mu_2 \in [\underline{\mu}, \overline{\mu}]$, $A = \{x : g_{\mu_1}(x) > g_{\mu_2}(x)\}$, $B = \{x : g_{\mu_1}(x) < g_{\mu_2}(x)\}$, $a = \sup_{x\in A} \psi(x)$, $b = \inf_{x\in B} \psi(x)$. By the monotone likelihood ratio property, $a \leq b$. Therefore,

$$\begin{aligned}
\phi(\mu_2) - \phi(\mu_1) &= \int \psi(x)(g_{\mu_2}(x) - g_{\mu_1}(x))d\,x \\
&\geq a\int_A (g_{\mu_2}(x) - g_{\mu_1}(x))dx + b\int_B (g_{\mu_2}(x) - g_{\mu_1}(x))d\,x \\
&= -a\int_B (g_{\mu_2}(x) - g_{\mu_1}(x))dx + b\int_B (g_{\mu_2}(x) - g_{\mu_1}(x))d\,x \\
&= (b-a)\int_B (g_{\mu_2}(x) - g_{\mu_1}(x))d\,x \\
&\geq 0,
\end{aligned}$$

where the second equality follows from

$$\int_A (g_{\mu_2}(x) - g_{\mu_1}(x))d\,x + \int_B (g_{\mu_2}(x) - g_{\mu_1}(x))d\,x = \int g_{\mu_2}(x)d\,x - \int g_{\mu_1}(x)d\,x = 1 - 1 = 0.$$

□

Note that if ψ is a non-increasing function of Y, $E[-\psi(Y)] = -E[\psi(Y)]$ is a non-decreasing function of μ, therefore $E[\psi(Y)]$ is a non-increasing function of μ.

Simple calculations show that the following parametric families of continuous density functions have the monotone likelihood property:

- Exponential distribution $\lambda e^{-\lambda x}$ (indexed by $\mu = 1/\lambda$);
- Normal distribution $(1/(\sigma\sqrt{2\pi}))e^{-(x-\mu)^2/(2\sigma^2)}$ indexed by μ (for fixed σ);
- Gamma distribution $x^{k-1}e^{-x/\theta}/(\theta^k\Gamma(k))$ indexed by θ (for fixed k).
- Uniform distribution $[a, b]$ indexed by a (for fixed b).

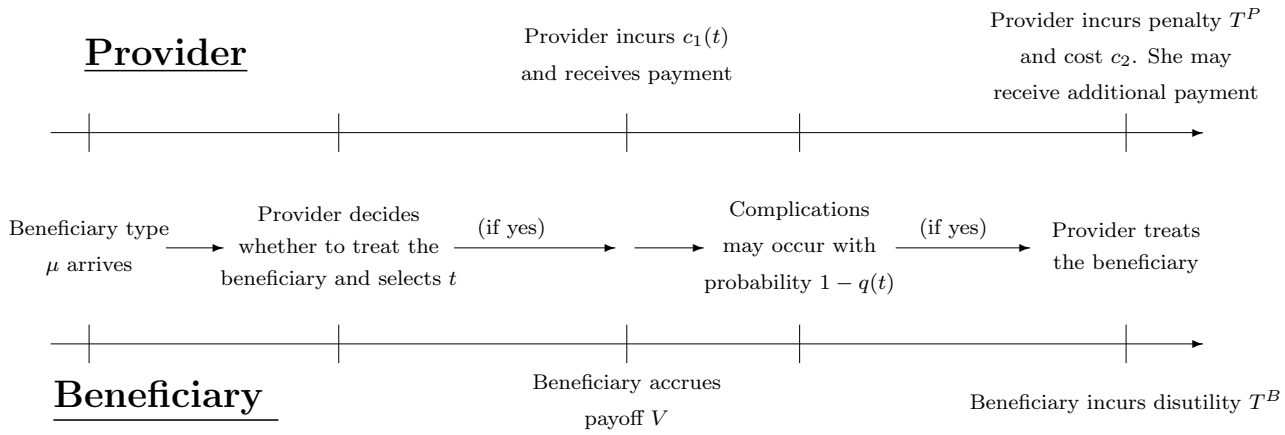


Figure 1.1: Sequence of events

1.3.2 Discussion

In this section we discuss our modeling framework and assumptions.

Treatment level

The model described above is a stylized way of formalizing the very complex process of patient admission and treatment selection considering financial and non-financial aspects of the decision making. The "treatment level," t, that we introduce is a measure of the intensity (not the quality) of the treatment implemented. Our premise is that providers face a vast array of treatment routes with different costs and different advantages and must decide how to treat the patient considering both costs incurred and potential benefits from the treatment options. Variable t hence illustrates the number of blood tests, x-rays/imaging procedures, exams, or specialist consultations that are selected. While this is certainly a crude way of modeling intricate treatment decisions, it enables us to capture the essential incentives and trade-offs of different payment systems.

Beneficiary

The payoff V experienced by beneficiaries for receiving treatment is assumed to be homogeneous over the population. This implies that, while some beneficiaries are more prone than others to require further costly treatment if developing complications, all stand to benefit the same amount from undergoing the procedure. For example, consider a knee replacement surgery episode. Patients in need of this procedure would see their quality of life improve by a similar amount upon completion (reduction of pain, improved mobility); however for various reasons (age, general health status, strength of support at home) not all incur the same treatment cost if the initial treatment fails.

The beneficiary payoff is zero when denied treatment, V when given treatment with a successful outcome, and $V - T^B$ when given treatment with a failed outcome. We assume that the beneficiary has no financial responsibility for the procedure, although including a fixed co-payment does not impact any of our findings. In our model, the beneficiary does not make any decision. As a result, our model does not rely on any risk-attitude assumption for the beneficiary. We denote the beneficiary's utility function from receiving payoff w as $U^B(w)$. The beneficiary's expected utility from receiving treatment with level t can then be written as $q(t)U^B(V) + (1 - q(t))U^B(V - T^B)$.

Insurer

We model the insurer as risk neutral. Hence its utility is given by the financial cost of reimbursing the provider for every beneficiary treated (which depends on the payment system). We assume risk neutrality for the insurer because the population size of beneficiaries insured is typically large and thus the insurer benefits from risk-pooling effects that make it immune to large variations in costs.

Provider

Unlike the insurer, the size of the beneficiary population served by a given provider is often not very large, and therefore the provider's costs may be subject to significant volatility. An outlier beneficiary with a very high cost may, in the bundled payment system, incur a large loss for the provider which may not be compensated by less costly patients because of the relatively small beneficiary population size, and this may cause a significant financial strain for the provider. As a result, we model the provider as risk averse. We consider that the provider's utility exhibits a constant absolute risk aversion (CARA) property [86]. That is, the provider's utility from a payoff w is given by:

$$U^P(w) = \frac{1}{\theta}\left(1 - e^{-\theta w}\right).$$

In this model, $\theta > 0$ is the risk-aversion coefficient. A payoff of zero provides no utility. The provider makes decisions so as to maximize her expected utility with respect to the treatment outcome of a beneficiary (success or failure)[4]. The case where the provider is risk neutral may be obtained by considering the limiting case of θ approaching zero as the utility function then tends to w.

We assume that the provider may reject patients after assessing the patient's expected complication cost, so the provider would only accept those patients expected to yield a positive utility. While emergency patients may not legally be denied treatment, for non-emergencies (which are the focus of this paper) physicians are free to choose which patients to serve. Indeed, [76] confirms that "Principle VI of the American Medical Association's (AMA) *Principles of Medical Ethics,* imply that no common law duty or ethical imperative exists (...) that requires a physician to treat every patient." According to [40], "Providers

[4]We note that the "Do No Harm" constraint that providers face is implicitly embedded in our formulation. We interpret "Do No Harm" as meaning that the beneficiary's expected utility under the treatment level selected by the provider cannot be lower than when she does not receive treatment. It is straightforward to show that such a constraint translates into a lower bound on the treatment level. Hence, our model captures this constraint through the lower bound $\underline{t}$.

of care also have many tools for risk selection. The most obvious one is to refuse to treat certain patients, or to refer more complex, expensive, or unwanted cases to other providers." There is a large volume of evidence showing that physicians do practice patient selection, sometimes referred to as "defensive medicine." In a study conducted by [106], 42 percent of responding physicians have restricted their practices to avoid risky procedures, patients with complex conditions, or those perceived to be litigious[5]. Often physicians avoid risky patients by unnecessarily referring them to other specialists. Specifically in the context of bundled payments, [39] explicitly states that patients deemed eligible for bundled payments for orthopedic surgery at a certain medical facility are selected based on medical criteria, such as a low enough Body-Mass-Index and "a lack of comorbidities increasing the likelihood of complications, such as diabetes or HIV." They quote a doctor who helped coordinate the program as saying that "We want ideally to have the healthiest patients possible for all surgery, especially the episode of care because we want to minimize the risk for infection, complications and readmissions." This is consistent with our assumption that the provider uses anticipated cost estimates to decide whether to accept a patient.

Finally, the provider's treatment level and selection decisions are made for each individual beneficiary. In other words the provider decisions for each beneficiary are independent of other beneficiaries.

Treatment costs and success probability

We assume that the marginal treatment cost increase goes up as the treatment level goes up (since presumably the provider selects the most cost-effective procedures among those that can be deemed "optional", first), whereas the probability of success improves less and

[5]This study found evidence of direct patient selection in a FFS environment, while in our model this type of patient selection does not take place under FFS. Indeed, in [106] the threat of malpractice liability plays a role in providers' clinical behavior: the main reason for patient selection is the fear of litigation, lack of confidence in liability insurance and burden of insurance premium. Our model does not capture liability insurance and the risk of malpractice litigation, to focus on incentives created by the payment system alone.

less with the increase in treatment level. Therefore, we model the first-stage cost $c_1(t)$ as increasing and convex in t, and the probability of success $q(t)$ as increasing and concave in t. For technical reasons due to risk aversion, we make the following slightly stronger assumptions. Below, $\kappa > 1$ denotes the reimbursement rate under the fee-for-service payment system (see more details in Section 1.4.1).

Assumption 1.2 *First-stage cost $c_1(t)$ is increasing in t and satisfies the inequality:*

$$c_1''(t) - \theta(\kappa - 1)c_1'(t)^2 > 0 \quad \textit{for all } t \in [\underline{t}, \overline{t}].$$

Assumption 1.3 *Probability of success $q(t)$ is increasing in t and satisfies the inequality:*

$$q''(t) + 2\theta q'(t)c_1'(t) < 0 \quad \textit{for all } t \in [\underline{t}, \overline{t}].$$

These assumptions ensure that the first-stage cost c_1 is *sufficiently convex* and the probability of success is *sufficiently concave*. Assumption 1.2 implies that c_1 is convex, that is, higher treatment intensity results in higher treatment costs, and that the marginal cost of extra procedures goes up with the treatment level at a sufficiently high rate. Assumption 1.3 implies that q is concave, that is, a more intense treatment level makes it more likely the treatment will be successful, but the positive effect of increasing the treatment intensity lessens as the treatment level gets higher. In particular, this assumption is consistent with the belief that providers would only intensify the treatment level when this would improve the chances of a positive outcome, even though this may not be justified from a cost-benefit standpoint. Thus, while we do consider financial incentives in treatment decisions, the model does not assume providers are purely driven by financial motives, by ruling out interventions that do not benefit patients.

We assume that the second-stage cost is independent of the first-stage treatment intensity. This assumption is valid when a failure of treatment in the first stage generates a need for a certain course of treatment independently of procedures undertaken in the first stage. For example, if a patient is re-admitted to a hospital after an episode of care, new imaging (x-ray,

MRI, etc.) is generally done to obtain the most recent information, even if imaging had been done in the first stage; new specialist consultations are ordered to have experts assess the current state of the patient even if the patient had such consultation in the original episode of care; a new hospital stay is necessary, regardless of how long the patient stayed in the hospital in the first stage. We recognize that our model does not apply to situations where it is possible to gradually increase the treatment intensity (i.e., if a physician may start at a low intensity, then in case of failure, try the next more intense treatment option, etc.).

In the main body of the paper, we assume that the probability of success does not depend on the patient type. Treatment failure may have a wide variety of causes, but for example it was shown that hospital readmissions are commonly caused by patients not understanding their discharge instructions (such as medications), lack of follow-up care, and a rushed discharge process [32]. These factors are independent of the patient's average second-stage treatment cost (i.e., the patient type) when a readmission does occur; therefore, as an approximation, we focus on the effect of measures taken in the original treatment on the probability of success of the outcome. Appendix A examines the effects of relaxing this assumption and making the success probability a function of the treatment intensity as well as the patient type (i.e., the probability of success is modeled as $q(t, \mu)$) when the provider is risk-neutral. We find that most of the analysis can be carried over to this case and the key managerial insights remain intact.

1.4 Payment Models and System Optimum

In this section we study fee-for-service, bundled payment, and the system optimum and we discuss our findings.

1.4.1 Fee-for-service

In the fee-for-service (FFS) payment system, the insurer reimburses the provider for every procedure or test done on the patient. This implies that the amount received from the insurer increases in t. In addition, the reimbursement must cover the treatment costs (otherwise the

provider would reject every patient). In reality, reimbursement levels are set through a complicated process (somewhat lacking transparency) involving negotiations between the insurer and the medical group representing the providers, and the negotiated rates and margins may vary significantly across providers and even depending on the procedure [94]. For modeling simplicity, we assume that the insurer pays an amount proportional to the treatment costs: the provider receives $\kappa c_1(t)$ in the first stage, and κc_2 in the second stage, where $\kappa > 1$. Thus the provider keeps a margin of $\kappa - 1$ for all procedures run on a beneficiary.

Consider a given beneficiary. If the beneficiary is accepted and treated at intensity t, the expected (with respect to the treatment outcome) utilities of the provider, insurer and beneficiary are:

$$\begin{aligned}
\pi^P(t) &= q(t)U^P((\kappa-1)c_1(t)) + (1-q(t))U^P((\kappa-1)(c_1(t)+c_2) - T^P) \\
&= \frac{1}{\theta} - \frac{1}{\theta}e^{-\theta(\kappa-1)c_1(t)}\left(q(t) + (1-q(t))e^{-\theta((\kappa-1)c_2 - T^P)}\right) \\
\pi^I(t) &= -\kappa\left(c_1(t) + (1-q(t))c_2\right) \\
\pi^B(t) &= q(t)U^B(V) + (1-q(t))U^B(V - T^B).
\end{aligned}$$

After determining the type μ of the beneficiary, if the beneficiary is accepted the provider selects the treatment level for this beneficiary so as to maximize her expected utility with respect to the second-stage cost, which can be written:

$$E_{c_2|\mu}\left[\pi^P(t)|\mu\right] = \frac{1}{\theta} - \frac{1}{\theta}e^{-\theta(\kappa-1)c_1(t)}\left(q(t) + (1-q(t))J_\mu\right)$$

where

$$J_\mu = E_{c_2|\mu}\left[e^{-\theta((\kappa-1)c_2 - T^P)}|\mu\right] = \int_{\underline{c}}^{\overline{c}} e^{-\theta((\kappa-1)c_2 - T^P)} g_\mu(c_2) d\, c_2.$$

Lemma 1.2 *J_μ is non-increasing in μ.*

Proof. We have

$$J_\mu = \int_{\underline{c}}^{\overline{c}} e^{-\theta((\kappa-1)c_2 - T^P)} g_\mu(c_2) d\, c_2 = e^{\theta T^P} \int_{\underline{c}}^{\overline{c}} e^{-\theta(\kappa-1)c_2} g_\mu(c_2) d\, c_2.$$

Under Assumption 1.1, we apply Lemma 1.1 to the function $\psi(x) = e^{-\theta(\kappa-1)x}$ and we find that $E[e^{-\theta(\kappa-1)c_2}]$ is non-increasing in μ, hence the result. □

We denote $\Delta c = c_1(\bar{t}) - c_1(\underline{t})$. The result below determines the provider's treatment strategy that maximizes her expected utility for a beneficiary of type μ.

Proposition 1.1 *For a beneficiary of type μ who receives treatment under FFS, the provider selects a treatment intensity*

$$t^{FFS}(\mu) = \begin{cases} \bar{t} & \text{if } \mu < \mu_1; \\ \underline{t} & \text{else,} \end{cases}$$

where μ_1 is uniquely defined as

$$J_{\mu_1} = 1 - \frac{1 - e^{-\theta(\kappa-1)\Delta c}}{1 - q(\underline{t}) - (1 - q(\bar{t}))e^{-\theta(\kappa-1)\Delta c}}. \tag{1.1}$$

Proof. The provider's expected utility for a type μ-patient is

$$E_{c_2|\mu}\left[\pi^P(t)|\mu\right] = \frac{1}{\theta} - \frac{1}{\theta}e^{-\theta(\kappa-1)c_1(t)}\left(q(t) + (1 - q(t))J_\mu\right)$$

where

$$J_\mu = \int_{\underline{c}}^{\overline{c}} e^{-\theta((\kappa-1)c_2 - T^P)} g_\mu(c_2) d\, c_2.$$

Taking the derivative with respect to t, we find

$$d\, E_{c_2|\mu}\left[\pi^P(t)|\mu\right]/d\, t = e^{-\theta(\kappa-1)c_1(t)}\left((\kappa-1)c_1'(t)(q(t) + (1 - q(t))J_\mu) + (J_\mu - 1)q'(t)/\theta\right).$$

If $J_\mu \geq 1$, because we also know that $\kappa \geq 1,\ c'(t) \geq 0,\ q'(t) \geq 0,\ 0 \leq q(t) \leq 1$ and $J_\mu \geq 0$, it follows that the provider's objective function is non decreasing in t, and thus the provider selects the highest possible treatment level $\bar{t}$. If $J_\mu < 1$, we need to evaluate the second derivative of the provider's expected utility:

$$\begin{aligned} d^2\, E_{c_2|\mu}\left[\pi^P(t)|\mu\right]/d\, t^2 = e^{-\theta(\kappa-1)c_1(t)}[&(\kappa-1)(c_1''(t) - \theta(\kappa-1)c_1'(t)^2)(q(t) + (1 - q(t))J_\mu) \\ &- (1 - J_\mu)q''(t)/\theta + 2(\kappa-1)c_1'(t)q'(t)(1 - J_\mu)]. \end{aligned}$$

By Assumption 1.2, we have $c_1''(t) - \theta(\kappa-1)c_1'(t)^2 \geq 0$. Since $q''(t) \leq 0$ it follows that $d^2\, E_{c_2|\mu}\left[\pi^P(t)|\mu\right]/d\,t^2 \geq 0$ and $E_{c_2|\mu}\left[\pi^P(t)|\mu\right]$ is therefore a convex function of t. Hence, it is maximized at either the lowest or the highest possible treatment levels $\underline{t}$ or $\bar{t}$, whichever leads to the highest value of the provider's objective function. We have

$$\begin{aligned} E_{c_2|\mu}\left[\pi^P(\bar{t})|\mu\right] > E_{c_2|\mu}\left[\pi^P(\underline{t})|\mu\right] &\Leftrightarrow e^{-\theta(\kappa-1)c_1(\bar{t})}\left(q(\bar{t}) + (1-q(\bar{t}))J_\mu\right) \leq e^{-\theta(\kappa-1)c_1(\underline{t})}\left(q(\underline{t}) + (1-q(\underline{t}))J_\mu\right) \\ &\Leftrightarrow e^{-\theta(\kappa-1)\Delta c} < \frac{q(\underline{t}) + (1-q(\underline{t}))J_\mu}{q(\bar{t}) + (1-q(\bar{t}))J_\mu} \\ &\Leftrightarrow J_\mu > \frac{q(\bar{t})e^{-\theta(\kappa-1)\Delta c} - q(\underline{t})}{1 - q(\underline{t}) - (1-q(\bar{t}))e^{-\theta(\kappa-1)\Delta c}} \\ &= 1 - \frac{1 - e^{-\theta(\kappa-1)\Delta c}}{1 - q(\underline{t}) - (1-q(\bar{t}))e^{-\theta(\kappa-1)\Delta c}} \end{aligned}$$

where in the right-hand side above, the denominator is clearly positive since q increasing implies $1 - q(\underline{t}) > 1 - q(\bar{t})$. It follows that

$$t^{FFS}(\mu) = \begin{cases} \bar{t} & \text{if } J_\mu \geq 1 - \frac{1-e^{-\theta(\kappa-1)\Delta c}}{1-q(\underline{t})-(1-q(\bar{t}))e^{-\theta(\kappa-1)\Delta c}} \\ \underline{t} & \text{else.} \end{cases}$$

By Lemma 1.2,

$$J_\mu \geq 1 - \frac{1 - e^{-\theta(\kappa-1)\Delta c}}{1 - q(\underline{t}) - (1-q(\bar{t}))e^{-\theta(\kappa-1)\Delta c}} \Leftrightarrow \mu \leq \mu_1,$$

where μ_1 is such that

$$J_{\mu_1} = 1 - \frac{1 - e^{-\theta(\kappa-1)\Delta c}}{1 - q(\underline{t}) - (1-q(\bar{t}))e^{-\theta(\kappa-1)\Delta c}}.$$

□

We observe that $(\kappa-1)c_2 - T^P$ is the added payoff when the treatment fails compared to a successful treatment. Hence $(1/\theta)(1-J_\mu)$ is the expected utility (with respect to c_2) of the added payoff due to a failed treatment. When $J_\mu > 1$, a failed treatment is expected to provide a *negative* added utility, thus the provider has every incentive to select the highest possible treatment intensity. Note that $J_\mu > 1$ implies $\mu < \mu_1$, so the result above confirms this intuition. If $J_\mu < 1$, a failed treatment is expected to provide a *positive* added utility, thus the provider faces a trade-off: treat little in the first stage to increase the chance of

failure and gain higher utility in the second stage, or treat intensely in the first stage to receive more payoff from the insurer's reimbursement in the first round, despite an increase in the chance of success. She selects the latter when the amplitude of utility differential is not too large, i.e. when

$$1 - J_\mu < 1 - J_{\mu_1} = \frac{1 - e^{-\theta(\kappa-1)\Delta c}}{1 - q(\underline{t}) - (1 - q(\bar{t}))e^{-\theta(\kappa-1)\Delta c}},$$

because the potential utility gain from first-stage failure is too low considering the chance of failure under the two extreme treatment levels and the added utility in the first stage from treating intensely. This occurs when (ceteris paribus) T^P is large, Δc is large, the chance of failure at $\underline{t}$ is small or the chance of failure at $\bar{t}$ is large.

The result below determines the effect of the expected second-stage cost (μ) on the provider utility. It illustrates that the incentives within the FFS system are not conducive to an efficient use of resources, and is one of the main motivation for designing a different payment system.

Proposition 1.2 *Under FFS, the provider's expected utility for a beneficiary of given type μ increases with μ. Hence, potentially costlier beneficiaries yield a higher expected provider utility.*

Proof. By Lemma 1.2, we have $J'_\mu \equiv \partial I_\mu / \partial \mu \leq 0$. Hence, on each domain ($\mu < \mu_1$ or $\mu > \mu_1$), we have

$$\partial E_{c_2|\mu}\left[\pi^P(t)|\mu\right] / \partial \mu = -\frac{1}{\theta} e^{-\theta(\kappa-1)c_1(t^{FFS}(\mu))}(1 - q(t^{FFS}(\mu)))J'_\mu \geq 0.$$

Moreover, from the proof of Proposition 1.1, it is clear that $E_{c_2|\mu}\left[\pi^P(t)|\mu\right]$ is continuous at the breakpoint $\mu = \mu_1$. The result thus follows. □

Since the beneficiaries with lower expected second-stage cost yield lower expected utility for the provider, it is possible that they would not generate an expected utility sufficiently high to motivate the provider to provide treatment. We refer to this outcome as "reverse patient selection". Reverse patient selection occurs when the reputation cost of a failed

treatment is so large that for beneficiaries with a very low expected second-stage cost, the potential loss from a failed treatment is too high and is not compensated by the potential gain from the insurer payment. This is formalized in the result below.

Proposition 1.3 *Under FFS, the provider may have incentives to implement reverse patient selection: if*

$$1 - e^{-\theta(\kappa-1)c_1(\bar{t})}\left(q(\bar{t}) + (1-q(\bar{t}))J_{\underline{\mu}}\right) < 0 \tag{1.2}$$

then the provider rejects beneficiaries of type $\mu \leq \mu^{FFS}$ where μ^{FFS} is such that

$$1 - e^{-\theta(\kappa-1)c_1(\bar{t})}\left(q(\bar{t}) + (1-q(\bar{t}))J_{\mu^{FFS}}\right) = 0.$$

Proof. We first show that the provider's expected utility at the slope breakpoint μ_1 is positive. This is because

$$\frac{1}{\theta} > \frac{1}{\theta}e^{-\theta(\kappa-1)c_1(t)} > 0$$

and $J_{\mu_1} < 1$, which implies that $q(t) + (1-q(t))J_{\mu} < 1$.

As a result, the provider may earn a negative expected utility by treating patients with a very low μ if the provider's expected utility is negative at $\underline{\mu}$. The cost threshold of reverse patient selection is then μ^{FFS} such that

$$\frac{1}{\theta} - \frac{1}{\theta}e^{-\theta(\kappa-1)c_1(\bar{t})}\left(q(\bar{t}) + (1-q(\bar{t}))J_{\mu^{FFS}}\right) = 0.$$

□

Because denial of treatment to potentially less costly patients is not a phenomenon generally observed in the current FFS system, we will assume in the remainder of the paper that condition (1.2) does not hold, i.e. the provider earns a non-negative expected utility even for the least costly beneficiary type ($\underline{\mu}$), so there is no (reverse) patient selection under FFS. Also, because in practice, under FFS, providers do not select a low treatment level to inflate the chance of treatment failure for clear ethical reasons, we will assume that $\overline{\mu} < \mu_1$ so that on the relevant domain $\mu \in [\underline{\mu}, \overline{\mu}]$, the provider selects $t^{FFS}(\mu) = \bar{t}$. Intuitively this

condition ensures that the penalty cost incurred by the provider, in case complications occur, outweighs the financial benefits. We summarize these two technical assumptions below. Note that these assumptions are made only to guide the selection of model parameters in a way that is aligned with practical observation; they have no impact on the technical aspects of our results.

Assumption 1.4 *We assume that model parameters are such that under FFS no beneficiary is denied treatment and treatment level is never at the minimum level. That is,*

$$1 - e^{-\theta(\kappa-1)c_1(\bar{t})}\left(q(\bar{t}) + (1-q(\bar{t}))J_{\underline{\mu}}\right) \geq 0, \qquad \overline{\mu} < \mu_1,$$

where μ_1 *is defined by* (1.1).

1.4.2 Bundled payment

In the bundled payment (BP) system, the provider receives a pre-set lump sum payment denoted by B to cover all services provided, including those in a potential second stage, should the first-stage treatment fail, regardless of the treatment intensity selected. The lump sum B is set at the average historical spending minus a required discount [77]. We denote by γ the discount rate, so that B is given by:

$$B = (1-\gamma)E_\mu\left[E_{c_2|\mu}[\kappa c_1(t^{FFS}(\mu)) + (1-q(t^{FFS}(\mu)))\kappa c_2|\mu]\right] = (1-\gamma)\kappa\left(c_1(\bar{t}) + (1-q(\bar{t}))E[\mu]\right). \tag{1.3}$$

Consider a given beneficiary. If the beneficiary is accepted and treated at intensity t, the expected (with respect to the treatment outcome) utilities of the provider, insurer and beneficiary are:

$$\begin{aligned}
\pi^P(t) &= q(t)U^P(B - c_1(t)) + (1-q(t))U^P(B - c_1(t) - c_2 - T^P) \\
&= \frac{1}{\theta} - \frac{1}{\theta}e^{-\theta(B-c_1(t))}\left(q(t) + (1-q(t))e^{\theta(c_2+T^P)}\right) \\
\pi^I(t) &= -B \\
\pi^B(t) &= q(t)U^B(V) + (1-q(t))U^B(V - T^B).
\end{aligned}$$

After determining the type μ of the beneficiary, if the beneficiary is accepted the provider selects the treatment level for this beneficiary so as to maximize her expected utility with respect to the second-stage cost, which can be written as

$$E_{c_2|\mu}\left[\pi^P(t)|\mu\right] = \frac{1}{\theta} - \frac{1}{\theta}e^{-\theta(B-c_1(t))}\left(q(t) + (1-q(t))I_\mu\right),$$

where

$$I_\mu = E_{c_2|\mu}\left[e^{\theta(c_2+T^P)}|\mu\right] = \int_{\underline{c}}^{\overline{c}} e^{\theta(c_2+T^P)} g_\mu(c_2) d\, c_2.$$

Lemma 1.3 *$I_\mu > 1$ and I_μ is non-decreasing in μ.*

Proof. We have

$$I_\mu = \int_{\underline{c}}^{\overline{c}} e^{\theta(c_2+T^P)} g_\mu(c_2) d\, c_2 = e^{\theta T^P} \int_{\underline{c}}^{\overline{c}} e^{\theta c_2} g_\mu(c_2) d\, c_2.$$

It is clear that $I_\mu > 1$. Under Assumption 1.1, we apply Lemma 1.1 to the function $\psi(x) = e^{\theta x}$ and we find that $E[e^{\theta c_2}]$ is non-decreasing in μ, hence the result. □

The result below determines the provider's treatment strategy that maximizes her expected utility for a beneficiary of type μ.

Proposition 1.4 *For a beneficiary of type μ who receives treatment under BP, the provider selects a treatment intensity $t^{BP}(\mu)$ such that*

$$t^{BP}(\mu) = \begin{cases} t_0 & \text{if } \underline{t} \le t_0 \le \overline{t}; \\ \underline{t} & \text{if } t_0 < \underline{t}; \\ \overline{t} & \text{if } t_0 > \overline{t}, \end{cases}$$

where t_0 is the unique solution of the equation

$$\theta c_1'(t)\left[1 - q(t) + \frac{1}{I_\mu - 1}\right] = q'(t). \tag{1.4}$$

Proof. The provider's expected utility for a type μ-patient is

$$E_{c_2|\mu}\left[\pi^P(t)|\mu\right] = \frac{1}{\theta} - \frac{1}{\theta}e^{-\theta(B-c_1(t))}\left(q(t) + (1-q(t))I_\mu\right)$$

where

$$I_\mu = \int_{\underline{c}}^{\overline{c}} e^{\theta(c_2+T^P)} g_\mu(c_2) d\, c_2.$$

Taking the derivative with respect to t, we find

$$\frac{d\, E_{c_2|\mu}\left[\pi^P(t)|\mu\right]}{d\, t} = e^{-\theta(B-c_1(t))}\left(-c_1'(t)(q(t)+(1-q(t))I_\mu) - \frac{1}{\theta}(1-I_\mu)q'(t)\right) \tag{1.5}$$

$$\begin{aligned}\frac{d^2\, E_{c_2|\mu}\left[\pi^P(t)|\mu\right]}{d\, t^2} &= e^{-\theta(B-c_1(t))}\left(-\theta c_1'(t)^2(q(t)+(1-q(t))I_\mu) - (1-I_\mu)q'(t)c_1'(t)\right. \\ &\quad \left. - c_1''(t)(q(t)+(1-q(t))I_\mu) - c_1'(t)q'(t)(1-I_\mu) - \frac{1}{\theta}(1-I_\mu)q''(t)\right) \\ &= -e^{-\theta(B-c_1(t))}\left[(1-I_\mu)\left(\frac{1}{\theta}q''(t) + 2c_1'(t)q'(t)\right) + \theta c_1'(t)^2(q(t)+(1-q(t))I_\mu) + \right. \\ &\quad \left. c_1''(t)(q(t)+(1-q(t))I_\mu)\right].\end{aligned}$$

The second bracketed term above is clearly non-negative. by Assumption 1.2, the last one is non-negative and the first one is positive. Hence, for a given μ, the provider's objective is a concave function of t. As a result, it is maximized at the only stationary point as long as this point lies in the feasible interval. Setting the first derivative of the provider's expected utility to zero leads to (1.4). □

We note that the treatment decision of the provider under BP depends not only on μ, but also on the entire distribution of the second-stage cost (through I_μ, an integral that depends on the density function of c_2). As a result, the specific distribution of the second-stage cost (and hence its variance) for this patient type influences the treatment level through the quantity I_μ. Our model ensures that the distribution of the second-stage cost is part of a given family of distributions indexed by a single parameter, μ, the expected value. By Assumption 1.1, the second-stage cost has a monotone likelihood ratio property, which ensures the monotonicity of the provider's expected utility with respect to μ, making μ a reasonable choice to model the patient type and provides a basis for selecting the treatment

levels. Therefore, while the treatment levels are functions of the entire distribution of the second-stage cost, in the remainder of this paper we continue to denote the treatment levels with the argument μ only.

Proposition 1.4 states that the provider may choose an intermediary treatment intensity contrasting with the analogous result under FFS, which states that all beneficiaries are treated at the same maximal treatment intensity.

The result below confirms that the BP system possesses the desirable property of treating more the beneficiaries expected to incur higher second-stage costs.

Proposition 1.5 *Under BP, potentially costlier beneficiaries require a higher treatment intensity.*

Proof. If there exists a $\underline{\mu} \leq \tilde{\mu} \leq \overline{\mu}$ for which $t^{BP}(\tilde{\mu}) = \overline{t}$, we show that for all $\tilde{\mu} \leq \mu \leq \overline{\mu}$, $t^{BP}(\mu) = \overline{t}$. To see this, note that if $t^{BP}(\tilde{\mu}) = \overline{t}$, then

$$q'(t) - \theta c_1'(t)\left(1 - q(t) + \frac{1}{I_{\tilde{\mu}} - 1}\right) \geq 0, \quad \forall t \in [\underline{t}, \overline{t}].$$

On the other hand, because $I_\mu - 1 > 0$ and increasing in μ by Lemma 1.3, it follows that

$$q'(t) - \theta c_1'(t)\left(1 - q(t) + \frac{1}{I_{\mu} - 1}\right) \geq 0, \quad \forall t \in [\underline{t}, \overline{t}], \forall \tilde{\mu} \leq \mu \leq \overline{\mu}.$$

Therefore, $t^{BP}(\mu) = \overline{t}$. Using a similar logic, it follows that if there exists a $\underline{\mu} \leq \check{\mu} \leq \overline{\mu}$ for which $t^{BP}(\check{\mu}) = \underline{t}$, we show that for all $\underline{\mu} \leq \mu \leq \check{\mu}$, $t^{BP}(\mu) = \underline{t}$.

For all other values of $\mu \in [\underline{\mu}, \overline{\mu}]$, t^{BP} is the solution to (1.4). Taking the derivative of (1.4), we have

$$\frac{d\, t^{BP}(\mu)}{d\, \mu}\Big[\theta c_1''\left(t^{BP}(\mu)\right)\left(1 - q\left(t^{BP}(\mu)\right) + \frac{1}{I_\mu - 1}\right) - \theta c'\left(t^{BP}(\mu)\right) q'\left(t^{BP}(\mu)\right) - q''\left(t^{BP}(\mu)\right)\Big] -$$
$$\theta c_1'\left(t^{BP}(\mu)\right) \frac{I_\mu'}{(I_\mu - 1)^2} = 0,$$

where by Lemma 1.3, $I_\mu' \equiv \partial\, I_\mu / \partial\, \mu \geq 0$. By assumption 1.3, because c_1 and q are increasing, we have

$$-q''\left(t^{BP}(\mu)\right) - \theta c'\left(t^{BP}(\mu)\right) q'\left(t^{BP}(\mu)\right) > -q''\left(t^{BP}(\mu)\right) - 2\theta c'\left(t^{BP}(\mu)\right) q'\left(t^{BP}(\mu)\right) > 0.$$

By Lemma 1.3, $I_\mu - 1 > 0$ and $I'_\mu \geq 0$. By Assumption 1.2, $c_1''\left(t^{BP}(\mu)\right) \geq 0$. It thus follows that $d\,t^{BP}(\mu)/d\,\mu \geq 0$ for all $\mu \in [\underline{\mu}, \overline{\mu}]$. □

Given that the second-stage treatment cost is not additionally compensated for under the BP system, the provider opts for a higher treatment intensity so as to increase the probability of success and reduce the risk of incurring additional charges for those beneficiaries who are on average costlier in case of complication. Hence, one would expect that that potentially costlier beneficiaries lead to a lower expected provider utility. The following result precisely shows that. Therefore, the provider may have incentives to deny treatment to the costliest beneficiaries. This confirms one of the criticisms of the BP system and illustrates a key difference with the FFS system under which costlier beneficiaries lead to a *higher* expected provider utility.

Proposition 1.6 *Under BP, the provider's expected utility for a beneficiary of given type μ is non-increasing with μ; hence the provider may have incentives to implement patient selection. Namely, if*

$$1 - e^{-\theta(B - c_1(t^{BP}(\overline{\mu})))}\left[q(t^{BP}(\overline{\mu})) + (1 - q(t^{BP}(\overline{\mu})))I_{\overline{\mu}}\right] < 0,$$

then the provider rejects beneficiaries of type $\mu \geq \mu^{BP}$ where μ^{BP} is such that

$$1 - e^{-\theta(B - c_1(t^{BP}(\mu^{BP})))}\left[q(t^{BP}(\mu^{BP})) + (1 - q(t^{BP}(\mu^{BP})))I_{\mu^{BP}}\right] = 0. \tag{1.6}$$

Proof. Clearly, if $t_0 < \underline{t}$ or $t_0 > \overline{t}$, we have $d\,t^{BP}(\mu)/d\,\mu = 0$ and the result holds. If $\underline{t} \leq t_0 \leq \overline{t}$ then taking the derivative of the provider's expected utility evaluated at $t^{BP}(\mu)$ for a given μ, with respect to μ, and using (1.4) we get

$$\begin{aligned}\frac{d\,E_{c_2|\mu}\left[\pi^P(t^{BP})|\mu\right]}{d\,\mu} &= -\frac{1}{\theta}e^{-\theta(B - c_1(t^{BP}))}\left[\frac{d\,t^{BP}(\mu)}{d\,\mu}(1 - I_\mu)\left(q'(t^{BP}) - \theta c_1'(t^{BP})\left[1 - q(t^{BP}) + \frac{1}{I_\mu - 1}\right]\right)\right. \\ &\quad \left. + (1 - q(t^{BP}))I'_\mu\right] \\ &= -\frac{1}{\theta}e^{-\theta(B - c_1(t^{BP}))}(1 - q(t^{BP}))I'_\mu \leq 0,\end{aligned}$$

where the second equality follows from (1.4), and the inequality follows from Lemma 1.3. This indicates that the provider earns a non-positive utility by treating patients with a very high μ if the provider utility is negative at $\overline{\mu}$. The cost threshold of patient selection is then μ^{BP} that leads to an expected utility equal to zero. □

When there is patient selection the provider rejects beneficiaries of type $\mu \in [\mu^{BP}, \overline{\mu}]$ and accepts beneficiaries of type $\mu \in [\underline{\mu}, \mu^{BP})$. Hence the expected number of beneficiaries that undergo treatment is $N' = N \cdot p$, where $p \in [0, 1]$, given by $p = \Pr(\mu < \mu^{BP}) = \int_{\underline{\mu}}^{\mu^{BP}} f(x)dx$, is the probability that a given beneficiary is not rejected.

1.4.3 Benchmark: the system optimum

[38, Chap. 4, p. 62] notes in the context of medical treatments that "the social planner is concerned with all incremental resources associated with treatment, whether borne by patients, providers, or insurers." When all agents are risk neutral, a natural way of defining the goal of a central planner is that of maximizing the total expected system payoff[6], comprising the beneficiary, insurer and provider. It is easy to see that for risk-neutral agents this is equivalent to finding a Pareto-optimal solution, that is, a solution such that no agent's expected payoff may be improved without impairing another agent's payoff. Then, since the payment system only impacts payment exchanges *internal* to the system, the central planner's goal is unaffected by the type of payment system – FFS or BP; it only depends on the treatment level and patient selection threshold decisions. Hence, coordination of the system aims at designing a payment system so that the treatment level and selection threshold decisions match those maximizing the total expected system payoff.

When one or more of the agents is risk averse, it is no longer clear what the central planner's goal should be, as pointed out in [49]. Indeed, the sum of the agents' expected

[6]One may consider that the system optimum is subject to a "Do No Harm" constraint, similarly to the constraint the providers face, as explained in Footnote 4. In this case, the beneficiary's expected utility under the treatment level selected by the central planner cannot be lower than when she does not receive treatment. As explained in the decentralized case, such a constraint translates into a lower bound on the treatment level. Hence, our model also captures this constraint at the system optimum through the lower bound $\underline{t}$.

utilities is not a good candidate for the central planner's objective because, for a given set of treatment and patient selection decision (*external* decisions), the sum of the agents' expected utilities depends on the *internal* allocation of payoff among agents, namely it *depends on the payment system.* This implies that it is impossible to define a certain set of external decisions – a system-optimal treatment level and selection threshold – as those that should be matched under any payment system if one wants to achieve coordination.

Observing that when all agents are risk neutral, the system-optimal decisions match the Pareto-optimal decisions, [49] propose using the concept of Pareto-optimality, widely used in group decision theory, to define coordination of a system with at least one risk-averse decision-maker. They suggest that the goal of a central planner is to make decisions in such a way that no agent's expected *utility* can be improved without impairing another agent's expected *utility.*

The use of Pareto optimality as the criterion in group decision-making dates to [120]. [120] considers " a group of individual decision-makers who must make a common decision under uncertainty, and who, as a result, will receive jointly a payoff to be shared among them". He analyzes "the decision process (...) when the members have diverse risk tolerances". This fits well with our framework. [11] discusses the "complex of services that center about the physician, private and group practice, hospitals, and public health" in the context of medical-care market in the presence of uncertainty and risk. He states that "the equilibrium is necessarily optimal in the following precise sense (due to V. Pareto): There is no other allocation of resources to services which will make all participants in the market better off." Therefore, we use the notion of Pareto-optimality in our paper to define the system optimum.

We consider the central planner's decision of determining for each beneficiary type whether the beneficiary should be treated and if so, at what level. A decision is said to be system-optimal if it is Pareto-optimal for the system consisting of the provider, insurer and beneficiary.

Proposition 1.7 *A treatment and patient selection decision is Pareto-optimal if and only*

if the system's total expected payoff is maximized.

Proof. This proof is adapted from [49]. Let S the set of decisions made by the central planner (treatment level and possibly patient selection level depending on μ). Let $w(S)$ the (uncertain) system payoff:

$$w(S) = w^P(S) + w^I(S) + w^B(S).$$

Because the insurer is risk neutral, the system agents' expected utilities for the set of decisions are:

$$\begin{aligned}\pi^P(S) &= E[U^P(w^P(S))] \\ \pi^I(S) &= E[w^I(S)] \\ \pi^B(S) &= E[U^B(w^B(S))].\end{aligned}$$

Consider S^C a coordinating set of decisions, i.e. decisions that lead to Pareto-optimal expected utilities. Suppose that S^C does not maximize the expected system payoff, i.e. there exists S^* such that $E[w(S^*)] > E[w(S^C)]$. Under decisions S^*, we internally allocate the payoff distribution or arrange side payments between the insurer and both the beneficiary and provider so that $w^P(S^*) = w^P(S^C)$ and $w^B(S^*) = w^B(S^C)$. Then the beneficiary and provider are indifferent between S^* and S^C (they get the same utility under each possible scenario, hence the same expected utility), and the insurer is left with the payoff

$$\begin{aligned}w^I(S^*) &= w(S^*) - w^P(S^*) - w^B(S^*) \\ &= w(S^*) - w^P(S^C) - w^B(S^C) \\ &= w(S^*) - w(S^C) + w^I(S^C).\end{aligned}$$

Hence

$$\pi^I(S^*) = E[w^I(S^*)] = E[w(S^*)] - E[w(S^C)] + E[w^I(S^C)] > E[w^I(S^C)] = \pi^I(S^C),$$

where the inequality follows from the assumption that $E[w(S^*)] > E[w(S^C)]$. Therefore the provider is better off with S^* than with S^C. This contradicts the Pareto-optimality of S^C.

For the reverse, suppose that S^* maximizes the expected system payoff. If S^* is not Pareto-optimal, there exists S^C that improves at least one agent's expected utility whithout reducing any other agent's expected utility. Hence,

$$\pi^P(S^C) + \pi^I(S^C) + \pi^B(S^C) > \pi^P(S^*) + \pi^I(S^*) + \pi^B(S^*).$$

Under decisions S^*, we internally allocate the payoff distribution or arrange side payments between the insurer and both the beneficiary and provider so that $w^P(S^*) = w^P(S^C)$ and $w^B(S^*) = w^B(S^C)$. It follows that $\pi^P(S^C) = \pi^P(S^*)$ and $\pi^B(S^C) = \pi^B(S^*)$. Therefore,

$$E[w^I(S^C)] = \pi^I(S^C) > \pi^I(S^*) = E[w^I(S^*)] = E[w(S^*)] - E[w(S^C)] + E[w^I(S^C)],$$

which implies $E[w(S^*)] - E[w(S^C)] < 0$, contradicting the fact that S^* maximizes the expected system payoff. □

This result implies that, as long as one agent (the insurer) is risk neutral, coordination may be achieved by maximizing the system's total expected payoff, regardless of the beneficiary and provider's specific utility functions. Intuitively, once the total expected payoff is maximized, it is possible to design payment exchanges internal to the system to ensure Pareto-optimality.

The expected (with respect to the treatment outcome) total system payoff from treating a beneficiary with second-stage cost c_2 at level t is

$$w^S(t) = -c_1(t) + (1 - q(t))(-T^P - c_2 - T^B) + V.$$

If the beneficiary is of type μ, it is optimal for the system to choose the treatment intensity that maximizes the system's total expected payoff, which can be written as:

$$E_{c_2|\mu}[w^S(t)|\mu] = -c_1(t) + (1 - q(t))(-T^P - \mu - T^B) + V.$$

The result below determines the central planner's treatment strategy that maximizes the system's total expected payoff for a beneficiary of type μ.

Proposition 1.8 *For a beneficiary of type μ who receives treatment, it is optimal for the system to select a treatment intensity $t^*(\mu)$ such that*

$$t^*(\mu) = \begin{cases} t_1 & \text{if } \underline{t} \le t_1 \le \bar{t}; \\ \underline{t} & \text{if } t_1 < \underline{t}; \\ \bar{t} & \text{if } t_1 > \bar{t}, \end{cases}$$

where t_1 is the unique solution of the equation

$$c_1'(t) = (T^P + \mu + T^B)q'(t). \tag{1.7}$$

Proof. Clearly, for a given μ, by Assumptions 1.2 and 1.3, the central planner's objective is a concave function of t, hence it is maximized at the only stationary point as long as this point lies in the feasible interval. □

The following result characterizes how the system optimum changes as the beneficiary's expected second stage treatment cost varies.

Proposition 1.9 *At the system optimum, potentially costlier beneficiaries require a higher treatment intensity.*

Proof. Clearly, if $t_1 < \underline{t}$ or $t_1 > \bar{t}$, we have $d\,t^*(\mu)/d\,\mu = 0$ and the result holds. If $\underline{t} \le t_1 \le \bar{t}$, taking the derivative of (1.7), we have

$$\frac{d\,t^*(\mu)}{d\,\mu}\Big[c_1''\left(t^*(\mu)\right) - (T^P + \mu + T^B)q''\left(t^*(\mu)\right)\Big] - q'\left(t^*(\mu)\right) = 0.$$

By Assumptions 1.2 and 1.3, and because q is increasing in t, we find that $d\,t^*(\mu)/d\,\mu \ge 0$. □

Note that costlier beneficiaries lead to lower system's total expected payoff (like BP, and unlike FFS, for the expected provider utility); hence it may not be in the benefit of the system, from a cost-benefit standpoint, to necessarily provide treatment to every single patient. For example, if a beneficiary has very high first-stage and expected second-stage

costs, the probability of success is low, failure penalties are high, and/or for the considered episode of care the utility to be obtained by the beneficiary upon receiving treatment is not very large, the costs may outweigh the benefits and the best decision from the perspective of the entire system would be to deny treatment to this beneficiary. The following result formalizes this argument.

Proposition 1.10 *The system's total expected payoff for a beneficiary of given type μ decreases with μ; hence it may be system-optimal to implement patient selection. Namely, if*

$$V - T^P - \overline{\mu} - T^B - c_1(t^*(\overline{\mu})) + (T^P + \overline{\mu} + T^B)q(t^*(\overline{\mu})) < 0, \tag{1.8}$$

then the total system payoff is maximized when beneficiaries of type $\mu \geq \mu^*$ are rejected, where

$$V - T^P - \mu^* - T^B - c_1(t^*(\mu^*)) + (T^P + \mu^* + T^B)q(t^*(\mu^*)) = 0. \tag{1.9}$$

Proof. Clearly, if $t_1 < \underline{t}$ or $t_1 > \bar{t}$, we have $d\,E_{c_2|\mu}\left[\pi^P(t)|\mu\right]/d\,\mu \leq 0$. If $\underline{t} \leq t_1 \leq \bar{t}$, taking the derivative of the expected total system payoff for a given μ, with respect to μ, and using (1.7) we get

$$\begin{aligned}\frac{d\,E_{c_2|\mu}\left[\pi^P(t)|\mu\right]}{d\,\mu} &= \frac{d\,t^*(\mu)}{d\,\mu}\Big[-c_1'\left(t^*(\mu)\right) + (T^P + \mu + T^B)q'\left(t^*(\mu)\right)\Big] - \left(1 - q\left(t^*(\mu)\right)\right) \\ &= -\left(1 - q\left(t^*(\mu)\right)\right) \leq 0.\end{aligned}$$

This indicates that the total expected system payoff may be negative when treating patients with a very high μ if the total expected system payoff is negative at $\overline{\mu}$. The expected second-stage cost threshold for patient selection is then μ^* that leads to a total expected system payoff equal to zero. □

This result implies, in particular, that the FFS system cannot be aligned with the system optimum because FFS may not lead to patient selection (only to reverse patient selection under certain conditions). The BP system optimal solution presents some similarities to the system optimum and we show in the next section that it may be possible to align the patient selection decision with that of the system optimum.

1.4.4 Discussion

Treatment intensity

In the following we compare the treatment levels under BP and FFS as well as the system optimum. We start by focusing on the case of a risk-neutral provider and we then consider risk-averse providers.

Proposition 1.11 *If the provider is risk neutral, for a beneficiary of type μ, the treatment levels under the different payment settings are ranked as follows:*

$$t^{BP}(\mu) \leq t^*(\mu) \leq t^{FFS}(\mu). \tag{1.10}$$

Proof. When the provider is risk neutral, it aims at maximizing its expected payoff

$$E_{c_2|\mu}\left[\pi^P(t)|\mu\right] = B - c_1(t) - (1 - q(t))(\mu + T^P).$$

This problem is identical to finding the system-optimal solution when $V = T^B = 0$. It follows that the optimal risk-neutral treatment level under BP is

$$\begin{cases} t_0^N & \text{if } \underline{t} \leq t_0^N \leq \bar{t}; \\ \underline{t} & \text{if } t_0^N < \underline{t}; \\ \bar{t} & \text{if } t_0^N > \bar{t}, \end{cases}$$

where t_0^N is the unique solution of the equation

$$c_1'(t) = (T^P + \mu)q'(t). \tag{1.11}$$

If $t_0^N < \underline{t}$, then by definition $t^*(\mu) \geq t^{BP}(\mu) = \underline{t}$ and the left inequality holds. If $t_0^N > \bar{t}$ then by concavity of the hospital's expected payoff we have $c_1'(t) \leq (T^P + \mu)q'(t)$ for all $t \in [\underline{t}, \bar{t}]$. Therefore, clearly $c_1'(t) \leq (T^P + \mu + T^B)q'(t)$ and $t^* = \bar{t}$ and the left inequality is tight. Finally if $\underline{t} \leq t_0^N \leq \bar{t}$ then we have

$$c_1'(t_1) = (T^P + \mu + T^B)q'(t_1), \qquad c_1'(t_0^N) = (T^P + \mu)q'(t_0^N).$$

Suppose $t_1 \leq t_0^N$. Because c_1 is convex,

$$0 \geq c_1'(t_1) - c_1'(t_0^N) = T^B\, q'(t_1) + (T^P + \mu)(q'(t_1) - q'(t_0^N)).$$

Since q is increasing, $T^B\, q'(t_1) > 0$, hence $q'(t_1) - q'(t_0^N) < 0$. This contradicts that q is concave. Thus $t_1 > t_0^N$.

Moreover, $t_0^N = t^{BP}(\mu) \geq \underline{t}$, so $t_1 > \underline{t}$. If $t_1 \leq \bar{t}$, then $t_1 = t^*(\mu)$, and the first inequality follows. If $t_1 > \bar{t}$, then $t^*(\mu) = \bar{t} \geq t^{BP}(\mu)$. So the left inequality holds for all cases. By Assumption 1.4, $t^{FFS}(\mu) = \bar{t}$ hence the second inequality is clear. □

The left inequality in (1.10) implies that the BP system in general achieves a lower treatment level than the system optimum, validating one of the main criticisms of the BP system, namely that it could lead to skimping on patient care to keep the costs down. This is because the treatment level selection by the provider does not take into account the beneficiary disutility from treatment failure, which is a factor in the system-optimal treatment level.

Because FFS only selects extreme treatment levels, while at the system optimum the treatment level is generally intermediate (i.e., in general $t^*(\mu) < t^{FFS}(\mu)$), the treatment level decision under FFS cannot in most cases achieve the system-optimal treatment level either. The lack of coordination can be attributed to the fact a single player (the insurer) bears all the risk under FFS. We observe a similar result in other contexts; [19] argues that when the risk is taken only by a single party, coordination cannot be achieved in many supply chains. The following result states that when risk aversion is sufficiently small, the finding obtained for risk-neutral providers continues to hold for risk-averse providers.

Corollary 1.1 *If the provider is risk averse and risk aversion is small*[7]*, for a beneficiary of type μ, the treatment levels under the different payment settings are ranked as follows:*

$$t^{BP}(\mu) \leq t^*(\mu) \leq t^{FFS}(\mu).$$

[7]Risk aversion being small is a sufficient, but not necessary, condition. In some cases, the inequalities hold for any risk-aversion coefficient (e.g., Figure 1.2b). In other cases (e.g., Figure 1.2a), there is a threshold which can be found by finding the smallest solution θ to the equality $t^{BP}(\mu) = t^*(\mu)$.

Proof. Since the first order Taylor expansion of I_μ as $\theta \to 0$ is

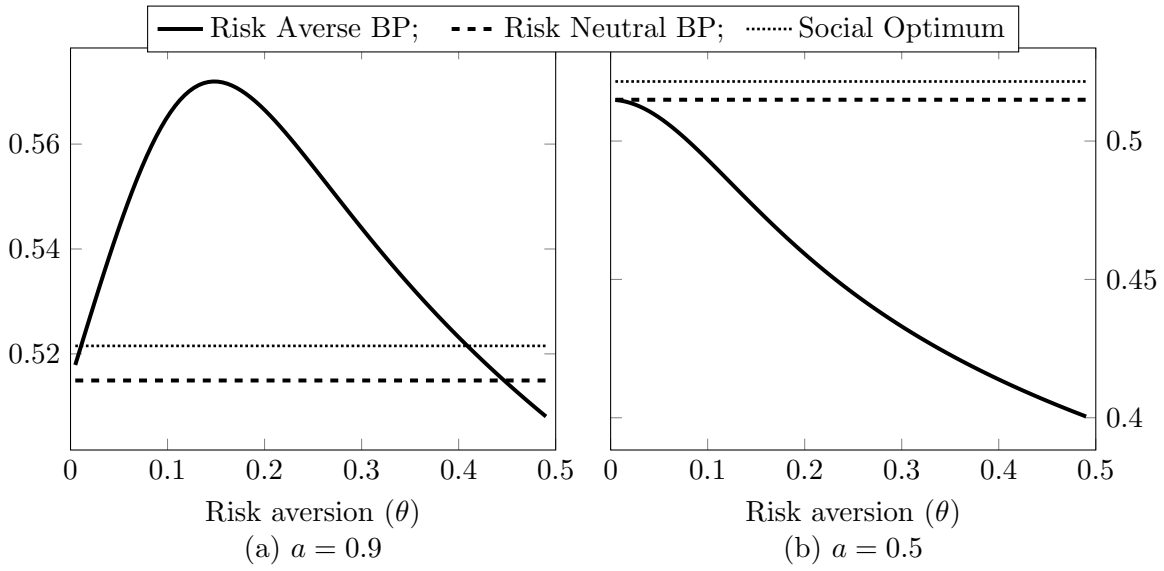
$$\int_{\underline{c}}^{\overline{c}} (1 + \theta(c_2 + T^P)) g_\mu(c_2) d\, c_2 = 1 + \theta(\mu + T^P),$$

the limit of the left hand side of (1.4) is

$$\lim_{\theta \to 0} \theta c_1'(t) \left[1 - q(t) + \frac{1}{I_\mu - 1} \right] = c_1'(t) \lim_{\theta \to 0} \frac{\theta}{I_\mu - 1} = c_1'(t) \frac{1}{\mu + T^P}.$$

Therefore as θ approaches zero, the (risk-averse) solution of (1.4) approaches the (risk-neutral) solution of (1.11).

Hence, the result follows from Proposition 1.11 after observing that the risk-neutral case corresponds to the limit of the risk-averse case for $\theta \to 0$, and that $t^{BP}(\mu)$ is clearly continuous in θ. □



(a) $a = 0.9$ (b) $a = 0.5$

Figure 1.2: Treatment level for a range of risk aversion of the provider for $q(t) = a - 0.4\, e^{-8\, t}, T^P = 1,\ T^B = 2,\ \mu = 20,\ s_\mu = 2,\ \underline{t} = 0,\ \overline{t} = 1,\ c_1(t) = 1 + 2\, t^4$, and $g_\mu(.)$ is a normal density function.

Note that for an arbitrary level of risk aversion, the treatment level under BP for a risk-averse provider may exceed the system-optimal treatment level, as illustrated in Figure 1.2.

In fact, it can be observed that the treatment level under BP may vary non-monotonically with risk aversion[8]; therefore a provider that is more risk averse does not necessarily treat with a higher intensity. As θ increases, the provider becomes more risk averse. Intuitively, the provider faces a trade-off: increasing the treatment level to improve the chance of success for the treatment, while incurring further first-stage costs, or decreasing the treatment level to reduce the first-stage cost despite a decrease in the chance of success. The curvature of the utility function and the success probability function contribute to determining which of these effects dominates the other. While θ is not too large, if the probability of success is large enough (as in Figure 1.2a for lower values of θ), the provider prioritizes the increase in the chance of success, but otherwise (as in Figure 1.2b), the provider aims at reducing the guaranteed first-stage costs. When θ is large, the first-stage cost increase caused by an increase in the treatment level may have an impact on the expected utility that is so large that it is not compensated by the benefit of increasing the probability of success, hence the treatment level decreases in θ.

As noted in the introduction, one of the concerns with a bundled payment mechanism is that it may lead providers to skimp on care, that is, to reduce the intensity of treatment in an effort to reduce cost. It follows from Corollary 1.1 that this concern is valid when providers are risk neutral or not very risk averse (and possibly also in other cases). Figure 1.2 illustrates that when providers are moderately risk averse, they may treat with either too much or too little intensity (compared to the system optimum).

Patient selection

As noted in Section 1.4.3, FFS may only lead to reverse patient selection while the system optimum may only have direct patient selection. As a result, the patient selection decisions

[8]It can be shown that the treatment level under BP for a risk-averse provider increases with θ iff $\theta \bar{I}_\mu - (I_\mu - 1)[1 + (1 - q(t^{BP}(\mu)))(I_\mu - 1)] \geq 0$, where $\bar{I}_\mu = \int_{\underline{c}}^{\overline{c}} (c_2 + T^P) e^{\theta(c_2 + T^P)} g_\mu(c_2) d\, c_2$ and $t^{BP}(\mu)$ solves (1.4). Note that I_μ and $\bar{I}_\mu$ depend on θ.

under the FFS setting and at the system optimum may not be aligned (unless none implements any kind of patient selection). However, as the next result illustrates, since BP leads to direct patient selection, it may be possible to align the patient selection outcomes under BP and the system optimum.

Proposition 1.12 *If there is patient selection at the system optimum, there exists a discount value γ_C such that the BP patient selection decision matches the system optimum.*

Proof. We want to show that there exists a coordinating bundled payment $B_C \geq 0$ such that $\mu^* = \mu^{BP}$. From (1.9) and (1.6), we find that this holds iff

$$1 - e^{-\theta(B_C - c_1(t^{BP}(\mu^*)))} \left[q(t^{BP}(\mu^*)) + (1 - q(t^{BP}(\mu^*)))I_{\mu^*}\right] = 0.$$

This equation can be rewritten as

$$e^{\theta\, B_C} = e^{\theta\, c_1(t^{BP}(\mu^*))} \left[q(t^{BP}(\mu^*)) + (1 - q(t^{BP}(\mu^*)))I_{\mu^*}\right].$$

There exists a solution $B_C \geq 0$ satisfying the above equation when the right-hand side above is greater than or equal to 1. We note that because $I_\mu > 1$ and $1 - q(t^{BP}(\mu^*)) \geq 0$, we have $q(t^{BP}(\mu^*)) + (1 - q(t^{BP}(\mu^*)))I_{\mu^*} \geq 1$, hence

$$e^{\theta\, c_1(t^{BP}(\mu^*))} \left[q(t^{BP}(\mu^*)) + (1 - q(t^{BP}(\mu^*)))I_{\mu^*}\right] \geq e^{\theta\, c_1(t^{BP}(\mu^*))} \geq 1.$$

As a result, there exists a coordinating bundled payment $B_C \geq 0$ given by

$$B_C = c_1(t^{BP}(\mu^*)) + \frac{1}{\theta} \ln\left(q(t^{BP}(\mu^*)) + (1 - q(t^{BP}(\mu^*)))I_{\mu^*}\right).$$

Setting γ_C such that $B = B_C$ in (1.3) leads to the result.

□

Hence, by carefully selecting the discount value (that is, the bundled payment value), the BP system may reach the system optimum in terms of patient selection. In other words, by adjusting B, the insurer can directly control the level of patient selection. For example, a very high B would generate no patient selection at all because any beneficiary would generate a positive utility, but a very low B would motivate the provider to reject every beneficiary.

Beneficiary population size

The previous analysis describes how the provider's risk aversion and the payment system help determine the decisions that the provider makes regarding every individual patient. However, it is also of interest to understand the amount of downside risk that the provider's average payoff is subject to in an aggregate way, from serving the entire beneficiary population. In this section, we measure the amount of downside risk that the provider's payoff is subject to *overall*, by considering the total payoff from serving a population of N beneficiaries and its variability.

A perceived shortcoming of the BP mechanism is the increased level of risk that it imposes on providers especially for those with low volume of patients for a certain episode. [109] argue that the main source of risk for providers is the variation in average per patient episode costs. In other Medicare initiatives, such as the Shared Savings Program, many patients participate (Accountable Care Organizations have at least 5000 enrollees), lowering the financial risk burdened by the providers due to random variations across individual beneficiaries. In contrast, an average medical provider participating in the BPCI experiment has between 100 and 200 cases for their *highest* volume episodes [77]; thus the losses imposed by a few costly beneficiaries may not be offset by less costly beneficiaries, and the average historical cost used to calculate the bundled payment value may significantly differ from the average cost in a given subsequent year. To address the issue of financial risk to the provider under BP, we consider the effect of the beneficiary population size N and we derive analytical bounds on the size of N that guarantees a minimum provider per-beneficiary average payoff with a certain probability. Such a minimum threshold is analogous to the notion of value-at-risk studied in the financial literature.

Recall that N is the size of the beneficiary population that refers to the provider for treatment. This beneficiary group can be viewed as a sample, extracted from a larger population, that consists of those beneficiaries who selected to receive care from this particular provider.

Suppose beneficiary i is of type μ^i and has a true second-stage cost c_2^i, for $i = 1, \ldots, N$. Without loss of generality we assume that $\mu^1 \leq \mu^2 \leq \cdots \leq \mu^N$. Under BP, the provider only accepts those beneficiaries for whom $\mu^i < \mu^{BP}$. Let N' be the highest index of the beneficiary types who are accepted by the provider; i.e., N' is equal to the largest j for which $\mu^j < \mu^{BP}$. Because the provider implements treatment level $t^{BP}(\mu^i)$, the true provider expected payoff (with respect to treatment outcome) of treating beneficiary $i = 1, \ldots, N'$ is $w_i^P = B - c_1(t^{BP}(\mu^i)) - T^P(1 - q(t^{BP}(\mu^i))) - (1 - q(t^{BP}(\mu^i)))c_2^i$, and the (sample) average of the provider's payoff for the treated beneficiaries, $\overline{w^P}$, is:

$$\overline{w^P} = \frac{1}{N'} \sum_{i=1}^{N'} w_i^P = B - \frac{1}{N'} \sum_{i=1}^{N'} [c_1(t^{BP}(\mu^i)) + (c_2^i + T^P)(1 - q(t^{BP}(\mu^i)))]. \tag{1.12}$$

Following the approach adopted by [49], in order to capture the risk faced by providers under the BP mechanism because of limited volume of cases for certain episodes and high levels of variability in treatment costs, we consider the provider's financial *risk exposure*, defined as the α-percentile of the total average payoff faced by the provider for some small α; i.e. the value ρ such that $\Pr\left(\overline{w^P} < \rho\right) = \alpha$.

Thanks to the Hoeffding's inequality [60], we find a relationship between ρ, α and the accepted beneficiary population size N'[9] . The next result formalizes this argument.

Proposition 1.13 *Let the size of the accepted population be*[10]

$$N' = \frac{(\zeta - \xi)^2}{2(B - \delta - \rho)^2} \ln\left(\frac{1}{\alpha}\right), \tag{1.13}$$

then the provider's risk exposure is at most ρ, that is, $Pr\left(\overline{w^P} < \rho\right) \leq \alpha$, where $\overline{w^P}$ is defined in (1.12) and

$$\zeta = c_1(t^{BP}(\mu^{BP})) + (\overline{c} + T^P)(1 - q(t^{BP}(\underline{\mu}))), \quad \xi = c_1(t^{BP}(\underline{\mu})) + (\underline{c} + T^P)(1 - q(t^{BP}(\mu^{BP})))$$

$$\delta = E_{c_2|E[c_2]\leq\mu^{BP}}\left[c_1(t^{BP}(E[c_2])) + (c_2 + T^P)(1 - q(t^{BP}(E[c_2])))|E[c_2] \leq \mu^{BP}\right].$$

[9]Note that because the coefficients c_2^i are not identically distributed and also due to the value of N' not being generally large, we cannot use the central limit theorem to approximate the probability above.

[10]This proposition provides the *sufficient* accepted population size to have a risk exposure of ρ for a given α.

Proof. Replacing the term for $\overline{w^P}$ from (1.12) in $\Pr\left(\overline{w^P} < \rho\right)$, we get:

$$\begin{aligned}
\Pr\left(\overline{w^P} < \rho\right) &= \Pr\left(B - \sum_{i=1}^{N'} \frac{c_1(t^{BP}(\mu^i)) + (c_2^i + T^P)(1 - q(t^{BP}(\mu^i))c_2^i)}{N'} < \rho\right) \\
&= \Pr\left(\sum_{i=1}^{N'} \frac{c_1(t^{BP}(\mu^i)) + (c_2^i + T^P)(1 - q(t^{BP}(\mu^i))c_2^i)}{N'} > B - \rho\right) \\
&= \Pr\left(\sum_{i=1}^{N'} \frac{c_1(t^{BP}(\mu^i)) + (c_2^i + T^P)(1 - q(t^{BP}(\mu^i))c_2^i)}{N'} - \delta > B - \rho - \delta\right)
\end{aligned}$$

Letting $\delta = E_{c_2|E[c_2]\leq\mu^{BP}}\left[c_1(t^{BP}(E[c_2])) + (\overline{c} + T^P)(1 - q(t^{BP}(E[c_2])))\right]$, and using [60, Theorem 2], we get

$$\Pr\left(\overline{w^P} < \rho\right) \leq e^{-\frac{2N'(B-\rho-\delta)^2}{(\zeta-\xi)^2}},$$

where $\zeta = c_1(t^{BP}(\mu^{BP})) + (\overline{c} + T^P)(1 - q(t^{BP}(\underline{\mu})))$ and $\xi = c_1(t^{BP}(\underline{\mu})) + (\underline{c} + T^P)(1 - q(t^{BP}(\mu^{BP})))$. The statement of the proposition is obtained by equating $\alpha = e^{-\frac{2N'(B-\rho-\delta)^2}{(\zeta-\xi)^2}}$ and solving for N'.

□

Note that in the above proposition $B-\delta$ is the provider's average net payoff from accepted beneficiaries and $\zeta-\xi$ is the expected gap in treatment cost between the highest- and lowest-cost beneficiaries accepted by the provider. So if this gap $(\zeta - \xi)$ is large or if the payoff threshold (ρ) is close to the provider's average payoff $(B - \delta)$ then the provider needs a large population of accepted beneficiaries to keep her financial risk (α) low. The following corollary directly results from Proposition 1.13.

Corollary 1.2 *Financial risk and risk exposure of the provider decrease when*

(a) the size of the accepted beneficiary population increases, or

(b) the cost gap between the most and least costly beneficiaries decreases.

This result indicates that the population size N is an important factor in determining the risk borne by providers. If a medical provider cannot attract a large enough patient population size for a certain episode of care, then she may not be able to efficiently risk-pool among the patients and potentially high-cost patients could pose a significant burden on the provider. Therefore the beneficiary population size, N, should be taken into account when considering the implementation of the BP mechanism.

1.5 Proposed Payment Schemes

To alleviate the drawbacks of the BP system while maintaining some of the advantages of the FFS system, such as the low risk borne by the provider, in this section we consider alternative payment mechanisms with the goal of aligning the treatment level and patient selection level selected by the provider to that of the system optimum and thereby fully coordinating this system. We first consider a hybrid payment, which is a combination of FFS and BP mechanisms, in Section 1.5.1. In Section 1.5.2 we analyze a stop-loss protection scheme, which modifies the BP model to limit the total provider cost. Both payment schemes are applicable in practice. The implementation of the hybrid payment system would be no more complicated than the BP mechanism currently tested, and some forms of the stop-loss protection model are readily being implemented in some Medicare programs [109].

1.5.1 A hybrid system

In this section we propose a hybrid payment (HP) system that is a combination of BP and FFS. Specifically, the insurer pays both a fixed amount B' to the provider (as in BP) as well as a variable amount (as in FFS) at each stage of treatment. Similar to the FFS system, the variable payment is proportional to the treatment cost: $\beta c_1(t)$ for the initial treatment, and βc_2 in case of complications. Because of the existence of a fixed payment we set the variable payment factor β to be less than one.

Note that the proposed hybrid system is equivalent to a BP mechanism adjusted for risk-sharing, in which the provider keeps only a fraction $\mathcal{F}$ of the savings if her treatment

costs are lower than the bundled payment B, and in return is only responsible for fraction $\mathcal{F}$ of losses if her treatment costs are higher than the bundled payment B, as long as $B' = \mathcal{F}B$ and $\beta = 1 - \mathcal{F}$. Such a risk-adjusted BP mechanism is analogous to the payment scheme intuitively proposed in [45] (but not analyzed quantitatively), where the provider is paid through a FFS system and yet is rewarded for spending reductions.

Consider a given beneficiary. If the beneficiary is accepted and treated at intensity t, the expected (with respect to the treatment outcome) utilities of the provider, insurer and beneficiary are:

$$\begin{aligned}
\pi^P(t) &= q(t)U^P(B' - (1-\beta)c_1(t)) + (1-q(t))U^P(B' - (1-\beta)c_1(t) - (1-\beta)c_2 - T^P) \\
&= \frac{1}{\theta} - \frac{1}{\theta}e^{-\theta(B'-(1-\beta)c_1(t))}\left(q(t) + (1-q(t))e^{\theta((1-\beta)c_2+T^P)}\right) \\
\pi^I(t) &= -B' - \beta[c_1(t) + (1-q(t))c_2] \\
\pi^B(t) &= q(t)U^B(V) + (1-q(t))U^B(V - T^B).
\end{aligned}$$

After determining the type μ of the beneficiary, if the beneficiary is accepted the provider selects the treatment level for this beneficiary so as to maximize her expected utility with respect to the second-stage cost, which can be written:

$$E_{c_2|\mu}\left[\pi^P(t)|\mu\right] = \frac{1}{\theta} - \frac{1}{\theta}e^{-\theta(B'-(1-\beta)c_1(t))}\left(q(t) + (1-q(t))L_\mu\right),$$

where

$$L_\mu = E_{c_2|\mu}\left[e^{\theta((1-\beta)c_2+T^P)}|\mu\right] = \int_{\underline{c}}^{\overline{c}} e^{\theta((1-\beta)c_2+T^P)}g_\mu(c_2)d\,c_2.$$

Lemma 1.4 *$L_\mu > 1$ and L_μ is non-decreasing in μ.*

Proof. We have

$$L_\mu = \int_{\underline{c}}^{\overline{c}} e^{\theta((1-\beta)c_2+T^P)}g_\mu(c_2)d\,c_2 = e^{\theta T^P}\int_{\underline{c}}^{\overline{c}} e^{\theta(1-\beta)c_2}g_\mu(c_2)d\,c_2.$$

It is clear that $L_\mu > 1$. Under Assumption 1.1, we apply Lemma 1.1 to the function $\psi(x) = e^{\theta(1-\beta)x}$ and we find that $E[e^{\theta(1-\beta)c_2}]$ is non-decreasing in μ, hence the result. □

The result below determines the provider's treatment strategy that maximizes her expected utility for a beneficiary of type μ.

Proposition 1.14 *For a beneficiary of type μ who receives treatment under the hybrid system, the provider selects a treatment intensity $t^{HP}(\mu)$ such that*

$$t^{HP}(\mu) = \begin{cases} t_2 & \text{if } \underline{t} \le t_2 \le \bar{t}; \\ \underline{t} & \text{if } t_2 < \underline{t}; \\ \bar{t} & \text{if } t_2 > \bar{t}, \end{cases}$$

where t_2 is the unique solution of the equation

$$\theta(1-\beta)c_1'(t)\left[1 - q(t) + \frac{1}{L_\mu - 1}\right] = q'(t). \tag{1.14}$$

Proof. The provider's expected utility for a type μ-patient is

$$E_{c_2|\mu}\left[\pi^P(t)|\mu\right] = \frac{1}{\theta} - \frac{1}{\theta}e^{-\theta(B'-(1-\beta)c_1(t))}\left(q(t) + (1-q(t))L_\mu\right)$$

where

$$L_\mu = \int_{\underline{c}}^{\bar{c}} e^{\theta((1-\beta)c_2 + T^P)} g_\mu(c_2)\, d\, c_2.$$

Taking the derivative with respect to t, we find

$$\begin{aligned}
\frac{d\, E_{c_2|\mu}\left[\pi^P(t)|\mu\right]}{d\, t} &= e^{-\theta(B'-(1-\beta)c_1(t))}\left(-(1-\beta)c_1'(t)(q(t) + (1-q(t))L_\mu) - \frac{1}{\theta}(1-L_\mu)q'(t)\right) \\
\frac{d^2\, E_{c_2|\mu}\left[\pi^P(t)|\mu\right]}{d\, t^2} &= e^{-\theta(B'-(1-\beta)c_1(t))}\left(-\theta(1-\beta)c_1'(t)^2(q(t) + (1-q(t))L_\mu) - (1-L_\mu)q'(t)(1-\beta)c_1'(t)\right. \\
&\quad \left. - (1-\beta)c_1''(t)(q(t) + (1-q(t))L\mu) - (1-\beta)c_1'(t)q'(t)(1-L_\mu) - \frac{1}{\theta}(1-L_\mu)q''(t)\right) \\
&= -e^{-\theta(B'-(1-\beta)c_1(t))}\left[(1-L_\mu)\left(\frac{1}{\theta}q''(t) + 2(1-\beta)c_1'(t)q'(t)\right)\right. \\
&\quad \left. +\theta(1-\beta)^2 c_1'(t)^2(q(t) + (1-q(t))L_\mu) + (1-\beta)c_1''(t)(q(t) + (1-q(t))I_\mu)\right].
\end{aligned}$$

The second bracketed term above is clearly non-negative. The last one is non-negative by Assumption 1.2. The first one is positive by Lemma 1.4 and Assumption 1.2 after noting that

$(1/\theta)q''(t) + 2(1-\beta)c_1'(t)q'(t) \leq (1/\theta)q''(t) + 2c_1'(t)q'(t)$. Hence, for a given μ, the provider's objective is a concave function of t. As a result, it is maximized at the only stationary point as long as this point lies in the feasible interval. Setting the first derivative of the provider's expected utility to zero leads to (1.14).

□

It is clear that the hybrid system shows much resemblance to the BP system and shares some of its analytical properties; in particular it exhibits the same incentives for patient selection, as noted in the result below (the proof is very similar to the proofs of Propositions 1.5 and 1.6 and is thus omitted).

Proposition 1.15 *Under the hybrid system, (i) potentially costlier beneficiaries require a higher treatment intensity; (ii) the provider's expected utility for a beneficiary of given type μ decreases with μ; hence the provider may have incentives to implement patient selection: namely, if*

$$1 - e^{-\theta(B' - (1-\beta)c_1(t^{HP}(\overline{\mu})))} \left[q(t^{HP}(\overline{\mu})) + (1 - q(t^{HP}(\overline{\mu})))L_{\overline{\mu}} \right] < 0,$$

then the provider rejects beneficiaries of type $\mu \geq \mu^{HP}$ where μ^{HP} is such that

$$1 - e^{-\theta(B' - (1-\beta)c_1(t^{HP}(\mu^{HP})))} \left[q(t^{HP}(\mu^{HP})) + (1 - q(t^{HP}(\mu^{HP})))L_{\mu^{HP}} \right] = 0. \tag{1.15}$$

We now show that when the provider is risk neutral, there exists a fraction β and a fixed payment B' that coordinate the decisions in the hybrid system to that of the system optimum (i.e., that lead to the same patient selection and treatment level as the system optimum for all patients).

Proposition 1.16 *When the provider is risk neutral, a hybrid system with $\beta = T^B/(T^B + T^P)$ and $B' = VT^P/(T^B + T^P)$ (i.e. $B' = V(1-\beta)$) aligns the patient selection and treatment intensity outcomes to those of the system optimum.*

Proof. Similarly to the proof of Proposition 1.11, we find that under the hybrid payment scheme, the treatment level satisfies (unless it lies at one of the extremes):

$$c_1'(t_2) = \left(\frac{T^P}{1-\beta} + \mu \right) q'(t_2).$$

Moreover, if

$$B' - (1-\beta)c_1(t^{HP}(\overline{\mu})) - (T^P + (1-\beta)\overline{\mu})(1 - q(t^{HP}(\overline{\mu}))) < 0,$$

then the provider rejects patients of type $\mu \geq \mu^{HP}$ where

$$B' - (1-\beta)c_1(t^{HP}(\mu^{HP})) - (T^P + (1-\beta)\mu^{HP})(1 - q(t^{HP}(\mu^{HP}))) = 0. \tag{1.16}$$

The system-optimal treatment level satisfies (unless it lies at one of the extremes):

$$c_1'(t_1) = (T^P + \mu + T^B)q'(t_1).$$

These two equations giving treatment levels are equivalent iff $T^P/(1-\beta) = T^P + T^B$, i.e. $\beta = T^B/(T^B + T^P)$.

From (1.9) and (1.16), we have

$$\mu^* = \frac{V - c_1(t^*(\mu^*))}{1 - q(t^*(\mu^*))} - T^P - T^B, \qquad \mu^{HP} = \frac{B'/(1-\beta) - c_1(t^{HP}(\mu^{HP}))}{1 - q(t^{HP}(\mu^{HP}))} - \frac{T^P}{1-\beta}.$$

These two quantities are equal iff

$$\frac{V - c_1(t^*(\mu^*))}{1 - q(t^*(\mu^*))} - T^P - T^B = \frac{B'/(1-\beta) - c_1(t^{HP}(\mu^{HP}))}{1 - q(t^{HP}(\mu^{HP}))} - \frac{T^P}{1-\beta}. \tag{1.17}$$

In this case, when $\beta = T^B/(T^B + T^P)$ we also have $t^{HP}(\mu^{HP}) = t^*(\mu^*)$, hence equation (1.17) can be rewritten

$$\begin{aligned}
B' &= (1-\beta)\left(V - (T^P + T^B)(1 - q(t^*(\mu^*))) + T^P \frac{1 - q(t^*(\mu^*))}{1-\beta}\right) \\
&= (1-\beta)\left(V + \left(\frac{T^P}{1-\beta} - (T^P + T^B)\right)(1 - q(t^*(\mu^*)))\right) \\
&= (1-\beta)\left(V + \left(\frac{T^P}{1 - T^B/(T^B + T^P)} - (T^P + T^B)\right)(1 - q(t^*(\mu^*)))\right) \\
&= V(1-\beta).
\end{aligned}$$

□

In the case of a risk neutral provider, even though FFS could not coordinate any of the decisions and BP could only align the incentives in terms of patient selection level, HP can

coordinate both the patient selection level and treatment intensity. This is because the HP system acts as a risk-sharing mechanism to distribute high-cost patients' risk between the provider and insurer.

When the provider is risk averse, this result no longer holds. There are two reasons for this. The first is mathematical: the way that the treatment levels are computed under the hybrid system and the system optimum are so structurally different that the solutions cannot match for all beneficiary types. The second is intuitive: the hybrid system is a mixture of a FFS system and BP system. Hence, the treatment level under HP lies between the BP and the FFS levels. However, when the provider is risk averse, the BP treatment level may *exceed* the system-optimal treatment level. Since the FFS treatment level is at the upper extreme, it follows that the HP treatment level also exceeds the system-optimal treatment level, hence no coordination is possible.

The following result investigates how a hybrid system with a menu of payment terms may coordinate the treatment level and patient selection to that of the system optimum as long as the BP treatment level does not exceed the system-optimal treatment level. By Corollary 1.1, this will for example be true when the risk aversion is small.

Proposition 1.17 *When the provider is risk averse and $t^{BP}(\mu) \leq t^*(\mu)$, there exists a cost share, β, and a bundled payment, B', that coordinate the provider's treatment level and patient selection decisions with the system optimum for a given beneficiary type.*

Proof. As shown in the proof of Proposition 1.14, the objective of the provider under HP is a concave function with the following first derivative.

$$\frac{d\, E_{c_2|\mu}\left[\pi^P(t)|\mu\right]}{d\, t} = \frac{e^{-\theta(B'-(1-\beta)c_1(t))}}{\theta}\left(q'(t) - \theta(1-\beta)c_1'(t)\left[1 - q(t) + \frac{1}{L_\mu - 1}\right]\right), \quad (1.18)$$

where

$$L_\mu = \int_{\underline{c}}^{\overline{c}} e^{\theta((1-\beta)c_2+T^P)} g_\mu(c_2) d\, c_2 = e^{\theta T^P} \int_{\underline{c}}^{\overline{c}} e^{\theta(1-\beta)c_2} g_\mu(c_2) d\, c_2.$$

We first show the existence of β that can coordinate the treatment levels, and then show there exists a B' to coordinate patient selection levels. We consider three cases: (1) $t^*(\mu) = \bar{t}$, (2) $t^*(\mu) = \underline{t}$, and (3) $\underline{t} < t^*(\mu) < \bar{t}$.

Case 1: $t^*(\mu) = \bar{t}$. If $\beta \to 1$ then (1.18) implies that the derivative of $E_{c_2|\mu}\left[\pi^P(t)|\mu\right]$ is non-negative for all $t \in [\underline{t}, \bar{t}]$ since $c_1'(t), q(t)$, and L_μ are all finite parameters. Therefore, $t^{HP}(\mu)\big|_{\beta\to 1} = \bar{t} = t^*(\mu)$.

Case 2: $t^*(\mu) = \underline{t}$. If $\beta \to 0$ then (1.18) is the same as the derivative of the hospital's utility under BP with bundled payment B'. Note that we assumed $t^{BP}(\mu) \leq t^*(\mu) = \underline{t}$. Since $t^*(\mu) = \underline{t}$ it follows $t^{HP}(\mu)\big|_{\beta\to 0} = t^{BP}(\mu) = \underline{t} = t^*(\mu)$.

Case 3: $\underline{t} < t^*(\mu) < \bar{t}$. In order to show that the there exists a β to coordinate the treatment levels in this case, we need to show the existence a β for which $t^*(\mu)$ is the root of (1.18), where $t^*(\mu)$ solves the following at the system optimum:

$$c_1'(t) = (T^P + \mu + T^B)q'(t).$$

Plugging $t^*(\mu)$ into (1.18) and using the property of the system optimum solution, we need to show the existence of β for which:

$$\frac{1}{T^P + \mu + T^B} - \theta(1-\beta)\left[1 - q(t^*(\mu)) + \frac{1}{L_\mu - 1}\right] = 0. \tag{1.19}$$

Thus, a hybrid system coordinates the treatment level decision iff there exists a $\beta \in [0,1]$ that satisfies (1.19) (note that β affects L_μ as well, so this equation is not linear in β). We note that clearly there is no β that satisfies (1.19) *for all* μ, hence it is impossible to design a hybrid payment scheme that coordinates simultaneously all beneficiary types.

When $\beta \to 1^-$, the left-hand-side of (1.19) equals $1/(T^P + \mu + T^B) > 0$.

When $\beta = 0$, $L_\mu = I_\mu$ and thus the left-hand-side of (1.19) equals:

$$\frac{1}{T^P + \mu + T^B} - \theta\left[1 - q(t^*(\mu)) + \frac{1}{I_\mu - 1}\right],$$

which is the derivative of the provider's utility under BP evaluated at $t^*(\mu)$. Hence, when $t^*(\mu) \geq t^{BP}(\mu)$, then the above expression is non-positive.

Because the left-hand-side of (1.19) is positive for $\beta = 1$ and non-positive for $\beta = 0$, while being continuous in β, it follows that there exists $\beta^* \in [0, 1]$ that satisfies (1.19) and thus that coordinates the treatment level decision.

We now show the existence of B' for which patient selection levels are coordinated. Recall that at the system optimum, if

$$V - T^P - \overline{\mu} - T^B - c_1(t^*(\overline{\mu})) + (T^P + \overline{\mu} + T^B)q(t^*(\overline{\mu})) < 0,$$

then the total system payoff is maximized when beneficiaries of type $\mu \geq \mu^*$ are rejected, where

$$V - T^P - \mu^* - T^B - c_1(t^*(\mu^*)) + (T^P + \mu^* + T^B)q(t^*(\mu^*)) = 0.$$

At HP, if

$$1 - e^{-\theta(B' - (1-\beta)c_1(t^{HP}(\overline{\mu})))} \left[q(t^{HP}(\overline{\mu})) + (1 - q(t^{HP}(\overline{\mu})))L_{\overline{\mu}}\right] < 0$$

then the provider rejects beneficiaries of type $\mu \geq \mu^{HP}$ where μ^{HP} is such that

$$1 - e^{-\theta(B' - (1-\beta)c_1(t^{HP}(\mu^{HP})))} \left[q(t^{HP}(\mu^{HP})) + (1 - q(t^{HP}(\mu^{HP})))L_{\mu^{HP}}\right] = 0.$$

We want to show that there exists a coordinating bundled payment $B'_C \geq 0$ such that $\mu^* = \mu^{HP}$. From the equations above, we find that this holds iff

$$1 - e^{-\theta(B'_C - (1-\beta)c_1(t^{HP}(\mu^*)))} \left[q(t^{HP}(\mu^*)) + (1 - q(t^{HP}(\mu^*)))L_{\mu^*}\right] = 0.$$

This equation can be rewritten as

$$e^{\theta\, B'_C} = e^{\theta(1-\beta)c_1(t^{HP}(\mu^*))} \left[q(t^{HP}(\mu^*)) + (1 - q(t^{HP}(\mu^*)))L_{\mu^*}\right].$$

There exists a solution $B'_C \geq 0$ satisfying the above equation when the right-hand side above is greater than or equal to 1. We note that because $L_\mu > 1$ and $1 - q(t^{HP}(\mu^*)) \geq 0$, we have $q(t^{HP}(\mu^*)) + (1 - q(t^{HP}(\mu^*)))L_{\mu^*} \geq 1$, hence

$$e^{\theta(1-\beta)c_1(t^{HP}(\mu^*))} \left[q(t^{HP}(\mu^*)) + (1 - q(t^{HP}(\mu^*)))L_{\mu^*}\right] \geq e^{\theta(1-\beta)c_1(t^{HP}(\mu^*))} \geq 1.$$

As a result, there exists a coordinating bundled payment $B'_C \geq 0$ given by

$$B'_C = (1-\beta)c_1(t^*(\mu^*)) + \frac{1}{\theta}\ln\left(q(t^*(\mu^*)) + (1-q(t^*(\mu^*)))L_{\mu^*}\right).$$

□

We note that when the provider is risk averse and $t^{BP}(\mu) > t^*(\mu)$, the hybrid system cannot coordinate decisions to those of the system optimum.

1.5.2 Stop-loss protection

A major drawback of the BP system is having a fixed reimbursement amount while the beneficiaries' cost varies *for a given beneficiary type* and *across beneficiary types*. Note that in our BP formulation, if a beneficiary is high-cost type ($\mu \geq \mu^{BP}$), then the provider has the option of not accepting the beneficiary. However, if a beneficiary is low-cost type ($\mu < \mu^{BP}$) and accepted for treatment, the full burden of the beneficiary's *actual* cost, which varies depending on the realization of c_2 given μ, is borne by the provider. Such variability is undesirable for the provider due to the potential existence of a few outlier cases which increase the risk borne by the provider.

The concern over the high-cost outlier cases has been acknowledged by CMS in other programs such as the Medicare Shared Savings Program and inpatient prospective payment system [109]. One remedy for this problem is implementing a stop-loss protection mechanism. Under our proposed stop-loss protection, the provider is only responsible for a beneficiary's second-stage costs below a certain threshold. Any realized second-stage costs over the pre-specified threshold are burdened by the insurer.

Let S be the stop-loss protection level for the readmission cost. Hence, the provider's cost for providing treatment level t is equal to $c_1(t)$ in case of success and $c_1(t)+\min\{c_2, S\}$ in case of failure. That is, the provider's expected total realized cost (with expectation taken over the treatment outcome) for providing treatment level t is equal to $c_1(t) + (1-q(t))\min\{c_2, S\}$. Therefore, under the BP mechanism with stop-loss protection, we can write the different agents' expected (with respect to the treatment outcome) utilities for a treated beneficiary

of type μ as

$$\begin{aligned}
\pi^P(t) &= q(t)U^P(B - c_1(t)) + (1 - q(t))U^P(B - c_1(t) - \min\{c_2, S\} - T^P) \\
&= \frac{1}{\theta} - \frac{1}{\theta}e^{-\theta(B-c_1(t))}\left(q(t) + (1 - q(t))e^{\theta(\min\{c_2,S\}+T^P)}\right) \\
\pi^I(t) &= -B - (1 - q(t))\left(c_2 - \min\{c_2, S\}\right) \\
\pi^B(t) &= q(t)U^B(V) + (1 - q(t))U^B(V - T^B).
\end{aligned}$$

After determining the type μ of the beneficiary, if the beneficiary is accepted the provider selects the treatment level for this beneficiary so as to maximize her expected utility with respect to the second-stage cost, which can be written:

$$\begin{aligned}
E_{c_2|\mu}\left[\pi^P(t)|\mu\right] &= E_{c_2|\mu}\left[\frac{1}{\theta} - \frac{1}{\theta}e^{-\theta(B-c_1(t))}\left(q(t) + (1 - q(t))e^{\theta(\min\{c_2,S\}+T^P)}\right)|\mu\right] \\
&= \frac{1}{\theta} - \frac{1}{\theta}e^{-\theta(B-c_1(t))}\left(q(t) + (1 - q(t))E_{c_2|\mu}\left[e^{\theta(\min\{c_2,S\}+T^P)}|\mu\right]\right) \\
&= \frac{1}{\theta} - \frac{1}{\theta}e^{-\theta(B-c_1(t))}\left(q(t) + (1 - q(t))M_\mu(S)\right),
\end{aligned}$$

where

$$M_\mu(S) = E_{c_2|\mu}\left[e^{\theta(\min\{c_2,S\}+T^P)}|\mu\right] = \int_{\underline{c}}^{S} e^{\theta(c_2+T^P)}g_\mu(c_2)d\,c_2 + e^{\theta(S+T^P)}\int_{S}^{\overline{c}} g_\mu(c_2)d\,c_2.$$

The next result shows that there may exist a payment system consisting of a BP mechanism augmented with a menu of stop-loss protection levels that coordinates the provider's treatment and selection decisions with the system optimum if the BP treatment level exceeds the system-optimal treatment level.

Proposition 1.18 *When $t^*(\mu) \leq t^{BP}(\mu)$, there exist a stop-loss protection level, S, and a bundled payment, B, that coordinate the provider's treatment level and patient selection decisions with the system optimum for a given beneficiary type if*

$$q(t^*(\mu)) \leq \frac{e^{\underline{c}+T^P}}{e^{\underline{c}+T^P} - 1} - \frac{1}{\theta(T^P + T^B + \mu)}. \tag{1.20}$$

Proof. For a given μ, the provider selects the treatment level $t \in [\underline{t}, \overline{t}]$ that maximizes

$$E_{c_2|\mu}\left[\pi^P(t)|\mu\right] = \frac{1}{\theta} - \frac{1}{\theta} e^{-\theta(B-c_1(t))}\left(q(t) + (1-q(t))M_\mu(S)\right)$$

where

$$M_\mu(S) = \int_{\underline{c}}^{S} e^{\theta(c_2+T^P)} g_\mu(c_2) d\, c_2 + e^{\theta(S+T^P)} \int_{S}^{\overline{c}} g_\mu(c_2) d\, c_2.$$

Taking the derivative of the provider's utility we have:

$$\frac{\partial\, E_{c_2|\mu}\left[\pi^P(t)|\mu\right]}{\partial\, t} = \frac{e^{-\theta(B-c_1(t))}}{\theta}\left[q'(t)M_\mu(S) - q'(t) - \theta q(t)c_1'(t) - (1-q(t))\theta c_1'(t)M_\mu(S)\right].$$

$$\begin{aligned}\frac{\partial^2\, E_{c_2|\mu}\left[\pi^P(t)|\mu\right]}{\partial\, t^2} = \frac{e^{-\theta(B-c_1(t))}}{\theta}\Big[&(q''(t) + 2\theta c_1'(t)q'(t))(M_\mu(S)-1) - \theta c_1''(t)q(t) - \theta c_1''(t)(1-q(t))M_\mu(S) \\ &-\theta^2(c_1'(t))^2 q(t) - \theta^2(c_1'(t))^2(1-q(t))M_\mu(S)\Big].\end{aligned}$$

Note that the first term in the bracket is negative due to Assumption 1.3, and the other terms are negative due to convexity of $c_1(t)$ and concavity of $q(t)$. So the solution of the first derivative is indeed the maximizer of the hospital's utility. Now let, (for clarity of exposition, because here μ is fixed, we write t^* instead of $t^*(\mu)$ below)

$$\left.\frac{\partial\, E_{c_2|\mu}\left[\pi^P(t)|\mu\right]}{\partial\, t}\right|_{t=t^*, S=\underline{c}} = e^{-\theta(B-c_1(t^*))}\left[-c_1'(t^*)(q(t^*) + (1-q(t^*))e^{\theta(\underline{c}+T^P)}) - \frac{1}{\theta}q'(t^*)(1-e^{\theta(\underline{c}+T^P)})\right] \tag{1.21}$$

and

$$\left.\frac{\partial\, E_{c_2|\mu}\left[\pi^P(t)|\mu\right]}{\partial\, t}\right|_{t=t^*, S=\overline{c}} = e^{-\theta(B-c_1(t^*))}\left[-c_1'(t^*)(q(t^*) + (1-q(t^*))I_\mu) - \frac{1}{\theta}q'(t^*)(1-I_\mu)\right]. \tag{1.22}$$

Note that (1.21) is negative due to (1.20). Furthermore, (1.22) is identical to the derivative with respect to t of the expected provider utility under BP (1.5) evaluated at $t = t^*$. As shown in the proof of Proposition 1.4, the expected provider utility under BP is a concave function of t, and its derivative is equal to zero at $t = t^{BP}$. When $t^* \leq t^{BP}$, concavity implies that the derivative of the expected provider utility under BP evaluated at $t = t^*$ is positive.

Therefore, there exists a stop-loss protection level for which the derivative of the provider's objective is zero.

We recall: at the system optimum, if

$$V - T^P - \overline{\mu} - T^B - c_1(t^*(\overline{\mu})) + (T^P + \overline{\mu} + T^B)q(t^*(\overline{\mu})) < 0,$$

then the total system payoff is maximized when beneficiaries of type $\mu \geq \mu^*$ are rejected, where

$$V - T^P - \mu^* - T^B - c_1(t^*(\mu^*)) + (T^P + \mu^* + T^B)q(t^*(\mu^*)) = 0,$$

which results in

$$\mu^* = \frac{V - c_1(t^*(\mu^*))}{1 - q(t^*(\mu^*))} - T^P - T^B. \tag{1.23}$$

In the stop-loss mechanism, if

$$1 - e^{-\theta(B - c_1(t^{SL}(\overline{\mu})))}\left[q(t^{SL}(\overline{\mu})) + (1 - q(t^{SL}(\overline{\mu})))M_{\overline{\mu}}(S)\right] < 0$$

then the provider rejects beneficiaries of type $\mu \geq \mu^{SL}$ where μ^{SL} is such that

$$1 - e^{-\theta(B - c_1(t^{SL}(\mu^{SL})))}\left[q(t^{SL}(\mu^{SL})) + (1 - q(t^{SL}(\mu^{SL})))M_{\mu^{SL}}(S)\right] = 0.$$

If we solve the above equation for B and use coordinating S^* (if exists) we have

$$B = \frac{1}{\theta}\, ln\left(q(t^*(\mu^{SL})) + \left(1 - q(t^*(\mu^{SL}))\right) M_{\mu^{SL}}(S^*)\right) + c_1(t^*(\mu^{SL})). \tag{1.24}$$

In order to be able to coordinate the patient selection level there should exist a non-negative B such that $\mu^* = \mu^{SL}$. From Equation (1.24), it is clear that B is non-negative ($M_{\mu^{SL}} > 1$). μ^* as appeared in Equation (1.23), can be replaced with μ^{SL} in Equation (1.24). □

The result above complements the coordination result of Proposition 1.17 as unlike the HP system, which could only coordinate the system when $t^{BP}(\mu) \leq t^*(\mu)$, Proposition 1.18 shows that it is possible to align the provider's incentives to the system optimum using a stop-loss mechanism when $t^{BP}(\mu) \geq t^*(\mu)$. This result shows the existence of an alternative payment mechanism that aligns the incentives of the provider and the system optimum by

sharing the risk of high-cost patients with the insurer. The hybrid mechanism considered in Section 1.5.1 was another way of sharing the risk between the provider and insurer. Under HP the insurer pays for a certain percentage of costs for *all* patients, while under the stop-loss mechanism the insurer participates in supporting the treatment costs of high-cost patients only. The hybrid system intends to allocate some of the risk borne by the provider due to cost uncertainty to the insurer, to give incentives to *raise* the treatment level to that of the system optimum even though it increases the deterministic first-stage cost. When the provider treats at an intensity that is already too high under BP, the stop-loss mechanism requires to give the provider incentives to *lower* the treatment level by reducing the second-stage related costs.

Finally, while the insurer provides financial support to the provider for costly patients in the stop-loss protection system, this mechanism may not necessarily increase the insurer's total payment. The lump sum payment B for a BP system modified with a stop-loss mechanism to achieve coordination with the system optimum is less than that in the BP mechanism without stop-loss protection, which would reduce the insurer's guaranteed payment to the provider.

We end this section by providing conditions under which neither the hybrid payment nor stop-loss protection scheme can coordinate the system. Given that Propositions 1.17 and 1.18 provide necessary and sufficient conditions for the coordinating payments, we have the following corollary.

Corollary 1.3 *There are no hybrid payment or stop-loss protection mechanisms that can coordinate the system if* both *of the following conditions hold:*

$$t^*(\mu) < t^{BP}(\mu), \qquad q(t^*(\mu)) > \frac{e^{\underline{c}+T^P}}{e^{\underline{c}+T^P}-1} - \frac{1}{\theta(T^P+T^B+\mu)}.$$

1.6 Numerical Analysis

In this section we present numerical experiments that address the motivating questions formulated in the introduction of the paper and that explore the differences in outcomes for the

various payment mechanisms analyzed. We use different performance measures to assess the various payment systems and compare them to one another. Note that in this entire section when we refer to the hybrid and stop-loss payments, we consider the *coordinating* mechanisms for those values of the parameters when these payments can coordinate the provider's decisions as illustrated in Propositions 1.16 and 1.18. In Section 2.6.1 we describe the input parameters used for this numerical study. Section 1.6.2 compares the average provider utilities and system payoffs under FFS, BP, and at the system optimum. Section 1.6.3 evaluates coordination issues by first characterizing regions for which each of the proposed contracts is coordinating and also the region for which no hybrid or stop-loss coordinating mechanism was found. We then assess the financial risk exposure borne by the provider under different payment mechanisms. Section 1.6.4 in the main body of the paper provides a summary of observations made from our numerical experiments.

1.6.1 Input parameters

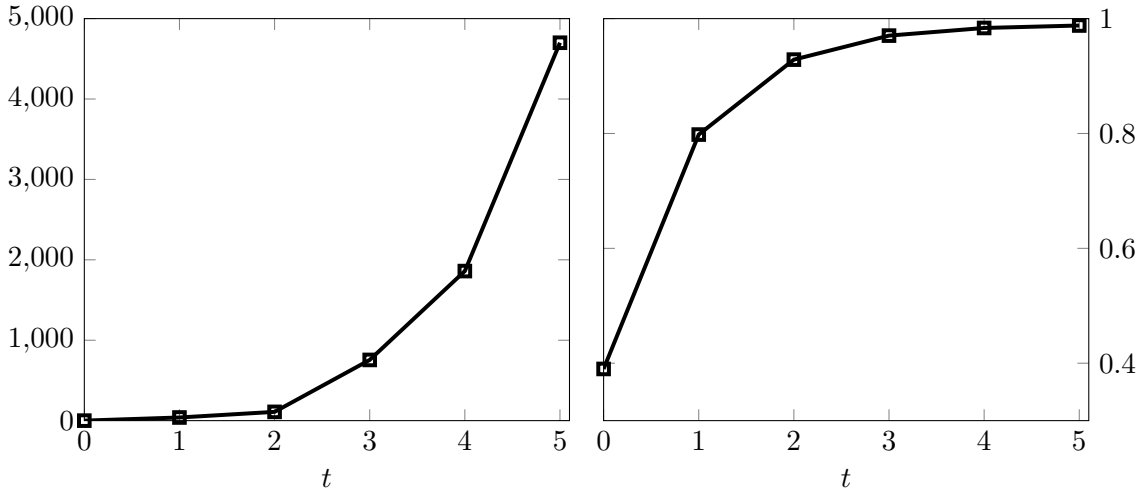
We use parameter values corresponding to a specific condition, GastroEsophageal Reflux Disease (GERD), obtained from the medical and health economics literature as described below. Based on [59], in a given episode of care, there are five different treatment levels; (1) routine visit and consultation, (2) daily dose prescription omeprazole 20 mg tablet for a period of 30 days, (3) manometry, (4) endoscopy, and (5) Laparoscopic Nissen Fundoplication (LNF) surgery, where each treatment level is cumulative (for example, level 2 is to prescribe the drug *and* do a routine consultation). Costs for different levels of treatment are obtained from [59] and summarized in Table 1.2. If complications occur, the average second-stage cost can be as low as \$109, which means the condition can be resolved by an office visit and a standard dose of omeprazole; thus we set $\underline{\mu} = 109$. Alternatively, the complications can be so severe that an Open Nissen Fundoplication (ONF) surgery is required. Using estimates for the cost of this procedure in [89] and a yearly rate of increase of 3% to approximate the cost of ONF in 1997 values, we estimate that the highest average second-stage cost $\overline{\mu}$ lies within \$8,000 – \$10,000. For the purpose of our numerical experiments, we set $\overline{\mu}$ at

\$9,414, but we run sensitivity analysis on this parameter and find that our observations remain unaffected. Finally, we assume that the expected second-stage cost (patient type) is uniformly distributed on $[\underline{\mu}, \overline{\mu}]$.

We set each patient's utility V as one unit of quality-adjusted-life-years (QALY). There is no consensus in the health economics literature on the value of a QALY, however [59] state that the health economics community considers \$50,000 – \$100,000 as an acceptable marginal cost-effectiveness (cost for an additional unit of QALY) range. We use similar values as proxies for the value of a QALY. Therefore, we vary V within the range \$25,000 – \$80,000, in order to evaluate the sensitivity of the outcomes with respect to this parameter, and we obtain similar results. [59] estimate the patient's disutility from complications, T^B, to be within [0.05 V, 0.5 V]; we vary the value of T^B in the same range but choose $T^B = 0.1V$ for the results in this section. Finally, in order to estimate the value of κ, we use the Medicare reimbursement data for LNF from [51] and compare them to the LNF procedure cost from [59], which leads, after adjusting for the year discrepancy, to $\kappa = 1.4$. We vary κ within the range 1.3 – 1.5 and find that our observations remain valid.

Due to a lack of data in prior work on the values of T^P and $q(t)$, we alter each of these parameters in a wide range to assess their impact on the numerical results. We assume that the provider's penalty from complications is not higher than the patient's disutility, and therefore vary T^P within $[0.01\,V,\, 0.1\,V]$ in $0.01\,V$ increments; we find that the value of T^P has little impact on the outcome. For our numerical experiments, we set T^P at $0.02\,V$. Finally, we assume that the probability of success follows a general exponential form $a - be^{-ct}$. We run the numerical experiments for different values of a, b, and c and find that different forms of the success probability, as described above, do not change the qualitative nature of the findings. Therefore, in this section[11] we present the results for the following success probability: $q(t) = 0.99 - 0.6e^{-1.14t}$. Under this probability model, the success likelihood is a small 39% if no action is taken, and a high 99% if the maximum treatment level is exerted.

[11] In section 1.6.3 we let parameter a vary.



(a) First Stage Cost, $c_1(t)$. (b) Success Probability, $q(t)$.

Figure 1.3: Parameter values for numerical experiments

Figure 1.3 illustrates the functional forms of $c_1(t)$ and $q(t)$. Note that this is consistent with Assumptions 1.2 and 1.3; we also check that with the parameter values that we selected, Assumption 1.4 is satisfied.

1.6.2 Total average provider utility and system payoff

Section 1.4 studies the average utilities for a given beneficiary type. In this section we compare the total average provider utility for all beneficiary types, Π^P(Equations (1.25) and (1.26)), and the total average system payoff, W^S(Equations (1.27)–(1.29)), under different payment schemes. Note that due to the possibility that coordinating hybrid and stop-loss protection payments do not exist, these two mechanisms are not studied in this subsection. More numerical analysis on hybrid and stop-loss payment schemes is carried out in the next subsection.

We use the law of total expectation in order to find Π^P for FFS and BP. Because the provider's utility at the system optimum depends on how the total system payoff is split

t	$\{1, 2, 3, 4, 5\}$
$c_1(t)$	{\$39, \$109, \$755, \$1,860, \$4,700}
V	varied in [\$25,000, \$80,000]
T^P	varied in $[0.01\,V,\, 0.1\,V]$, set at $0.02V$
T^B	varied in $[0.05V, 0.5V]$, set at $0.1\,V$
κ	varied in $[1.3, 1.5]$, set at 1.4
$\underline{\mu}$	varied in [\$100, \$200], set at \$109
$\overline{\mu}$	varied in [\$8,000, \$10,000], set at \$9,414
μ	random variable with uniform distribution in $[\underline{\mu}, \overline{\mu}]$
$\underline{c}$	0.001 percentile of the exponential with mean $\underline{\mu}$
$\overline{c}$	0.999 percentile of the exponential with mean $\overline{\mu}$
$c_2\|\mu$	random variable with exponential distribution with average μ

Table 1.2: Parameter values for numerical experiments

among the agents of the system, we do not depict Π^P at the system optimum.

$$\Pi^P_{FFS} = E_\mu\Big[E_{c_2|\mu}\left[\pi^P(\bar{t})|\mu\right]\Big] = \frac{1}{\theta} - \frac{1}{\theta}e^{-\theta(\kappa-1)c_1(\bar{t})}\left(q(\bar{t}) + (1-q(\bar{t}))E_\mu\left[J_\mu\right]\right), \tag{1.25}$$

$$\begin{aligned}\Pi^P_{BP} =& E_\mu\Big[E_{c_2|\mu}\left[\pi^P(t^{BP}(\mu))|\mu\right]\Big] \\ =& \int_{\underline{\mu}}^{\mu^{BP}} \left[\frac{1}{\theta} - \frac{1}{\theta}e^{-\theta(B-c_1(t^{BP}(x)))}\left(q(t^{BP}(x)) + (1-q(t^{BP}(x)))I_x\right)\right] f(x)dx.\end{aligned} \tag{1.26}$$

Similarly, we can find the total expected system payoff, W^S, under FFS, BP, and the system-optimal (denoted by SO) decisions.

$$W^S_{FFS} = V - c_1(\bar{t}) - (1-q(\bar{t}))\left(T^B + T^P + E[\mu]\right), \tag{1.27}$$

$$W^S_{BP} = \int_{\underline{\mu}}^{\mu^{BP}} \left[V - c_1(t^{BP}(x)) - (1-q(t^{BP}(x)))(T^B + T^P + x)\right] f(x)dx, \tag{1.28}$$

$$W^S_{SO} = \int_{\underline{\mu}}^{\mu^*} \left[V - c_1(t^*(x)) - (1-q(t^*(x)))(T^B + T^P + x)\right] f(x)dx. \tag{1.29}$$

Figure 1.4 illustrates the average system payoff in \$1000s under BP, FFS and at the system optimum for $V = \$25,000$ as $1-\gamma$ and θ change. Note that the proportion of past cost reimbursed by the provider under the BP mechanism, $1-\gamma$, is proportional to the bundled payment value, B, as $B = (1-\gamma)\kappa\left(c_1(\bar{t}) + (1-q(\bar{t}))E[\mu]\right)$ from (1.3).

Figure 1.4(a) plots the average system payoff from (1.27), (1.28), and (1.29), respectively, for increasing values of the bundled payment value. By definition the system optimum leads to the highest system payoff. The system payoff initially increases under BP; this is due to the fact that in this range, a higher value of B motivates the provider to accept a larger pool of beneficiaries for treatment, up to the point when $1-\gamma = 0.7$. At that point, the entire patient population is selected for treatment. For $1-\gamma \geq 0.7$, the decrease in insurer's payoff as B increases is offset by the increase in the provider's payoff and as the result the system payoff under BP is constant.

Figure 1.4(b) plots the average system payoff for increasing values of the provider's risk aversion. From Propositions 1.1 and 1.8 and Assumption 1.4, it is clear that the system-optimal and the FFS treatment levels are independent of θ. Hence, the total system payoff is also independent of θ under SO and FFS, as can be observed in Figure 1.4(b). Under BP, the treatment level does depend on θ, in a way that varies according to μ. More importantly, the patient selection level changes with θ. We find (not shown in this figure) that the patient selection threshold for BP changes sharply for θ near 0.06: beyond this value of the risk-aversion parameter, more and more patients are denied treatment (more than system-optimum), which adversely impacts the system payoff, resulting in a sharp drop.

Figure 1.5(a) and (b) show the effect of the bundled payment value and the provider's risk aversion, respectively, on the average provider utility, by plotting the provider's total expected utility in \$1000s under FFS and BP from (1.25) and (1.26). Clearly, under BP the provider's utility increases with the value of B as illustrated in Figure 1.5(a) since the provider's payoff increases in B, while FFS is not affected by changes to the bundled payment value.

Under FFS, the treatment level (and hence the provider payoff) does not vary with θ.

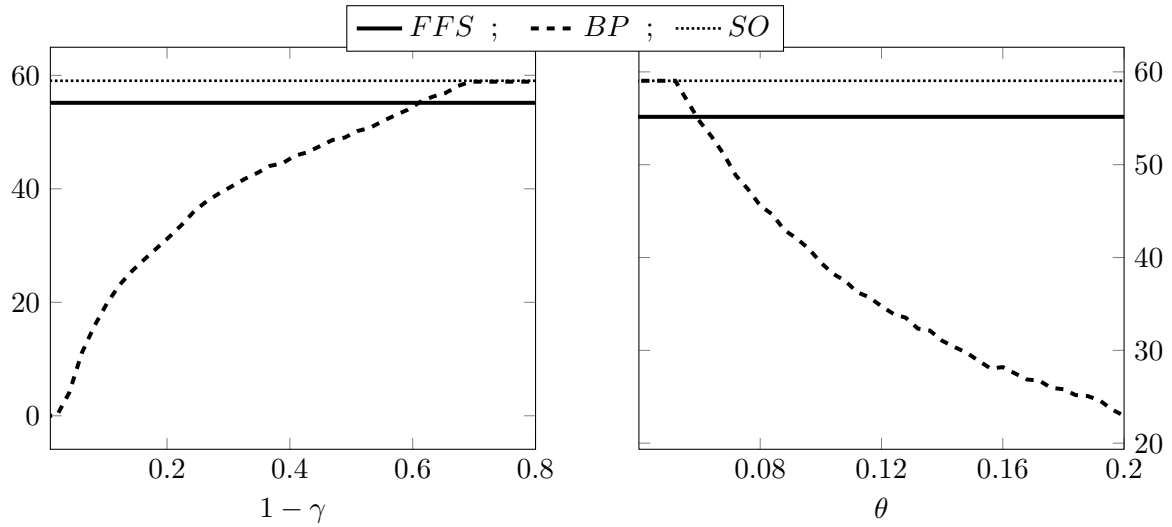
However, for a fixed payment level and hence a fixed payoff, the provider's utility function decreases in θ. Therefore, we observe in Figure 1.5(b) that the average provider's utility under FFS decreases in θ. Under BP, in addition to the utility function decreasing in θ, the treatment level also changes with θ. In this case, for most values of μ, the treatment level goes up in θ to lower the chance of complications. This leads to a higher first-stage cost which negatively impacts the provider utility; therefore, the average provider utility decreases faster under BP than under FFS. The curvature changes around $\theta = 0.13$ because the selection threshold drops at that point and the patient pool used to average the provider utility is reduced for higher values of θ.

It is interesting to observe that when providers are not too risk averse and if B is sufficiently large, both the system and providers themselves benefit from BP as compared to FFS. This is because they no longer treat at the highest possible level, the gain due to the difference between treatment costs and bundled payment exceeds the margin they earn under FFS. Unsurprisingly, the performance of BP critically depends on the value of B (or γ) and θ. For lower values of B and higher values of θ, BP can lead to a lower average system payoff than FFS because of low reimbursement from the insurer leading to suboptimal treatment levels and intense patient selection.

One key observation from Figures 1.4 and 1.5 is that while BP can be an effective mechanism to increase the provider's utility and/or system payoff compared to FFS, its performance is extremely sensitive to the choice of the bundled payment value, B, and the degree of risk aversion θ. For certain values of B and θ, it can be a fairly effective tool, otherwise, its performance can even be dominated by that of FFS, mainly due to patient selection. In Figure 1.4, the whole population has been selected under the system optimum due to our choice of parameters.

1.6.3 Coordinating mechanisms

We start this section by investigating coordination under the proposed mechanisms, namely hybrid payment and stop-loss. Figure 1.6 shows the effect of the risk-aversion parameter, θ,

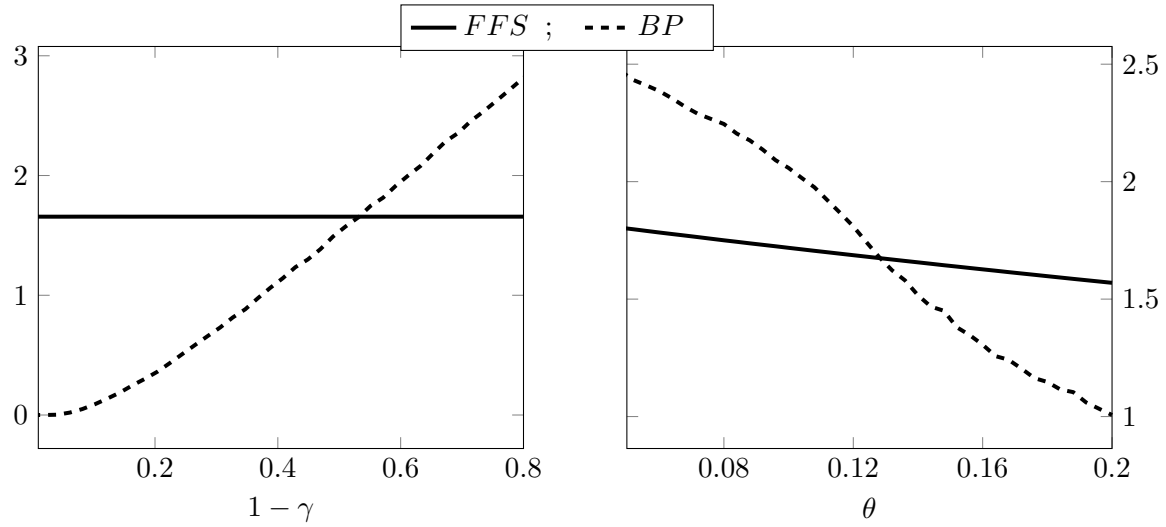


(a) Given a Fixed Risk Aversion Measure, $\theta = 0.14$ (b) Given a Fixed Payment, $\gamma = 0.8$

Figure 1.4: Average system payoff in \$1000s for $V = \$60,000$

and the constant parameter of the success probability function, a, on the payment scheme that would coordinate the supply chain. We allow a to vary in $[0.2, 1]$ to guarantee that the probability of failure $q(t)$ is between zero and one for our set of treatment levels. Based on this figure, for smaller values of θ, HP coordinates. Note that this is consistent with the result of Corollary 1.1, which states that when provider's risk aversion is small enough, $t^{BP}(\mu) \leq t^*(\mu)$ and therefore the HP mechanism coordinates (Proposition 1.17). For large enough values of θ, $t^{BP}(\mu) \geq t^*(\mu)$ and condition (1.20) is satisfied so the stop-loss mechanism coordinates, consistent with Proposition 1.18. For moderate values of θ and large enough values of a (leading to large values of the function q), where $t^{BP}(\mu) \geq t^*(\mu)$ but condition (1.20) is violated, no hybrid or stop-loss coordinating mechanism can be found.

As highlighted earlier, one of the provider's main concerns under BP is the downside risk resulting from high-cost patients. Therefore, in this section we turn our attention to the risk exposure measure introduced in Section 1.4.4. More specifically, for a fixed α and ρ, we find the patient population size that guarantees $\Pr\left(\overline{w^P} < \rho\right) = \alpha$, where $\overline{w^P}$ is defined in



(a) Given a Fixed Risk Aversion Measure, $\theta = 0.14$ (b) Given a Fixed Payment, $\gamma = 0.5$

Figure 1.5: Average provider utility in \$1000s for $V = \$60,000$

(1.12) for BP. We use a similar measure for HP and stop-loss (SL), and compare the patient population size that guarantees a certain risk exposure to that of the BP mechanism.

Table 1.3 illustrates the performance of the BP system compared to the coordinating mechanisms proposed in this paper in terms of the population size required to limit the provider's risk exposure. Table 1.3(a) compares the ratio of the population size of the BP to HP systems ($\frac{N'(\text{BP})}{N'(\text{HP})}$ where N' is obtained from (1.13)) to reach a certain level of risk exposure under each payment mechanisms. As demonstrated in Figure 1.6, HP is coordinating only for small values of θ; therefore we set the parameter values as follows: $\alpha = 0.02$, $\theta \to 0$ and $\rho_{BP} = 0.9(B - \delta_{BP})$, and we modify ρ accordingly for HP, where δ is the conditional expected provider payoff, obtained from Proposition 1.13 for BP and can be obtained similarly for HP. Note that a ratio of greater than 1 indicates a higher risk borne by the provider under the BP mechanism than HP. In general we find that (a) the patient population size is highly dependent on the value of B under BP, and (b) the patient population size that guarantees a certain risk exposure under HP is a small fraction of the one needed for BP.

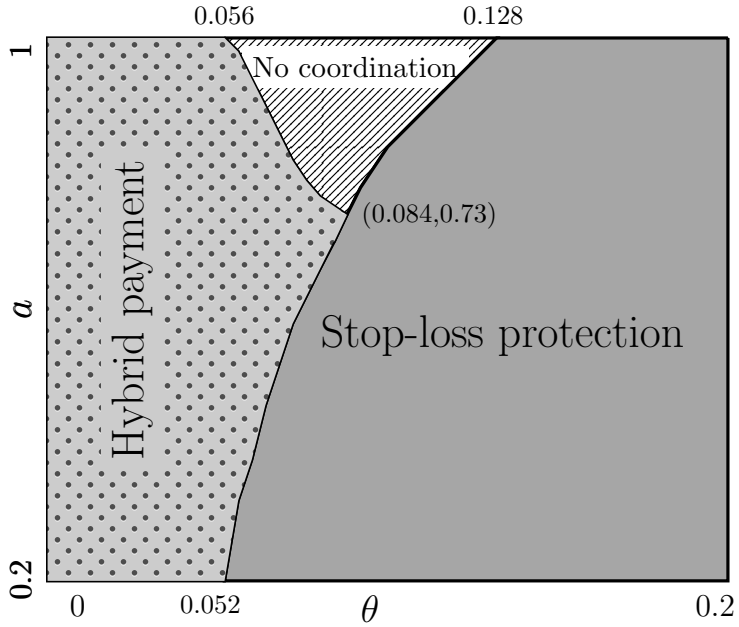


Figure 1.6: Coordination by stop-loss mechanism or hybrid payment scheme or none for $V = \$60,000$, $T^P = 0.02V$ and $\mu = 8,000$

In Table 1.3(b), we perform a similar comparison with the stop-loss (SL) payment mechanism. We notice from Figure 1.6 that the SL and HP payments cannot both coordinate for the same parameter values. Therefore, we choose a different parameter set of parameters: $\alpha = 0.4$ and $\theta = 0.2$, to assure that SL is a coordinating payment mechanism. Also, in order to make a fair comparison we limit the of range bundled payment values to an interval near the value of the coordinating bundled payment under SL. (Otherwise, a larger value of B would heavily benefit BP while the coordinating bundled payment value under SL is relatively small.) Under the selected set of parameters, $1 - \gamma$ for SL is 0.1225 thus we vary $1 - \gamma \in [0.06, 0.18]$ for BP. Based on Table 1.3(b), the stop-loss population size also is less than the BP population size needed to meet a certain level of risk exposure, given that the bundled payment under BP is comparable to that of SL. This confirms that SL helps lower the provider's risk exposure as compared with BP.

In conclusion, both the coordinating SL and HP perform much better than BP overall in

	$1-\gamma$					
	0.65	0.7	0.75	0.8	0.85	0.9
$\frac{N'(\text{BP})}{N'(\text{HP})}$	40.3	35.1	29.1	26.0	22.0	19.9

(a) $V = \$25,000$, $T^P = 0.01V$

	$1-\gamma$					
	0.06	0.08	0.1	0.12	0.14	0.16
$\frac{N'(\text{BP})}{N'(\text{SL})}$	18.7	7.4	3.8	2.3	1.5	1.1

(b) $V = \$45,000$, $T^P = 0.01V$

Table 1.3: Patient population size N needed for provider risk exposure of $\rho = 0.9(B - \delta)$

terms of provider risk exposure, and especially for smaller values of the BP bundle payment.

1.6.4 Summary of findings from the numerical experiments

The numerical experiments address the motivating questions formulated in the introduction and explore the differences in outcomes for the various payment mechanisms presented above. We present here the main conclusions.

Based on our extensive numerical experiments we make the following observations.

Observation 1: Performance of BP in terms of the system payoff and provider utility is highly dependent on the value of B and the resulting degree of patient selection, with larger values of B favoring the BP system. Furthermore, the BP system performs poorly in terms of risk imposed on the provider.

Observation 2: Higher risk aversion by the provider degrades the performance of both FFS and BP mechanisms. The system payoff and provider utility are significantly reduced for larger values of risk aversion under the BP system, mainly due to an increasing level of patient selection.

Observation 3: The hybrid and stop loss payment systems are particularly more effective at reducing the downside risk (patient population size) compared to BP when B is small.

Observation 4: We can find a coordinating mechanism (hybrid or stop-loss protection) for the majority of parameter values. For moderate risk aversion and high success probabilities no hybrid or stop-loss coordinating mechanism may be found. Otherwise, one of the proposed contracts coordinates. The hybrid payment system typically coordinates for lower risk aversions and the stop-loss mechanism coordinates for higher risk aversions.

1.7 Concluding Remarks

This paper is one of the first attempts at using a model-based approach for evaluating the performance of a variety of payment systems for healthcare services, including fee-for-service (FFS) and bundled payment (BP). The BP system is widely seen as a promising direction for reform due to its potential for re-aligning incentives, re-allocating risks and promoting quality of care over volume. While most experts agree that FFS is not a sustainable system, some point out that BP could have some unintended consequences which could negatively affect both patients and providers. Even though it is still too early to draw complete conclusions from the BPCI pilot program that CMS is currently running in selected facilities, our analysis sheds some light on some of the questions raised by proponents and opponents of this new type of payment system.

Our findings on FFS confirm the broad understanding that while FFS does not generally lead to patient selection and does not impose any risk on providers, it provides incentives for excessive treatment intensity and thus a high cost for the insurer. We find that the bundled payment system performance is extremely sensitive to the payment value for the bundle and the provider's risk aversion; practical implementation of the system should involve detailed evaluation of providers' risk attitudes and careful selection of the payment value. Depending on the provider's risk aversion level and other factors, BP could lead to suboptimal patient selection, treating more or less intensively than would be desirable for the system, a lower system payoff than FFS, and to an extremely high financial risk borne by the provider which increases the chance of bankruptcy and could lower the number of providers. This could

have seriously damaging long-term consequences such as reduced access to care, quality of care, and care availability.

Interestingly, we obtain that some fairly minor modifications of the bundled payment system could go a long way toward improving its performance and reducing its shortcomings. A combination of FFS and BP, so-called hybrid payment system in our paper, could markedly improve most performance measures – including provider utility, provider risk, and system payoff – without imposing significant implementation hurdles. A stop-loss mechanism could also improve the BP performance by spreading risks that are otherwise concentrated on the provider. In fact our results show that, when carefully designed, one of these two payment schemes can indeed fully coordinate the decisions of the provider with the system optimum in most, but not all, cases.

Our findings suggests that the current FFS system can be improved upon without sacrificing the quality of healthcare, but the proposed BP system should be very carefully implemented to avoid creating new issues, possibly using simple adjustments. Further research using empirical results from the BPCI pilot program should test whether these findings are confirmed by the data collected. A modeling approach can also be used to introduce more refined models specific to certain providers (home health agencies, inpatient rehabilitation facilities, nursing homes, etc.) and investigate whether the findings obtained are robust with respect to the type of care considered.

Chapter 2

REFERENCE PRICING FOR HEALTHCARE SERVICES

2.1 Introduction

The cost of a given medical procedure varies widely not just across the nation, but also across medical providers within the same geographic area [28, 82]. For example, the maximum price charged by a provider for a knee replacement in Atlanta, GA is over 6 times that charged by the lowest-priced provider. For an MRI in Columbus, OH, the ratio is 6.65 [28]. The price charged for a procedure does not generally reflect the quality of the care provided [81]. Rather, price variation results from variation in provider market power in their negotiations with insurers, extent of provider competition within a geographic area, type of facility offering the procedure, the lack of pricing transparency, and who is footing the bill – Medicare, Medicaid, private insurer or patient [28, 93].

The current payment system does little to incentivize patients to be price-conscious in their selection of a provider. Usually, patients pay out-of-pocket either a fixed co-payment on a procedure, or a co-insurance, that is, a fraction of the billed charges, subject to a maximum yearly out-of-pocket. For an expensive procedure, the co-insurance may exceed the yearly out-of-pocket and thus patients pay the same amount regardless of what the provider actually charges the insurer. The same is true in case of a fixed co-payment. When the patient pays a co-insurance and the procedure price is relatively low, so that the patient's share of the cost does not exceed the maximum yearly out-of-pocket, the patient pays a variable amount, proportional to the price charged, but the amount tends to be small and so less consequential for patients. Therefore, patients have limited incentives to select a less expensive provider, and as a result providers have no motivation to control the price they charge, and every incentive to increase prices.

To better align incentives and control rising healthcare costs, a different payment system called *reference pricing* (RP) has been proposed, involving cost-sharing with patients. Medical procedure price variations that are not linked to quality of service or patient health outcomes indicate that insurers may be able to reduce spending by incentivizing patients to use lower-priced providers [1]. Reference pricing has long been used for pharmacy benefits. The California public employees' retirement system (CalPERS) has recently applied it to inpatient knee and hip replacements[2] [118]. The idea of reference pricing is to set a "reference price" as the upper limit of charges to be reimbursed by the insurer. If a patient selects a provider charging the reference price or less (i.e., a "value-based" provider), she pays a co-payment or co-insurance as under the current payment system. However, if the patient selects a provider charging more than the reference price (i.e., a "non-value-based" provider), she has to pay the full portion of the charge above the reference price (not applicable towards a yearly maximum out-of-pocket), in addition to the co-payment or co-insurance from the portion below the reference price.

While reference pricing has some clear advantages in the way it aligns incentives, it may also have some unintended consequences. Proponents of this payment system argue that it both provides patients with incentives to make a price-conscious provider selection, and providers with incentives to reduce their prices if they want to maintain their market share with price-sensitive patients [91]. Hence, in the long-run reference pricing could help reduce the current trend of unwarranted rising prices for medical procedures. However, critics argue that making patients bear a larger share of the cost could reduce patient welfare, especially if the reference price is set low. In addition, the quality of care could decrease under reference pricing as a way for providers to lower their own costs and maintain a sufficient profit margin

[1] Under the current implementation of reference pricing only the hospitals' prices are considered and not the prices of the other healthcare providers.

[2] For a joint replacement surgery, the price charged is usually high enough so that for patients normally paying a co-insurance, the patient's cost share would exceed the maximum yearly out-of-pocket. Hence patients are subject to a fixed payment (their maximum yearly out-of-pocket) for this type of procedure in the current payment system.

when their prices are lowered [88].

Our goal is to analyze the reference pricing payment scheme and its effects on all agents involved – competing medical providers, insurer, and patients. We propose to answer the following research questions: (1) How do competing price-setting medical providers react to reference pricing, i.e., does reference pricing reduce prices overall? (2) How should the insurer set the reference price? (3) Are each of the agents considered (the medical providers, the insurer, the patients) and the entire system better or worse off under reference pricing relative to either a system where the patient pays a fixed amount (e.g., co-payment) or a variable amount (e.g., co-insurance), based on some performance measure?

We consider an insurer with a network including multiple differentiated competing medical providers who offer a given medical procedure, and a population of heterogeneous patients seeking treatment for this procedure. Medical providers set their prices, and patients then select a provider based on both monetary and non-monetary factors according to a discrete choice model. The insurer's payment system may impose on patients either a fixed co-payment or a variable co-insurance. In a *fixed payment* (FP) system, the patient is financially responsible for a fixed amount regardless of the provider selected, and the insurer reimburses the remainder. In a *variable payment* (VP) system, the patient pays a fraction of the price charged, while the insurer covers the rest. Under reference pricing, the insurer first sets the reference price; each medical provider then selects the price charged for the procedure. In particular, the price selected determines whether the provider is considered value-based or not.

This paper introduces a new model of reference pricing payment system that considers competition among differentiated medical providers. Our model incorporates heterogeneous patients influenced by monetary and non-monetary motives. We derive optimal pricing strategies for the providers and optimal provider selection for patients under a fixed payment system, a variable payment system, and a reference pricing payment. We also investigate the reference price selection made by the insurer. We compare the competing providers' pricing strategies and the utilities of the involved agents, as well as the patients' out-of-

pocket cost across the different payment schemes. We find that under certain conditions reference pricing results in higher provider and system-wide utility compared to variable and fixed payment models. We also show that the patient out-of-pocket is lower under reference pricing than under a variable payment. We find that the expected patient utility is higher under the reference pricing scheme while the insurer also gains higher utility under the reference pricing scheme for a large set of parameters. Moreover, as the cost of treatment decreases, the providers prefer reference pricing over a variable payment system for larger set of payment parameters. We observe that increasing the differentiation of the providers in terms of their attributes (availability of technology, accessibility, etc.) can affect the preference of providers regarding the payment models in different directions depending on the provider's attributes.

2.2 Literature Review

Four streams of research informed and inspired the present paper: research that studies healthcare payment systems; research that focuses on reference pricing in the pharmaceutical market; research that investigates reference pricing for healthcare services; and research that models discrete consumer choice.

The first stream of research evaluates how payment systems other than the traditional fee-for-service most commonly used presently in the US may perform. These new payment systems, often inspired by the Affordable Care Act, aim at realigning incentives to improve patient health outcomes and curb costs. Since the fee-for-service system is not sustainable, experts estimate that CMS should move towards alternative payment systems for at least 75% of their payments by 2022 [41]. A well-promoted new payment system is the bundled payment model under which a fixed lump sum payment is provided for a given episode of care regardless of the number of tests and procedures implemented and also regardless of possible complications. [3] study the bundled payment model and compare it with fee-for service. They show that bundled payment remove incentives to over treat but could generate suboptimal patient selection. both of these models could lead to sub optimal treatment and patient

selection decisions. They propose a hybrid payment system and a stop-loss mechanism as alternatives that can align the provider decisions to that of the system optimum. [55] look into the implementation of bundled payment and investigate how the bundles and payments should be selected and adopted by CMS. [9] compare fee-for-service, pay-for-performance and bundled payment when the chance of readmission is a result of both the patient and the provider's efforts. They show that fee-for-service fails to provide incentives for the provider and patients to exert readmission reduction effort, while pay-for-performance is generally more effective at reducing readmissions than bundled payments. [54] compare the readmission rate, wait time and patient welfare under a fee-for-service and bundled payment scheme considering a queuing perspective. They characterized the conditions under which one of the payment schemes can be dominated for the three performance measures.

The second stream of research focuses more specifically on the use of reference pricing as a payment system for pharmaceuticals. Reference pricing has been mostly used, and studied, as a price-control mechanism in the pharmaceutical market. [72] provide a review of the literature on reference pricing for pharmaceuticals. Reference pricing was first implemented in Germany in 1989, soon followed by other European countries as well as Australia, New Zealand, and British Columbia in the 1990s [17]. In this context, the reference pricing payment system may apply to clusters of drugs with the same active chemical ingredients, drugs with related active chemical ingredients that are pharmacologically equivalent, or more broadly to drugs with comparable therapeutic effects. These categories may include generic drugs, original drugs with and without patent protection. In the pharmaceutical market, reference pricing is believed to improve price competition, make demand more price-elastic, and hence reduce expenditures. However, critics argue that the inclusion of on-patent drugs in effect removes patent protection and could negatively impact research and development efforts by pharmaceutical firms. In addition, reference pricing for drugs that are not perfect substitutes (in terms of side effects and/or efficacy) introduces difficult trade-offs for patients who must choose between a lower out-of-pocket or a better-suited drug. [31] empirically study the effect that reference pricing has had in different countries on pharmaceutical innovation,

new compound availability, and price competition. [69] analytically investigate the problem of price competition in pharmaceuticals. They consider the problem of pharmacy benefit managers (PBMs) as a third-party group between drug manufacturers and the insurer. PBMs select the prices that should be charged to the insurers. [69] study tiered co-payment pricing for generic and original brands with and without patent protection, and they show that a unique Nash Equilibrium exist on the price decisions. [111], look into the pricing strategies and contracts for the pharmaceutical industry with tiered co-payment pricing for different groups of drugs. They analyze a general risk-sharing contract between the manufacturer and the insurer for new and innovative drugs, as well as coupon programs offered by original drug manufacturers to reduce the patient's out-of-pocket.

The third stream of research investigates the use of reference pricing for healthcare services. Reference pricing has been implemented for healthcare services in the US only recently, and on a very limited scale. In fact [88] calls it the "sleeper in healthcare payment reform" due to its limited application. He points out that the first proposal for implementing reference pricing in healthcare was by Prof. Alain Enthoven back in 1977 as consumer choice health plan [42], and it has been ignored until very recently. The literature focusing on analytically studying the effect of reference pricing for healthcare services is sparse. [36] propose an algorithm for a mixed-integer program to solve the insurer's problem of selecting value-based providers when the reference price is exogenous and the demand parameters are uncertain. The few experiments with reference pricing in healthcare have generated some empirical studies of the effect of reference pricing for medical procedures. The Safeway grocery store chain started to implement reference pricing for colonoscopy in 2009 for its employees [91] after observing a significant variation in pricing among different providers. It expanded its reference pricing program to laboratory tests and advance imaging. Lowe's Home Improvement chain also has adopted reference pricing for interventional cardiology and cardiac surgery [91]. The California Public Employees' Retirement System (CalPERS), with 1.3 million insured people is responsible for a major implementation of reference pricing for medical services. CalPERS started using reference pricing in 2011 for knee and hip

replacement [21]. They observed a 26.3% decrease in prices charged to CalPERS members from 2010 to 2011 without significant increase of the average consumer cost, accompanied by a significant shift of demand from non-value-based facilities towards value-based ones. The implementation resulted in $2.8 million in savings for CalPERS in one year and $0.3 million in lower cost sharing for CalPERS members, due both to market growth at value-based facilities and price reductions at non-value based facilities [21, 90]. [119] point out that if the application of reference pricing is expanded to other inpatient and ambulatory services, the potential savings could increase, since hip and knee surgery only account for 1.6% of total spendings. They specify potential candidates procedures suitable for reference pricing and they estimate roughly a 5% decrease in total expenditure after implementing reference pricing to these additional conditions. [47] examine empirically the potential savings from using reference pricing on a several types of health services that have fairly uniform protocols. While they find that implementing reference pricing for these medical procedures could save 1.6% of health care spending, they warn that the value of the reference price is critical for the success of the reference pricing scheme. Furthermore, savings could be significantly reduced if value-based providers currently charging below the reference price raise their price up to the reference price.

2.3 Modeling Framework

We now introduce our framework for modeling various payment structures, and establish measures to assess their performance from the lens of various stakeholders in our model. Consider one insurer and n competing and profit-maximizing providers who belong to the insurer's network and offer a medical procedure to a population of m patients covered by the insurer. The cost split between patients and the insurer to pay for the procedure depends on the specific payment scheme. Our oligopolist setting allows us to simultaneously examine the effect of competition among providers as well as the role of different payment mechanisms and their performance. Table 2.1 in the Appendix summarizes the notations used for the parameters and variables used in our model.

n	population size of the providers
m	population size of the patients
A_j	the non-idiosyncratic payoff that every patient receives when obtaining care at provider j
c	treatment cost of provider j
O_j	out-of-pocket cost for the patient obtaining care at provider j
η_{ij}	idiosyncratic attributes of the provider for provider j when visited by patient i
F	fixed amount that patients pay under the fixed payment model
$\overline{p}^K$	maximum price that providers can set under payment model $K = F, R, V$
p_j	price that provider j selects; has support $[0, \overline{p}]$
p^*	reference price value set by the insurer
γ	price sensitivity of patients
λ	portion of the payment that patient is responsible for
δ	fraction of the patient population average utility that affects the utility of the insurer
$S_j(P)$	probability that provider j is chosen by a randomly selected patient when providers price at P
$Q_j(P)$	provider j's market share when providers price at P
U_{ij}	utility that patient i gains when receiving care at provider j
V_j	utility that provider j gains when offering a procedure
W	utility of the insurer

Table 2.1: Notations

As mentioned earlier, in this paper we aim to inform policy discussion by analyzing each payment mechanism from the lens of patients, providers and insurer. Therefore, we next study performance measures for each of these stakeholders.

2.3.1 Patients

Patients are heterogeneous in terms of their provider preferences. They select a provider according to a multinomial logit model. They also have the option of not seeking treatment. The patient utility gained when selecting a certain provider results from monetary and non-monetary factors. We assume that all patients have the same price sensitivity γ. We denote

O_j to be the out-of-pocket cost for the patient, when obtaining care from provider j. We denote A_j the non-idiosyncratic payoff that every patient receives when obtaining care at provider j, exclusive of price considerations. A_j captures general attributes such as comfort level, availability of advanced technologies, quality and quantity of staff, etc. We denote η_{ij} the patient-specific payoff that patient i receives when obtaining care from provider j, exclusive of price considerations. η_{ij} captures idiosyncratic attributes such as distance to the patient's residence, ease of access from the patient's residence, patient's familiarity with the facility and/or doctor performing the procedure, etc. Hence the utility U_{ij} that patient i gains when receiving care at provider j can be expressed as

$$U_{ij} = A_j - \gamma O_j + \eta_{ij}, \quad i = 1, \ldots, m, \ j = 1, \ldots, n, \tag{2.1}$$

and the utility of patient i from not seeking treatment is

$$U_{i0} = u_0 + \eta_{i0}.$$

Utility of not seeking treatment has a fixed component u_0, as well as an idiosyncratic error term. Note that O_j may depend on the price charged by provider j and on the payment system. We assume that A_j is fixed and known, and that the η_{ij}s are independent identically distributed, following a standardized Gumbel (or type-I extreme value) distribution with cumulative distribution function form $exp(-exp(-\eta))$ and mean and variance of 0.5772 (Euler constant) and $\pi^2/6$ respectively. This form of distribution for error terms results in a multinomial logit (MNL) choice model for patients when making a selection across different providers.[3]

Thus, the probability that a randomly selected patient chooses provider j is given by

$$S_j(P) = \frac{e^{A_j - \gamma O_j}}{e^{u_0} + \sum_{k=1}^{n} e^{A_k - \gamma O_k}} \in (0, 1], \quad j = 1, \ldots, n,$$

[3]For details of derivation see section 3.10 of [110].

where $P = (p_1, \ldots, p_n)$ denotes the vector of provider prices and the probability of seeking no treatment is

$$S_0 = 1 - \sum_{j=1}^{n} S_j.$$

Notice that O_j may depend on price p_j but we suppress the notation for ease of exposition.

The overall expected patients utility is

$$E[U] = m\left(\sum_{j=1}^{n} U_{ij} S_j(P) + U_{i0} S_0\right). \tag{2.2}$$

Note that the error terms have a constant mean, and thus we can drop the error term from Equation (2.2) without loss of generality.

2.3.2 Providers

There are n competing providers in the insurer network. We assume providers incurs treatment cost c for the procedure [4] [5]. Provider j selects the price p_j it wishes to charge for the medical procedure (for $j = 1, \ldots, n$). The provider selects its price so as to maximize the utility it gains from offering the procedure, given by

$$V_j(P) = (p_j - c)Q_j(P), \quad j = 1, \ldots, n,$$

where $Q_j(P)$ is provider j's market share for the procedure, given by

$$Q_j(P) = mS_j(P) = m\frac{e^{A_j - \gamma O_j}}{e^{u_0} + \sum_{k=1}^{n} e^{A_k - \gamma O_k}}, \quad j = 1, \ldots, n. \tag{2.3}$$

Note that providers always select a price that is larger than their cost ($p_j > c$) to ensure a positive profit.

---

[4] Reference pricing is mainly implemented for the standard conditions that should result in similar cost across different hospitals such as hip and knee replacement, MRI, colonoscopy, etc.

[5] Most of the results of this paper hold true for unequal costs of treatment across different providers if ordering providers based on their non-price attributes result in the same ordering for treatment cost ($A_1 \leq \ldots \leq A_n \rightarrow c_1 \leq \ldots \leq c_n$). Recall that parameter A_j captures attributes such as comfort level, quality of care, etc. that are not specific to a given patient. A provider able to improve such characteristics typically incurs more costs.

2.3.3 Insurer

The insurer sets the parameters of the payment model to maximize its utility. We assume that the insurer utility stems from financial considerations, namely the cost of covering the population for a procedure, as well as from its mission to maintain the health of its beneficiaries. A similar modeling framework of insurer utility has been used in the healthcare operations and health economics literature, including [122], [9], [112], and [111]. Hence we model the insurer utility as a combination of the expected patient population utility and the insurer's direct cost of coverage. Thus, the utility of the insurer is given by

$$W = \delta E[U] - \sum_{j=1}^{n} (p_j - O_j) Q_j(P),$$

where $E[U]$ is the expected utility of the patient population (which depends in particular on the payment system).

2.4 Payment Systems

We analyze three types of payment systems. First, we consider a fixed payment system where the patient pays a fixed amount (F) when undergoing the medical procedure. This situation is closest to the current payment system for many cases. It occurs in practice when the patient is subject to a fixed co-payment. It may also occur when the patient is subject to a co-insurance with a yearly maximum out-of-pocket and the range of prices for the procedure is high enough to cause the patient to meet the maximum out-of-pocket regardless of the provider selection (e.g., joint replacement surgery). Second, we examine the reference pricing scheme, where the patient pays a fixed amount and, if selecting a provider charging above the reference price, also pays the entire portion above the reference price in addition. Third, we investigate a co-insurance mechanism without maximum out-of-pocket [6] or what we call

[6] In practice, this situation occurs when the patient is subject to a co-insurance with or without maximum out-of-pocket, as long as the out-of-pocket maximum is large compared to the procedure prices and the patient is far from meeting this yearly maximum, so the patient actual cost is proportional to the price charged.

variable payment, where the patient is responsible for a given fraction of the price charged by the provider.

2.4.1 Fixed payment

We start by analyzing the decisions under the payment system where the patient pays a fixed amount $F \in [0, \overline{p}^{FP}]$ for the procedure regardless of the provider selected, where $\overline{p}^{FP}$ is the maximum price that the providers can charge for a given condition.

Proposition 2.1 *Under a fixed payment, providers always charge the highest possible price* $\overline{p}^{FP}$.

Proof. Under a fixed payment, when using provider j, the patient out-of-pocket $O_j = F$ is independent of p_j. Therefore, each provider's market share is also independent of p_j. It follows that the provider utility is monotonically increasing in p_j, hence the optimal price is $p_j = \overline{p}^{FP}$. □

The result of Proposition 2.1 stems from the fact that when the patient pays a fixed amount irrespective of the provider selected and the price that the provider charges, the price has no adverse effect on the provider's market share. Since the insurer covers the remainder charges not paid by the patient, high prices result in higher income with no downside, thus providers have no reason to limit the price they charge. This situation is an example of moral hazard, where the patients make decisions without having to bear the consequences of these decisions. It clearly illustrates the issue of patients not having any "skin in the game" which misaligns incentives and leads to rising prices. This situation motivates the need for a modified payment system where providers would have incentives to control their prices.

We next derive the utilities of the patients, providers and insurer. The fraction of patients who select provider j is

$$S_j^{FP} = \frac{e^{A_j - \gamma F}}{e^{u_0} + \sum_{k=1}^{n} e^{A_k - \gamma F}}, \quad j = 1, \ldots, n, \tag{2.4}$$

and the proportion of patients electing not to obtain the procedure is

$$S_0^{FP} = 1 - \sum_{j=1}^{n} S_j^{FP} = \frac{e^{u_0}}{e^{u_0} + \sum_{k=1}^{n} e^{A_k - \gamma F}}.$$

Thus, provider j's market share is $Q_j^{FP} = mS_j^{FP}, \quad j = 1, \ldots, n$. From Equation 2.4 it is clear that the providers with better non-price attributes, as quantified by A_j for provider j, attract more market share.

Using Equation (2.2), we next evaluate the patients overall expected utility as

$$E[U^{FP}] = m \sum_{j=1}^{n} (A_j - \gamma F) S_j^{FP} + m u_0 S_0^{FP}.$$

Provider j's utility is then given by

$$V_j^{FP} = (\overline{p}^{FP} - c) Q_j^{FP}, \quad j = 1, \ldots, n,$$

and the insurer's utility is

$$\begin{aligned} W^{FP} &= m\delta \left(\sum_{j=1}^{n} (A_j - \gamma F) S_j^{FP} + u_0 S_0^{FP} \right) - m \sum_{j=1}^{n} (\overline{p}^{FP} - F) S_j^{FP} \\ &= m\delta \left(\sum_{j=1}^{n} A_j S_j^{FP} - \gamma F (1 - S_0^{FP}) + u_0 S_0^{FP} \right) - m(\overline{p}^{FP} - F)(1 - S_0^{FP}). \end{aligned} \tag{2.5}$$

Proposition 2.2 shows how the insurer decides on the patients' cost share F.

Proposition 2.2 *Under the fixed payment model, if $\gamma\delta \geq 1$, the optimal value of the patients' fixed payment (F^*) is either 0 or $\overline{p}^{FP}$, such that*

$$F^* = \begin{cases} 0 & \text{if } \tau \leq \frac{\overline{p}^{FP}}{\delta}; \\ \overline{p}^{FP} & \text{Otherwise.} \end{cases}$$

where $\tau = \frac{\sum_{j=1}^{n} A_j \Delta S_j^{FP} + u_0 \Delta S_0^{FP}}{\gamma\delta(1 - S_0^{FP}(\overline{p}^{FP})) - (1 - S_0^{FP}(0))}$, $\Delta S_j^{FP} = S_j^{FP}(\overline{p}^{FP}) - S_j^{FP}(0)$, and $\Delta S_0^{FP} = S_0^{FP}(\overline{p}^{FP}) - S_0^{FP}(0)$.

If $\gamma\delta < 1$ and $\left(\delta u_0 + \overline{p}^{FP}\right)\left(1 - \frac{e^{u_0}}{e^{u_0} + \sum_{k=1}^{n} e^{A_k}}\right) < \frac{1 - \gamma\delta}{\gamma}$ then

$$F^* = \begin{cases} \hat{F} & \text{if } 0 \leq \hat{F} \leq \overline{p}^{FP}; \\ 0 & \text{if } \hat{F} < 0; \\ \overline{p} & \text{if } \hat{F} > \overline{p}^{FP}, \end{cases}$$

where $\hat{F}$ is the unique solution of the equation

$$\gamma\delta S_0^{FP}(F) \sum_{j=1}^{n} A_j S_j^{FP}(F) = (1 - S_0^{FP}(F)) \left(1 - \gamma\delta + \gamma S_0^{FP}(F) \left(\delta\gamma F + \delta u_0 + \overline{p}^{FP} - F\right)\right).$$

If $\gamma\delta < 1$ and $\left(\delta u_0 + \overline{p}^{FP}\right) \left(1 - \frac{e^{u_0}}{e^{u_0} + \sum_{k=1}^{n} e^{A_k}}\right) \geq \frac{1-\gamma\delta}{\gamma}$,

$$F^* = \underset{F}{Argmax}\{W^{FP}(F = 0), W^{FP}(F = \overline{p}^{FP}), W^{FP}(F = \hat{F})\}.$$

Proof. First notice that as the value of fixed payment (F) increases, the probability of seeking treatment becomes lower such that for large values of F, insurer's utility converges to $m\delta u_0$. From Equation 2.5 we have

$$\frac{\partial W^{FP}}{\partial F} = m \left(-\gamma\delta S_0^{FP} \sum_{j=1}^{n} A_j S_j^{FP} + (1 - S_0^{FP}) \left(1 - \gamma\delta + \underbrace{\gamma S_0^{FP} \left(\delta\gamma F + \delta u_0 + \overline{p}^{FP} - F\right)}_{B} \right) \right).$$

From the expression above,

$$\frac{\partial B}{\partial F} = \gamma S_0^{FP} \left(\gamma \left(\gamma\delta F + \delta u_0 + \overline{p}^{FP} - F\right) + \gamma\delta - 1\right).$$

This expression confirms that for $\gamma\delta \geq 1$, B is increasing in F and as the result the sign of the first derivative of W^{FP} can only change from negative to positive. This indicates that W^{FP} is quasi-convex in F. Thus, the optimal value of F falls on the boundaries. In other words if $\gamma\lambda \geq 1$, we have:

$$F^* = \begin{cases} 0 & \text{if } \tau \leq \frac{\overline{p}^{FP}}{\delta}; \\ \overline{p}^{FP} & \text{Otherwise,} \end{cases}$$

where $\tau = \frac{\sum_{j=1}^{n} A_j \Delta S_j^{FP} + u_0 \Delta S_0^{FP}}{\gamma\delta(1-S_0^{FP}(\overline{p}^{FP}))-(1-S_0^{FP}(0))}$.

Note that $\Delta S_j^{FP} = S_j^{FP}(\overline{p}^{FP}) - S_j^{FP}(0)$, and $\Delta S_0^{FP} = S_0^{FP}(\overline{p}^{FP}) - S_0^{FP}(0)$.

If $\gamma\delta < 1$ and $(\overline{p}^{FP} + \delta u_0)(1 - S_0^{FP}(0)) \leq \frac{1-\gamma\lambda}{\gamma}$, then it is easy to see that $\frac{\partial B}{\partial F} \leq 0$. This means that W^{FP} can only change sign from positive to negative resulting in W^{FP} being quasi-concave. In this case F^* is

$$F^* = \begin{cases} \hat{F} & \text{if } 0 \leq \hat{F} \leq \overline{p}^{FP}; \\ 0 & \text{if } \hat{F} < 0; \\ \overline{p} & \text{if } \hat{F} > \overline{p}^{FP}, \end{cases}$$

where $\hat{F}$ is the unique solution of the equation

$$\gamma\delta S_0^{FP}(F) \sum_{j=1}^{n} A_j S_j^{FP}(F) = (1 - S_0^{FP}(F)) \left(1 - \gamma\delta + \gamma S_0^{FP}(F) \left(\delta\gamma F + \delta u_0 + \overline{p}^{FP} - F\right)\right). \tag{2.6}$$

If $\gamma\delta < 1$ and $(\overline{p}+\delta u_0)(1-S_0^{FP}(0)) > \frac{1-\gamma\lambda}{\gamma}$ then utility of the insurer can be quasi-concave or convex in F and

$$F^* = \underset{F}{Argmax}\{W^{FP}(F=0), W^{FP}(F=\overline{p}^{FP}), W^{FP}(F=\hat{F})\}$$

□

Proposition 2.2 describes situations where the patient fixed payment is on the boundary (0 or $\overline{p}^{FP}$) or it has an intermediate value.

2.4.2 Reference pricing

In the reference pricing payment system, the insurer first sets the reference price p^*. Providers then select the price they charge for the procedure, $p_j,\ j = 1, \ldots, n$. If $p_j \leq p^*$, provider j is a "value-based" provider, otherwise it is a "non-value-based" provider. Finally, patients select the provider according to the discrete choice model described in Section 2.3.1. We proceed by backwards induction, considering first the patients' choice of a provider, then the provider's price selection.

We first analyze the patients' choice of a provider, for a given set of provider prices and a given reference price. Since the patients all pay the same fixed co-payment regardless of the provider selected (in addition to other costs when choosing a non-value-based provider), this fixed co-payment has no effect on their choice of provider. Thus, without loss of generality, we set the fixed co-payment to zero. Under reference pricing, the patient's out-of-pocket cost when selecting provider j is thus $O_j = (p_j - p^*)^+ = \max\{0, p_j - p^*\}$. Hence, using (2.3), the patients choose their providers in such a way that provider j's proportion of patients is

$$S_j^{RP}(P) = \frac{e^{A_j - \gamma(p_j - p^*)^+}}{e^{u_0} + \sum_{k=1}^{n} e^{A_k - \gamma(p_k - p^*)^+}}, \quad j = 1, \ldots, n, \tag{2.7}$$

the proportion of patients electing not to obtain the procedure is

$$S_0^{RP}(P) = \frac{e^{u_0}}{e^{u_0} + \sum_{k=1}^{n} e^{A_k - \gamma(p_k - p^*)^+}},$$

and provider j's market share is $Q_j^{RP}(P) = mS_j^{RP}(P), \quad j = 1, \ldots, n$.

We next evaluate the patients overall expected utility as

$$E[U^{RP}] = m \sum_{j=1}^{n} (A_j - \gamma(p_j - p^*)^+) S_j^{RP}(P) + mu_0 S_0^{RP}(P).$$

Each provider selects its price via Nash competition to maximize its utility given by

$$V_j^{RP}(P) = (p_j - c) m S_j^{RP}(P), \quad j = 1, \ldots, n,$$

while anticipating the patients' reaction through $S_j^{RP}(P)$. The insurer's utility is

$$W^{RP} = \delta E[U^{RP}] - \sum_{j=1}^{n} p^* Q_j^{RP}(P)$$
$$= \delta E[U^{RP}] - mp^*(1 - S_0^{RP}(P)).$$

Before proceeding we formalize a condition on the upper bound of prices. We later show that this condition guarantees the existence of a pure Nash equilibrium within the analysis of reference pricing and variable payment models. This condition is similar to that introduced in [8].

Assumption 2.1 *Let $\overline{p}_j^K$ be the best response of provider j in payment model $K = RP, VP$, when all competing providers price at p^{max} tends to ∞. Then, provider j captures less than 50% of the market share when it prices at $\overline{p}_j^K$ while the competing providers price at p^{max}.*

This condition resembles the result in classic economics literature with a monopoly firm and a linear price-demand function where the monopoly firm's optimal market share is less than 50% . We note that when all competing providers set their prices at ∞, then provider j would effectively be the only provider in the market and gain its maximum market share. Note that with this assumption in place, the market share of provider j is always less than 50% when pricing at $\overline{p}_j^K$, regardless of the competing providers' prices. Later we prove that $\overline{p}_j^K$ can be used as the upper bound on the equilibrium prices.

We now derive the following technical lemma.

Lemma 2.1 *$S_j^{RP}(P)$ is continuous in p_j. Its partial derivative with respect to p_j is continuous everywhere except at p^*, and we have*

$$\frac{\partial S_j^{RP}}{\partial p_j} = \begin{cases} 0 & \text{if } p_j < p^* \\ -\gamma(1 - S_j^{RP}(P))S_j^{RP}(P) & \text{if } p_j > p^*. \end{cases}$$

Proof. Let $\varphi(p_j) = e^{A_j - \gamma(p_j - p^*)^+}$ and $a = e^{u_0} + \sum_{k \neq j} e^{A_k - \gamma(p_k - p^*)^+}$, where a is independent

of p_j. Then $S_j^{RP}(P) = \varphi(p_j)/(a + \varphi(p_j))$. Moreover,

$$\varphi'(p_j) = \begin{cases} 0 & \text{if } p_j < p^* \\ -\gamma\varphi(p_j) & \text{if } p_j > p^*. \end{cases}$$

Therefore,

$$\begin{aligned} \frac{\partial S_j^{RP}}{\partial p_j} &= \frac{\varphi'(p_j)(a + \varphi(p_j)) - \varphi'(p_j)\varphi(p_j)}{(a + \varphi(p_j))^2} = \frac{a\varphi'(p_j)}{(a + \varphi(p_j))^2} \\ &= \begin{cases} 0 & \text{if } p_j < p^* \\ \frac{-\gamma a\varphi(p_j)}{(a+\varphi(p_j))^2} & \text{if } p_j > p^* \end{cases} \\ &= \begin{cases} 0 & \text{if } p_j < p^* \\ -\gamma[1 - S_j^{RP}(P)]S_j^{RP}(P) & \text{if } p_j > p^*. \end{cases} \end{aligned}$$

This concludes the proof of Lemma 2.1. □

As described in Lemma 2.1, for a non-value-based provider, a price increase leads to a decrease in market share, while the market share of a value-based provider is not sensitive to its price choice. Intuitively, for a value-based provider patients are not responsible for any out-of-pocket and as a result are indifferent to the price charged by the provider. On the other hand, an increase in price for a non-value-based provider results in higher patient out-of-pocket which disincentivizes patients from selecting the provider.

We show in Lemma 2.2 that the best response of provider j to any feasible pricing strategy of the competing providers is bounded by $\overline{p}_j^{RP}$.

Lemma 2.2 *In the RP payment system, under Assumption 2.1, the best response price of provider $j \in N$ to any competing provider prices has an upper bound of $\overline{p}_j^{RP}$.*

Proof. We start the proof by showing how Assumption 2.1 is implemented for the case of reference pricing. Using Lemma 2.1, we have

$$\frac{\partial V_j^{RP}}{\partial p_j} = mS_j^{RP}(P) + m(p_j - c)\frac{\partial S_j^{RP}}{\partial p_j}$$
$$= \begin{cases} mS_j^{RP}(P) > 0 & \text{if } p_j < p^* \\ mS_j^{RP}(P)[1 - \gamma(p_j - c)(1 - S_j^{RP}(P))] & \text{if } p_j > p^*. \end{cases} \tag{2.8}$$

The best response of provider j when the competing providers price at p^{max} can then be computed by solving the following equation for p_j.

$$\begin{cases} p_j = p^* & \text{if } \tau < 0 \\ 1 - \gamma(p_j - c)(1 - S_j^{RP}(p_j, P_{-j} = p^{max})) = 0 & \text{if } \tau \geq 0, \end{cases} \tag{2.9}$$

where $\tau = \frac{\partial V_j^{RP}}{\partial p_j}\Big|_{p_j=(p^*)^+, P_{-j}=p^{max}} = mS_j^{RP}(p_j = p^*, P_{-j} = p^{max})$ $\left[1 - \gamma(p^* - c)\left(1 - S_j^{RP}(p_j = p^*, P_{-j} = p^{max})\right)\right]$ and P_{-j} is the vector of prices for all the providers excluding provider j. Note that the sign of τ determines a value-based or non-value-based provider when $P_{-j} = p^{max}$. Using the expression for S_j^{RP} from (2.7) for $\tau \geq 0$, the best response of the providers j when the competing provider price at $p^{max} \to \infty$ solves for

$$\frac{e^{A_j - \gamma(\overline{p}_j^{RP} - p^*)}}{\gamma(\overline{p}_j^{RP} - c) - 1} = e^{u_0}.$$

Notice that as p^{max} increases, the market share of provider j increases at a given price. Thus, provider j attracts the maximum market share when $p^{max} \to \infty$. Note that if provider j is value-based, then $\overline{p}_j^{RP} = p^*$.

The statement of Assumption 2.1 requiring the market share of provider j when competing providers price at p^{max} to be less than or equal to 0.5, can then be written in mathematical form for a non-value-based provider,

$$\frac{e^{A_j - \gamma(\overline{p}_j^{RP} - p^*)}}{e^{u_0} + \sum_{i \neq j} e^{A_i - \gamma(p^{max} - p^*)} + e^{A_j - \gamma(\overline{p}_j^{RP} - p^*)}} \leq 0.5 \xrightarrow{p^{max} \to \infty} A_j - \gamma(\overline{p}_j^{RP} - p^*) \leq u_0$$

where the second inequality is achieved as $p^{max} \to \infty$.

From Equation (2.9) we know that under reference pricing the best response of a value-based provider is p^* and for $p_j > p^*$

$$\frac{\partial V_j^{RP}}{\partial p_j} = mS_j^{RP}(P)[1 - \gamma(p_j - c)(1 - S_j^{RP}(P))].$$

By Assumption 2.1, if $p_j > \overline{p}_j^{RP}$, $1 - S_j^{RP}(P) > 0.5$ for any vector of prices, since the market share of provider j is maximized when the competing providers price at maximum possible price level p^{max}, which itself is less than 0.5 by Assumption 2.1. Also the first order condition of provider j sets $\gamma(\overline{p}_j^{RP} - c) = \frac{1}{1 - S_j^{RP}(p_j = \overline{p}_j^{RP}, P_{-j} = p^{max})} \leq 2$. Thus, if $p_j > \overline{p}_j^{RP}$, $\gamma(p_j - c) > 2$. We can then conclude that for a $p_j > \overline{p}_j^{RP}$, $\frac{\partial V_j^{RP}}{\partial p_j} < 0$ and the best response of the non-value-based provider cannot exceed $\overline{p}_j^{RP}$ bound. □

Lemma 2.2 shows that providers only become worse off if they price above $\overline{p}_j^{RP}$. This property constructs a closed action set for providers' prices, which later helps us prove the existence of a unique Nash equilibrium.

We now focus on the providers' choice of price, for a given reference price p^*. We assume that all the providers make their pricing decisions simultaneously in a Nash game. Note that the provider's price determines whether it is a value-based provider (if $p_j \leq p^*$) or a non-value-based provider (if $p_j > p^*$).

Proposition 2.3 *If either (i) $p^* \leq c$, or (ii) $p^* > c$ and*

$$\frac{e^{A_j}}{e^{u_0} + \sum_{k=1}^{n} e^{A_k}} + \frac{1}{\gamma(p^* - c)} > 1, \tag{2.10}$$

then provider j chooses to be non-value-based.

Proof. We start with some preliminary results. Using (2.8)

$$\left.\frac{\partial V_j^{RP}}{\partial p_j}\right|_{p_j = (p^*)^+} = mS_j^{RP}(p_j = p^*, P_{-j})[1 - \gamma(p^* - c)(1 - S_j^{RP}(p_j = p^*, P_{-j}))], \tag{2.11}$$

where P_{-j} represents the vector of prices of all providers except provider j, and the equality follows from Lemma 2.1.

Moreover, since $e^{u_0} + \sum_{k \neq j} e^{A_k - \gamma(p_k - p^*)^+} \leq e^{u_0} + \sum_{k \neq j} e^{A_k}$, and since the function $x/(x + e^{A_j})$ is monotonically increasing in x, we have

$$\frac{e^{u_0} + \sum_{k \neq j} e^{A_k - \gamma(p_k - p^*)^+}}{e^{u_0} + e^{A_j} + \sum_{k \neq j} e^{A_k - \gamma(p_k - p^*)^+}} \leq \frac{e^{u_0} + \sum_{k \neq j} e^{A_k}}{e^{u_0} + e^{A_j} + \sum_{k \neq j} e^{A_k}}.$$

Observing that the left-hand-side above equals $1 - S_j^{RP}(p_j = p^*, P_{-j})$, we obtain

$$1 - S_j^{RP}(p_j = p^*, P_{-j}) \leq \frac{e^{u_0} + \sum_{k \neq j} e^{A_k}}{e^{u_0} + \sum_{k=1}^{n} e^{A_k}} = 1 - \frac{e^{A_j}}{e^{u_0} + \sum_{k=1}^{n} e^{A_k}}. \tag{2.12}$$

We now prove the statement of the proposition.

If $p^* \leq c$, since $S_j^{RP} \in (0, 1]$ it is straightforward that expression (2.11) is positive. If $p^* > c$ and condition (2.10) holds, then inequality (2.12) implies that $1 - S_j^{RP}(p_j = p^*, P_{-j}) < 1/(\gamma(p^* - c))$, and thus expression (2.11) is positive. Thus $\frac{\partial V_j^{RP}}{\partial p_j}\Big|_{p_j = (p^*)^+} > 0$. In addition, we know from (2.8) that $\frac{\partial V_j^{RP}}{\partial p_j} > 0$ for $p_j < p^*$. Therefore, provider j can gain higher utility with $p_j > p^*$ than with $p_j \leq p^*$. As a result provider j chooses to be non-value-based. □

We now show a property of value-based providers.

Proposition 2.4 *A value-based provider prices at the reference price, p^*.*

Proof. It can be seen from Equation (2.8) that provider j's utility is monotonically increasing in p_j over the domain $p_j \leq p^*$. □

This result is consistent with our findings in the case of a fixed payment. When the patient chooses a value-based provider, she pays the same fixed co-payment (which we have assumed to be zero) regardless of the value-based provider selected. Hence, the actual price charged by the provider (up to the reference price) has no effect on patient choice and thus on the provider's market share. Yet, the price has an effect on the provider's revenue, thus value-based providers prefer setting their prices at the maximum level, that is, at the reference price.

The following result explains how a given set of non-value-based providers jointly determine their prices at the Nash equilibrium.

Theorem 2.1 *Consider as given the set of non-value-based providers, $\mathcal{N}$. At equilibrium, the prices of the non-value-based providers are the unique solution of the system of first-order condition equations:*

$$1 - \gamma(p_j - c)(1 - S_j^{RP}(P)) = 0, \quad \forall j \in \mathcal{N}, \qquad p_i = p^* \quad \forall i \notin \mathcal{N}.$$

Proof.

As shown from (2.8) in the proof of Proposition 2.3, on the domain $p_j > p^*$, we have

$$\frac{\partial V_j^{RP}}{\partial p_j} = m S_j^{RP}(P)[1 - \gamma(p_j - c)(1 - S_j^{RP}(P))]. \tag{2.13}$$

Moreover, using Lemma 2.1, we find after some simplifications, on the domain $p_j > p^*$,

$$\frac{\partial^2 V_j^{RP}}{\partial p_j^2} = -m\gamma S_j^{RP}(P)[1 - S_j^{RP}(P)] \left[2 + \gamma(p_j - c)(2S_j^{RP}(P) - 1)\right]. \tag{2.14}$$

Using (2.7), on the domain $p_j > p^*$, we have

$$\begin{aligned} 2 + \gamma(p_j - c)(2S_j^{RP}(P) - 1) &= 2 + \gamma(p_j - c)\frac{e^{A_j - \gamma(p_j - p^*)} - e^{u_0} - \sum_{k \neq j} e^{A_k - \gamma(p_k - p^*)^+}}{e^{u_0} + \sum_{k=1}^n e^{A_k - \gamma(p_k - p^*)^+}} \\ &= \frac{[2 - \gamma(p_j - c)](e^{u_0} + \sum_{k \neq j} e^{A_k - \gamma(p_k - p^*)^+}) + [2 + \gamma(p_j - c)]e^{A_j - \gamma(p_j - p^*)}}{e^{u_0} + \sum_{k=1}^n e^{A_k - \gamma(p_k - p^*)^+}}. \end{aligned}$$

The denominator of the expression above is clearly always positive. The partial derivative with respect to p_j of the numerator is

$$-\gamma(e^{u_0} + \sum_{k \neq j} e^{A_k - \gamma(p_k - p^*)^+}) - \gamma(1 + \gamma(p_j - c))e^{A_j - \gamma(p_j - p^*)},$$

which is negative since the price selected by provider j must exceed the cost. Therefore, expression $2 + \gamma(p_j - c)(2S_j^{RP}(P) - 1)$ can only change sign at most once, from positive to negative, as p_j gets larger. Hence, $\frac{\partial^2 V_j^{RP}}{\partial p_j^2}$ can only change sign at most once, from negative to positive, as p_j gets larger. Moreover, as p_j tends to infinity, the provider utility V_j^{RP}

tends to zero as a result of a disappearing market share. Thus the provider utility V_j^{RP} is quasi-concave in p_j on the domain $p_j > p^*$. In particular, a non-value-based provider j finds its optimal price (assuming all other prices are fixed) by solving for the first order condition:

$$1 - \gamma(p_j - c)(1 - S_j^{RP}(P)) = 0. \tag{2.15}$$

Note that equation above results in a p_j such that $\gamma(p_j - c) > 1$, since $1 - S_j^{RP}(P) < 1$. Moreover, from Lemma 2.2, $p_j \leq \overline{p}_j^{RP}$.

To show the existence of such equilibrium we utilize Debreu Fan Glicksberg theorem based on which, if for all players (providers) the set of actions is a non-empty, convex, and compact set (in our case $P_j \in [p^*, \overline{p}_j^{RP}]\ \forall\ j \in NVB$),

and utility of players is a continuous function which is quasi-concave on the action set; there exists a Nash Equilibrium for the game [33, 44, 52].

Based on [115], to show uniqueness we can show

$$\left|\frac{\partial^2 V_j}{\partial^2 p_j}\right| > \sum_{\substack{i \neq j \\ i \in N}} \left|\frac{\partial^2 V_j}{\partial p_j \partial p_i}\right|, \quad \forall j \in N.$$

Using Equation (2.14), for $j \in N$, we have

$$\left|\frac{\partial^2 V_j}{\partial p_j^2}\right| = m\gamma S_j^{RP}(P)[1 - S_j^{RP}(P)]|2 + \gamma(p_j - c)(2S_j^{RP}(P) - 1)|.$$

Moreover, using (2.13), we obtain that for $i \neq j,\ i, j \in N$,

$$\frac{\partial^2 V_j}{\partial p_j \partial p_i} = m\frac{\partial S_j^{RP}}{\partial p_i}[1 + \gamma(p_j - c)(2S_j^{RP}(p) - 1)].$$

Using (2.7) we find

$$\frac{\partial S_j^{RP}}{\partial p_i} = \begin{cases} 0 & \text{if } p_i < p^* \\ \gamma S_j^{RP}(P) S_i^{RP}(P) & \text{if } p_i > p^*. \end{cases}$$

Thus,

$$\sum_{\substack{i\neq j\\ i\in N}} \left|\frac{\partial^2 V_j}{\partial p_j \partial p_i}\right| = \sum_{\substack{i\neq j\\ i\in N}} m\gamma S_j^{RP}(P) S_i^{RP}(P) |1+\gamma(p_j-c)(2S_j^{RP}(P)-1)|$$

$$= m\gamma S_j^{RP}(P) |1+\gamma(p_j-c)(2S_j^{RP}(P)-1)| \sum_{\substack{i\neq j\\ i\in N}} S_i^{RP}(P)$$

$$\leq m\gamma S_j^{RP}(P) |1+\gamma(p_j-c)(2S_j^{RP}(P)-1)| (1-S_J^{RP}(P)),$$

where the last inequality follows from observing that $S_J^{RP}(P) + \sum_{\substack{i\neq j\\ i\in N}} S_i^{RP}(P) \leq 1$. Thus, to prove uniqueness, it remains to show that

$$|2+\gamma(p_j-c)(2S_j^{RP}(P)-1)| > |1+\gamma(p_j-c)(2S_j^{RP}(P)-1)|,$$

which holds as long as $1+\gamma(p_j-c)(2S_j^{RP}(P)-1) > -1/2$.

Using (2.15), we find that

$$1+\gamma(p_j-c)(2S_j^{RP}(P)-1) = \gamma(p_j-c)-1 = \frac{1}{1-S_j^{RP}(P)} - 1 \geq 0.$$

The uniqueness thus follows.

Hence the unique Nash equilibrium is in the interior range of price, and is the solution to the system of FOC equations. Also, due to quasi-concavity and existence of the global maxima in the interior domain of price, the solution to the FOC is a pure Nash equilibrium. □

Theorem 2.1 provides a way to find the equilibrium prices when the set of non-value-based providers is known. Specifically, the prices can be found by solving a system of equations where the prices of value-based providers is set to p^*. We show this system of equations has a unique solution. To use this theorem and find the equilibrium prices, we need to know the set of non-value-based providers. Proposition 2.3 provides a sufficient condition to know whether a given provider chooses to be non-value-based. However, to obtain the set of non-value-based providers, we need further characterization of the provider decision.

Proposition 2.5 *There exists $j \in \{0, 1, \ldots, n\}$ such that provider k is non-value-based iff $k > j$.*

Proof. We start by showing a technical lemma.

Lemma 2.3 *Provider j is value-based if and only if $1-\gamma(p^*-c)(1-S_j^{RP}(p_j = p^*, P_{-j})) \leq 0$.*

Recall from (2.8) that a provider's utility is monotonically increasing to the left of p^*, with a slope discontinuity at p^*. Moreover, when $p_j > p^*$, we have

$$\frac{\partial V_j^{RP}}{\partial p_j} = mS_j^{RP}(P)[1 - \gamma(p_j - c)(1 - S_j^{RP}(P))].$$

We observe that the sign of the expression above is given by the sign of $\psi(P) \equiv 1 - \gamma(p_j - c)(1 - S_j^{RP}(P))$. Taking the derivative of this expression with respect to p_j (on the domain when $p_j > p^*$), using Lemma 2.1, we find

$$\begin{aligned}
\frac{\partial \psi}{\partial p_j} &= -\gamma(1 - S_j^{RP}(P)) + \gamma(p_j - c)\frac{\partial S_j^{RP}}{\partial p_j} \\
&= -\gamma(1 - S_j^{RP}(P)) - \gamma^2(p_j - c)[1 - S_j^{RP}(P)]S_j^{RP}(P) \\
&= -\gamma(1 - S_j^{RP}(P))[1 + \gamma(p_j - c)S_j^{RP}(P)] \\
&\leq 0,
\end{aligned}$$

where the last inequality follows from the fact that provider j must price above its cost to make a profit. Thus ψ is monotonically decreasing in p_j. Hence it takes the value 0 at most once, and if it does, it goes from being positive to being negative on the domain $p_j > p^*$. As a result, $\frac{\partial V_j^{RP}}{\partial p_j}$ also takes the value 0 at most once, and if it does, it goes from being positive to being negative on the domain $p_j > p^*$. Furthermore, it is easy to observe that $\lim_{p_j \to \infty} V_j^{RP}(P) = 0$, and thus V_j^{RP} is not monotonically increasing in p_j on the domain $p_j > p^*$. Therefore, V_j^{RP} is a unimodal function of p_j: it increases in p_j on the domain $p_j < p^*$, and it is either monotonically decreasing on the domain $p_j > p^*$ (when its derivative at $(p^*)^+$ is non-positive) or it is increasing and then decreasing on the domain $p_j > p^*$ (when its derivative at $(p^*)^+$ is positive), with a derivative equal to zero at exactly one point

$p_j > p^*$. In the former case, provider j maximizes its profit by selecting $p_j = p^*$ (i.e., it is value-based); in the latter case, provider j maximizes its profit by selecting $p_j > p^*$ (i.e., it is non-value-based). Using (2.11), this concludes the proof of Lemma 2.3.

We now prove the result of the Proposition. Assume that provider j is non-value-based. We want to show that any provider ranked $k > j$ is also non-value-based. By induction, it suffices to show that provider $j+1$ is non-value based. We proceed by contradiction: suppose that provider $j+1$ is value-based, that is, $p_{j+1} = p^*$. We have

$$1-\gamma(p^*-c_{j+1})(1-S^{RP}_{j+1}(p_{j+1}=p^*,P_{-j})) = 1-\gamma(p^*-c_{j+1})\frac{e^{u_0}+\sum_{k\neq j+1}e^{A_k-\gamma(p_k-p^*)^+}}{e^{u_0}+e^{A_{j+1}}+\sum_{k\neq j+1}e^{A_k-\gamma(p_k-p^*)^+}}.$$

The expression above has the same sign as

$$\delta \equiv e^{u_0}+e^{A_{j+1}}+\sum_{k\neq j+1}e^{A_k-\gamma(p_k-p^*)^+}-\gamma(p^*-c_{j+1})\left(e^{u_0}+\sum_{k\neq j+1}e^{A_k-\gamma(p_k-p^*)^+}\right).$$

Because $c_{j+1} \geq c$, we obtain

$$\delta \geq e^{u_0}+e^{A_{j+1}}+\sum_{k\neq j+1}e^{A_k-\gamma(p_k-p^*)^+}-\gamma(p^*-c_j)\left(e^{u_0}+\sum_{k\neq j+1}e^{A_k-\gamma(p_k-p^*)^+}\right).$$

Provider j is non-value-based, so $p_j > p^*$ and thus

$$\delta > e^{u_0}+e^{A_{j+1}}+\sum_{k\neq j+1}e^{A_k-\gamma(p_k-p^*)^+}-\gamma(p_j-c_j)\left(e^{u_0}+\sum_{k\neq j+1}e^{A_k-\gamma(p_k-p^*)^+}\right).$$

Provider j selects its price p_j such that

$$1-\gamma(p_j-c)(1-S^{RP}_j(P))=0,$$

that is,

$$\frac{e^{u_0}+\sum_k e^{A_k-\gamma(p_k-p^*)^+}}{e^{u_0}+\sum_{k\neq j}e^{A_k-\gamma(p_k-p^*)^+}}=\gamma(p_j-c).$$

Therefore,

$$\delta > e^{u_0}+e^{A_{j+1}}+\sum_{k\neq j+1}e^{A_k-\gamma(p_k-p^*)^+}-\frac{e^{u_0}+\sum_{k\neq j+1}e^{A_k-\gamma(p_k-p^*)^+}}{e^{u_0}+\sum_{k\neq j}e^{A_k-\gamma(p_k-p^*)^+}}\left(e^{u_0}+\sum_k e^{A_k-\gamma(p_k-p^*)^+}\right).$$

Note that the ratio in the expression above can be written as

$$\frac{e^{u_0}+\sum_{k\neq j+1} e^{A_k-\gamma(p_k-p^*)^+}}{e^{u_0}+\sum_{k\neq j} e^{A_k-\gamma(p_k-p^*)^+}} = \frac{e^{A_j-\gamma(p_j-p^*)^+}+e^{u_0}+\sum_{k\neq j,j+1} e^{A_k-\gamma(p_k-p^*)^+}}{e^{A_{j+1}}+e^{u_0}+\sum_{k\neq j,j+1} e^{A_k-\gamma(p_k-p^*)^+}} < 1,$$

where the inequality follows from $A_j - \gamma(p_j - p^*)^+ < A_j \leq A_{j+1}$. Therefore, we have

$$\delta > e^{u_0} + e^{A_{j+1}} + \sum_{k\neq j+1} e^{A_k-\gamma(p_k-p^*)^+} - e^{u_0} - \sum_{k} e^{A_k-\gamma(p_k-p^*)^+} = 0.$$

Hence, $1-\gamma(p^*-c_{j+1})(1-S^{RP}_{j+1}(p_{j+1}=p^*,P_{-j}))>0$, which contradicts Lemma 2.3 given that provider $j+1$ is value-based. Therefore, provider $j+1$ is non-value based. □

We use the result of Proposition 2.5 along with Theorem 2.1 below to propose a process with at most n steps for obtaining the set of non-value based providers and their prices at equilibrium.

Algorithm 2.1 Step 0. *Initialize* $\mathcal{N} = \{1,\ldots,n\}$ *and* $j = 1$

Step 1. *Solve the system of equations given in Theorem 2.1 for the set* $\mathcal{N}$ *of non-value-based providers. If* $p_j > p^* \quad \forall j \in \mathcal{N}$*, stop. Else, go to Step 2.*

Step 2. $\mathcal{N} \to \mathcal{N} \setminus \{j\}$, $j \to j+1$. *If* $\mathcal{N} = \emptyset$*, stop. Else, go to Step 1.*

Algorithm 2.1 involves solving at most n systems of equations to iteratively test possible candidates for the set of non-value-based providers. The process starts by testing the case where all providers are non-value-based, and checking if there exist prices solving the first-order conditions that are indeed above the reference price. If not, according to Proposition 2.5, we remove the lowest-ranked provider and make it a value-based provider. We iterate the process until we find the set of non-value-based providers for which the optimal prices are all above the reference price (or until no candidate is left to be non-value-based). Upon completion of the algorithm we identify the set of non-value-based providers as well as the prices they select in equilibrium. Using the prices that are solution of this system of equations, we can find the utility of each provider, the patients, and the insurer.

Finally the following proposition illustrates how the set of non-value-based providers changes with the reference price.

Proposition 2.6 *As the reference price increases, the number of non-value-based providers decreases(or remains constant).*

Proof. Being a value-based or non-value-based provider depends on the sign of $1-\gamma(p^*-c)(1-S_j^{RP}(p_j=p^*,P_{-j}))$ as demonstrated in Lemma 2.3. From this expression it is easy to see that as the value of p^* increases this expression can only chagne sign from positive to negative. This means that by increasing the value of p^* only the non-value-based providers can become value-based. □

In general, as the reference price increases, value-based providers gain a higher margin while maintaining a high level of market share. Thus, non-value-based providers may choose to become value-based. Notice that providers with better non-price attributes (higher A) require a larger reference price increase to become value-based, as they offer a high appeal to consumers which can justify charging above the reference price.

2.4.3 Variable payment

Next we propose and study a *variable payment*. In this system the patient is responsible for a portion (λ) of the price charged, while the remainder is paid by the insurer. Hence the patient has an incentive to select a lower-priced provider. Under the variable payment system, the patient's out-of-pocket cost when selecting provider j is $O_j=\lambda p_j$. Hence, using (2.3), the patients choose their providers in such a way that provider j's proportion of patients is

$$S_j^{VP}(P)=\frac{e^{A_j-\gamma\lambda p_j}}{e^{u_0}+\sum_{k=1}^{n}e^{A_k-\gamma\lambda p_k}}, \tag{2.16}$$

and thus provider j's market share is $Q_j^{VP}(P)=mS_j^{VP}(P),\quad j=1,\dots,n.$

We next evaluate the patients' overall expected utility as

$$E[U^{VP}]=m\sum_{j=1}^{n}(A_j-\gamma\lambda p_j)S_j^{VP}(P)+mu_0S_0^{VP}(P).$$

Each provider decides on its price to maximize its utility given by

$$V_j^{VP}(P) = m(p_j - c)S_j^{VP}(P), \quad j = 1, \ldots, n,$$

while anticipating the patients' reaction.

The insurer's utility is

$$W^{VP} = \delta E[U^{VP}] - (1-\lambda)\sum_{j=1}^{n} p_j Q_j^{VP}(P). \tag{2.17}$$

Similar to the RP system, we show in the following lemma that under a variable payment system, the prices that providers select at equilibrium are bounded from above.

Lemma 2.4 *In the VP payment system, under Assumption 2.1 and for $\lambda \in (0, 1]$, the best response price of the provider $j \in N$ to any competing provider prices has an upper bound of $\overline{p}_j^{VP}$.*

Proof. We can find the best response of provider j when the competing providers price at $p^{max} \to \infty$ by solving

$$\frac{e^{A_j - \gamma\lambda\overline{p}_j^{VP}}}{\gamma\lambda(\overline{p}_j^{VP} - c) - 1} = e^{u_0}.$$

It can be seen that the market share of provider j is at its maximum level at a given price when the price of the competing providers increase to ∞. The statement of Assumption 2.1 for the variable payment can be written as

$$\frac{e^{A_j - \gamma\lambda\overline{p}_j^{VP}}}{e^{u_0} + \sum_{i\neq j} e^{A_i - \gamma\lambda p^{max}} + e^{A_j - \gamma\lambda\overline{p}_j^{VP}}} \leq 0.5 \xrightarrow{p^{max}\to\infty} A_j - \gamma\lambda\overline{p}_j^{VP} \leq u_0$$

Where the last inequality is reached as $p^{max} \to \infty$.

Similar to the reference pricing case, we can then show that for any $p_j > \overline{p}_j^{VP}$, $\frac{\partial V_j^{VP}}{\partial p_j} < 0$ under Assumption 2.1 and for any feasible value of price for the competing providers. Hence, provider j never prices higher than $\overline{p}_j^{VP}$. □

Lemma 2.4 shows that providers only become worse off if they price above $\overline{p}_j^{VP}$. This property constructs a closed action set for providers' prices ($p_j^{VP} \in [c, \overline{p}_j^{VP}]$), which later helps us prove the existence of a unique Nash equilibrium.

Proposition 2.7 shows the existence of a unique Nash equilibrium for provider prices in the case of a variable payment system and describes how to obtain these prices.

Proposition 2.7 *At equilibrium, the provider prices are the unique solution of the system of equations:*

$$1 - \gamma\lambda(p_j - c)(1 - S_j^{VP}(P)) = 0, \quad j = 1, \ldots, n.$$

Proof. Similar to Lemma 2.1, we have

$$\frac{\partial S_j^{VP}}{\partial p_j} = -\lambda\gamma S_j^{VP}(P)(1 - S_j^{VP}(P)). \tag{2.18}$$

Therefore,

$$\frac{\partial V_j^{VP}}{\partial p_j} = m S_j^{VP}(P)\left(1 - \gamma\lambda(p_j - c)(1 - S_j^{VP}(P))\right) \tag{2.19}$$

$$\frac{\partial^2 V_j^{VP}}{\partial p_j^2} = -m\gamma\lambda S_j^{VP}(P)(1 - S_j^{VP}(P))\left(2 + \gamma\lambda(p_j - c)(2S_j^{VP}(P) - 1)\right). \tag{2.20}$$

Similar to the proof of Theorem 2.1, we can show that the second derivative above can only change sign at most once from negative to positive as p_j gets larger. Moreover, as p_j tends to infinity, the provider utility $V_j^{VP}(P)$ tends to zero as a result of a disappearing market share. Thus the provider utility is quasi-concave in p_j. Therefore a provider j finds its optimal price (assuming all other prices are fixed) by solving for the first order condition

$$1 - \gamma\lambda(p_j - c)(1 - S_j^{VP}(P)) = 0. \tag{2.21}$$

We establish existence of the equilibrium similarly to the proof of Theorem 2.1 using the closed interval of $[c, \overline{p}_j^{VP}]$. Also for the uniqueness similar argument as in Theorem 2.1 holds. In other words, using similar argument we can show that

$$\left|\frac{\partial^2 V_j^{VP}}{\partial^2 p_j}\right| > \sum_{\substack{i\neq j \\ i\in N}} \left|\frac{\partial^2 V_j^{VP}}{\partial p_j \partial p_i}\right|, \quad \forall j \in N,$$

where

$$\left|\frac{\partial^2 V_j^{VP}}{\partial p_j^2}\right| = m\gamma\lambda S_j^{VP}(P)[1 - S_j^{VP}(P)]|2 + \gamma\lambda(p_j - c)(2S_j^{VP}(P) - 1)|,$$

and

$$\sum_{\substack{i\neq j \\ i\in N}} \left|\frac{\partial^2 V_j^{VP}}{\partial p_j \partial p_i}\right| = \sum_{\substack{i\neq j \\ i\in N}} m\gamma\lambda S_j^{VP}(P) S_i^{VP}(P)|1 + \gamma\lambda(p_j - c)(2S_j^{VP}(P) - 1)|.$$

Hence the unique Nash equilibrium is the solution to the system of FOC equations. □

Using the prices that are solution of this system of equations, we can find the utility of each provider, the patients, and the insurer. Next we discuss some of the properties of the utility functions under the different payment models and compare and contrast the models for the case of two providers.

2.5 Policy Implications and Payment Comparisons - Special Case: Two Providers

In this section, we derive insightful properties of the equilibrium outcomes under the three different payment regimes, and explore their implications on different stakeholders to guide policy decisions in the case of two providers. As it is evident from best response and equilibrium expressions, the provider's expected profit functions depends on the vector of other provider's actions that involves analysis of, in general, non-concave functions which do not show super- or sub-modular properties. Therefore we focus in this section on the case of two providers to improve tractability and derive some analytical insights. This special case captures the effect of provider competition as well as the possibility of having both value-based and non-value-based providers under reference pricing, while maintaining the tractability of our model. We show in Section 2.6 that the key analytical insights we obtain in this special case continue to hold true for the case of n providers.

2.5.1 Preliminaries

We first derive insightful properties of the various payment schemes and investigate their implications on different stakeholders that are relevant from a policy perspective. In particular, we are interested in exploring how patients, providers, and the insurer are impacted under each payment regime.

Before we present the main results, Let

$$F_j\left(p_j, p_i\right) = e^{A_j - \gamma\lambda p_j} - \left(\gamma\lambda(p_j - c) - 1\right)\left(e^{u_0} + e^{A_i - \gamma\lambda p_i}\right) \quad j = 1, 2 \quad \text{and} \quad i \neq j,$$

and

$$G_j\left(p_j, p_i\right) = e^{A_j - \gamma(p_j - p^*)} - \left(\gamma(p_j - c) - 1\right)\left(e^{u_0} + e^{A_i - \gamma(p_i - p^*)}\right) \quad j = 1, 2 \quad \text{and} \quad i \neq j.$$

Next, we prove the following two lemmas, which state some technical properties of the providers pricing strategies that will be used in our subsequent analysis in this section.

Lemma 2.5 *$\left.\frac{\partial F_j}{\partial p_j}\right|_{P^{VP}} < 0$, where P^{VP} is a vector of equilibrium prices of providers i and j under the variable payment system. Similarly for a non-value-based provider under RP, $\left.\frac{\partial G_j}{\partial p_j}\right|_{P^{RP}} < 0$, where P^{RP} is the equilibrium prices of providers i and j under the reference pricing system system. Similar results hold for patient out of pocket ($\left.\frac{\partial F_j}{\partial O_j}\right|_{O^{VP}} < 0$ and $\left.\frac{\partial G_j}{\partial O_j}\right|_{O^{RP}} < 0$).*

Proof. Based on Proposition 2.7, the provider prices under the variable payment at equilibrium are the unique solution of the system of equations:

$$1 - \gamma\lambda\left(p_j - c\right)\left(1 - S_j^{VP}\left(p_j, p_i\right)\right) = 0, \quad j = 1, 2 \quad \text{and} \quad i \neq j. \tag{2.22}$$

Using the expression for $S_j^{VP}(P)$ from Equation (2.16) we can rewrite the left hand side of the FOC condition above for a general value for p_j as

$$F_j\left(p_j, p_i(p_j)\right) = e^{A_j - \gamma\lambda p_j} - (\gamma\lambda(p_j - c) - 1)\left(e^{u_0} + e^{A_i - \gamma\lambda p_i(p_j)}\right), \;\; j = 1, 2 \;\text{ and }\; i \neq j, \quad (2.23)$$

where $p_i(p_j)$ is the best response of provider i when provider j prices at p_j. Taking the derivative of the expression above with respect to p_j results in

$$\frac{\partial F_j}{\partial p_j} = -\gamma\lambda\left(e^{u_0} + \sum_{k=1}^{2} e^{A_k - \gamma\lambda p_k}\right) + \gamma\lambda\left(\gamma\lambda(p_j - c) - 1\right)\frac{\partial p_i}{\partial p_j} e^{A_i - \gamma\lambda p_i(p_j)} \quad (2.24)$$

where $\frac{\partial p_i}{\partial p_j}$ can be computed using the expression for the best response of provider i to p_j, we have

$$-\frac{\partial p_i}{\partial p_j}\left(e^{u_0} + \sum_{k=1}^{2} e^{A_k - \gamma\lambda p_k}\right) + \left(\gamma\lambda(p_i(p_j) - c) - 1\right) e^{A_j - \gamma\lambda p_j} = 0.$$

This implies that $\frac{\partial p_i}{\partial p_j} > 0$. In equilibrium providers price at mutual best responses. For ease of notation we consider $S_j^{VP}(P^{VP}) = S_j^{VP}(p_j^{VP}, p_i^{VP})$, where p_k^{VP} is the hospital k price at equilibrium ($k = i, j$). Hence, using Equation 2.24 for hospitals' best responses results in $\left(\gamma\lambda(p_k^{VP} - c) - 1\right) = S_k^{VP}(P^{VP})/(1 - S_k^{VP}(P^{VP})$ for $k = i, j$ at equilibrium. We then have

$$\frac{\partial p_i}{\partial p_j} = \left(\gamma\lambda(p_i(p_j) - c) - 1\right) S_j^{VP}\left(p_j, p_i(p_j)\right) \overset{\text{At Equilibrium}}{=} \frac{S_i^{VP}(P)\, S_j^{VP}(P)}{1 - S_i^{VP}(P)},$$

where the first equality is true based on Equation 2.16 and the second equality is derived by using Equation 2.21. We can now replace the expression derived for $\frac{\partial p_i}{\partial p_j}$ in Equation 2.24 and evaluate $\left.\frac{\partial F_j}{\partial p_j}\right|_{P^{VP}}$.

$$\left.\frac{\partial F_j}{\partial p_j}\right|_{P^{VP}} = \gamma\lambda\left(e^{u_0}+\sum_{k=1}^{2}e^{A_k-\gamma\lambda p_k^{VP}}\right)\left(-1+\frac{(S_i^{VP}(P^{VP})S_j^{VP}(P^{VP}))^2}{(1-S_i^{VP}(P^{VP}))(1-S_j^{VP}(P^{VP}))}\right)$$

$$= \gamma\lambda\left(e^{u_0}+\sum_{k=1}^{2}e^{A_k-\gamma\lambda p_k^{VP}}\right)\left(-S_0^{VP}(P^{VP})-S_j^{VP}(P^{VP})+\frac{(S_i^{VP}(P^{VP})S_j^{VP}(P^{VP}))^2}{(1-S_i^{VP}(P^{VP}))(1-S_j^{VP}(P^{VP}))}-S_i^{VP}(P^{VP})\right)$$

$$= \gamma\lambda\left(e^{u_0}+\sum_{k=1}^{2}e^{A_k-\gamma\lambda p_k^{VP}}\right)\left(-S_0^{VP}(P^{VP})-S_j^{VP}(P^{VP})+\frac{S_i^{VP}(P^{VP})\left(\overbrace{-1+S_i^{VP}(P^{VP})+S_j^{VP}(P^{VP})}^{<0}\overbrace{-S_i^{VP}(P^{VP})S_j^{VP}(P)(1-S_j^{VP}(P^{VP}))}^{<0}\right)}{(1-S_i^{VP}(P^{VP}))(1-S_j^{VP}(P^{VP}))}\right)$$

$$< 0, \qquad j=1,2 \text{ and } i\neq j.$$

Similarly for G_j we have

$$\frac{\partial G_j}{\partial p_j} = -\gamma\left(e^{u_0}+\sum_{k=1}^{2}e^{A_k-\gamma(p_k-p^*)}\right)+\gamma\left(\gamma(p_j-c)-1\right)\frac{\partial p_i^{RP}}{\partial p_j}e^{A_i-\gamma(p_i(p_j)-p^*)},\ \ j=1,2 \text{ and } i\neq j$$

where

$$\frac{\partial p_i^{RP}}{\partial p_j} = \left(\gamma(p_i(p_j)-c)-1\right)S_j^{RP}(p_j,p_i(p_j)) \overset{\text{At Equilibrium}}{=} \frac{S_i^{RP}(P^{RP})S_j^{RP}(P^{RP})}{1-S_i^{RP}(P^{RP})}.$$

Similar to the previous case then we can show that $\left.\frac{\partial G_j}{\partial p_j}\right|_{P^{RP}} < 0$. We can extend this proof for the case of different treatment costs across different providers.

We can obtain similar result for patient out-of-pocket. Notice that we can rewrite the expressions of functions F_j and G_j based on the patient's out-of-pocket by a simple change of variables.

$$F_j\left(O_j,O_i\right) = e^{A_j-\gamma O_j}-\left(\gamma(O_j-\lambda c)-1\right)\left(e^{u_0}+e^{A_i-\gamma O_i}\right) \quad j=1,2 \quad \text{and} \quad i\neq j.$$

and,

$$G_j\left(O_j, O_i\right) = e^{A_j - \gamma O_j} - \left(\gamma(O_j + p^* - c) - 1\right)\left(e^{u_0} + e^{A_i - \gamma O_i}\right) \quad j = 1, 2 \quad \text{and} \quad i \neq j.$$

Then we can show that $\left.\frac{\partial F_j}{\partial O_j}\right|_{O^{VP}} < 0$ and $\left.\frac{\partial G_j}{\partial O_j}\right|_{O^{VP}} < 0$, where O^{VP} and O^{RP} are vectors of patients' out-of-pocket from visiting hospital j at equilibrium under the variable and reference pricing payment systems respectively.

The proof follows similar steps. Note that at equilibrium $\gamma(O_k - \lambda c) = \frac{1}{1 - S_k^{VP}(O^{VP})}$ and $\gamma(O_k + p^* - c) = \frac{1}{1 - S_k^{RP}(O^{RP})}$ for $k = i, j$.

□

Lemma 2.5 indicates that the left hand side of the FOC for provider j is decreasing in its own price and patient out-of-pocket at equilibrium prices under both reference pricing and variable payment schemes.

2.5.2 Patients

The primary objective of this section is to evaluate the outcomes of payment models on patient utilities. We note that provider prices determine patient utilities through the patient out-of-pocket amount. Hence we start by investigating how provider prices vary as the values of payment parameters (i.e., λ in the variable payment and p^* in reference pricing) change.

Proposition 2.8 *(a) Under the variable payment system each provider's equilibrium price is decreasing in λ. (b) Given a set of value-based and non-value-based providers, increasing the value of reference price p^* results in higher provider prices.*

Proof. Under the variable payment provider j price is selected by solving the following equation (FOC)

$$e^{A_j - \gamma\lambda p_j^{VP}} - \left(\gamma\lambda(p_j^{VP} - c) - 1\right)\left(e^{u_0} + e^{A_i - \gamma\lambda p_i^{VP}}\right) = 0, \quad j = 1, 2 \quad \text{and} \quad i \neq j. \tag{2.25}$$

Taking the derivative of the FOC condition for provider j with respect to λ under the variable payment scenario, we have

$$
\begin{aligned}
&-\gamma\left(p_j^{VP}+\lambda\frac{\partial p_j^{VP}}{\partial\lambda}\right)\left(e^{u_0}+\sum_{k=1}^{2}e^{A_k-\gamma\lambda p_k^{VP}}\right)+\gamma c\left(e^{u_0}+e^{A_i-\gamma\lambda p_i^{VP}}\right)\\
&+\gamma\left(\gamma\lambda(p_j^{VP}-c)-1\right)\left(p_i^{VP}+\lambda\frac{\partial p_i^{VP}}{\partial\lambda}\right)e^{A_i-\gamma\lambda p_i^{VP}}=0,\quad j=1,2 \text{ and } \ i\neq j.
\end{aligned}
$$

Using Equations 2.16 and 2.21, this can be written as

$$
\begin{aligned}
&\gamma\left(e^{u_0}+\sum_{k=1}^{2}e^{A_k-\gamma\lambda p_k^{VP}}\right)\\
&\underbrace{\left(-\left(p_j^{VP}+\lambda\frac{\partial p_j^{VP}}{\partial\lambda}\right)+c\left(1-S_j^{VP}(P)\right)+\left(p_i^{VP}+\lambda\frac{\partial p_i^{VP}}{\partial\lambda}\right)\frac{S_i^{VP}(P)S_j^{VP}(P)}{1-S_j^{VP}(P)}\right)}_{=0}=0. \quad (2.26)
\end{aligned}
$$

Similar expression then can be derived when taking derivative of FOC condition for provider i with respect to λ, based on which

$$
p_i^{VP}+\lambda\frac{\partial p_i^{VP}}{\partial\lambda}=c\left(1-S_i^{VP}(P)\right)+\left(p_j^{VP}+\lambda\frac{\partial p_j^{VP}}{\partial\lambda}\right)\frac{S_i^{VP}(P)S_j^{VP}(P)}{1-S_i^{VP}(P)}.
$$

Expression above can be replaced into Equation 2.26 and simplified to get

$$
\begin{aligned}
&\lambda\frac{\partial p_j^{VP}}{\partial\lambda}\left(\overbrace{-1+\frac{\left(S_i^{VP}(P)S_j^{VP}(P)\right)^2}{\left(1-S_i^{VP}(P)\right)\left(1-S_j^{VP}(P)\right)}}^{<0}\right)\\
&\underbrace{+p_j^{VP}\left(-1+\frac{\left(S_i^{VP}(P)S_j^{VP}(P)\right)^2}{\left(1-S_i^{VP}(P)\right)\left(1-S_j^{VP}(P)\right)}\right)+c\left(1-S_j^{VP}(P)\right)+c\frac{\left(1-S_i^{VP}(P)\right)S_i^{VP}(P)S_j^{VP}(P)}{1-S_j^{VP}(P)}}_{B}=0.
\end{aligned}
$$

(2.27)

First we focus on expression B. We know that provider prices larger than its cost to be profit making. Assuming equal treatment cost $c = c = c$ across different providers then, $p_j^{VP} \geq c$ we can conclude

$$B < p_j^{VP}\left(\frac{-\left(1-S_i^{VP}(P)\right)\left(1-S_j^{VP}(P)\right)+\left(S_i^{VP}(P)S_j^{VP}(P)\right)^2+\left(1-S_i^{VP}(P)\right)\left(\left(1-S_j^{VP}(P)\right)^2+\left(1-S_i^{VP}(P)\right)^2 S_i^{VP}(P)S_j^{VP}(P)\right)}{\left(1-S_i^{VP}(P)\right)\left(1-S_j^{VP}(P)\right)}\right)$$

$$= \frac{p_j^{VP}S_j^{VP}(P)}{\left(1-S_i^{VP}(P)\right)\left(1-S_j^{VP}(P)\right)}\left[-1+S_j^{VP}(P)+S_i^{VP}(P)\left\{2+S_i^{VP}(P)S_j^{VP}(P)-S_j^{VP}(P)-2S_i^{VP}(P)+(S_i^{VP}(P))^2\right\}\right]$$

$$= \frac{p_j^{VP}S_j^{VP}(P)}{\left(1-S_i^{VP}(P)\right)\left(1-S_j^{VP}(P)\right)}\left[\underbrace{\left(-1+S_i^{VP}(P)+S_j^{VP}(P)\right)}_{<0}\left\{1-S_i^{VP}(P)\left(1-S_i^{VP}(P)\right)\right\}\right] < 0$$

Using $B < 0$ in Equation 2.27 results in $\frac{\partial p_j^{VP}}{\partial \lambda} < 0$.

For the proof of part b of this proposition, consider that the changes in the value of p^* is small enough such that the set of value-based and non-value-based providers are not going to change as we vary p^*. In general, we can show that the derivative of the provider j market share with respect to p^* has the following form.

$$\frac{\partial S_j^{RP}}{\partial p^*} = \begin{cases} \gamma S_j^{RP}(P)\sum\limits_{k\in NVB}\left(\frac{\partial p_k}{\partial p^*}-1\right)S_k^{RP}(P) & \text{if } p_j < p^* \\ \gamma S_j^{RP}(P)\left(\sum\limits_{k\in NVB}\left(\frac{\partial p_k}{\partial p^*}-1\right)S_k^{RP}(P)+1-\frac{\partial p_j}{\partial p^*}\right) & \text{if } p_j > p^* \end{cases}$$

With two providers in the insurer's network we have similar expression, where $\frac{\partial p_i}{\partial p^*} = 1 \ \ \forall i \in VB$. We can then observe that if both providers are value-based market share of the providers will not be sensitive to the value of p^*.

Lets look at the effect of p^* on a non-value-based provider utility and price choice. We have

$$\frac{\partial V_j^{RP}}{\partial p^*} = m\frac{S_j^{RP}(P)}{1-S_j^{RP}(P)}\left(\sum_{\substack{K\in NVB\\ k\neq j}}\left(\frac{\partial p_k}{\partial p^*}-1\right)S_k^{RP}(P)+1-S_j^{RP}(P)\right).$$

In the case of having two providers, with $k \neq j$ depending on whether or not provider k is value-based we have

$$\frac{\partial V_j^{RP}}{\partial p^*} = \begin{cases} mS_j^{RP}(P) > 0 & \text{if } p_k < p^* \\ \\ m\frac{S_j^{RP}(P)}{1-S_j^{RP}(P)}\left(\left(\frac{\partial p_k}{\partial p^*}-1\right)S_k^{RP}(P)+1-S_j^{RP}(P)\right) & \text{if } p_k > p^* \end{cases} \tag{2.28}$$

Considering the FOC for provider j and taking the derivative of the corresponding equation with respect to p^* results in

$$\frac{\partial p_j}{\partial p^*} = \begin{cases} S_j^{RP}(P) > 0 & \text{if } p_k < p^* \\ \\ \frac{S_j^{RP}(P)}{1-S_j^{RP}(P)}\left(\left(\frac{\partial p_k}{\partial p^*}-1\right)S_k^{RP}(P)+1-S_j^{RP}(P)\right) & \text{if } p_k > p^*. \end{cases}$$

If the competing provider is value-based, provider j always price higher as the value of p^* increases. We can also show that if the competing provider k is non-value-based $\frac{\partial p_j}{\partial p^*} > 0$. This stems from expanding the expression above for non-value-based provider k to have

$$\underbrace{\frac{\partial p_j}{\partial p^*}\left(\frac{1-S_k^{RP}(P)-S_j^{RP}(P)+S_k^{RP}(P)S_j^{RP}(P)(1-S_k^{RP}(P)S_j^{RP}(P))}{S_j^{RP}(P)}\right)}_{>0} =$$

$$(S_k^{RP}(P))^2(2-S_k^{RP}(P)-S_j^{RP}(P))+S_j^{RP}(P)(1-S_k^{RP}(P))+1-2S_k^{RP}(P) >$$

$$(S_k^{RP}(P))^2+1-2S_k^{RP}(P) > 0$$

Thus $\frac{\partial p_j}{\partial p^*} > 0$.

□

This result is intuitive; as the patient cost share (λ) increases, patients bear a larger portion of the price, and thus providers lower their prices to attract more of the market share. Similarly, part (b) of Proposition 2.8 shows providers' reaction to changes in the value of reference price, p^*, under the reference pricing system. Strictly speaking, increasing the value of reference price results in increasing prices for a provider. Note that even if the provider is value-based increasing p^* results in value-based providers increasing their prices to the new reference price level.

Proposition 2.8 illustrates changes in provider prices as the payment parameters change. Patients' utilities are determined by the out-of-pocket amount $O_j^{VP} = \lambda p_j$ and $O_j^{RP} = (p_j - p^*)^+$ under the variable and reference pricing schemes, respectively. The following proposition shows the net effect of changing payment parameters on the patients' out-of-pocket.

Proposition 2.9 *(a) Under the variable payment model the patient out-of-pocket from visiting provider j in equilibrium is increasing in λ. (b) Under the reference pricing model the patient out-of-pocket from visiting provider j in equilibrium is non-increasing in p^*.*

Proof. Provider j out-of-pocket is determined by solving the following equation (FOC)

$$e^{A_j-\gamma O_j^{VP}} - \left(\gamma(O_j^{VP} - \lambda c) - 1\right)\left(e^{u_0} + e^{A_i-\gamma O_i^{VP}}\right) = 0, \quad j = 1, 2 \quad \text{and} \quad i \neq j. \tag{2.29}$$

Taking the derivative of the FOC for provider j with respect to λ under the variable payment scenario, we have

$$\begin{aligned} &-\gamma \frac{\partial O_j^{VP}}{\partial \lambda}\left(e^{u_0} + \sum_{k=1}^{2} e^{A_k-\gamma O_k^{VP}}\right) + \gamma c\left(e^{u_0} + e^{A_i-\gamma O_i^{VP}}\right) \\ &+\gamma\left(\gamma(O_j^{VP} - \lambda c) - 1\right)\frac{\partial O_i^{VP}}{\partial \lambda} e^{A_i-\gamma O_i^{VP}} = 0, \quad j = 1, 2 \text{ and } \ i \neq j. \end{aligned}$$

Using Equations 2.16 and 2.21, this can be written as

$$\gamma\left(e^{u_0}+\sum_{k=1}^{2}e^{A_k-\gamma O_k^{VP}}\right)\underbrace{\left(-\frac{\partial O_j^{VP}}{\partial\lambda}+c\left(1-S_j^{VP}(P)\right)+\frac{\partial O_i^{VP}}{\partial\lambda}\frac{S_i^{VP}(P)S_j^{VP}(P)}{1-S_j^{VP}(P)}\right)}_{=0}=0 \tag{2.30}$$

Similar expression then can be derived when taking derivative of FOC condition for provider i with respect to λ, based on which

$$\frac{\partial O_i^{VP}}{\partial\lambda}=c\left(1-S_i^{VP}(P)\right)+\frac{\partial O_j^{VP}}{\partial\lambda}\frac{S_i^{VP}(P)S_j^{VP}(P)}{1-S_i^{VP}(P)}.$$

Expression above can be replaced into Equation 2.30 and simplified to get

$$\frac{\partial O_j^{VP}}{\partial\lambda}\left(\overbrace{-1+\frac{\left(S_i^{VP}(P)S_j^{VP}(P)\right)^2}{\left(1-S_i^{VP}(P)\right)\left(1-S_j^{VP}(P)\right)}}^{<0}\right)+\underbrace{c\left(1-S_j^{VP}(P)\right)+c\frac{\left(1-S_i^{VP}(P)\right)S_i^{VP}(P)S_j^{VP}(P)}{1-S_j^{VP}(P)}}_{>0}=0. \tag{2.31}$$

Equation 2.31 results in $\frac{\partial O_j^{VP}}{\partial\lambda}>0$. Clearly, this is true for the special case of having equal costs across different providers c.

For the second part of the proposition, we are assuming here that the change in the value of p^* is in a small interval such that the set of value-based and non-value-based providers are not changing as the value of p^* varies. Provider j out-of-pocket is determined by solving the following equations (FOC) for non-value-based providers

$$e^{A_j-\gamma O_j^{RP}}-\left(\gamma(O_j^{RP}+p^*-c)-1\right)\left(e^{u_0}+e^{A_i-\gamma O_i^{RP}}\right)=0,\quad i\ \&\ j=1,2\quad\text{and}\quad i\neq j. \tag{2.32}$$

Note that if provider j is value based, the patient out-of-pocket is zero and does not change as long as the value of the p^* varies in a sufficiently small interval. This proof is similar to the proof of Proposition 2.9. Following similar steps we first Take the derivative

of the FOC for non-value-based provider j with respect to p^* under the reference pricing payment scenario, we have

$$
\begin{aligned}
&-\gamma\frac{\partial O_j^{RP}}{\partial p^*}\left(e^{u_0}+\sum_{k=1}^{2}e^{A_k-\gamma O_k^{RP}}\right)-\gamma\left(e^{u_0}+e^{A_i-\gamma O_i^{RP}}\right)\\
&+\gamma\left(\gamma(O_j^{RP}+p^*-c)-1\right)\frac{\partial O_i^{RP}}{\partial p^*}e^{A_i-\gamma O_i^{RP}}=0,\quad j=1,2 \text{ and } \ i\neq j.
\end{aligned}
$$

Using Equations 2.7 and 2.9, this can be written for equilibrium as

$$
\gamma\left(e^{u_0}+\sum_{k=1}^{2}e^{A_k-\gamma O_k^{RP}}\right)\underbrace{\left(-\frac{\partial O_j^{RP}}{\partial p^*}-\left(1-S_j^{RP}(P)\right)+\frac{\partial O_i^{RP}}{\partial p^*}\frac{S_i^{RP}(P)S_j^{RP}(P)}{1-S_j^{RP}(P)}\right)}_{=0}=0 \quad (2.33)
$$

Note that if provider i is value-based, then $\frac{\partial O_i^{RP}}{\partial p^*}=0$ and based on the expression above $\frac{\partial O_j^{RP}}{\partial p^*}<0$. Although if provider i is non-value-based, similar expression then can be derived when taking derivative of FOC condition for provider i with respect to p^*

$$
\frac{\partial O_i^{RP}}{\partial p^*}=-\left(1-S_i^{RP}(P)\right)+\frac{\partial O_j^{RP}}{\partial p^*}\ \frac{S_i^{RP}(P)S_j^{RP}(P)}{1-S_i^{RP}(P)}.
$$

Expression above can be replaced into Equation 2.33 and simplified to get

$$
\frac{\partial O_j^{RP}}{\partial p^*}\left(\overbrace{-1+\frac{\left(S_i^{RP}(P)S_j^{RP}(P)\right)^2}{\left(1-S_i^{RP}(P)\right)\left(1-S_j^{RP}(P)\right)}}^{<0}\right)-\underbrace{\left(1-S_j^{RP}(P)\right)-\frac{\left(1-S_i^{RP}(P)\right)S_i^{RP}(P)S_j^{RP}(P)}{1-S_j^{RP}(P)}}_{<0}=0. \quad (2.34)
$$

Equation 2.34 results in $\frac{\partial O_j^{RP}}{\partial p^*}<0$. □

Note that changing the patients' cost share fraction of payment has a direct effect on the patient's out-of-pocket, as well as an indirect effect through provider prices. On the one hand, rising λ increases the patient out-of-pocket. On the other hand, rising λ results in

lower provider prices. Hence, the effect of λ on patient out-of-pocket dominates the increase in prices, and a larger portion of payment results in larger out-of-pocket for the patients in equilibrium. Moreover, if a provider is non-value-based an increase of the reference price lowers or does not affect the out-of-pocket for a patient selecting this provider. Clearly if the provider is value-based, the patient out-of-pocket is zero and does not change as the value of p^* increases. These results are subject to the assumption of varying p^* in a sufficiently small interval such that sets of value-based and non-value-based providers remain fixed. Note that increasing the value of the reference price has a direct effect on the value of the patient out-of-pocket, as well as an indirect effect through the provider prices. In other words, rising p^* decreases the value of the patient out-of-pocket while increasing the provider prices as shown in Proposition 2.8 (b). Hence, the effect of reference price on patient out-of-pocket dominates the increase in prices, and a larger reference price results in smaller out-of-pocket for the patients in equilibrium.

We now proceed to compare the different payment schemes. We first compare the equilibrium prices of providers. Proposition 2.1 states that the fixed payment model results in the maximum pricing level $\overline{p}^{FP}$ for all providers. The following result compares the pricing strategy under reference pricing and variable payment.

Proposition 2.10 *There exists a threshold λ_j^* for hospital j such that, provider j prices higher under the variable payment system than under the reference pricing scheme iff $\lambda < \lambda_j^*$, where λ_j^* is given by*

$$e^{A_i - \gamma\lambda_j^*\hat{p}} - \left(\gamma\lambda_j^*(\hat{p}-c)-1\right)\left(e^{u_0} + e^{A_j - \gamma\lambda_j^* p_j^{VP}(\hat{p})}\right) = 0,$$

where $\hat{p}$ solves

$$e^{A_i - \gamma\hat{p}} - \left(\gamma(\hat{p}-c)-1\right)\left(e^{u_0 - \gamma p^*} + e^{A_j - \gamma p_j^{RP}(\hat{p})}\right) = 0,$$

where $p_j^{VP}(\hat{p})$ and $p_j^{RP}(\hat{p})$ are the best response prices of provider j when provider i prices at $\hat{p}$ under the variable payment and reference pricing systems respectively.

Proof. To prove the statement of the proposition we first investigate the behavior of the provider at the extreme points when $\lambda \to 0$ and $\lambda \to 1$.

First we compare the equilibrium prices for provider j under the two payment models as $\lambda \to 0$. As $\lambda \to 0$, the FOC equation for the provider under the variable payment does not have a solution if p_j is bounded. Also we know that $\frac{\partial p_j^{VP}}{\partial \lambda} < 0$ so as the λ takes its minimum value price should get larger. Since $p_j^{VP} > c + 1/\gamma\lambda$ for the FOC of provider j to have a solution for p_j, $p_j \to \infty$. Note that $\overline{p}_j^{VP}$ also is pushed to infinity as $\lambda \to 0$. Since we allow only a finite value for $\overline{p}_j^{VP}$ as λ varies $\in (0, 1]$, we instead consider a sufficiently small value of cost share ($\underline{\lambda}$) that makes the price under the variable payment model to exceed that of the reference pricing scheme.

In other words, under reference pricing, if provider j is value-based, then its price equals $p^* < p_j^{VP}$ for a small enough value of λ. If provider j is non-value-based, it chooses its price according to equation (2.15), such that

$$1 - \gamma(p_j^{RP} - c)(1 - S_j^{RP}(P)) = 0.$$

We know that p_j^{RP} that solves equation above should be finite. Thus, for a small enough value of λ, $p_j^{RP} < p_j^{VP}$. Therefore, there exists a small enough value for cost share ($\underline{\lambda}$), under which $p_j^{VP} \geq p_j^{RP} \quad \forall j$.

We now show that the reverse is true when $\lambda \to 1$. We start by comparing F_i and G_i for provider i at a given equal price $\hat{p}$ under both payment models, where

$$F_i\left(p_i = \hat{p}, p_j^{VP}\right) = e^{A_i - \gamma\hat{p}} - (\gamma(\hat{p} - c) - 1)\left(e^{u_0} + e^{A_j - \gamma p_j^{VP}}\right)$$

$$G_i\left(p_i = \hat{p}, p_j^{RP}\right) = e^{A_i - \gamma\hat{p}} - (\gamma(\hat{p} - c) - 1)\left(e^{u_0 - \gamma p^*} + e^{A_j - \gamma p_j^{RP}}\right).$$

Note that at equilibrium the provider sets these statements to zero to find the optimal price. As can be seen in the statements in order to compare F_i and G_i we need to first compare the best response of the competing provider (p_j^{RP} and p_j^{VP}) when provider i sets the price to

$\hat{p}$ under both payment models. The best response of provider j under the variable payment is computed by solving the following equation for p_j^{VP}.

$$F_j\left(p_j^{VP}, p_i = \hat{p}\right) = e^{A_j - \gamma p_j^{VP}} - \left(\gamma(p_j^{VP} - c) - 1\right)\left(e^{u_0} + e^{A_i - \gamma\hat{p}}\right) = 0$$

For the reference pricing payment model provider j can either be value-based or non-value based. If provider j is non-value-based it sets $G_j = 0$ to find the best response and $p_j^{RP} > p^*$.

$$G_j\left(p_j^{RP}, p_i = \hat{p}\right) = e^{A_j - \gamma p_j^{RP}} - \left(\gamma(p_j^{RP} - c) - 1\right)\left(e^{u_0 - \gamma p^*} + e^{A_i - \gamma\hat{p}}\right) = 0.$$

By comparing G_j to F_j it is easy to see that

$$\frac{e^{A_j - \gamma p_j^{VP}}}{\gamma(p_j^{VP} - c) - 1} > \frac{e^{A_j - \gamma p_j^{RP}}}{\gamma(p_j^{RP} - c) - 1}$$

$\frac{e^{A_j - \gamma x}}{\gamma(x-c)-1}$ is a decreasing function of x as long as $x > c$ (which is the case for equilibrium prices), from the inequality above we can conclude that $p_j^{VP} < p_j^{RP}$. In other words, the best response of provider j when provider i prices equally under the two payments is to price lower for the variable payment model. Below we discuss three cases that may occur depending on the providers being value-based or non-value-based.

Case 1. Both providers are non-value-based: We can now compare F_i and G_i under a fixed price $\hat{p}$ for both payments. Since we showed that $p_j^{VP} < p_j^{RP}$, it is easy to see that $G_i\left(\hat{p}, p_j^{RP}\right) > F_i\left(\hat{p}, p_j^{VP}\right)$ at any given $\hat{p} > \frac{1}{\gamma} + c$, and for $\hat{p} \leq \frac{1}{\gamma} + c$, $F_i\left(\hat{p}, p_j^{VP}\right) \geq G_i\left(\hat{p}, p_j^{RP}\right) > 0$ and cannot be zero. Note that since the providers are non-value-based $\hat{p}$s that we consider should be greater than p^*. Thus if $\frac{1}{\gamma} + c < p^*$, $G_i > F_i$ for all feasible values of $\hat{p} > p^*$.

Also from Proposition 2.5, $\partial\ F_i/\partial\ p_i|_{p_i = p_i^{VP}} < 0$, and for $p_i > p^*$, $\partial\ G_i/\partial\ p_i|_{p_i = p_i^{RP}} < 0$. Hence, when provider i is non-value-based at the points where $F_i\left(p_i^{VP}, p_j^{VP}\right) = G_i\left(p_i^{RP}, p_j^{RP}\right) = 0$ (which result in optimum prices under both payments), $p_i^{RP} > p_i^{VP}$. Note that if $\hat{p} \leq \frac{1}{\gamma} + c$, both $G_i\left(\hat{p}, p_j^{RP}\right)$ and $F_i\left(\hat{p}, p_j^{VP}\right)$ are strictly positive and cannot be zero. In other words,

$G_i\left(\hat{p}, p_j^{RP}\right)$ and $F_i\left(\hat{p}, p_j^{VP}\right)$ can only be zero when $\hat{p} > \frac{1}{\gamma} + c$ under which $G_i\left(\hat{p}, p_j^{RP}\right) > F_i\left(\hat{p}, p_j^{VP}\right)$, resulting in $p_i^{RP} > p_i^{VP}$.

Case 2. Provider i is value-based while provider j is non-value-based: $G_i\left(p_i^{RP}, p_j^{RP}\right)$ is never zero in this case and the provider i chooses p^* regardless of provider j's price choice. Moreover, as stated in Lemma 2.1, for a value-based provider $\frac{\partial V_i^{RP}}{\partial p_i}\Big|_{p_i=(p^*)^+} < 0$ which is equivalent to $G_i\left(p_i = (p^*)^+, p_j^{RP}\right) < 0$ at $p_i = (p^*)^+$.

If p^* is also chosen under the variable payment for provider i, we have already shown that the best response of provider j to the fix price of provider i under the two payment is such that $p_j^{VP} < p_j^{RP}$. Hence,

$$\begin{aligned} F_i(p_i = p^*, p_j^{VP}) &= e^{A_i - \gamma p^*} - (\gamma(p^* - c) - 1)\left(e^{u_0} + e^{A_j - \gamma p_j^{VP}}\right) \\ &< e^{A_i - \gamma p^*} - (\gamma(p^* - c) - 1)\left(e^{u_0 - \gamma p^*} + e^{A_j - \gamma p_j^{RP}}\right) \\ &= G_i\left(p_i = (p^*)^+, p_j^{RP}\right) < 0 \end{aligned}$$

Since $F_i\left(p_i = p^*, p_j^{VP}\right) < 0$ and F_i is monotonically decreasing in p_i we can conclude p_i^{VP} which makes $F_i\left(p_i^{VP}, p_j^{VP}\right) = 0$ is such that $p_i^{VP} < p^*$.

Case 3. Provider i is non-value-based while provider j is value-based: If provider j is value-based it selects p^* for any given price of provider i. Moreover, for a value-based provider $\frac{\partial V_j^{RP}}{\partial p_j}\Big|_{p_j=(p^*)^+} < 0$ which is equivalent to $G_j\left(p_j = (p^*)^+, p_i = \hat{p}\right) < 0$. Comparing the expressions for $G_j\left((p^*)^+, \hat{p}\right)$ and $F_j\left(p_j^{VP}, \hat{p}\right)$ then results in

$$\frac{e^{A_j - \gamma p_j^{VP}}}{\gamma(p_j^{VP} - c) - 1} > \frac{e^{A_j - \gamma p^*}}{\gamma(p^* - c) - 1}.$$

Thus $p_j^{VP} < p^*$. Given this, similar to Case 1. we can show that $p_i^{VP} < p_i^{RP}$.

Case 4. Both providers are value-based: We can now focus on the case where both providers are value-based under reference pricing, which results in both providers pricing at p^*. We know that if a provider is value-based, its utility's derivative at $(p^*)^+ < 0$. Using the expression from Equation 2.11 in the case of having two value-based providers results in

$$\gamma(p^* - c) > \frac{e^{u_0} + e^{A_i} + e^{A_j}}{e^{u_0} + e^{A_i}}.$$

Without loss of generality let the prices of the two providers in equilibrium under the variable payments be p_i^{VP} and p_j^{VP} such that $p_j^{VP} \geq p_i^{VP}$. Then, we can show that

$$e^{u_0+A_j-\gamma\lambda p_j^{VP}} + e^{A_i+A_j-\gamma\lambda p_j^{VP}} < e^{u_0+A_j} + e^{A_i+A_j-\gamma\lambda p_i^{VP}},$$

which with some algebra results in

$$\gamma\lambda(p_j^{VP} - c) = \frac{e^{u_0} + e^{A_i-\gamma\lambda p_i^{VP}} + e^{A_j-\gamma\lambda p_j^{VP}}}{e^{u_0} + e^{A_i-\gamma\lambda p_i^{VP}}} < \frac{e^{u_0} + e^{A_i} + e^{A_j}}{e^{u_0} + e^{A_i}} < \gamma(p^* - c).$$

Note that the equality above is derived from the FOC of provider under the variable payment. From the expression above then it is easy to see that as $\lambda \to 1$, $p_j^{VP} < p^*$. Moreover, since $p_i^{VP} < p_j^{VP}$, $p_i^{VP} < p^*$.

So we were able to show that in general as $\lambda \to 1$, providers price lower under the variable payment.

From Proposition 2.8 we know that $\frac{\partial p_i^{VP}}{\partial \lambda} < 0$ for equal treatment costs. Moreover equilibrium price of providers under reference pricing is not sensitive to the value of λ and is fixed as λ changes. Thus for equal treatment costs, there can only be one value of λ (λ^*) that makes the equilibrium prices equal under the two payments and after this point $p_i^{VP} < p_i^{RP}$ (since we showed this is true as $\lambda \to 1$) and before λ^*, $p_i^{RP} < p_i^{VP}$. □

Proposition 2.10 shows that for larger values of λ a given provider prices lower under the variable payment. This is due to the fact that as λ increases, patients are responsible for larger portion of the price and hence are more price sensitive. Thus, in order for the provider to attract patients the price should be lower.

While, Proposition 2.10 indicates that the provider prices can be higher under either payment mechanism depending on the value of λ, the following proposition states that, the patients always have lower out-of-pocket under the reference pricing scheme.

Proposition 2.11 *The patient's out-of-pocket for visiting a given hospital is higher under the variable payment than under the reference pricing scheme.*

Proof. Under the reference pricing scheme, if the provider is value-based, the patient out-of-pocket is zero and always less than that of the variable payment regardless of the value of λ. Note that the out-of-pocket of the patient under the variable payment always obtain a positive value since based on the FOC of the provider j, $O_j^{VP} > \frac{1}{\gamma} + \lambda c$ at the best response.

If a provider is non-value-based, we start by comparing F_i and G_i for provider i at a given equal patient out-of-pocket $\hat{O}$ under both payment models, where

$$F_i\left(O_i = \hat{O}, O_j^{VP}\right) = e^{A_i - \gamma\hat{O}} - \left(\gamma(\hat{O} - \lambda c) - 1\right)\left(e^{u_0} + e^{A_j - \gamma O_j^{VP}}\right),$$

$$G_i\left(O_i = \hat{O}, O_j^{RP}\right) = e^{A_i - \gamma\hat{O}} - \left(\gamma(\hat{O} + p^* - c) - 1\right)\left(e^{u_0} + e^{A_j - \gamma O_j^{RP}}\right).$$

At equilibrium provider i sets these statements to zero to find the optimal price. Note that $O_j^{VP} = \lambda p_j^{VP}$, and $O_j^{RP} = p_j - p^*$. As can be seen in the statements in order to compare F_i and G_i we need to first compare the best response of the competing provider (resulting in O_j^{RP} and O_j^{VP}) when provider i has patient out-of-pocket of $\hat{O}$ under both payment models. The best response of provider j under the variable payment is computed by solving the following equation for O_j^{VP}.

$$F_j\left(O_j^{VP}, O_i = \hat{O}\right) = e^{A_j - \gamma O_j^{VP}} - \left(\gamma(O_j^{VP} - \lambda c) - 1\right)\left(e^{u_0} + e^{A_i - \gamma\hat{O}}\right) = 0$$

For the reference pricing payment model provider j can either be value-based or non-value based. If provider j is value-based, $O_j^{RP} = 0 < O_j^{VP}$. If provider j is non-value-based on the other hand, it sets $G_j = 0$ to find the best response and $O_j^{RP} > 0$.

$$G_j\left(O_j^{RP}, O_i = \hat{O}\right) = e^{A_j - \gamma O_j^{RP}} - \left(\gamma(O_j^{RP} + p^* - c) - 1\right)\left(e^{u_0} + e^{A_i - \gamma\hat{O}}\right) = 0.$$

By comparing G_j to F_j it is easy to see that in order to have

$$\frac{e^{A_j-\gamma O_j^{VP}}}{\gamma(O_j^{VP}-\lambda c)-1}=\frac{e^{A_j-\gamma O_j^{RP}}}{\gamma(O_j^{RP}+p^*-c)-1},$$

$O_j^{RP} < O_j^{VP}$ (considering that $p^* \geq c$). In other words, regardless of whether or not a provider is value-based the best response of provider j when provider i have equal patient out-of-pocket under the two payments is such that the patient out-of-pocket is lower under the reference pricing compared to the variable payment model.

We can now compare F_i and G_i under a fixed patient out-of-pocket $\hat{O}$ for both payments. Since we showed that $O_j^{RP} < O_j^{VP}$, it is easy to see that $G_i\left(\hat{O}, O_j^{RP}\right) < F_i\left(\hat{O}, O_j^{VP}\right)$ at any given $\hat{O} > \frac{1}{\gamma}+\lambda c$, and for $\hat{O} \leq \frac{1}{\gamma}+\lambda c$, $F_i\left(\hat{O}, O_j^{VP}\right) \leq G_i\left(\hat{O}, O_j^{RP}\right) > 0$ and cannot be zero.

Also from Proposition 2.5, $\partial\ F_i/\partial\ O_i|_{O_i=O_i^{VP}} < 0$, and for $p_i > p^*$, $\partial\ G_i/\partial\ O_i|_{O_i=O_i^{RP}} < 0$. Hence, when provider i is non-value-based at the points where $F_i\left(O_i^{VP}, O_j^{VP}\right) = G_i\left(O_i^{RP}, O_j^{RP}\right) = 0$ (which result in optimum prices under both payments), $O_i^{RP} < O_i^{VP}$. Note that if $\hat{O} \leq \frac{1}{\gamma}+\lambda c$, both $G_i\left(\hat{O}, O_j^{RP}\right)$ and $F_i\left(\hat{O}, O_j^{VP}\right)$ are strictly positive and cannot be zero. In other words, $G_i\left(\hat{O}, O_j^{RP}\right)$ and $F_i\left(\hat{O}, O_j^{VP}\right)$ can only be zero when $\hat{O} > \frac{1}{\gamma}+\lambda c$ under which $G_i\left(\hat{O}, O_j^{RP}\right) < F_i\left(\hat{O}, O_j^{VP}\right)$, resulting in $O_i^{RP} < O_i^{VP}$. □

Note that under the reference pricing scheme, value-based providers visits result in zero patient out-of-pocket. Hence the non-value-based providers have to provide low enough patient out-of-pocket to remain competitive in the market. On the other hand, for the variable payment case all providers collect a positive amount of patient out-of-pocket, which we show is greater than the out-of-pocket amount they can collect under the reference pricing.

2.5.3 Providers

In this section we evaluate the outcomes of payment models on provider utilities. We initially investigate the effect of insurer's decisions (p^* in the reference pricing scheme and λ in the variable payment system) on the provider's utility which enables us to draw comparison of these two models from the provider's perspective. The next proposition studies the effect of λ and p^* on the provider utility under the variable payment and reference pricing respectively.

Proposition 2.12 *Under the variable payment the utility of each provider in equilibrium is monotonically decreasing in the value of λ, while for a given set of value-based and non-value-based providers, utility of each provider in equilibrium is increasing the value of reference price (p^*).*

Proof. The expression for the utility of the provider under the variable payment is

$$V_j^{VP} = m(p_j^{VP} - c)S_j^{VP},$$

where at equilibrium can be written as

$$V_j^{VP} = m(p_j^{VP} - c)\left(1 - \frac{1}{\gamma\lambda(p_j^{VP} - c)}\right).$$

Taking the derivative of this expressin with respect to λ results in

$$\frac{\partial V_j^{VP}}{\partial \lambda} = m\left(\frac{\partial p_j^{VP}}{\partial \lambda} + \frac{1}{\gamma\lambda^2}\right). \tag{2.35}$$

Notice that $\frac{1}{\gamma\lambda} = (p_j^{VP} - c)(1 - S_j)$ at equilibrium when the costs of the two providers are equal. Moreover, we can now use the expression derived in Equation (2.27) for $\frac{\partial p_j^{VP}}{\partial \lambda}$. Replacing these expressions in Equation 2.35 and simplifying results in

$$\begin{aligned}
\frac{\partial V_j^{VP}}{\partial \lambda} &= \frac{m}{\lambda}\left(-p_j^{VP}S_j^{VP}(P) - c\left(\frac{(1-S_i^{VP}(P))\left((1-S_j^{VP}(P))^2 + S_i^{VP}(P)S_j^{VP}(P)(1-S_i^{VP}(P))\right)}{-(1-S_i^{VP}(P))(1-S_j^{VP}(P)) + (S_i^{VP}(P)S_J^{VP}(P))^2} + 1 - S_j^{VP}(P)\right)\right) \\
&< \frac{-mc}{\lambda}\left(1 + \underbrace{\frac{(1-S_i^{VP}(P))\left((1-S_j^{VP}(P))^2 + S_i^{VP}(P)S_j^{VP}(P)(1-S_i^{VP}(P))\right)}{-(1-S_i^{VP}(P))(1-S_j^{VP}(P)) + (S_i^{VP}(P)S_J^{VP}(P))^2}}_{B}\right),
\end{aligned}$$

where the inequality above comes from the fact that $p_j^{VP} > c$. Expression B is clearly negative due to negative denominator. If we can show that $B > -1$ then the proof of proposition is complete.

Note that

$$1 - S_i^{VP}(P) - S_j^{VP}(P) > (1 - S_i^{VP}(P) - S_j^{VP}(P))S_i^{VP}(P)(1 - S_i^{VP}(P))$$
$$1 - S_i^{VP}(P) - S_j^{VP}(P) + S_i^{VP}(P)S_j^{VP}(P) - S_i^{VP}(P)(1 - S_i^{VP}(P))^2 > (S_i^{VP}(P))^2 S_j^{VP}(P)$$

multiplying both sides by $-S_j^{VP}(P)$

$$-S_j^{VP}(P)(1 - S_j^{VP}(P)) + S_i^{VP}(P)S_j^{VP}(P)\left(1 - S_j^{VP}(P) + (1 - S_i^{VP}(P))^2\right) < -(S_i^{VP}(P)S_j^{VP}(P))^2$$

with some algebra results in

$$(1 - S_i^{VP}(P))\left((1 - S_j^{VP}(P))^2 + S_i^{VP}(P)S_j^{VP}(P)(1 - S_i^{VP}(P))\right)$$
$$< (1 - S_i^{VP}(P))(1 - S_j^{VP}(P)) + (S_i^{VP}(P)S_J^{VP}(P))^2$$
$$B > -1$$

For the second part of this proposition, notice that if the competing provider is value-based, the non-value-based provider always gain higher utility as p^* increases from Equation (2.28).

This means that $\frac{\partial V_j^{RP}}{\partial p^*}$ is also positive for a non-value-based provider j. It is easy to see that if both providers are value-based also $\frac{\partial V_j^{RP}}{\partial p^*} > 0$. It remains to investigate the case where provider j is value-based while the competing provider is non-value-based. In this case we have

$$\frac{\partial V_j^{RP}}{\partial p^*} = mS_j^{RP}(P)\left(1 + \gamma(p^* - c)\left(\frac{\partial p_k}{\partial p^*} - 1\right)S_k^{RP}(P)\right)$$

From our previous derivations we know that $\frac{\partial p_k}{\partial p^*} = S_k^{RP}(P)$ for this scenario. Thus we have

$$\frac{\partial V_j^{RP}}{\partial p^*} = mS_j^{RP}(P)\left(1 - \gamma(p^* - c)(1 - S_k^{RP}(P))S_k^{RP}(P)\right) > mS_j^{RP}(P)\left(1 - \gamma(p^* - c)(1 - S_k^{RP}(P))\right) > 0$$

Note that if provider k is non-value-based, then the derivative of the provider utility with respect to its price at p^* should be positive resulting in the inequality above to hold. Thus, increasing p^* results in increase in provider utility. □

Proposition 2.12, illustrates changes in provider utilities as the payment parameters change. Note that the payment parameters are determined by the insurer and is considered to be exogenous. On one hand, as the portion of patient payment under the variable payment increases, the provider utility decreases. This is due to the negative effect of λ on the providers market share. As the value of λ decreases patients become less price sensitive and consequently providers can price higher while maintaining their market share level. This results in higher utility for the providers. On the other hand, increasing the value of p^* increases the provider utility. This is due to the increase in market share of providers since fewer patients choose not to seek treatment and allowing providers to charge higher prices as the value of p^* increases.

Having established the effect of change in payment parameters on the provider utility, we can now proceed to compare the different payment schemes against one another in terms of provider utility. The following result compare the provider utility under reference pricing and the variable payment.

Proposition 2.13 *There exists a threshold $\hat{\lambda}_j$ for provider j such that, provider j gains higher utility under the variable payment system than under the reference pricing scheme iff $\lambda < \hat{\lambda}_j$, where $\hat{\lambda}_j$ is given by*

$$\hat{\lambda}_j = \frac{1}{1 + \gamma \left(p_j^{VP}(\hat{\lambda}_j) - p_j^{RP} \right)}$$

Proof. Utilities of providers under the two payment model have the following form in equilibrium.

$$V_j^{VP} = (p_j^{VP} - c) \left(1 - \frac{1}{\gamma \lambda (p_j^{VP} - c)} \right)$$

$$V_j^{RP} = (p_j^{RP} - c) \left(1 - \frac{1}{\gamma (p_j^{RP} - c)} \right)$$

Since both the price and equilibrium market share obtain lower values under the variable payment, when $\lambda \geq \lambda^*$ provider's utility also become lower under the variable payment.

We showed in Proposition 2.12, that the utility of provider is monotonically decreasing in λ. Thus there exists $\hat{\lambda} < \lambda^*$, that sets the utility of the provider equal under the two payment models and solves for

$$\hat{\lambda}\left(1+\gamma\left(p_j^{VP}(\hat{\lambda})-p_j^{RP}\right)\right)-1=0. \tag{2.36}$$

In order to show the existence of a solution to Equation (2.36) we first prove that the left hand side of Equation (2.36), is monotonically decreasing in λ. Taking the derivative of this statement with respect to λ results in

$$1-\gamma p_j^{RP}+\gamma\left(p_j^{VP}+\lambda\frac{\partial p_j^{VP}}{\partial\lambda}\right).$$

From Equation (2.27), we know that when costs are equal across different providers

$$p_j^{VP}+\lambda\frac{\partial p_j^{VP}}{\partial\lambda}=c\left(\frac{(1-S_i^{VP}(P))\left((1-S_j^{VP}(P))^2+(1-S_i^{VP}(P))S_i^{VP}(P)S_j^{VP}(P)\right)}{(1-S_i^{VP}(P))(1-S_j^{VP}(P))-(S_i^{VP}(P)S_j^{VP}(P))^2}\right).$$

Moreover, from the first order condition equation under reference pricing (Equation (2.15)), $1-\gamma p_j^{RP}<-\gamma c$ for a non-value-based provider. Thus we have

$$\begin{aligned}&1-\gamma p_j^{RP}+\gamma\left(p_j^{VP}+\lambda\frac{\partial p_j^{VP}}{\partial\lambda}\right)<\\&-\gamma c\left(1-\frac{(1-S_i^{VP}(P))\left((1-S_j^{VP}(P))^2+(1-S_i^{VP}(P))S_i^{VP}(P)S_j^{VP}(P)\right)}{(1-S_i^{VP}(P))(1-S_j^{VP}(P))-(S_i^{VP}(P)S_j^{VP}(P))^2}\right)<0,\end{aligned}$$

where the second inequality can be proved with some algebra. Hence, the left hand side of Equation (2.36), is monotonically decreasing in λ. Moreover, as $\lambda \to 1$ since $p_j^{VP} < p_j^{RP}$, it is easy to see that the left hand side has a negative value. On the other hand, as λ gets a very small value near zero $p_j^{VP} \to \overline{p}_j^{VP}$ ($\overline{p}_j^{VP}$ itself increases as λ decrease). Then the left

hand side statement will be $\gamma O_j^{VP} - 1 + o(\lambda(1 - \gamma p_j^{RP}))$, where the last term has a small value due to the small value of λ and as the result $O_j^{VP} - 1 > \gamma\lambda c > 0$. Thus, there exists a value of $\lambda = \hat{\lambda}_j$ for provider j that sets the left hand size of Equation (2.36) to zero. □

Proposition 2.13 shows that for larger values of λ there is a higher likelihood that a given provider gains larger utility under the reference pricing scheme. Note that as the value of λ increases providers price lower to attract some market share which shrinks their utility as well compared to the reference pricing case based on Proposition 2.13. Although, for smaller values of λ providers price higher under the variable payment while maintaining a high level of market share due to lower price sensitivity of patients. This results in higher utility for the providers under the variable payment. Based on the exogenous values of λ and p^* we can determine which one of the payment models is preferred from each provider's perspective.

2.6 Numerical Study

In this section we present numerical experiments that address the motivating questions formulated in the introduction of the paper and that explore the differences in outcomes for the various payment mechanisms analyzed. For our numerical study we relax some of the assumptions in our analysis as follows: (i) we consider several (more than two) competing providers to confirm our comparison results from the previous section, (ii) we endogenize the decision of the insurer under the different payment models to determine optimal values for λ, p^* and F, and (iii) we address policy questions regarding the impact of different payment models on the utilities and decisions of patients, providers, and the insurer. Note that after the insurer decision is made regarding its cost share, providers make their competitive price choices based on the insurer's decision, and finally patients select providers based on their characteristics and prices. In Section 2.6.1 we describe the input parameters used for this numerical study. Section 2.6.2 verifies our results for the case of several providers in the market. Section 2.6.3 studies the insurer decision, and Section 2.6.4 investigates the policy implications of implementing different payment models. Finally, Section 2.6.5 provides a summary of observations made from our numerical experiments.

2.6.1 Input parameters

We use parameter values corresponding to hip and knee (total) replacement, obtained from the medical and health economics literature as described below. CalPERS has adopted reference pricing for hip and knee replacement as it is a standard procedure resulting in similar per treatment provider cost in different facilities. Moreover since it is an elective procedure, patients can decide when and where to have the treatment and can shop for a provider with lower price. The providers within the CalPERS network consist of 41 value-based, and 72 non-value-based providers [21]. For the purpose of our numerical analysis we let n get the values of $2, 5, 10, 20, 50$, and 100, where the last value resembles the case of CalPERS network. For the figures presented in this section we set $n = 5$. Also we consider the population of patients seeking treatment for knee and hip replacement from these providers to be $m = 1000$.

CalPERS imposes a deductible and a 20% co-insurance up to a \$3000 maximum out-of-pocket as well as reference pricing at \$30,000 level [21]. For an expensive condition like knee and hip replacement the \$3000 maximum out-of-pocket is met by all patients regardless of the provider choice [47]. Thus, the co-insurance does not affect the patient's choice and the provider pricing decisions. Hence, with out loss of generality we eliminate the co-insurance from the reference pricing model.

There is a considerable variation in the prices that providers negotiate for knee and hip replacement. In fact data suggest that it can vary between \$15,000-\$110,000 [70]. The healthcare blue book estimate the *fair* price of knee replacement for the provider to be about \$27,000, which includes the anesthesia fee for about 2 hours and 30 minutes, physician fee for postoperative care, and provider services for implant or device and provider stay. Given that the average cost of the providers per treatment can depend on many factors, we allow the providers cost to vary in $[\$10,000, \$20,000]$. For the purpose of our numerical illustrations, we set $c = \$1.5(\times 10^4)$. Note that the reference pricing is implemented for the conditions with little variation in terms of quality of the outcome due to a standard care-delivery protocols

such as knee and hip surgeries, and MRI. We let the non-price attributes of the provider j (A_j) to have a form $a + \sigma_j b$, where σ varies across different providers as shown in Table 2.2. The variation in the provider attributes are due to comfort, accessibility, technology availability, quality, etc. There are general quality metrics publicly available through Center for Medicare and Medicaid Services (CMS) [25].

We set u_0 to be zero for the results in this section although we vary u_0 within $[0, E[A_j]]$ and the results show very little sensitivity to this value. Our results are sensitive to the values of γ and δ. Due to lack of data for these parameters we alter each of them in a wide range. Varying these parameters affects the insurer decisions on p^*, λ, and F. In most of what follows we use $\delta = 1$ and $\gamma = 2.5$.

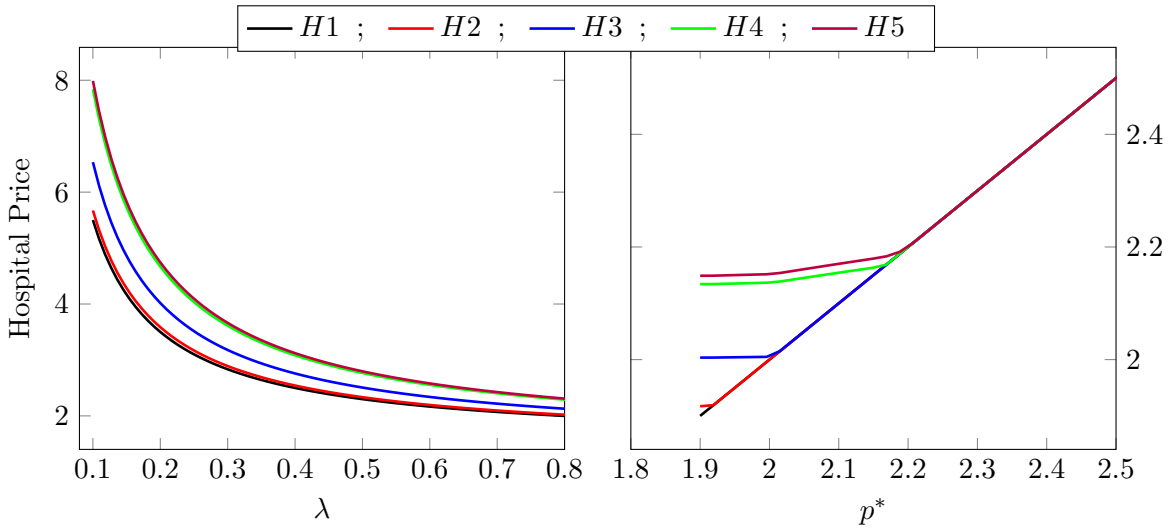
The value of the parameters that we have used is summarized in Table 2.2.

n	$\{2,5,10,20,50,100\}$ set at 5
m	1000
c	varied in $[\$1, \$2](\times 10^4)$, set at $\$1.5(\times 10^4)$
a	average value for providers' non-price attributes (set to 10×10^4
σ	$-n/2, \ldots, n/2 \in \mathbb{Z}$, if n is even; $-(n-1)/2, \ldots, (n-1)/2 \in \mathbb{Z}$, if n is odd
b	varied in $[1\ ,\ 10]$, set at 5×10^4
A	$a + \sigma b$
u_0	varied in $[0, mean(A)]$, set at 0
γ	varied in $[0.1, 3]$, set at 2.5
δ	varied in $[0.25, 1]$, set at 1

Table 2.2: Notations

2.6.2 Case of n providers

In this section we confirm the results in Section 2.5 for the case of more than two providers. Figure 2.1, shows how the provider pricing strategy is affected by insurer's decision under the



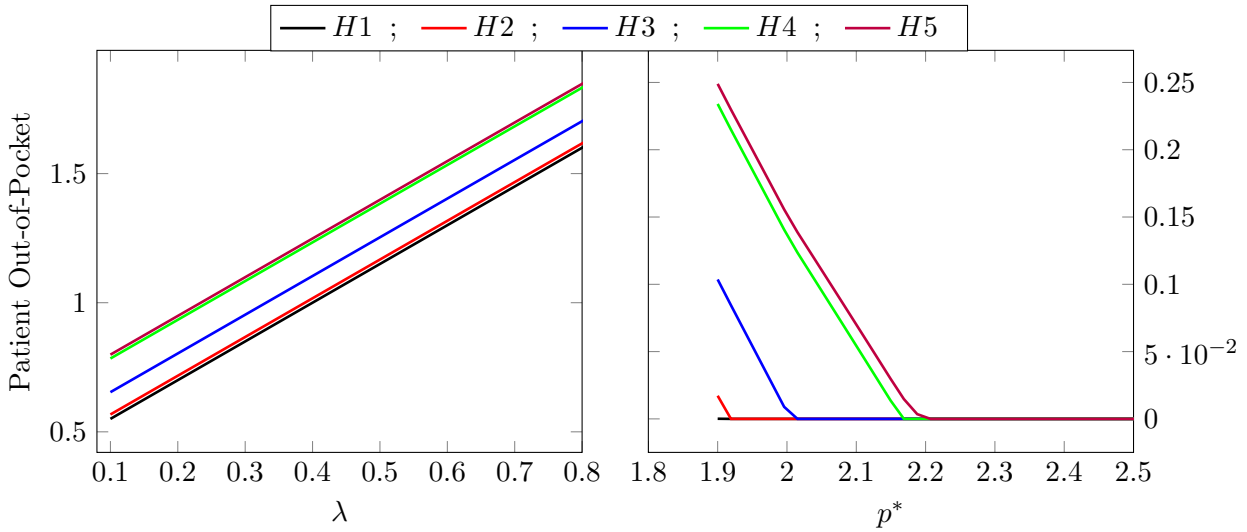
(a) Variable patient payment. (b) Reference pricing.

Figure 2.1: Hospital pricing strategy in \$10,000s for $n = 5$

variable payment and reference pricing for the case of five providers. We get similar result for different values of n. These plots confirm out results that were obtained for two providers in Proposition 2.6 (a) and (b)). Namely, we verified this result for a range of parameters discussed in Table 2.2.

Likewise, we study the effect of insurer's decision regarding the payment model parameters on patient out-of-pocket in Figure 2.2, where Figures 2.2(a) and 2.2(b) confirm the results in Proposition 2.9 (a) and (b) respectively. Notice that in Figure 2.2(b) patient out-of-pocket is zero as the provider becomes value-based when p^* increases. We verified this result for a range of parameters discussed in Table 2.2.

Figure 2.3 demonstrates the effect of parameters of the model on provider utility, which confirm the results in Proposition 2.12. In this figure since Provider 1 has a very low non-price attribute it fails to attract much of the market share even for smaller values of λ and large values of p^*.



(a) Variable patient payment. (b) Reference pricing.

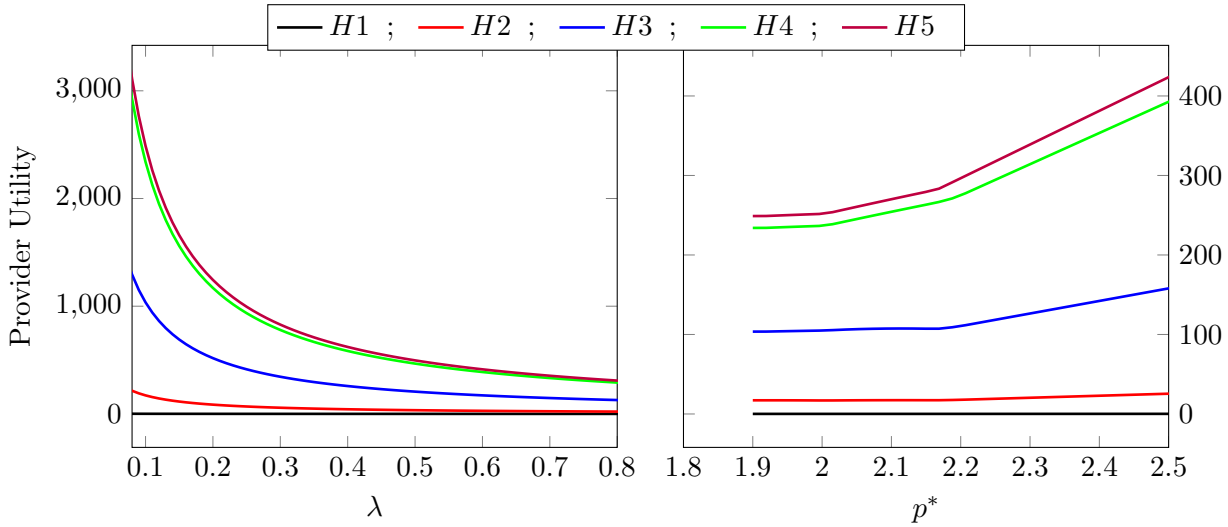
Figure 2.2: Patient out-of-pocket in \$10,000s for $n = 5$

In our numerical results we can confirm the results of Propositions 2.10, 2.11, and 2.13 for the case of having more than two providers. In other words, our comparison results structure for price, patient out-of-pocket and provider utility still hold with more than two providers.

2.6.3 Insurer decision

In this section we endogenize the decision of the insurer regarding the parameters of the model. Note that the insurer's objective include both the insurer's payment as well as the expected utility of the patients. Figure 2.4 shows how the utility of the insurer is affected by the parameters of the model. As can be seen, under the variable payment the insurer's utility is a smooth curve resulting in an intermediate value for λ that maximizes the insurer's utility. Although under the reference pricing, changing the value of p^* changes the set of value-based and non-value-based providers that creates kinks in the utility of the provider.

Figure 2.5 shows the sensitivity of the insurer's decision (λ under variable payment and

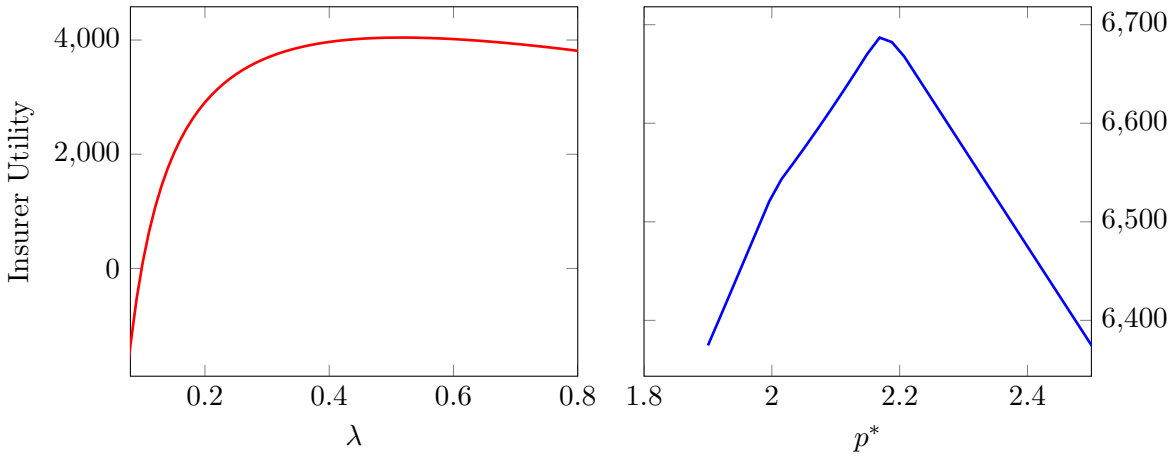


(a) Variable patient payment. (b) Reference pricing.

Figure 2.3: Provider Utility for $n = 5$

p^* under reference pricing) to the provider cost (c). In Figure 2.5(a), note that when the value of c is nearly zero, the prices of hospitals decrease to attract more market share. The patients are insensitive to the very low prices and the insurer can benefit by setting $\lambda = 1$. As the value of c gets larger providers price at a higher level and patients become more price sensitive. This forces the insurer to collaborate more in reimbursing the providers to keep the patients utility at a higher level. Thus the value of λ starts decreasing as c increases. Figure 2.5(b), illustrates a more intuitive behavior in that as the value of c increases, p^* or reference price value also increases. This is due to the increase in provider prices which leads to the insurer to increase its share of the reimbursement to keep the patient's expected utility at a higher level. Note we consider $c \leq p^*$ throughout our analysis.

Finally, Figure 2.6 demonstrates the sensitivity of the insurer decision to non-price attributes of the providers, A, as described in Table 2.2. It can be seen in Figure 2.6(a), increasing the variation of the non-price attributes leads to choosing higher values of λ.



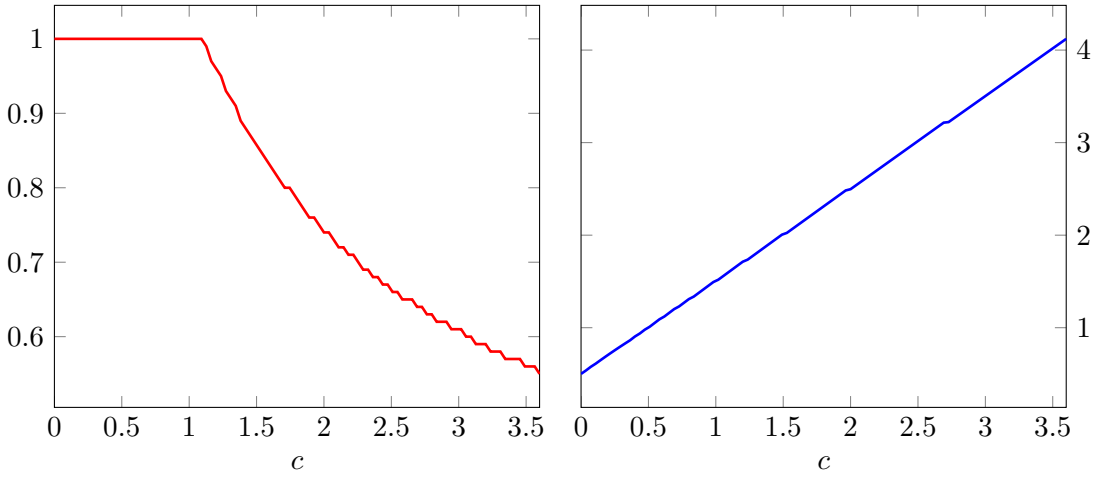
(a) Variable patient payment. (b) Reference pricing.

Figure 2.4: Insurer's utility for $n = 5$

Strictly speaking, when the patients face providers that are highly different in terms of the non-priced attributes, the insurer sets the portion of the payment that patients are responsible for at a higher level. This way the patients who seek very high attributes in providers have the option but they will be responsible for a large portion of the payment and higher out-of-pocket. Similarly, under the reference pricing scheme, as the variation increases insurer is less willing to contribute to the payment and as the result p^* decreases (Figure 2.6(b)).

2.6.4 Policy implications

In this section we investigate how implementing different payment policies can affect the decision making and outcomes for the involved stakeholders. To begin, we examine the effect of different payment models on patients expected utility. We utilized a large set of parameters to study this effect and overall we observed that the patients are better off under the reference pricing scheme for the providers with equal costs and for the numerical

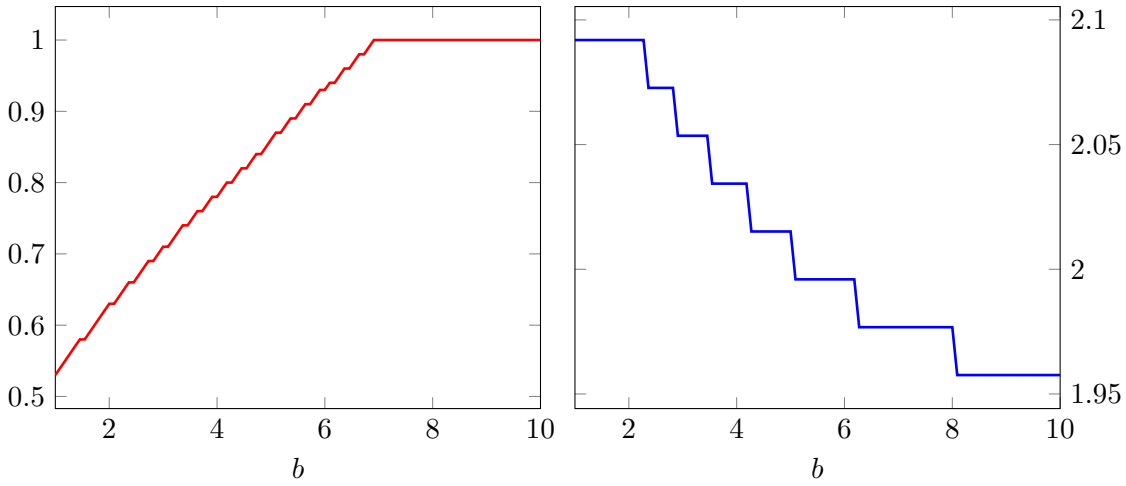


(a) Optimal choice of λ. (b) Optimal choice of p^*.

Figure 2.5: Payment model parameters sensitivity to cost for $n = 5$

values that we choose based on Table 1. In Figure 2.7 we vary the treatment cost and non-price attributes of the providers to demonstrate the dominance of reference pricing from the patient's perspective. In Figure 2.7(a) notice that the patients' expected utility is decreasing in c under both payment models. In Figure 2.7(b) the expected patient utility is increasing in variation of non-price attributes. In other words, as the providers differentiate more in A, the patients become better off. Note that patients are always better off under the reference pricing scheme.

Figure 2.8 shows how the pricing strategies of the providers are affected by the cost of treatment and the non-price attributes of the providers. Figure 2.8(a) illustrates that as the cost of treatment increases the value of λ_j^* (the point where providers price equally under RP and VP) increases for $j = 1, \ldots, 5$. This means that if an insurer chooses $\lambda > \lambda_j^*$ for provider j, then this provider prices higher under reference pricing compared to the variable payment (shaded area). The reverse holds true if $\lambda \leq \lambda_j^*$. Note that as c increases providers



(a) Optimal choice of λ. (b) Optimal choice of p^*.

Figure 2.6: Payment model parameters sensitivity to A for $n = 5$

need to increase their prices as well to be profitable. Although under the reference pricing scheme the non-value-based providers are competing with value-based providers and can only increase their prices to a certain extent. Hence the value of λ that makes the provider prices equal under the two payments should increase (remember as λ increases the provider prices decrease under VP). Of course if λ increases past this point the price under variable payment further decreases for provider j resulting in $p_j^{VP} < p_j^{RP}$.

In Figure 2.8(b), we study the effect of higher variation in the values of A on the pricing decisions under the two payment models. As the variation in non-price attributes increases the providers with lower non-price attributes have a higher disadvantage and lower their prices to attract more of the market share. Although, under reference pricing the these providers (that are mostly value-based) have more protection (zero patient out-of-pocket), and can still attract market share despite of their low non-price attributes. Thus as the providers become more differentiated λ^* decreases, to allow for these providers to price

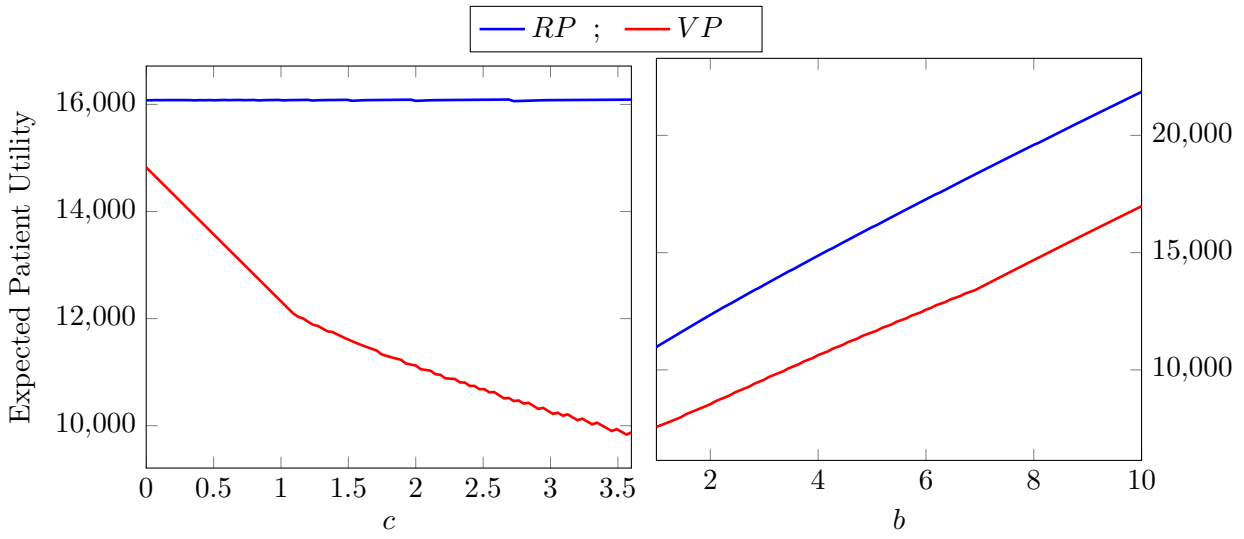


Figure 2.7: Effect of cost and non-price attributes of the providers on the expected patient utility

higher under the variable payment (at the RP level). Note that for providers with higher non-priced attributes increasing b result in higher market share and they can price higher. Although under the reference pricing this increase in price is limited due to competition with the value-based providers. Therefore, as b increases the value of λ^* for these providers increase to force providers to price lower under the variable payment and match the RP prices.

Next we consider the effect of different payment models on the utility of providers. Figure 2.9 shows the changes of $\hat{\lambda}_j$ for $j = 1, \ldots, 5$ (value of λ that makes the provider utilities equal under RP and VP), as the treatment cost and the variation in non-price attributes of the providers change. In Figure 2.9(a) as c increases the value of $\hat{\lambda}_j$ increases for all providers. Similar to the previous figure, this phenomenon can be explained by the limitation of reference pricing for non-value-based providers to increase their prices and their competition with value-based providers.

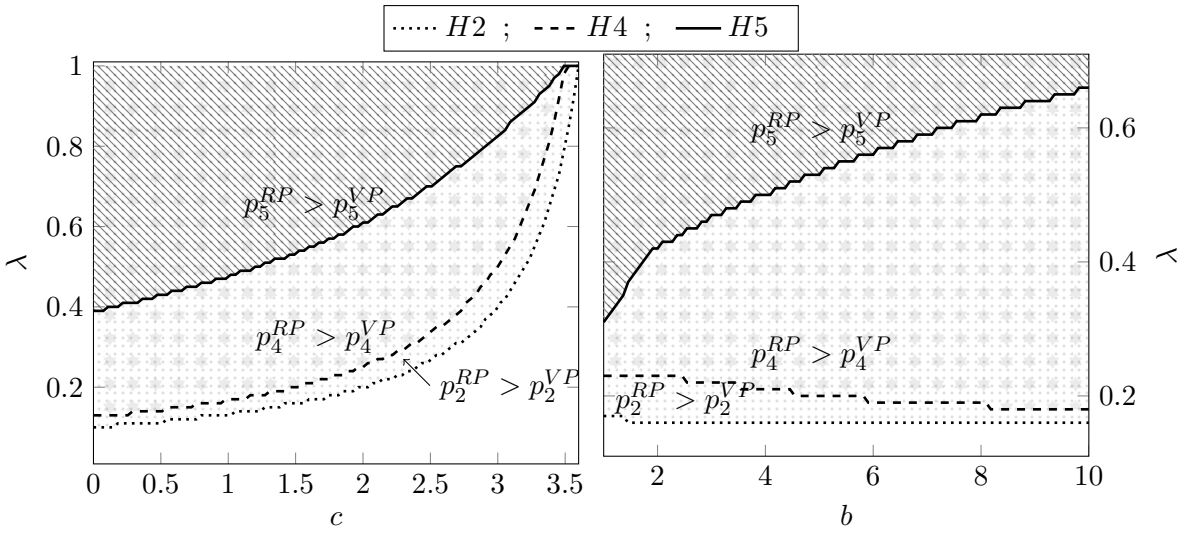


Figure 2.8: Effect of cost and non-price attributes of the providers on prices under RP and VP

In Figure 2.9(b), for providers with lower non-price attribute and for small values of b as the providers become more differentiated on their non-price attributes, they can price differentiate more easily under the variable payment which can benefit all providers. Although after a certain threshold if providers become too differentiated, under the reference pricing system the value-based providers (with lower non-price attributes) have an advantage in attracting market share. On the other hand, for providers with high value of non-price attributes increasing the variation always increases $\hat{\lambda}$, since these providers if categorized as non-value-based have a disadvantage under reference pricing as b increases due to competition with the value-based providers.

Finally, let's investigate the effect of the payment models on the insurer's utility as c and A vary. In Figure 2.10 the insurer utility is higher under the reference pricing scheme as the cost and non-price attributes of the provider change. Under reference pricing the insurer payment is a fixed amount resulting in lower risk for insurer. Moreover, patient expected utility plays a role in the utility of insurer and as mentioned before patients are in

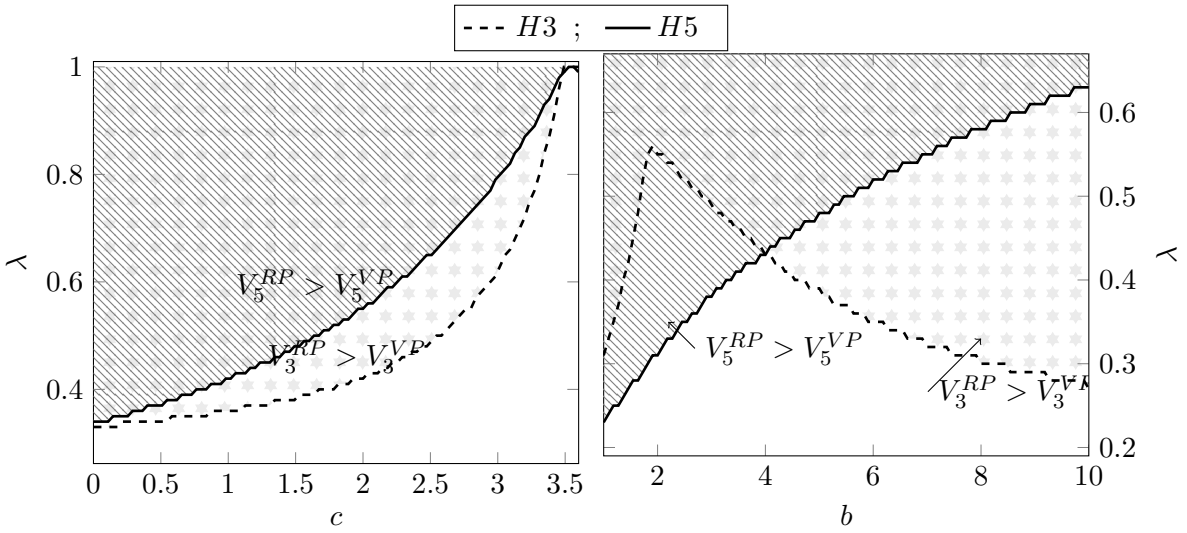


Figure 2.9: Effect of cost and non-price attributes of the providers on hospital utilities under RP and VP

general better off under the reference pricing scheme. Although there are instances where the insurer utility can be higher under the variable payment when λ is high. Notice that with high values of λ the portion of payment that the insurer is responsible is lower and we can come up with sets of parameters that can make the insurer utility higher under the variable payment scheme.

2.6.5 Summary of observations

Observation 1: Patients have lower out-of-pocket under the reference pricing scheme compared to the variable payment. Also the expected patient utility is higher under the reference pricing.

Observation 2: As treatment cost increases the set of values of λ that makes providers better off under the reference pricing decreases.

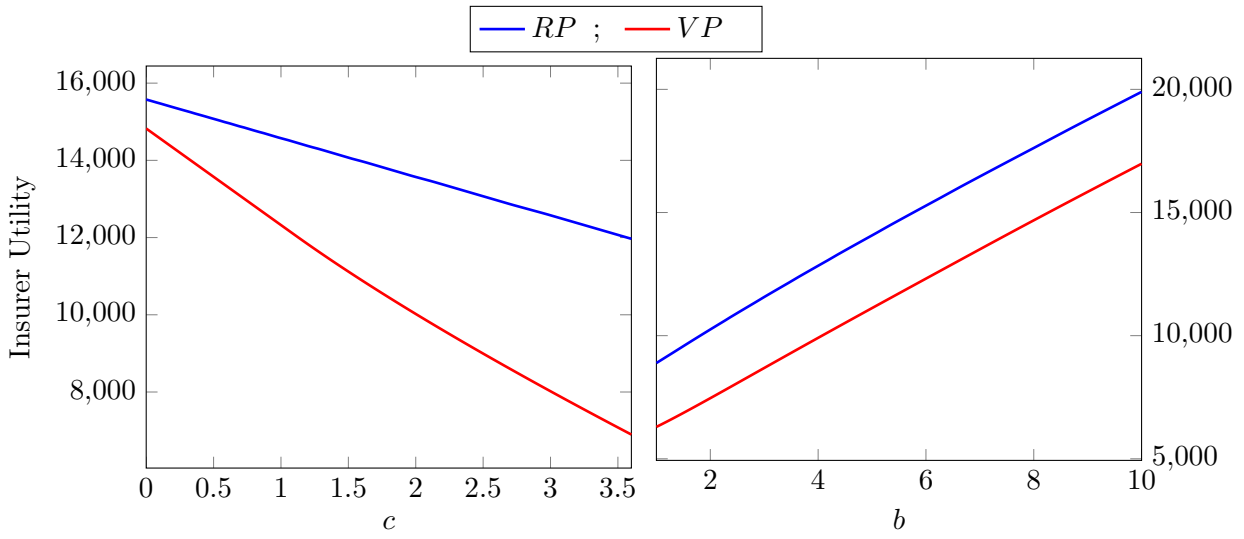


Figure 2.10: Effect of cost and non-price attributes of the providers on the insurer's utility

Observation 3: As the providers become more differentiated the set of values of λ that makes provider better off under the reference pricing can increase or decrease depending on the non-price attributes of the providers.

Observation 4: Insurer utility is higher under reference pricing for over 74% of our parameter set instances (we considered 24,000 different sets of parameters). With small values of λ insurer utility is always higher under the reference pricing.

2.7 Concluding Remarks

This chapter is one of the first attempts to study the use of reference pricing for healthcare procedures. Reference pricing has been widely implemented for pharmaceutical reimbursements, while it has been ignored for reimbursing the healthcare procedures. The current reimbursement models fail to achieve lowering the cost while maintaining high level of quality and offering treatment opportunities to beneficiaries as mentioned in previous chapter.

Reference pricing can incentivize healthcare providers to lower their costs by indirectly affecting their market share through a cost sharing mechanism with patients.

We characterize the equilibrium price and market share under reference pricing scheme as well as the current status quo, fixed payment, and variable payment models, where the fixed payment system represents the co-payment and co-insurance when maximum out-of-pocket is met and the variable payment system can be described as a co-insurance with no maximum out-of-pocket.

Our findings on the fixed payment models confirm that this payment scheme creates no incentive for the patients to choose better priced hospitals while incentivizing the providers to select extremely high prices. Reference pricing and variable payment systems on the other hand can create incentives for the patient to shop around for better-priced hospitals while taking quality considerations into account, which can result in lower hospital prices. We further compare these two payments to see which one can make each one of the involved agents better off. For the case of two providers we analytically show that 1) patients always have lower out-of-pocket under the reference pricing, 2) providers preference between these two schemes depend on the parameters of the payment model. In fact we can characterize the values for parameter models that make providers better off under one payment vs. the other.

Next we extend all our results numerically to case of n providers and study the effect of cost and non-price attributes of the providers on their pricing strategies and utilities. We observe that as treatment cost increases the set of values of λ that makes providers better off under the reference pricing decreases. Also as the hospitals become more differentiated the set of values of λ that makes provider better off under the reference pricing can increase or decrease depending on the non-price attributes of the providers. Moreover, we investigate the insurer decision regarding the parameters of different payment models and we find that for most of the parameter sets reference pricing induce higher insurer utility. In fact the implementation of reference pricing can make all the involved agents better off compared to the variable payment for a large set of parameters.

Chapter 3

IMPACT OF FACILITY LAYOUT AND PATIENT FLOW IN OUTPATIENT CLINICS ON PHYSICIANS' BEHAVIOR

3.1 Introduction

In this work we study a change of layout in an outpatient clinic. Traditionally exam rooms consist of a consult space with a desk and a computer, and an exam space with a bed and other examining equipments. The problem with the traditional layout is that it occupies the exam space (bed and equipments) for the entire visit, despite the low utilization of this section. Shadowing and observations done within the outpatient clinic indicate that on average less than 20% of the physician time with patients is spent for examination. Moreover, there are patients who do not need the exam space. In fact, most of the returning patients only require follow up consultation. In order to create an engaging and inviting environment for the patients and healthcare providers, while increasing the utilization of the exam space, the outpatient clinic has transformed some of its exam rooms to what we call *the flexible suite* in 2010 on General Internal Medicine (GIM) floor. Each flexible suite consists of two consult rooms and a shared exam room (Figure 3.1).

The flexible suite, on one hand, can enrich the physician-patient interaction, and on the other hand has a potential of increasing the utilization of the examination facilities by incorporating a shared exam room. A flexible suite is either shared between two physicians, where each physician has a consult room and they share the exam room, or the whole suite is dedicated to one physician who can multi-task. Multitasking in our context refers to a physician serving more than one customers simultaneously. Note that while a patient is preparing before and after an exam, physician has some idle time to start initial consultation of the next patient in the queue and the flexible suite can facilitate multitasking for physicians.

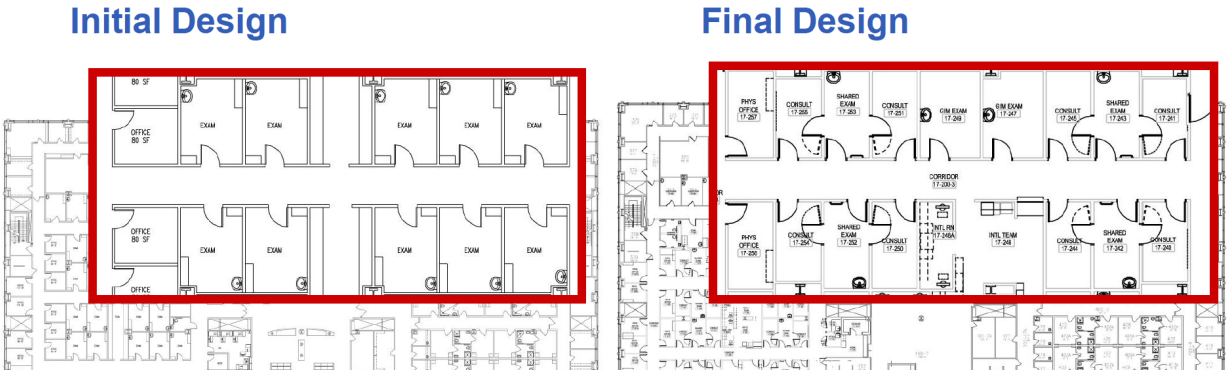


Figure 3.1: Layout change of exam rooms at the outpatient clinic

The cleaning time of the exam space after examination can create a significant idle time for physicians and delay for patients [73]. Flexible suites can help reduce this time, since the physicians can use the consult rooms while the exam room is being cleaned. Separating consult and exam rooms for a visit can also have positive implications on patient satisfaction. The examining equipment in the consulting room increases anxiety in the outpatient clinic patients. In addition, the new layout also provides higher flexibility for team-based care. Figure 3.1 shows how the layout of rooms have changed compared to the traditional exam room. As can be seen in this figure after the implementation of the change there exists both flexible suites and traditional rooms on a given floor.

All the patient visits in our study are appointment based. The patient-physician portion of an outpatient appointment can be split into three major stages: 1) interaction between the patient and a healthcare provider to get the history of the patient which involves extensive use of electronic medical records (EMR) as well as the initial consultation, 2) examination of the patient, and 3) checkout process by providing the patients pertinent information and necessary discharge instructions about their condition. Based on the three stages of physician visit, within a flexible suite the patients usually start in a consult room, then move to the shared exam room and finally move back to the consult room before discharge. In a flexible suite there may be additional wait time in the consultation room before transitioning to

the shared exam room, because of the service time variability. Found on a survey that was conducted in 2012 at the outpatient clinic, nurses seem to be neutral to the change of the layout, while the physicians prefer the new layout and medical assistants like the flexible suite model the most.

Our goal is to analyze the effect of this new layout on the patients and healthcare providers. We propose to answer the following research questions: How does the new layout affect (1) the time physicians spend with patients, (2) the patient wait time until first seen, and (3) the patients' length of stay.

3.2 Literature Review

Our work is related to two main streams of research: 1) behavioral operations specifically related to the healthcare providers, and 2) facility and process design.

The first stream of research studies how behavior of healthcare providers can change as a result of multitasking, load, social pressure, learning, fatigue, etc. There is a vast body of literature studying the impact of workload on the service time of providers. [35] provide a comprehensive review of literature related to the effect of load on these mechanisms and efficiency and productivity of service providers. The empirical evidence in the literature indicates different effects of load on server's time. [12] study the effect of increase in arrival rate on patient wait time and length of stay (LOS), where they show these measures increase as the result of increase in workload. Opposite effect of load on service time is observed in [46]. Some of the empirical literature report non-monotone impact of load on service time. [67] observe that the healthcare providers speed up as the system becomes more congested until the providers become overworked where they start slowing down (U-shaped effect on service time). This result is as opposed to observation of [107] within the context of restaurant industry, and [14] where the servers exert inverted U-shaped response to the increase in load. Finally, [15] characterize N-shaped service time response to load with two tipping points.

Part of the load for healthcare providers can be due to simultaneously serving multiple

patients (multitasking). The adverse effect of multitasking on productivity has been studied in psychology literature [97]. This adverse effect is mainly due to the mental setup cost necessary after an interruption in treatment. The effect of multitasking on server's performance is studied in [10]: the authors report a non-monotonic behavior, with small multitasking resulting in lower throughput and excessive multitasking rising to opposite effect for a recruiting firm. Similar result is observed by [68] using evidence from the emergency departments.

The behavioral effect of load and multitasking is also studied in analytical literature. Effect of congestion on speeding up and rework is analytically studied in [26], where they identified the scenarios where speeding up can help reduce congestion as well as the scenarios where it results in rework. [34] design a queuing model with load-dependent service rate, to study the effect of load and overwork on speeding up/down. [56] study collaborative processing and multitasking in a given network and their effect on unavoidable idleness of bottleneck. [22] investigate queuing models when multitasking is allowed and provide bounds on the wait time in the system.

Within the second stream of research, our paper is most closely related to [102], where the authors study the effect of queue pooling in an emergency department (ED) on wait time of patients and LOS. Since the change of the queue design from pooled system to dedicated queue system is only implemented in part of the ED in their study, they implement a difference-in-differences model to show that the wait time and LOS of the patients can be reduced by implementing a dedicated queue system for each provider, due to creating sense of ownership of the queue and resources for individual providers. Using an experimental design to test the effect of queue pooling, [100] find that pooled queue, and poor visibility of the queue can slow down the servers.

3.3 Hypothesis Development

The change in the layout of outpatient department at the outpatient clinic can have severe impact on the hospital costs associated with physician service time, patient wait time, and

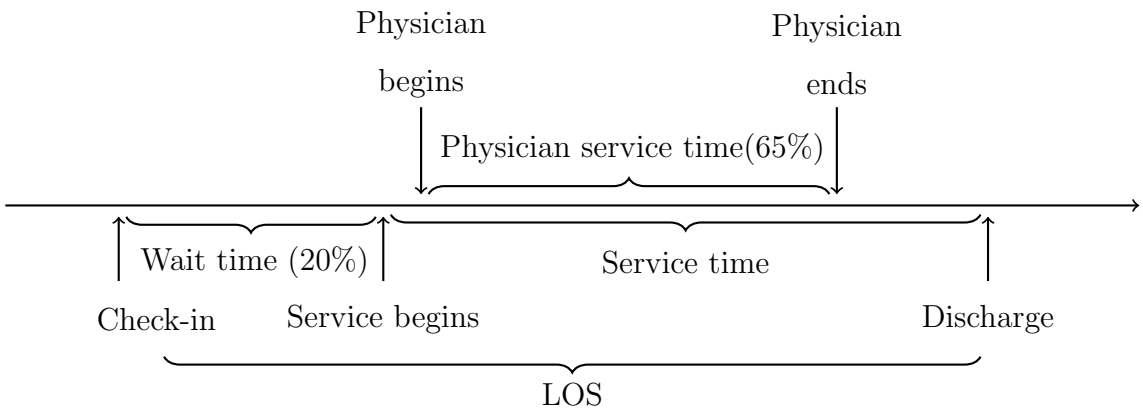


Figure 3.2: Timeline for patient flow in the outpatient clinic

length of stay. Thus studying the effect of layout change on these outcomes allows us to develop managerial insights regarding the use of flexible suites for outpatient clinics. Notice that within the new layout physicians may have an incentive to select different sets or sequence of treatments for the patients. The number of tests and procedures requested by the physician can directly impact the physician service time, the wait time, and LOS. Besides the physicians, patients' overall service time is influenced by the other providers involved in healthcare delivery such as nurses, medical assistants, etc.

Figure 3.2 illustrates the patient flow process at the outpatient department of the outpatient clinic and clarifies the definition of the dependent variables that we are interested to study in this section. LOS is defined as the time patients spend in system from check-in until discharge. Each patient starts the process with an initial wait time before the service starts which on average takes 20% of the patient LOS. Part of the service is offered by physicians which contributes to the *physician service time* (on average 65% of LOS). Interactions with nurses, medical assistants, and additional wait time may also occur during the service time.

Most of the outpatient clinics in the US contain traditional exam rooms where the physician is assigned to a patient in a room that includes both the consult and exam spaces. On the contrary, under the flexible suite layout a physician can address more than one patient

at a time by moving back and forth between rooms of a given suite (multitasking), which we observe in the outpatient clinic dataset as well. Moreover, in the case of assigning two physicians to one suite, physicians face competition for the exam room. The multitasking and competition both can impact the time that a given physician spends with patients as well as the wait time and LOS. Also, the flexible suite can allow for more efficiency in the system through reducing the cleaning time in between patients. In addition, the new layout may impact the set and/or sequence of treatments physicians choose for a given patient. We refer to this potential phenomenon as the behavioral effect.

Given no behavioral effect we expect the new layout to create extra transition time between the rooms. Additionally, multitasking and competition over the exam room, introduce additional waiting for the patients during the physician visit which can be offset by the strategic response of the provider to reduce the service time and an overall U-shaped response of the service time to multitasking [68]. We expect that our first order model has a linear form (similar to [102]), and add non-linearity to the service loads (which represent multitasking) in our robustness checks. The additional pressure on the physicians as a result of competition and multitasking as well as the travel time can incentivize physicians to choose different sets of treatment to reduce the service time of the patients (for example avoiding exams, tests and imaging as much as possible) [14]. The treatment reduction effect *may* overcome the effect of extra wait and transition time on physician service time. Hence we develop two competing hypotheses for the physician service time.

Hypothesis 1a *Physicians service time is shorter in a flexible suite as opposed to a traditional exam room.*

Hypothesis 1b *Physicians service time is longer in a flexible suite as opposed to a traditional exam room.*

The patient wait time is also affected by the physician service time. Although other

factors can impact the patient wait time, including the cleaning time necessary in between the patients and availability of other healthcare providers such as nurses, medical assistants, etc. Thus we predict the following competing hypotheses for the patient wait time.

Hypothesis 2a *Patients wait time is shorter in a flexible suite as opposed to a traditional exam room.*

Hypothesis 2b *Patients wait time is longer in a flexible suite as opposed to a traditional exam room.*

Finally, we develop a hypothesis for the length of stay which includes the initial wait time as well as the physician service time and the time in other stages of treatment that patient might be involved in (such as nurse, intake, and lobby). Clearly, LOS is affected by the patient wait time and the physician service time which can move in different directions after the intervention, and hence we develop a competing hypothesis for the patient LOS.

Hypothesis 3a *Length of Stay is shorter in a flexible suite as opposed to a traditional exam room.*

Hypothesis 3b *Length of Stay is longer in a flexible suite as opposed to a traditional exam room.*

In the following sections we investigate these hypotheses in more detail.

3.4 Data Description and Definition

In this paper we have before and after intervention dataset that includes all the traditional exam rooms and flexible suites information in the floor that we study. We have data for 33 traditional exam rooms before implementing the change, and for 16 flexible suites and 9 traditional exam rooms after the intervention. This structure of the change allows us to implement difference-in-differences (DID) model with a control group of 9 rooms without a change, and treatment group of the traditional rooms that were transformed to flexible suites. Note that there might be unobserved policy changes after the intervention. Taking the difference in differences of the outcomes cancels out all the unobserved changes that were implemented to both groups of rooms.

Patients are scheduled Monday to Friday, 9:00 am - 7:00 pm. We only focus on the floor of the clinic taking care of General Internal Medicine related conditions. We have over a year worth of data at the patient level including the before and after implementation of the layout change. We have three weeks of data before the intervention and 36 weeks after. We eliminate observations with the final stage of "no show", "pending", and "auto-discharge", which are either due to data entry errors or not useful for the current research questions. Notice that the outpatient clinic that we study is a destination medical center and most often the patients who are scheduled for a visit show up for the appointments. In fact only about 6% of the original dataset contained no shows. We checked the data to make sure there is no systematic error before eliminating these observations. We also only focus on physicians who treat more than 10 patients over the course of study. Additionally, we exclude the patient visits that do not involve a physician, for example if patients only see a nurse during their visits. This data cleaning results in a dataset with 1,281 observations before and 15,518 observations after the layout change. In order to capture the potential behavioral effect we only focus on physicians who work both before and after the intervention results in $9,061$ observations, $8,361$ of which correspond to after implementing the change. About 40% of the observations are dropped after taking this subset of data. Note that we have an unbalanced

panel data with 22 physicians, each of which has 19 to 1302 treatments throughout the course of study.

Our data from the outpatient clinic includes information about the appointment type, physician ID, patient ID, appointment time, check-in time of the patients, room numbers, and the time stamps for the stages of treatment that the patients go through for every patient visit. Although, our data does not include the time stamps for the movement of patients within one flexible suite. We are interested in studying the effect of layout change ($Room$) on outcomes such as $PhysicianServiceTime$, $PatientWaitTime$, and LOS. As illustrated in Figure 3.2, part of patient service includes interaction with the physician, which is captured in $PhysicianServiceTime$. We can infer the value for $PhysicianServiceTime$ from the duration of the stages with the physician. $PatientWaitTime$ describes the initial wait time of the patient from the check-in until the beginning of service. LOS is the entire time a given patient is in the system, from check-in until discharge (Figure 3.2).

We also measure the effect of load. We consider three different loads that can potentially affect the outcomes of interest. These include $ClinicLoad$, $PhysicianQueueLoad$, and $PhysicianServiceLoad$. $ClinicLoad_i$ measures the overall number of patients in the clinic upon arrival of patient i including patient i. Notice that some of the resources like nurses are shared among all the physicians. Thus, clinic load can have an impact on the outcomes like patient wait time and LOS. $PhysicianQueueLoad_i$ represents the number of patients in a queue of a given physician upon beginning of service for patient i. Note that physicians can only observe their queue length when they are not serving patients. Thus the observed queue length before starting the service for patient i can have an impact on the service time for patient i. By lagging the queue load the way we defined the variable, we avoid the simultaneity bias when studying the effect of intervention on the provider service time. We observe that providers are allowed to visit multiple patients at the same time: The load allowed before the start of service can impact the service time, while the service time cannot impact the load prior to the start of service. $PhysicianServiceLoad_i$ captures the number of patients who are simultaneously served by a given physician when the service of patient i begins. When

multitasking the physician can take advantage of the time that patients need for preparation for exam to start serving another patient. For wait time estimation we define the load of queue and service differently. For the waiting we define variable $ArrivalQueueLoad_i$ to describe the number of patients in the queue, and $ArrivalServiceLoad$ for the number of patients who are being served upon arrival of patient i. We count the number of patients in the queue and with the physician upon check-in which directly impacts the patient wait time, while the wait time of a patient cannot impact the load prior to check-in.

The patients move through different stages of treatment which may include, initial waiting, intake, extra waiting in the lobby, physician, and nurse that can be repeated over one visit. The number of stages for treatment of a given patient is captured with $NumStage$. Notice that all the patients go through different stages where they can interact with different healthcare providers including nurses, physicians, or medical assistants.

Room represents whether a room utilized to serve patient i is among the treatment (traditional rooms that were converted into flexible suites after intervention) or control (exam rooms that remained unchanged after the intervention) group of rooms and gets a binary value (=1 for treatment group, =0 control group). Moreover, the variable $Intervention$ describes whether a given patient visits the clinic after the intervention or not (=1 after and =0 before intervention).

In case of sharing the flexible suite with multiple physicians, it is possible that they compete over the shared resource which is the shared exam room. Competition in the flexible suite can affect outcomes and hence we also control for $Competition$. Notice that $Competition$ gets a binary value which can only be one in the flexible suite when two physicians use consultation rooms of the suite at the same time. Notice that we only want to capture the effect of competition over the shared exam room in the flexible suite.

In addition to the variables described above we control for the following variables. We control for $ApptCategory_i$. The original dataset contains the detailed appointment types. We classify them in broader categories through our communications with the outpatient clinic. Most of the appointments are the subsequent visits (SV) or the returning patients. Multi-

system evaluation appointment (ME) category includes the exams most often referred by the providers within the outpatient clinic. External multi-system evaluations (EME) refers to the self-referred and external-referred examinations. Limited exam (LE) is also included as major appointment type which includes minor exams and consultations. The rest of the appointment types are categorized as Others and do not have significant frequency.

We also control for a few time-dependent variables such as $DayOfWeek_i$ and $TimeOfDay_i$, when patient i is visited. $TimeOfDay$ is classified into morning (9:00 am - 12:00 pm), afternoon (12:00 pm - 3:00 pm), and late afternoon (after 3:00 pm) depending on the physician begin time of service for a given patient. We control for potential variation in individual characteristics of physicians by including $PhysicianID$. Table 3.1 summarizes a list of variable definitions.

Table 3.2 illustrates the descriptive statistics for the variables of interest before and after implementation of the flexible suite layout. Notice the changes in our variables of interest: The average physician service time after the intervention is approximately 70 minutes while this value is about 50 minutes before the intervention. The average wait time before the intervention is about 25 minutes, and after the layout change is about 20 minutes. Also we can observe that the average LOS of patients in the traditional exam rooms after the intervention is considerably larger than the other scenarios. Note that using the DID model we will study the relative change in the outcomes of interest resulting from the intervention. In our future research, we are going to focus on subsets of control group of rooms that are used by similar physicians before and after intervention, to address more detailed questions on how the intervention is affecting each group of rooms.

As Table 3.3 illustrates, the majority of appointments are the subsequent visits (over 45% in all categories) followed by the multi-system evaluation. Limited exams and external multi-system evaluations account for smaller portion of the appointments. Additionally, we can observe that most of the appointments are scheduled later in the afternoon. Competition only occurs in 1.36% of observations within the flexible suites after intervention. Hence, we eliminate competition effect in our models.

Variable	Definition
	Main Dependent Variables†
PhysicianServiceTime	Duration of service conducted by a physician
PatientWaitTime	Wait time of patient before the service starts
LOS	Length of stay: total time a patient spends in the system
	Independent and Control Variables
	Numerical variables
ClinicLoad	Number of all the patients in the clinic upon check-in of a patient
PhysicianQueueLoad	Number of patients in the queue upon the beginning of service for a patient
PhysicianServiceLoad	Number of patients simultaneously served as a patient starts her physician service
ArrivalQueueLoad	Number of patients in the queue observed upon arrival of a patient
ArrivalServiceLoad	Number of patients simultaneously served upon arrival of a patient
NumStage	Number of different stages of treatment for a patient
	Binary variables
Room	Room type that a patient is assigned to (= 1 treatment group, = 0 control group)
Intervention	Indicator of layout change implementation for a patient visit (= 1 after implementation, = 0 before implementation)
Competition	Competition between different physicians at the same flexible suite as a patient starts service (=0 no competition, =1 competition)
	Categorical variables
ApptCategory	Appointment category of a patient (SV, ME, EME, LE, and Other)
DayOfWeek	Day of week for a patient visit
TimeOfDay	Time of day for a patient visit
PhysicianID	ID for physician serving a patient

†All times are in minutes

Table 3.1: Variable Definition

	Treatment Group		*Control Group*	
	Before	After	Before	After
Variable	mean(s.e.)†	mean(s.e.)	mean(s.e.)	mean(s.e.)
PhysicianServiceTime	56.1(1.41)	65.9(0.43)	48.8(2.48)	73.9(0.73)
PatientWaitTime	25.3(1.52)	18.86(0.35)	24.7(2.66)	22.2(0.65)
LOS	98.8(2.09)	99.4(0.6)	87.1(4.11)	113.2(0.98)
ClinicLoad	19.8(0.36)	12.6(0.08)	20.7(0.6)	12.4(0.13)
PhysicianQueueLoad	0.24(0.023)	0.17(0.005)	0.28(0.05)	0.17(0.009)
PhysicianServiceLoad	1.2(0.02)	1.4(0.006)	1.17(0.03)	1.3(0.01)
ArrivalQueueLoad	0.24(0.02)	0.17(0.005)	0.26(0.05)	0.17(0.008)
ArrivalServiceLoad	0.44(0.02)	0.57(0.008)	0.46(0.05)	0.49(0.01)
NumStage	3.1(0.05)	2.7(0.01)	2.7(0.08)	2.8(0.02)
N	526	5,789	151	2,595

†Standard error of the mean in parentheses

Table 3.2: Summary statistics of patient visit information for numerical variables

	Treatment Group		*Control Group*	
Variable	Before	After	Before	After
Competition(=1)	0	1.36	0	0
ApptCategory: SV	44.7	47.6	56.3	46
ApptCategory: ME	41.8	42.8	34.4	42.6
ApptCategory: EME	2.9	2.4	0	5.2
ApptCategory: LE	4.9	6	9.3	5.4
DayOfWeek:Mon	14.3	17.4	19.2	21.9
DayOfWeek:Tue	17.7	20.5	17.9	26.8
DayOfWeek:Wed	25.5	21.8	17.2	14.7
DayOfWeek:Thu	23.2	22.4	36.4	21.4
TimeOfDay: Morning	19.2	20.6	23.1	15.8
TimeOfDay: Afternoon	29.8	30.3	36.4	31

Table 3.3: Percentage of sample for categorical and binary variables

By investigating the contingency table corresponding to room type and physicians we observe association between these two variables. Although, the fixed effect model should capture much of this association.

We perform a log transformation of all our time variables and compute the correlation between the numerical variables. The log transformations are indicated by L in the beginning of the names of variables. We add a small constant to our variables before transformation to avoid the natural log of zero. Table 3.4 demonstrates the correlations for the continuous variables. We observe that the correlations between the predictors to be used for a given model are small which suggests we should not be concerned with multicollinearity. Consider that the arrival queue and service load is not used with the physician queue and service load simultaneously in one model.

	1	2	3	4	5	6	7	8	9
1- LPhysicianServiceTime	1								
2- LPatientWaitTime	−0.17*	1							
3- LLOS	0.73*	0.13*	1						
4- ClinicLoad	−0.12*	0.14*	−0.07*	1					
5- PhysicianQueueLoad	−0.11*	0.10*	−0.05*	0.08*	1				
6- PhysicianServiceLoad	0.01	0.14*	0.08*	0.18*	0.08*	1			
7- ArrivalQueueLoad	−0.12*	0.17*	0.02	0.13*	0.47*	0.23*	1		
8- ArrivalServiceLoad	−0.07*	0.06*	−0.01	0.31*	0.08*	0.55*	0.1*	1	
9- numStage	0.35*	-0.34*	0.45*	-0.09*	-0.10*	-0.05*	−0.1	−0.13	1

Note: *p<0.0001

Table 3.4: Correlation matrix for continuous variables

The negative correlation between the patient wait time and the physician service time can be a result of a behavior change of providers for patients who have waited for a long time and are impatient. Physicians may want to treat these cases faster. In fact we observe that

for a subset of the data where the patient wait time is less than 10 minutes, the physician service time is positively correlated with the patient wait time.

In the following section we build econometrics models to address our hypotheses.

3.5 Econometric Specification

Before moving on to the econometric models, we take a look at the panel data for physician service time. Figure 3.3 shows the log of physician service time variations over different time stamps. Each plot demonstrates the service time for a given physician over time. The ProviderIDs are included at the top of the figure starting from the bottom-left plot and ending in top-right. The time stamps contain the phase of implementation of change (before or after) followed by the date ID and the time that physician begins service. Based on this graph we can see that there is no trend or seasonality involved in the outcomes. We select four physicians for illustration purposes in this figure, although we do not observe seasonality for the other 18 physicians either. We use Lagrange multiplier tests to study the time fixed effects in the models describing our hypotheses and we do not observe significance of such effects. Hence, we do not include time fixed-effects in our model.

In order to choose the correct econometric model, it is also interesting to look at the physician heterogeneity. Figure 3.4 shows physician heterogeneity when looking at the PatientWaitTime. We observe the presence of heterogeneity in PhysicianServiceTime and LOS as well. Using the Wooldridge's test for unobserved individual effects also confirms the existence of unobserved effects. Hence, in order to control for time-invariant heterogeneity among physicians, we introduce physician level fixed-effects in our model.

We use the following difference-in-differences model to test Hypotheses 1a and 1b.

$$\begin{aligned} LPhysicianServiceTime_{ijt} =& \alpha_0 + \alpha_1\ Room_{ij} + \alpha_2\ Intervention_t + \alpha_3\ Room_{ij} \times Intervention_t + \\ & \gamma\ \mathbf{L'_{ijt}} + \delta\ \mathbf{C_{ijt}} + \zeta \mathbf{PhysicianID_i} + \epsilon_{ijt}. \end{aligned} \tag{3.1}$$

In Equation (3.1), α_3 is the difference-in-differences estimator, indicating whether the

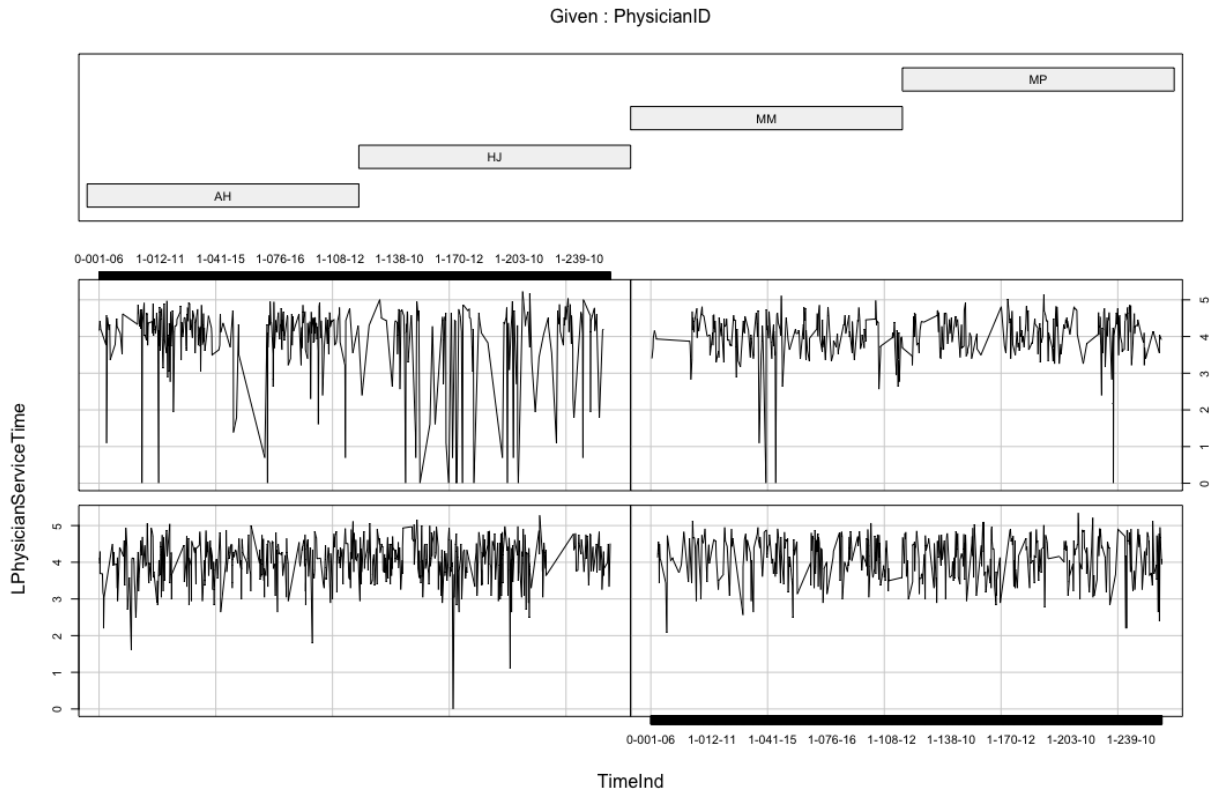


Figure 3.3: Investigating trend and seasonality for physician service time

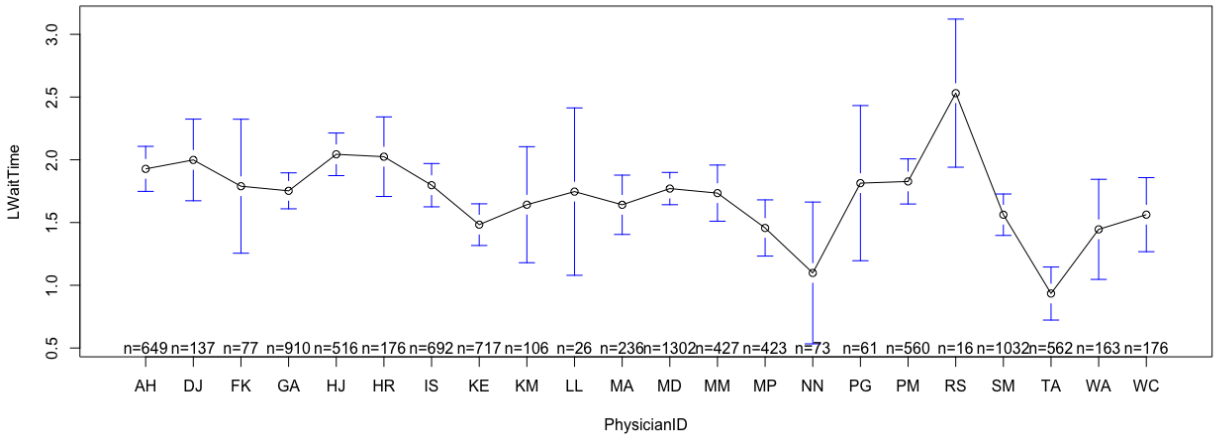


Figure 3.4: Physician heterogeneity in terms of wait time after the intervention

expected mean change in the physician service time from before to after was different in the two groups of rooms (treatment and control). $\mathbf{L'_{ijt}}$ includes the physician service and queue loads. For the regression corresponding to the physician duration we exclude $ClinicLoad_i$, since the physicians service durations are mainly affected by their own service and queue loads. $\mathbf{C_{ijt}}$ represents the controls including the $DayOfWeek$, $TimeOfDay$, $ApptCategory$, and $Numstage$. We also include $PhysicianID$ which is a vector of providers' fixed-effects in our model.

In order to investigate Hypotheses 2a and 2b, we introduce the following difference-in-differences model

$$\begin{aligned} LPatientWaitTime_{ijt} =& \beta_0 + \beta_1\ Room_{ij} + \beta_2\ Intervention_t + \beta_3\ Room_{ij} \times Intervention_t + \\ & \gamma\ \mathbf{L_{ijt}} + \delta\ \mathbf{C_{ijt}} + \zeta \mathbf{PhysicianID_i} + \epsilon_{ijt}, \end{aligned} \tag{3.2}$$

where β_3 is the difference-in-differences estimator, indicating whether the expected mean change in the patient wait time from before to after was different in the two groups. $\mathbf{L_{ijt}}$ includes clinic, and arrival queue and service loads. Notice that in Equation 3.2, we include the clinic load, since it can impact the load of the nurses, receptionists, medical assistants, etc. which has an effect on the initial wait time.

Finally we study the effect of the change in layout through a difference-in-differences model for LOS using similar model as Equation 3.2

$$\begin{aligned} LLOS_{ijt} =& \theta_0 + \theta_1\ Room_{ij} + \theta_2\ Intervention_t + \theta_3\ Room_{ij} \times Intervention_t + \\ & \gamma\ \mathcal{L}_{\mathbf{ijt}} + \delta\ \mathbf{C_{ijt}} + \zeta \mathbf{PhysicianID_i} + \epsilon_{ijt}, \end{aligned} \tag{3.3}$$

where θ_3 is the difference-in-differences estimator, indicating whether the expected mean change in the LOS from before to after was different in the two groups. Also $\mathcal{L}_{\mathbf{ijt}}$ is the clinic and provider queue and service loads.

We test the fixed-effects model against the random-effects model for our hypotheses using a Hausman test and verify that the fixed-effects model outperforms the random-effects model.

Next we derive results from the models proposed in this section and interpret the implications of layout change in the outpatient clinics.

3.6 Base Results

We begin this section by testing the validity of the results of models proposed in the previous section. First, we test for stationarity using the augmented Dickey-Fuller test for a unit root, and we observe stationarity in all the three models that we introduced (Figure 3.3). Next we implement Breusch-Pagan test for heteroskedasticity and observe evidence of heteroskedasticity in all three models. Finally, we find evidence of temporal dependence using the Wooldridge's test for serial correlation in the models testing our hypotheses. In order to deal with the heteroskedasticity and serial correlation we use Driscoll and Kraay robust estimators of the covariance matrix of coefficients. This is a non-parametric estimator of covariance matrix which is robust to the serial and cross-sectional correlations as well as heteroskedasticity. Moreover this estimator of covariance matrix is suitable for unbalanced panels and the fixed-effects models that we propose. Table 3.5 describes the fixed-effects models with robust estimate of coefficients capturing the layout change in the outpatient department of the outpatient clinic.

The interaction term describes the difference-in-differences coefficient. We observe a significant 22% decrease in the difference in average physician service time for treatment and control group of rooms after the implementation of the layout change. Strictly speaking, after the layout change in the control group of rooms the difference in provider service time is reduced. Thus the change in layout of the controlled rooms results in lower average provider service time in the flexible units. This provides strong support for Hypothesis 1, indicating that the behavioral effect of the new layout dominates the extra transition time in the flexible suites and results in strategic physicians to lower the time they spend with the patients.

On the contrary, the layout change does not have a significant impact on the wait time of patients. Remember that while the physicians may work faster in flexible suites there are other healthcare providers that are involved in delivering care and can diminish the effect of

	Dependent variable:		
	LPhysicianServiceTime	LWaitTime	LLOS
	(1)	(2)	(3)
Room	0.032 (0.063)	0.045 (0.169)	0.025 (0.047)
Intervention	0.368*** (0.061)	−0.454*** (0.162)	0.125*** (0.046)
Room×Intervention	−0.219*** (0.075)	−0.048 (0.196)	−0.093* (0.053)
ClinicLoad		0.024*** (0.004)	−0.001 (0.001)
PhysicianQueueLoad	−0.060*** (0.011)		0.022** (0.011)
PhysicianServiceLoad	0.046*** (0.013)		0.117*** (0.010)
ArrivalQueueLoad		0.585*** (0.051)	
ArrivalServiceLoad		−0.111** (0.050)	
NumStage	−0.009 (0.009)	−0.532*** (0.039)	0.077*** (0.006)
AptCategoryEME	0.523*** (0.047)	0.020 (0.156)	0.284*** (0.029)
AptCategoryME	0.520*** (0.037)	−0.033 (0.104)	0.303*** (0.020)
AptCategoryOTHER	−0.363** (0.142)	1.815*** (0.346)	0.136 (0.149)
AptCategorySV	−0.202*** (0.039)	0.771*** (0.114)	−0.167*** (0.023)
Day.of.weekMon	−0.002 (0.019)	0.120 (0.084)	0.077*** (0.017)
Day.of.weekTue	−0.002 (0.018)	0.223*** (0.080)	0.071*** (0.016)
Day.of.weekWed	−0.019 (0.019)	0.039 (0.080)	0.029* (0.017)
Day.of.weekThu	−0.014 (0.019)	0.057 (0.074)	0.037** (0.017)
TimeofDayMorning	−0.018 (0.018)	0.005 (0.069)	0.012 (0.015)
TimeofDayLateAfternoon	−0.023* (0.013)	−0.110** (0.053)	−0.005 (0.011)
Adjusted R^2	0.323	0.155	0.307
Note:			*p<0.1; **p<0.05; ***p<0.01

Table 3.5: Fixed-effects models

the lower physician time on wait time of the patients.

Additionally, we observe that the average LOS has a 9.3% reduction in the treatment and control group of rooms after the intervention at the significance level of 10%. Note that physician time accounts for 65% of LOS on average and hence can have significant effect on the overall LOS. Table 3.5 also illustrates effects of load, number of stages, and other control variable on the outcomes of interest.

Notice the effect of different appointment categories on the outcomes of interest. The subsequent visits (SV- returning patients) and Other (immunization, nurse, education services, etc.) take up significantly smaller portion of the physician time compared to the comprehensive examinations like ME and EME. The overall length of stay is affected by the physician time and we observe that the comprehensive exams have longer LOS than the subsequent visits. Also, subsequent visits and Other appointment categories tend to wait longer in the system due to their low urgency.

3.7 Explanation of Observations and Robustness Checks

We observed in the previous section that the intervention can have significant effect on the provider time spent with patients. In this section, we explore other possible explanations for this reduction in the provider time and some robustness checks for the base model.

One possible explanation of reduction in provider time after the intervention can be due to reduction in intensity of care. As observed in Table 3.4, there is a strong and significant correlation between the physician time and the number of stages for treating patients. In practice the physicians very often prescribe lab tests and imaging that can be done during one patient visit. Prescribing such tests can increase the number of stages patients are involved in and the time physicians spend with patients.

The additional pressure due to the new layout, may incentivize the physicians to skimp on care by means of reducing or eliminating some tests and consequently the number of stages. Thus, we can test whether the intervention has significantly impacted the number of stages. As mentioned in the data description section the number of stages can be two

to five, where the initial wait time is the first stage and a physician is always involved in a patient visit. We use a conditional logit model for the likelihood of having 2 to 4 stages of treatment. The logit models are conditioned on the PhysicianID to account for the physician fixed-effects. Table 3.6 summarizes the effect of intervention on the likelihood of number of stages.

	Dependent variable:			
	NumStage=2	NumStage=3	NumStage=4	NumStage=5
	(1)	(2)	(3)	(4)
Room	0.011 (0.126)	0.006 (0.429)	−0.102 (0.185)	0.068 (0.406)
Intervention	0.143 (0.118)	1.185*** (0.404)	−0.602*** (0.179)	−0.937** (0.402)
Room×Intervention	−0.064 (0.145)	−0.058 (0.443)	0.077 (0.206)	0.232 (0.448)
ClinicLoad	0.005* (0.002)	−0.007 (0.004)	−0.006* (0.004)	0.006 (0.008)
PhysicianQueueLoad	−0.007 (0.030)	−0.149* (0.084)	0.113* (0.064)	0.140 (0.142)
PhysicianServiceLoad	−0.044 (0.029)	−0.074 (0.058)	0.016 (0.050)	0.536*** (0.103)
AptCategoryEME	−0.567** (0.281)	−0.110 (0.131)	0.109 (0.118)	0.010 (0.274)
AptCategoryME	0.489*** (0.141)	−0.184** (0.081)	0.036 (0.076)	−0.026 (0.160)
AptCategoryOTHER	2.047*** (0.278)	−1.835** (0.755)	−2.902** (1.402)	−1.531 (1.870)
AptCategorySV	2.165*** (0.136)	−4.130*** (0.222)	−3.250*** (0.141)	−3.943*** (0.383)
Day.of.weekMon	−0.103** (0.052)	0.288*** (0.100)	−0.063 (0.084)	−0.317 (0.200)
Day.of.weekTue	−0.003 (0.047)	0.134 (0.100)	−0.038 (0.083)	−0.263 (0.197)
Day.of.weekWed	−0.003 (0.046)	0.284*** (0.102)	−0.149* (0.089)	−0.156 (0.204)
Day.of.weekThu	−0.001 (0.043)	0.112 (0.104)	−0.034 (0.087)	−0.007 (0.199)
TimeofDayMorning	−0.004 (0.042)	0.053 (0.089)	0.016 (0.079)	0.120 (0.167)
TimeofDayLateAfternoon	−0.021 (0.034)	0.075 (0.062)	−0.047 (0.055)	−0.097 (0.127)
R^2	0.264	0.180	0.198	0.051
Max. Possible R^2	0.999	0.902	0.942	0.420
Log Likelihood	-30,516.240	-9,591.774	-11,903.310	-2,224.210
Wald Test (df = 16)	1,934.500***	460.580***	738.110***	180.210***
LR Test (df = 16)	2,765.008***	1,791.165***	1,991.946***	473.157***
Score (Logrank) Test (df = 16)	2,489.791***	1,321.141***	1,557.033***	382.292***
Note:				*p<0.1; **p<0.05; ***p<0.01

Table 3.6: Conditional logistic regression for number of stages

As Table 3.6 illustrates, $Room \times Intervention$ is not statistically significant for the different number of stages, which suggests that the difference in likelihood of having a certain number of stages between the treatment and control group of rooms has not changed after

the implementation of intervention. Thus the effect that is observed in basic results is not due to reducing stages of care.

Next we check the robustness of our model by excluding the NumStage and PhysicianServiceLoad, which do not significantly change in terms of their averages before and after intervention. Our base results remain very robust to this change of regressors with $\alpha_3 = -0.218,\ p < 0.01$, $\beta_3 = -0.055,\ p > 0.1$, and $\theta_3 = -0.09,\ p < 0.1$ (magnitude of change 0.1%, 0.7%, and 0.3% respectively). Reduce regressors in Table 3.7 presents this result.

In order to account for any other possible change within the clinic after the implementation of the flexible suite we modify the time of the study after intervention to 3, 6, and 9 months. We observe that base result remains robust to the shorter time frames with respect to the coefficient of difference-in-differences . Table 3.7 summarizes the values of these coefficients for different time frames.

	Dependent variable:		
	α_3	β_3	θ_3
	(1)	(2)	(3)
Base Model	-0.219^{***} (0.075)	-0.048 (0.196)	-0.093^{*} (0.053)
Reduce Regressors	-0.218^{***} (0.075)	-0.055 (0.193)	-0.09^{*} (0.053)
9 months after	-0.193^{***} (0.058)	-0.003 (0.239)	-0.086^{*} (0.047)
6 months after	-0.185^{***} (0.059)	0.075 (0.252)	-0.113^{**} (0.049)
3 months after	-0.181^{***} (0.069)	0.082 (0.264)	-0.126^{**} (0.056)
Nonlinear Model	-0.225^{***} (0.074)	-0.046 (0.195)	-0.096^{*} (0.052)
Note:			*p<0.1; **p<0.05; ***p<0.01

Table 3.7: Robustness checks for difference-in-differences coefficient

Recall that the nonlinear effect of multitasking is observed in several studies (including

[107] and [68]). As a robustness check of our linear model, we add non-linearity to the service loads using a quadratic model that includes $PhysicianServiceLoad^2$ for Models 3.1 and 3.3, and $ArrivalServiceLoad^2$ for Model 3.2. Comparing the adjusted R^2 does not show any improvement in goodness-of-fit after adding the quadratic term. Moreover our difference-in-differences coefficient remained robust to this change as illustrated in Table 3.7.

3.8 Conclusion and Future Directions

In this paper we study a recent implementation of layout change at an outpatient clinic. The new layout allows for multitasking and can have behavioral effects on the healthcare providers. We use a difference-in-differences model to capture the effect of the intervention. In addition to the new layout we also study the effect of different loads in the system, number of stages, appointment types, etc.

We find a significant 22% decrease in the difference in average physician service time for treatment and control group of rooms after the implementation of the layout change. This indicates that the behavioral effect of the new layout dominates the extra travel time in the flexible units and results in strategic physicians to lower the time they spend with the patients. We did not observe a significant impact on the wait time of patients as a result of intervention. This is due to involvement of multiple healthcare providers that can diminish the effect of lower physician time on wait time of the patients. Additionally, we observe that the average LOS has a 9.3% reduction in the treatment and control group of rooms after the intervention at the significance level of 10%. This potential reduction of provider service time and LOS can help hospitals reduce their associated costs in practice under the flexible suite layout.

We report the results that are robust to heteroskedastisity, serial and cross-sectional correlation using a non-parametric robust covariance matrix estimator a la Driscoll and Kraay for panel models. Also we perform robustness checks based on the severity of visits (where the number of stages is an indication of severity), eliminating regressors, considering subsets of data for different time frames, and using bootstrap sampling for standard errors

and nonlogged outcomes. Our results remained robust to these checks.

For a future extension of this work we will study the effect of layout on readmissions as well as the learning of physicians. Notice that the number of flexible unit visits that a physicians have, may impact how fast and efficient they can treat patients.

BIBLIOGRAPHY

[1] Payment reform: Bundled episodes vs. global payments: A debate between francois de brantes and robert berenson, 2012.

[2] R. Abelson and S. Cohen. Sliver of Medicare doctors get big share of payouts. *The New York Times.*, April 2014. Published on April 9, 2014.

[3] E. Adida, H. Mamani, and S. Nassiri. Bundled payment vs. fee-for-service: Impact of payment scheme on performance. *Management Science (forthcoming)*, 2016.

[4] Elodie Adida, Hamed Mamani, and Shima Nassiri. Bundled payment vs. fee-for-service: Impact of payment scheme on performance. *Management Science*, 2016.

[5] V. Agrawal and I. Bellos. The potential of servicizing as a green business model. Georgetown McDonough School of Business Research Paper, 2013.

[6] M. Aksoy-Pierson, G. Allon, and A. Federgruen. Price competition under mixed multinomial logit demand functions. *Management Science*, 59(8):1817–1835, 2013.

[7] G. Allon and A. Federgruen. Competition in service industries. *Operations Research*, 55(1):37–55, 2007.

[8] G. Allon and A. Federgruen. Competition in service industries with segmented markets. *Management Science*, 55(4):619–634, 2009.

[9] Dimitrios Andritsos and Christopher S Tang. Incentive programs for reducing readmissions when patient care is co-produced. *Available at SSRN 2666215*, 2015.

[10] Sinan Aral, Erik Brynjolfsson, and Marshall Van Alstyne. Information, technology, and information worker productivity. *Information Systems Research*, 23(3-part-2):849–867, 2012.

[11] K. J. Arrow. Uncertainty and the welfare economics of medical care. *The American economic review*, pages 941–973, 1963.

[12] Phillip V Asaro, Lawrence M Lewis, and Stuart B Boxerman. The impact of input and output factors on emergency department throughput. *Academic Emergency Medicine*, 14(3):235–242, 2007.

[13] B. Ata, B. L. Killaly, T. L. Olsen, and R. P. Parker. On hospice operations under Medicare reimbursement policies. *Management Science*, 59(5):1027–1044, 2013.

[14] Robert J Batt and Christian Terwiesch. Doctors under load: An empirical study of state-dependent service times in emergency care. *The Wharton School, the Univ ersity of Pennsylvania, Philadelphia, PA*, 19104, 2012.

[15] Jillian A Berry Jaeker and Anita L Tucker. Past the point of speeding up: The negative effects of workload saturation on efficiency and patient severity. *Management Science*, 2016.

[16] T. Brady, A. Davies, and D. M. Gann. Creating value by delivering integrated solutions. *International Journal of Project Management*, 23(5):360–365, 2005.

[17] K.R. Brekke, I. Königbauer, and O.R. Straume. Reference pricing of pharmaceuticals. *Journal of Health Economics*, 26(3):613–642, 2007.

[18] J. Burns. Bundled payments. *Hospitals and Health Networks/AHA*, 87(4):26–31, 2013.

[19] G. P. Cachon. Supply chain coordination with contracts. In S. Graves and T. de Kok, editors, *Handbook in Operations Research and Management Science: Supply Chain Management*. Elsevier, 2003.

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

[21] California Public Employees' Retirement System. Hips and knees reference based pricing evaluation, 2013.

[22] Fernanda Campello, Armann Ingolfsson, and Robert A Shumsky. Queueing models of case managers. *Management Science*, 2016.

[23] D. W. Carlton and J. M. Perloff. *Modern Industrial Organization*. Scott, Foresman/Little, Brown Higher Education, 1990.

[24] G. M. Carter, J. P. Newhouse, and D. A. Relles. How much change in the case mix index is DRG creep? *Journal of Health Economics*, 9(4):411–428, 1990.

[25] Center for Medicare and Medicaid Services. Hospital compare, 2013.

[26] Carri W Chan, Galit Yom-Tov, and Gabriel Escobar. When to use speedup: An examination of service systems with returns. *Operations Research*, 62(2):462–482, 2014.

[27] W. O. Cleverley, J. O. Cleverley, and P. H. Song. *Essentials of health care finance*. Jones & Bartlett Publishers, 2010.

[28] Z. Cooper, S. Craig, M. Gaynor, and J. Van Reenen. The price ain't right? Hospital prices and health spending on the privately insured. *NBER Working paper 21815*, 2015.

[29] D. Cowling. Calpers reference pricing program for hip or knee replacement, 2013.

[30] P. Dahl, M. Horman, T. Pohlman, and M. Pulaski. Evaluating design-build-operate-maintain delivery as a tool for sustainability. In *Construction Research Congress*, pages 1–10, 2005.

[31] P.M. Danzon and J.D. Ketcham. *Reference pricing of pharmaceuticals for Medicare: Evidence from Germany, the Netherlands and New Zealand*, volume 7. 2004.

[32] Dartmouth Atlas Project and PerryUndem Research & Communications. *The Revolving Door: A Report on U.S. Hospital Readmissions*. Robert Wood Johnson Foundation, February 2013.

[33] Gerard Debreu. A social equilibrium existence theorem. *Proceedings of the National Academy of Sciences*, 38(10):886–893, 1952.

[34] Mohammad Delasay, Armann Ingolfsson, and Bora Kolfal. Modeling load and overwork effects in queueing systems with adaptive service rates. *Operations Research*, 64(4):867–885, 2016.

[35] Mohammad Delasay, Armann Ingolfsson, Bora Kolfal, and Kenneth Schultz. Load effect on service times. 2015.

[36] V. Denoyel, A. Thiele, and L. Alfandari. Optimal facility in-network selection for healthcare payers under reference pricing. Working paper, 2015.

[37] A. Dobson and J. E. Da Vanzo. Medicare payment bundling: Insights from claims data and policy implications, 2013.

[38] D. Dranove. Measuring costs. In Frank A. Sloan, editor, *Valuing Health Care: Costs, Benefits, and Effectiveness of Pharmaceuticals and Other Medical Technologies*, chapter 4. Cambridge University Press, 1996.

[39] L. Dyrda. 10 steps to negotiate smart bundled payment deals for orthopedic surgery. Becker's Spine Review, June 11 2012.

[40] R. P. Ellis and J. G. Fernandez. Risk selection, risk adjustment and choice: Concepts and lessons from the Americas. *International Journal of Environmental Research and Public Health*, 10(11):5299–5332, 2013.

[41] E. Emanuel, N. Tanden, S. Altman, S. Armstrong, D. Berwick, F. de Brantes, M. Calsyn, M. Chernew, J. Colmers, D. Cutler, T. Daschle, P. Egerman, B. Kocher, A. Milstein, E. Oshima Lee, J.D. Podesta, U. Reinhardt, M. Rosenthal, J. Sharfstein, S. Shortell, A. Stern, P.R. Orszag, and T. Spiro. A systemic approach to containing health care spending. *New England Journal of Medicine*, 367(10):949–954, 2012.

[42] Alain C Enthoven. Consumer-choice health plan. 1978.

[43] Alain C Enthoven. The history and principles of managed competition. *Health affairs*, 12(suppl 1):24–48, 1993.

[44] Ky Fan. Fixed-point and minimax theorems in locally convex topological linear spaces. *Proceedings of the National Academy of Sciences of the United States of America*, 38(2):121, 1952.

[45] J. Feder. Bundle with care—Rethinking Medicare incentives for post-acute care services. *New England Journal of Medicine*, 369(5):400–401, 2013.

[46] IT Franks and RJ Sury. The performance of operators in conveyor-paced work. *International Journal of Production Research*, 5(2):97–112, 1966.

[47] P. Fronstin and M. C. Roebuck. Reference pricing for health care services: A new twist on the defined contribution concept in employment-based health benefits. *EBRI Issue Brief*, (398), 2014.

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

[49] X. Gan, S. P. Sethi, and H. Yan. Coordination of supply chains with risk-averse agents. *Production and Operations Management*, 13(2):135–149, 2004.

[50] A. Gawande. Overkill. *The New Yorker.*, May 2015. Published on May 11, 2015.

[51] L. B. Gerson, A. S. Robbins, A. Garber, J. Hornberger, and G. Triadafilopoulos. A cost-effectiveness analysis of prescribing strategies in the management of gastroesophageal reflux disease. *The American Journal of Gastroenterology*, 95(2):395–407, 2000.

[52] Irving L Glicksberg. A further generalization of the kakutani fixed point theorem, with application to nash equilibrium points. *Proceedings of the American Mathematical Society*, 3(1):170–174, 1952.

[53] J. A. Guajardo, M. A. Cohen, S.-H. Kim, and S. Netessine. Impact of performance-based contracting on product reliability: An empirical analysis. *Management Science*, 58(5):961–979, 2012.

[54] Pengfei Guo, Christopher S Tang, Yulan Wang, and Ming Zhao. The impact of reimbursement policy on patient welfare, readmission rate and waiting time in a public healthcare system: Fee-for-service vs. bundled payment. *anderson.ucla.edu*, 2016.

[55] Diwakar Gupta and Mili Mehrotra. Bundled payments for healthcare services: Proposer selection and information sharing. *Operations Research*, 63(4):772–788, 2015.

[56] Itai Gurvich and Jan A Van Mieghem. Collaboration and multitasking in networks: Architectures, bottlenecks, and capacity. *Manufacturing & Service Operations Management*, 17(1):16–33, 2014.

[57] A. Y. Ha. Supplier-buyer contracting: Asymmetric cost information and cutoff level policy for buyer participation. *Naval Research Logistics*, 48(1):41–64, 2001.

[58] William L Healy, Adam J Rana, and Richard Iorio. Hospital economics of primary total knee arthroplasty at a teaching hospital. *Clinical Orthopaedics and Related Research®*, 469(1):87–94, 2011.

[59] G. R. Heudebert, R. Marks, C. M. Wilcox, and R. M. Centor. Choice of long-term strategy for the management of patients with severe esophagitis: A cost-utility analysis. *Gastroenterology*, 112:1078–1086, 1997.

[60] W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

[61] D. C. Hsia, W. M. Krushat, A. B. Fagan, J. A. Tebbutt, and R. P. Kusserow. Accuracy of diagnostic coding for Medicare patients under the prospective-payment system. *The New England Journal of Medicine*, 318(6):352–355, 1988.

[62] P. J. Huckfeldt, N. Sood, J. J. Escarce, D. C. Grabowski, and J. P. Newhouse. Effects of Medicare payment reform: Evidence from the home health interim and prospective payment systems. *Journal of Health Economics*, 34:1–18, 2014.

[63] P. S. Hussey, A. W. Mulcahy, C. Schnyer, and E. C. Schneider. Closing the quality gap: Revisiting the state of the science (vol. 1: Bundled payment: Effects on health care spending and quality). Technical report, Agency for Healthcare Research and Quality (US), 2012.

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

[65] S. H. Jain and E. Besancon. Reimbursement: Understanding how we pay for health care. In *An Introduction to Health Policy*, pages 179–187. Springer, New York, 2013.

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

[67] Diwas S Kc and Christian Terwiesch. Impact of workload on service time and pa-

tient safety: An econometric analysis of hospital operations. *Management Science*, 55(9):1486–1498, 2009.

[68] Diwas Singh Kc. Does multitasking improve performance? evidence from the emergency department. *Manufacturing & Service Operations Management*, 16(2):168–183, 2013.

[69] P. Kouvelis, Y. Xiao, and N. Yang. Pbm competition in pharmaceutical supply chain: Formulary design and drug pricing. *Manufacturing & Service Operations Management*, 17(4):511–526, 2015.

[70] A Lechner, Rebecca Gourevitch, P Ginsburg, and PROVIDINGINSIGHTSTHAT CONTRIBUTETOBETTERHEA LTHPOLICY. The potential of reference pricing to generate health care savings: lessons from a california pioneer. *Center for Studying Health System Change. HSC Research Brief*, 30, 2013.

[71] D. K. Lee and S. A. Zenios. An evidence-based incentive system for Medicare's end-stage renal disease program. *Management Science*, 58(6):1092–1105, 2012.

[72] G. López-Casasnovas and J. Puig-Junoy. Review of the literature on reference pricing. *Health policy*, 54(2):87–123, 2000.

[73] Jill Maben, Clarissa Penfold, Glenn Robert, and Peter Griffiths. Evaluating a major innovation in hospital design: Workforce implications and impact on patient and staff experiences of all single room hospital accommodation report of phase 1 findings for haciric nnru, june 2012, 2012.

[74] R. Mayes. The origins, development, and passage of Medicare's revolutionary prospective payment system. *Journal of the History of Medicine and Allied Sciences*, 62(1):21–55, 2007.

[75] M. McClellan. Hospital reimbursement incentives: An empirical analysis. *Journal of Economics & Management Strategy*, 6(1):91–128, 1997.

[76] J. M. McKoy. Obligation to provide services: A physician-public defender comparison. *Virtual Mentor / AMA Journal of Ethics*, 8(5):332–334, 2006.

[77] R. Mechanic and C. Tompkins. Lessons learned preparing for Medicare bundled payments. *New England Journal of Medicine*, 367(20):1873–1875, 2012.

[78] R. E. Mechanic and S. H. Altman. Payment reform options: Episode payment is a good place to start. *Health Affairs*, 28(2):w262–w271, 2009.

[79] MedPAC. *Approaches to bundling payment for post-acute care*, chapter 3, pages 57–88. Medicare Payment Advisory Commission, June 2013.

[80] J. P. Newhouse. Reimbursing health plans and health providers: Efficiency in production versus selection. *Journal of Economic Literature*, pages 1236–1263, 1996.

[81] J.P. Newhouse, A.M. Garber, R.P. Graham, M.A. McCoy, M. Mancher, and A. Kibria. *Variation in Health Care Spending: Target Decision Making, not Geography*. National Academies Press, 2013.

[82] D. Newman, S.T. Parente, E. Barrette, and K. Kennedy. Prices for common medical services vary substantially among the commercially insured. *Health Affairs*, 35(5):923–927, 2016.

[83] Office of Inspector General and Office of Evaluation and Inspections. Medicare hospital prospective payment system: How DRG rates are calculated and updated, 2001.

[84] E. L. Plambeck and S. A. Zenios. Performance-based incentives in a dynamic principal-agent model. *Manufacturing & Service Operations Management*, 2(3):240–263, 2000.

[85] A. Powell, S. Savin, and N. Savva. Physician workload and hospital reimbursement: Overworked physicians generate less revenue per patient. *Manufacturing & Service Operations Management*, 14(4):512–528, 2012.

[86] John W. Pratt. Risk aversion in the small and in the large. *Econometrica*, 32(1/2):122–136, 1964.

[87] B. Pyenson, S. Connor, K. Fitch, and B. Kinzbrunner. Medicare cost in matched hospice and non-hospice cohorts. *Journal of Pain and Symptom Management*, 28(3):200–210, 2004.

[88] U.E. Reinhardt. The sleeper in health care payment reform. *The New York Times.*, 2013. Published on August 2, 2013.

[89] K. F. Richards, K. H. Fisher, J. H. Flores, and B. J. Christensen. Laparoscopic Nissen fundoplication: cost, morbidity, and outcome compared with open surgery. *Surgical Laparoscopy Endoscopy and Percutaneous Techniques*, 6(2):140–143, 1996.

[90] J.C. Robinson and T.T. Brown. Increases in consumer cost sharing redirect patient volumes and reduce hospital prices for orthopedic surgery. *Health Affairs*, 32(8):1392–1397, 2013.

[91] J.C. Robinson and K. McPherson. Payers test reference pricing and centers of excellence to steer patients to low-price and high-quality providers. *Health Affairs*, 31(9):2028–2036, 2012.

[92] E. Rosenthal. The $2.7 trillion medical bill. *The New York Times.*, June 2013. Published on June 1, 2013.

[93] E. Rosenthal. The $2.7 trillion medical bill. *The New York Times.*, June 2013. Published on June 1, 2013.

[94] E. Rosenthal. As hospital prices soar, a single stitch tops $500. *The New York Times.*, December 2013. Published on December 2, 2013.

[95] E. Rosenthal. Patients' costs skyrocket; specialists' incomes soar. *The New York Times.*, January 2014. Published on January 18, 2014.

[96] M. B. Rosenthal, R. Fernandopulle, H. R. Song, and B. Landon. Paying for quality: Providers' incentives for quality improvement. *Health Affairs*, 23(2):127–141, 2004.

[97] Joshua S Rubinstein, David E Meyer, and Jeffrey E Evans. Executive control of cognitive processes in task switching. *Journal of Experimental Psychology: Human Perception and Performance*, 27(4):763, 2001.

[98] M. Schaffhauser-Linzatti, A. Zeileis, and M. S. Rauner. Effects of the austrian performance-oriented inpatient reimbursement system on treatment patterns: Illus-

trated on cases with knee-joint problems. *Central European Journal of Operations Research*, 17(3):293–314, 2009.

[99] A. Shleifer. A theory of yardstick competition. *The RAND Journal of Economics*, pages 319–327, 1985.

[100] Masha Shunko, Julie Niederhoff, and Yaroslav Rosokha. Humans are not machines: The behavioral impact of queueing design on service time. *Management Science*, 2017.

[101] M. Shwartz and M. L. Lenard. Improving economic incentives in hospital prospective payment systems through equilibrium pricing. *Management Science*, 40(6):774–787, 1994.

[102] Hummy Song, Anita L Tucker, and Karen L Murrell. The diseconomies of queue pooling: An empirical investigation of emergency department length of stay. *Management Science*, 61(12):3032–3053, 2015.

[103] N. Sood, P. J. Huckfeldt, J. J. Escarce, D. C. Grabowski, and J. P. Newhouse. Medicare's bundled payment pilot for acute and postacute care: Analysis and recommendations on where to begin. *Health Affairs*, 30(9):1708–1717, 2011.

[104] N. Sood, P. J. Huckfeldt, D. C. Grabowski, J. P. Newhouse, and J. J. Escarce. The effect of prospective payment on admission and treatment policy: Evidence from inpatient rehabilitation facilities. *Journal of Health Economics*, 32(5):965–979, 2013.

[105] Bruce A Strombom, Thomas C Buchmueller, and Paul J Feldstein. Switching costs, price sensitivity and health plan choice. *Journal of Health economics*, 21(1):89–116, 2002.

[106] D. Studdert, M. Mello, W. Sage, C. DesRoches, J. Peugh, K. Zapert, and T. Brennan. Defensive medicine among high-risk specialist physicians in a volatile malpractice environment. *The Journal of the American Medical Association*, 293(21):2609–2617, 2005.

[107] Tom Fangyun Tan and Serguei Netessine. When does the devil make work? an em-

pirical study of the impact of workload on worker productivity. *Management Science*, 60(6):1574–1593, 2014.

[108] M. W. Toffel. *Contracting for servicizing*. Harvard Business School, 2008.

[109] C. Tompkins, G. Ritter, R. Mechanic, J. Perloff, and J. Chapman. Analysis of financial risk and risk mitigation option in the Medicare Bundled Payment for Care Improvement program. Technical report, The Schneider Institutes for Health Policy and The Heller School for Social Policy and Management, Brandeis University, 2012.

[110] Kenneth Train. *Discrete choice methods with simulation*. Cambridge university press, 2003.

[111] V.-A. Truong and D. Yao. Analytical models for designing pharmaceutical contracts. Technical report, Columbia University, 2013.

[112] Van-Anh Truong. Optimal selection of medical formularies. *Journal of Revenue and Pricing Management*, 13(2):113–132, 2014.

[113] Steve J Ubl and Richard J Price. Counterpoint: joint implant prices are not the principal forces driving up the cost of joint replacement surgery. *The Journal of arthroplasty*, 2016.

[114] S. Ülkü, L. B. Toktay, and E. Yücesan. Risk ownership in contract manufacturing. *Manufacturing & Service Operations Management*, 9(3):225–241, 2007.

[115] X. Vives. *Oligopoly Pricing. Old Ideas and New Tools*. MIT Press, Cambridge, 1999.

[116] Z. K. Weng. The power of coordinated decisions for short-life-cycle products in a manufacturing and distribution supply chain. *IIE Transactions*, 31(11):1037–1049, 1999.

[117] A. L. White, M. Stoughton, and L. Feng. Servicizing: The quiet transition to extended product responsibility. Technical report, Tellus Institute, Boston, 1999.

[118] C. White and M. Eguchi. Reference pricing: A small piece of the health care price and quality puzzle. *National Institute for Health Care Reform Research Brief*, (18), 2014.

[119] Chapin White, Megan Eguchi, et al. Reference pricing: a small piece of the health care price and quality puzzle. Technical report, Mathematica Policy Research, 2014.

[120] R. Wilson. The theory of syndicates. *Econometrica:*, pages 119–132, 1968.

[121] R. Yaesoubi and S. D. Roberts. Payment contracts in a preventive health care system: A perspective from Operations Management. *Journal of Health Economics*, 30(6):1188–1196, 2011.

[122] Peter Zweifel, Friedrich Breyer, and Mathias Kifmann. *Health economics*. Springer Science & Business Media, 2009.

Appendix A

PATIENT TYPE-DEPENDENT PROBABILITY OF SUCCESS

In this section we examine scenarios where the probability of complication depends not only on the treatment level in the first stage but also on the patient type. That is, the success probability is stated as $q(t, \mu)$. To keep the model tractable and derive analytical results, we assume that the provider is risk neutral and maximizes her expected payoff. Interestingly, most of the analysis implemented in the paper and all of the managerial insights carry over to this case so long as the following two assumptions hold.

Assumption AC1 *The success probability $q(t, \mu)$ has the following properties:*

1. $\dfrac{\partial q(t,\mu)}{\partial \mu} \leq 0.$

2. $\dfrac{\partial^2 q(t,\mu)}{\partial t \partial \mu} \geq 0.$

The first part of the assumption states that the success probability is lower for potentially costlier patients. Therefore this assumption states that a less costly patient is more likely to be healthier and therefore may have a higher probability of success compared to costlier patient for the same treatment level. The second part of the assumption states that marginal increase in the treatment level is more effective for costlier patients supposedly since these are the patients who are in more need of the treatment in the first place.

We note that in this case the provider's and system's payoffs are

$$\pi^P(t) = (\kappa - 1)c_1(t) + (1 - q(t,\mu))(-T^P + (\kappa - 1)c_2), \quad \text{(provider's payoff under FFS).} \tag{A.1}$$

$$\pi^P(t) = B - c_1(t) + (1 - q(t,\mu))(-T^P - c_2), \quad \text{(provider's payoff under BP).} \tag{A.2}$$

$$\pi^P(t) = B' + (\beta - 1)c_1(t) + (1 - q(t,\mu))(-T^P + (\beta - 1)c_2), \quad \text{(provider's payoff under HP).} \tag{A.3}$$

$$\pi^S(t) = -c_1(t) + (1 - q(t,\mu))(-T^P - c_2 - T^B) + V, \quad \text{(system's payoff).} \tag{A.4}$$

Under Assumption AC1 the two key results of our paper continue to hold true as stated in Propositions AC1 and AC2 below

Proposition AC1 *For a patient of type μ, the treatment levels under the different payment settings are ranked as follows:*

$$t^{BP}(\mu) \leq t^*(\mu) \leq t^{FFS}(\mu).$$

Proposition AC2 *A hybrid system with $\beta = T^B/(T^B + T^P)$ and $B' = VT^P/(T^B + T^P)$ (i.e. $B' = V(1 - \beta)$) aligns the patient selection and treatment intensity outcomes to those of the system optimum.*

Propositions AC1 and AC2 are the equivalents of Propositions 1.11 and 1.16, respectively. In the remainder of the this section we state and derive the results required for the proof of these statements.

A.0.1 Proofs

Proof of Proposition AC1. Using Lemmas AC1, AC2, and AC3 below, the proof is similar to the proof of Proposition 1.11. □

Proof of Proposition AC2. Using Lemmas AC3 and AC4 below, the proof is similar to the proof of Proposition 1.16. □

Lemma AC1 *Under the FFS mechanism defined in (A.1), the optimal treatment intensity is* $t^{FFS} = \bar{t}$.

Proof. Using Assumption 1.4, the proof is similar to the proof of Proposition 1.1 when $\theta \to 0$. □

Lemma AC2 *Under the BP mechanism defined in (A.2),*

(a) the optimal treatment intensity is given by

$$t^{BP}(\mu) = \begin{cases} t_0 & \text{if } \underline{t} \leq t_0 \leq \bar{t}; \\ \underline{t} & \text{if } t_0 < \underline{t}; \\ \bar{t} & \text{if } t_0 > \bar{t}, \end{cases}$$

(b) costlier beneficiaries require a higher treatment intensity, and

(c) the provider may have incentives to implement patient selection.

Proof.

Part (a): The proof is similar to the proof of Proposition 1.4 when $\theta \to 0$.

Part (b): The proof is similar to the proof of Proposition 1.5 when $\theta \to 0$ with the qualification that taking derivative of (A.2) with respect to μ results in

$$\frac{\partial q(t,\mu)}{\partial t} + \left(T^P + \mu\right) \frac{\partial^2 q(t,\mu)}{\partial t\, \partial \mu} = \frac{d\ t^{BP}(\mu)}{d\ \mu} \underbrace{\left(c_1''(t) - (T^P + \mu) \frac{\partial^2 q(t,\mu)}{\partial t^2} \right)}_{\geq 0}.$$

The necessary and sufficient condition for the claim to hold true is then, $\frac{\partial q(t,\mu)}{\partial t} + \left(T^P + \mu\right) \frac{\partial^2 q(t,\mu)}{\partial t\, \partial \mu} \geq 0$. This condition is satisfied considering Assumption AC1.

Part (c): The proof is similar to the proof of Proposition 1.6 when $\theta \to 0$ with the qualification that taking the derivative of the provider's expected utility for a given μ, with respect to μ we have

$$\begin{aligned}\frac{d\, E_{c_2|\mu}\left[\pi^P(t)|\mu\right]}{d\,\mu} &= \frac{d\, t^{BP}(\mu)}{d\,\mu}\Big[-c_1'\left(t^{BP}(\mu)\right) + (T^P+\mu)\frac{\partial q(t(\mu),\mu)}{\partial t}|_{t=t^{BP}}\Big] \\ &\quad - \Big[1 - q\left(t^{BP}(\mu)\right)\Big] + (T^P+\mu)\left(\frac{\partial q(t^{BP}(\mu),\mu)}{\partial \mu}\right) \\ &= -\Big[1 - q\left(t^{BP}(\mu)\right)\Big] + (T^P+\mu)\left(\frac{\partial q(t^{BP}(\mu),\mu)}{\partial \mu}\right) < 0.\end{aligned}$$

The last inequality holds based of Assumption AC1. Therefore, the provider's expected utility for a beneficiary of a given type μ decreases with μ, and the provider may have an incentive to implement patient selection, if

$$B - c_1(t^{BP}(\overline{\mu})) - \left(T^P + \overline{\mu}\right)\left(1 - q(t^{BP}(\overline{\mu}), \overline{\mu})\right) < 0.$$

□

Lemma AC3 *Under the system optimum case defined in (A.4),*

(a) the optimal treatment intensity is given by

$$t^*(\mu) = \begin{cases} t_1 & \text{if } \underline{t} \le t_1 \le \overline{t}; \\ \underline{t} & \text{if } t_1 < \underline{t}; \\ \overline{t} & \text{if } t_1 > \overline{t}, \end{cases}$$

(b) costlier beneficiaries require a higher treatment intensity, and

(c) the central planner may have incentives to implement patient selection.

Proof. The proof is similar to the proof of Lemma AC2. □

Lemma AC4 *Under the HP mechanism defined in (A.3),*

(a) the optimal treatment intensity is given by

$$t^{HP}(\mu) = \begin{cases} t_2 & \text{if } \underline{t} \leq t_2 \leq \bar{t}; \\ \underline{t} & \text{if } t_2 < \underline{t}; \\ \bar{t} & \text{if } t_2 > \bar{t}, \end{cases}$$

(b) costlier beneficiaries require a higher treatment intensity, and

(c) the provider may have incentives to implement patient selection.

Proof. The proof is similar to the proof of Lemma AC2. □