Management Science
manuscript MS-14-01857

Bundled Payment vs. Fee-for-Service: Impact of Payment Scheme on Performance

Elodie Adida
School of Business Administration, University of California at Riverside, elodie.goodman@ucr.edu
Hamed Mamani
Foster School of Business, University of Washington, Seattle, hmamani@uw.edu
Shima Nassiri
Foster School of Business, University of Washington, Seattle, shiman@uw.edu

Healthcare reimbursements in the US have been traditionally based upon a fee-for-service (FFS) scheme, providing incentives for high *volume* of care, rather than *efficient* care. The new healthcare legislation tests new payment models that remove such incentives, such as the bundled payment (BP) system. We consider a population of patients (beneficiaries). The provider may reject patients based on the patient's cost profile, and selects the treatment intensity based on a risk-averse utility function. Treatment may result in success or failure, where failure means that unforeseen complications require further care. Our interest is in analyzing the effect of different payment schemes on outcomes such as the presence and extent of patient selection, the treatment intensity, the provider's utility and financial risk, and the total system payoff. Our results confirm that FFS provides incentives for excessive treatment intensity and results in suboptimal system payoff. We show that BP could lead to suboptimal patient selection and treatment levels that may be lower or higher than desirable for the system, with a high level of financial risk for the provider. We also find that the performance of BP is extremely sensitive to the bundled payment value and to the provider's risk aversion. The performance of both BP and FFS degrades when the provider becomes more risk averse. We design two payment systems, hybrid payment and stop-loss mechanisms, that alleviate the shortcomings of FFS and BP and may induce system optimum decisions in a complementary manner.

Key words: healthcare, payment models, bundled payment, fee-for-service, coordination
History: June 17, 2016

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 (Feder, 2013).

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. Rosenthal (2013a) 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." Rosenthal (2014) 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. Abelson and Cohen (2014) 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

[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.

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 (Burns, 2013). Since complications and readmissions after discharge do not lead to further reimbursements from the payer, providing high quality of care from the start of treatment increases the chances of making a profit for the healthcare provider (MedPAC, 2013). In addition, BP removes incentives to implement unnecessary procedures and hence is expected to lower costs (MedPAC, 2013). 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 (Feder, 2013). 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 (Burns, 2013). 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 (MedPAC, 2013; Mechanic and Tompkins, 2012). 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 (Jain and Besancon, 2013). 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 (Mayes, 2007). 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 (Hussey et al., 2012); 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 (Office of Inspector General and Office of Evaluation and Inspections, 2001), as opposed to BPCI, where the reimbursement is based on the historical cost at this specific provider (Mechanic and Tompkins, 2012). 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

[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.

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 paper 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 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.

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, 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. McClellan (1997) 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. Huckfeldt et al. (2014) 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, Sood et al. (2013) 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, Lee and Zenios (2012) 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. Hussey et al. (2012) 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. Rosenthal et al. (2004) 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, Sood et al. (2011) 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. Dobson and Da Vanzo (2013) 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, Shleifer (1985) 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. Newhouse (1996) 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, Ata et al. (2013) 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 (Plambeck and Zenios, 2000): 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, Fuloria and Zenios (2001) 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. Yaesoubi and Roberts (2011) 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, Jiang et al. (2012) study optimal contracts between a purchaser and a provider, where performance is achieved when a waiting-time target is met.

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., Carlton and Perloff (1990, chapt. 9, 10), Weng (1999), and Ha (2001)). 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,

[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 Powell et al. (2012) 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.

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 (Dahl et al., 2005; Brady et al., 2005). 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 (Agrawal and Bellos, 2013; White et al., 1999; Toffel, 2008). 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 (Guajardo et al., 2012).

3. Model

3.1. Modeling framework

In this section we outline the model framework and its assumptions. Table 1 in the Appendix 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 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 $g_\mu(.)$.

ASSUMPTION 1. The family of conditional probability distribution functions $g_\mu(.)$ has the monotone likelihood ratio property.

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. Appendix B provides some details about 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).

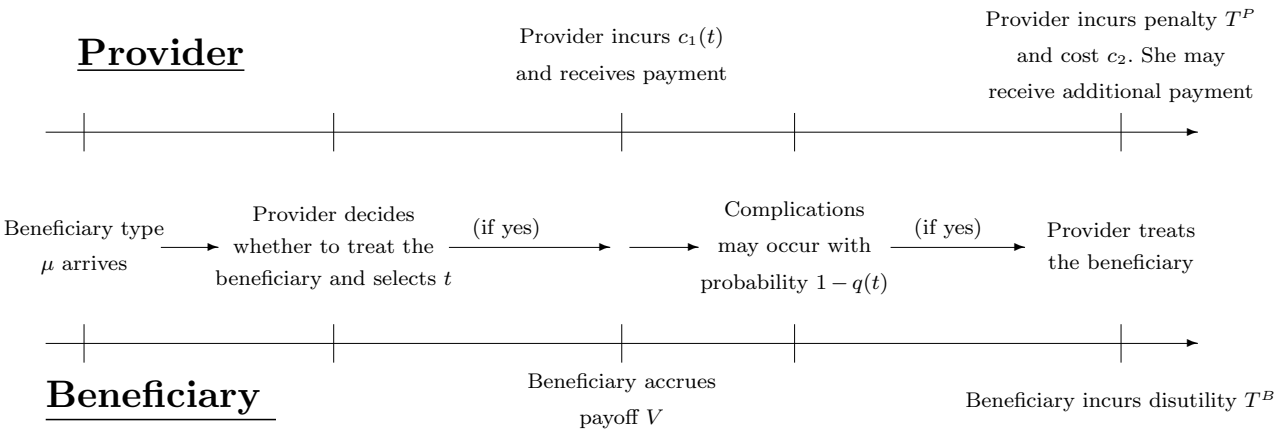


Figure 1 Sequence of events

3.2. Discussion

In this section we discuss our modeling framework and assumptions.

3.2.1. 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.

3.2.2. 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 $q(t)U^B(V) + (1 - q(t))U^B(V - T^B)$.

3.2.3. 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.

3.2.4. 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 (Pratt, 1964). 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, McKoy (2006) 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 Ellis and Fernandez (2013), "Providers 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 Studdert et al. (2005), 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, Dyrda (2012) 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.

[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 $\underline{t}$.

[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 Studdert et al. (2005) 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.

3.2.5. 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 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 4.1).

ASSUMPTION 2. First-stage cost $c_1(t)$ is increasing in t and satisfies the inequality:

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

ASSUMPTION 3. Probability of success $q(t)$ is increasing in t and satisfies the inequality:

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

These assumptions ensure that the first-stage cost c_1 is *sufficiently convex* and the probability of success is *sufficiently concave*. Assumption 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 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 (Dartmouth Atlas Project and PerryUndem Research & Communications, 2013). 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 F 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.

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.

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 (Rosenthal, 2013b). 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. J_μ is non-increasing in μ.

Proofs for technical results are provided in Appendix E.

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. 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

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

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 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.

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 3. Under FFS, the provider may have incentives to implement reverse patient selection: if

$$1-e^{-\theta(\kappa-1)c_1(\bar{t})}\left(q(\bar{t})+(1-q(\bar{t}))J_{\underline{\mu}}\right)<0 \tag{2}$$

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

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

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 (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 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,

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

where μ_1 is defined by (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 (Mechanic and Tompkins, 2012). We denote by γ the discount rate, so that B is given by:

$$B=(1-\gamma)E_\mu\left[E_{c_2|\mu}[\kappa c_1(t^{FFS}(\mu))+(1-q(t^{FFS}(\mu)))\kappa c_2|\mu]\right]=(1-\gamma)\kappa\left(c_1(\bar{t})+(1-q(\bar{t}))E[\mu]\right). \tag{3}$$

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

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

PROPOSITION 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

$$\theta c_1'(t)\left[1-q(t)+\frac{1}{I_\mu-1}\right] = q'(t). \tag{4}$$

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, 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 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 5. Under BP, potentially costlier beneficiaries require a higher treatment intensity.

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 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

$$1-e^{-\theta(B-c_1(t^{BP}(\mu^{BP})))}\left[q(t^{BP}(\mu^{BP}))+(1-q(t^{BP}(\mu^{BP})))I_{\mu^{BP}}\right]=0. \tag{5}$$

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.

4.3. Benchmark: the system optimum

Dranove (1996, 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,

[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

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 Gan et al. (2004). Indeed, the sum of the agents' expected 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, Gan et al. (2004) 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 Wilson (1968). Wilson (1968) 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. Arrow (1963) 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.

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 $\underline{t}$.

PROPOSITION 7. A treatment and patient selection decision is Pareto-optimal if and only if the system's total expected payoff is maximized.

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 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

$$c_1'(t) = (T^P + \mu + T^B)q'(t). \tag{6}$$

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

PROPOSITION 9. At the system optimum, potentially costlier beneficiaries require a higher treatment intensity.

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 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

$$V - T^P - \overline{\mu} - T^B - c_1(t^*(\overline{\mu})) + (T^P + \overline{\mu} + T^B)q(t^*(\overline{\mu})) < 0, \tag{7}$$

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

$$V - T^P - \mu^* - T^B - c_1(t^*(\mu^*)) + (T^P + \mu^* + T^B)q(t^*(\mu^*)) = 0. \tag{8}$$

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.

4.4. Discussion

4.4.1. 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 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:

$$t^{BP}(\mu) \leq t^*(\mu) \leq t^{FFS}(\mu). \tag{9}$$

The left inequality in (9) 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; Cachon (2003) 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. 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:

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

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 2 in the Appendix. 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 2a for lower values of θ), the provider prioritizes the increase in the chance of success, but otherwise (as in Figure 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 that this concern is valid when providers are risk neutral or not very risk averse (and possibly also in other cases). Figure 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).

4.4.2. Patient selection

As noted in Section 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 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.

[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 2b in the Appendix). In other cases (e.g., Figure 2a in the Appendix), there is a threshold which can be found by finding the smallest solution θ to the equality $t^{BP}(\mu) = t^*(\mu)$.

[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 $\bar{I}_\mu = \int_{\underline{c}}^{\bar{c}} (c_2 + T^P) e^{\theta(c_2 + T^P)} g_\mu(c_2) dc_2$ and $t^{BP}(\mu)$ solves (4). Note that I_μ and $\bar{I}_\mu$ depend on θ.

PROPOSITION 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.

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.

4.4.3. 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. Tompkins et al. (2012) 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 (Mechanic and Tompkins, 2012); 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:

$$\overline{w^P} = \frac{1}{N'}\sum_{i=1}^{N'} w_i^P = B - \frac{1}{N'}\sum_{i=1}^{N'}[c_1(t^{BP}(\mu^i)) + (c_2^i + T^P)(1 - q(t^{BP}(\mu^i)))]. \tag{10}$$

Following the approach adopted by Gan et al. (2004), 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 (Hoeffding, 1963), we find a relationship between ρ, α and the accepted beneficiary population size N'[9] . The next result formalizes this argument.

Proposition 13. Let the size of the accepted population be[10]

$$N' = \frac{(\zeta - \xi)^2}{2(B - \delta - \rho)^2}\ \ln\left(\frac{1}{\alpha}\right), \tag{11}$$

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 (10) and

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

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 13.

Corollary 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.

[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.

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 5.1. In Section 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 (Tompkins et al., 2012).

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 Feder (2013) (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 3. $L_\mu > 1$ and L_μ is non-decreasing in μ.

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

PROPOSITION 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

$$\theta(1-\beta)c_1'(t)\left[1-q(t)+\frac{1}{L_\mu - 1}\right] = q'(t). \tag{12}$$

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 5 and 6 and is thus omitted).

PROPOSITION 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 $\mu \ge \mu^{HP}$ where μ^{HP} is such that

$$1 - e^{-\theta(B'-(1-\beta)c_1(t^{HP}(\mu^{HP})))}\left[q(t^{HP}(\mu^{HP})) + (1-q(t^{HP}(\mu^{HP})))L_{\mu^{HP}}\right] = 0. \tag{13}$$

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 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.

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, this will for example be true when the risk aversion is small.

PROPOSITION 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.

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.

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 (Tompkins et al., 2012). 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 18. When $t^*(\mu) \le t^{BP}(\mu)$, 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

$$q(t^*(\mu)) \le \frac{e^{\underline{c}+T^P}}{e^{\underline{c}+T^P}-1} - \frac{1}{\theta(T^P+T^B+\mu)}. \tag{14}$$

The result above complements the coordination result of Proposition 17 as unlike the HP system, which could only coordinate the system when $t^{BP}(\mu) \le t^*(\mu)$, Proposition 18 shows that it is possible to align the provider's incentives to the system optimum using a stop-loss mechanism when $t^{BP}(\mu) \ge t^*(\mu)$. 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 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 17 and 18 provide necessary and sufficient conditions for the coordinating payments, we have the following corollary.

Corollary 3. There are no hybrid payment or stop-loss protection mechanisms that can coordinate the system if *both* of the following conditions hold:

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

6. Summary of findings from the numerical experiments

Appendix D provides the details of numerical experiments that address the motivating questions formulated in the introduction and that 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.

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.

References

Abelson, R., S. Cohen. 2014. Sliver of Medicare doctors get big share of payouts. *The New York Times.* Published on April 9, 2014.

Agrawal, V., I. Bellos. 2013. The potential of servicizing as a green business model. Georgetown McDonough School of Business Research Paper.

Arrow, K. J. 1963. Uncertainty and the welfare economics of medical care. *The American economic review* 941–973.

Ata, B., B. L. Killaly, T. L. Olsen, R. P. Parker. 2013. On hospice operations under Medicare reimbursement policies. *Management Science* **59**(5) 1027–1044.

Brady, T., A. Davies, D. M. Gann. 2005. Creating value by delivering integrated solutions. *International Journal of Project Management* **23**(5) 360–365.

Burns, J. 2013. Bundled payments. *Hospitals and Health Networks/AHA* **87**(4) 26–31.

Cachon, G. P. 2003. Supply chain coordination with contracts. S. Graves, T. de Kok, eds., *Handbook in Operations Research and Management Science: Supply Chain Management*. Elsevier.

Carlton, D. W., J. M. Perloff. 1990. *Modern Industrial Organization*. Scott, Foresman/Little, Brown Higher Education.

Dahl, P., M. Horman, T. Pohlman, M. Pulaski. 2005. Evaluating design-build-operate-maintain delivery as a tool for sustainability. *Construction Research Congress*. 1–10.

Dartmouth Atlas Project, PerryUndem Research & Communications. 2013. *The Revolving Door: A Report on U.S. Hospital Readmissions*. Robert Wood Johnson Foundation.

Dobson, A., J. E. Da Vanzo. 2013. Medicare payment bundling: Insights from claims data and policy implications. URL `http://www.aha.org/research/reports/12bundling.shtml`.

Dranove, D. 1996. Measuring costs. Frank A. Sloan, ed., *Valuing Health Care: Costs, Benefits, and Effectiveness of Pharmaceuticals and Other Medical Technologies*, chap. 4. Cambridge University Press.

Dyrda, L. 2012. 10 steps to negotiate smart bundled payment deals for orthopedic surgery. Becker's Spine Review. URL `http://www.beckersspine.com/orthopedic-spine-practices-improving-profits/item/12220-10-steps-to-negotiate-smart-bundled-payment-deals-for-orthopedic-surgery`.

Ellis, R. P., J. G. Fernandez. 2013. Risk selection, risk adjustment and choice: Concepts and lessons from the Americas. *International Journal of Environmental Research and Public Health* **10**(11) 5299–5332.

Feder, J. 2013. Bundle with care—Rethinking Medicare incentives for post-acute care services. *New England Journal of Medicine* **369**(5) 400–401.

Fuloria, P. C., S. A. Zenios. 2001. Outcomes-adjusted reimbursement in a health-care delivery system. *Management Science* **47**(6) 735–751.

Gan, X., S. P. Sethi, H. Yan. 2004. Coordination of supply chains with risk-averse agents. *Production and Operations Management* **13**(2) 135–149.

Gerson, L. B., A. S. Robbins, A. Garber, J. Hornberger, G. Triadafilopoulos. 2000. A cost-effectiveness analysis of prescribing strategies in the management of gastroesophageal reflux disease. *The American Journal of Gastroenterology* **95**(2) 395–407.

Guajardo, J. A., M. A. Cohen, S.-H. Kim, S. Netessine. 2012. Impact of performance-based contracting on product reliability: An empirical analysis. *Management Science* **58**(5) 961–979.

Ha, A. Y. 2001. Supplier-buyer contracting: Asymmetric cost information and cutoff level policy for buyer participation. *Naval Research Logistics* **48**(1) 41–64.

Heudebert, G. R., R. Marks, C. M. Wilcox, R. M. Centor. 1997. Choice of long-term strategy for the management of patients with severe esophagitis: A cost-utility analysis. *Gastroenterology* **112** 1078–1086.

Hoeffding, W. 1963. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association* **58**(301) 13–30.

Huckfeldt, P. J., N. Sood, J. J. Escarce, D. C. Grabowski, J. P. Newhouse. 2014. Effects of Medicare payment reform: Evidence from the home health interim and prospective payment systems. *Journal of Health Economics* **34** 1–18.

Hussey, P. S., A. W. Mulcahy, C. Schnyer, E. C. Schneider. 2012. Closing the quality gap: Revisiting the state of the science (vol. 1: Bundled payment: Effects on health care spending and quality). Tech. rep., Agency for Healthcare Research and Quality (US).

Jain, S. H., E. Besancon. 2013. Reimbursement: Understanding how we pay for health care. *An Introduction to Health Policy*. Springer, New York, 179–187.

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

Lee, D. K., S. A. Zenios. 2012. An evidence-based incentive system for Medicare's end-stage renal disease program. *Management Science* **58**(6) 1092–1105.

Mayes, R. 2007. 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.

McClellan, M. 1997. Hospital reimbursement incentives: An empirical analysis. *Journal of Economics & Management Strategy* **6**(1) 91–128.

McKoy, J. M. 2006. Obligation to provide services: A physician-public defender comparison. *Virtual Mentor / AMA Journal of Ethics* **8**(5) 332–334.

Mechanic, R., C. Tompkins. 2012. Lessons learned preparing for Medicare bundled payments. *New England Journal of Medicine* **367**(20) 1873–1875.

MedPAC. 2013. *Approaches to bundling payment for post-acute care*, chap. 3. Medicare Payment Advisory Commission, 57–88.

Newhouse, J. P. 1996. Reimbursing health plans and health providers: Efficiency in production versus selection. *Journal of Economic Literature* 1236–1263.

Office of Inspector General and Office of Evaluation and Inspections. 2001. Medicare hospital prospective payment system: How DRG rates are calculated and updated. URL `http://oig.hhs.gov/oei/reports/oei-09-00-00200.pdf`.

Plambeck, E. L., S. A. Zenios. 2000. Performance-based incentives in a dynamic principal-agent model. *Manufacturing & Service Operations Management* **2**(3) 240–263.

Powell, A., S. Savin, N. Savva. 2012. Physician workload and hospital reimbursement: Overworked physicians generate less revenue per patient. *Manufacturing & Service Operations Management* **14**(4) 512–528.

Pratt, John W. 1964. Risk aversion in the small and in the large. *Econometrica* **32**(1/2) 122–136.

Richards, K. F., K. H. Fisher, J. H. Flores, B. J. Christensen. 1996. Laparoscopic Nissen fundoplication: cost, morbidity, and outcome compared with open surgery. *Surgical Laparoscopy Endoscopy and Percutaneous Techniques* **6**(2) 140–143.

Rosenthal, E. 2013a. The $2.7 trillion medical bill. *The New York Times.* Published on June 1, 2013.

Rosenthal, E. 2013b. As hospital prices soar, a single stitch tops $500. *The New York Times.* Published on December 2, 2013.

Rosenthal, E. 2014. Patients' costs skyrocket; specialists' incomes soar. *The New York Times.* Published on January 18, 2014.

Rosenthal, M. B., R. Fernandopulle, H. R. Song, B. Landon. 2004. Paying for quality: Providers' incentives for quality improvement. *Health Affairs* **23**(2) 127–141.

Shleifer, A. 1985. A theory of yardstick competition. *The RAND Journal of Economics* 319–327.

Sood, N., P. J. Huckfeldt, J. J. Escarce, D. C. Grabowski, J. P. Newhouse. 2011. Medicare's bundled payment pilot for acute and postacute care: Analysis and recommendations on where to begin. *Health Affairs* **30**(9) 1708–1717.

Sood, N., P. J. Huckfeldt, D. C. Grabowski, J. P. Newhouse, J. J. Escarce. 2013. The effect of prospective payment on admission and treatment policy: Evidence from inpatient rehabilitation facilities. *Journal of Health Economics* **32**(5) 965–979.

Studdert, D., M. Mello, W. Sage, C. DesRoches, J. Peugh, K. Zapert, T. Brennan. 2005. Defensive medicine among high-risk specialist physicians in a volatile malpractice environment. *The Journal of the American Medical Association* **293**(21) 2609–2617.

Toffel, M. W. 2008. *Contracting for servicizing*. Harvard Business School.

Tompkins, C., G. Ritter, R. Mechanic, J. Perloff, J. Chapman. 2012. Analysis of financial risk and risk mitigation option in the Medicare Bundled Payment for Care Improvement program. Tech. rep., The Schneider Institutes for Health Policy and The Heller School for Social Policy and Management, Brandeis University.

Weng, Z. K. 1999. The power of coordinated decisions for short-life-cycle products in a manufacturing and distribution supply chain. *IIE Transactions* **31**(11) 1037–1049.

White, A. L., M. Stoughton, L. Feng. 1999. Servicizing: The quiet transition to extended product responsibility. Tech. rep., Tellus Institute, Boston.

Wilson, R. 1968. The theory of syndicates. *Econometrica:* 119–132.

Yaesoubi, R., S. D. Roberts. 2011. Payment contracts in a preventive health care system: A perspective from Operations Management. *Journal of Health Economics* **30**(6) 1188–1196.

Appendix A: Notation

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 Notations

Appendix B: Monotone Likelihood Ratio Property

Definition 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 $\mu_1 \le \mu_2 \in [\underline{\mu}, \overline{\mu}]$, the ratio $g_{\mu_2}(x)/g_{\mu_1}(x)$ is a non-decreasing function of x.

Lemma 4. 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 $\mu_1 \le \mu_2 \in [\underline{\mu}, \overline{\mu}]$, $A = \{x : g_{\mu_1}(x) > g_{\mu_2}(x)\}$, $B = \{x : g_{\mu_1}(x) < g_{\mu_2}(x)\}$, $a = \sup_{x\in A} \psi(x)$, $b = \inf_{x\in B} \psi(x)$. By the monotone likelihood ratio property, $a \le b$. Therefore,

$$\phi(\mu_2) - \phi(\mu_1) = \int \psi(x)(g_{\mu_2}(x) - g_{\mu_1}(x))dx$$

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

where the second equality follows from

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

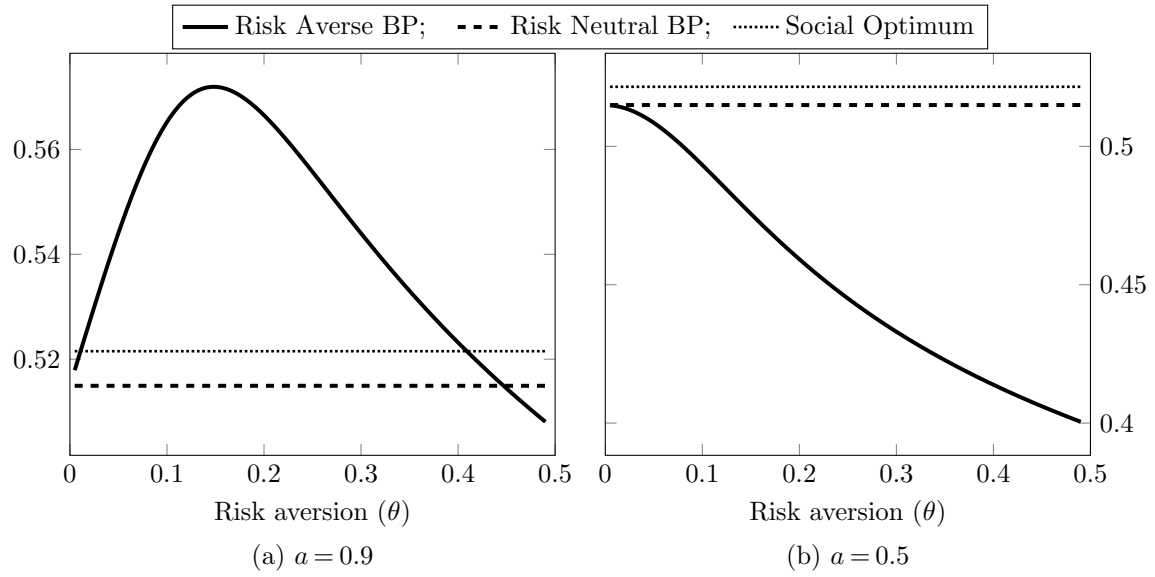
□

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).

Appendix C: Additional Figure



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

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

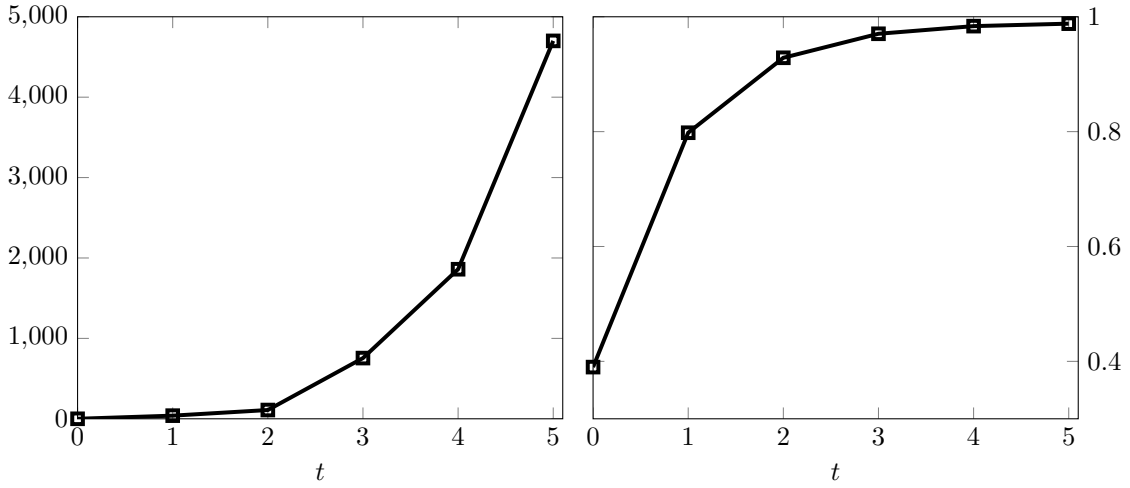
Appendix D: Numerical Analysis

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 16 and 18. In Section D.1 we describe the input parameters used for this numerical study. Section D.2 compares the average provider utilities and system payoffs under FFS, BP, and at the system optimum. Section D.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 6 in the main body of the paper provides a summary of observations made from our numerical experiments.

D.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 Heudebert et al. (1997), 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 Heudebert et al. (1997) and summarized in Table 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 Richards et al. (1996) 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 Heudebert et al. (1997) 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. Heudebert et al. (1997) 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



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

Figure 3 Parameter values for numerical experiments

reimbursement data for LNF from Gerson et al. (2000) and compare them to the LNF procedure cost from Heudebert et al. (1997), 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. Figure 3 illustrates the functional forms of $c_1(t)$ and $q(t)$. Note that this is consistent with Assumptions 2 and 3; we also check that with the parameter values that we selected, Assumption 4 is satisfied.

D.2. Total average provider utility and system payoff

Section 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 (15) and (16)), and the total average system payoff, W^S(Equations (17)–(19)), 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

[11] In section D.3 we let parameter a vary.

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 2 Parameter values for numerical experiments

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

$$\Pi^P_{FFS} = E_\mu\Big[E_{c_2|\mu}\left[\pi^P(\bar{t})|\mu\right]\Big] = \frac{1}{\theta} - \frac{1}{\theta}e^{-\theta(\kappa-1)c_1(\bar{t})}\left(q(\bar{t}) + (1-q(\bar{t}))E_\mu\left[J_\mu\right]\right), \tag{15}$$

$$\begin{aligned}\Pi^P_{BP} =& E_\mu\Big[E_{c_2|\mu}\left[\pi^P(t^{BP}(\mu))|\mu\right]\Big] \\ =& \int_{\underline{\mu}}^{\mu^{BP}} \left[\frac{1}{\theta} - \frac{1}{\theta}e^{-\theta(B-c_1(t^{BP}(x)))}\left(q(t^{BP}(x)) + (1-q(t^{BP}(x)))I_x\right)\right] f(x)dx.\end{aligned} \tag{16}$$

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

$$W^S_{FFS} = V - c_1(\bar{t}) - (1-q(\bar{t}))\left(T^B + T^P + E[\mu]\right), \tag{17}$$

$$W^S_{BP} = \int_{\underline{\mu}}^{\mu^{BP}} \left[V - c_1(t^{BP}(x)) - (1-q(t^{BP}(x)))(T^B + T^P + x)\right] f(x)dx, \tag{18}$$

$$W^S_{SO} = \int_{\underline{\mu}}^{\mu^*} \left[V - c_1(t^*(x)) - (1-q(t^*(x)))(T^B + T^P + x)\right] f(x)dx. \tag{19}$$

Figure 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 (3).

Figure 4(a) plots the average system payoff from (17), (18), and (19), 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.

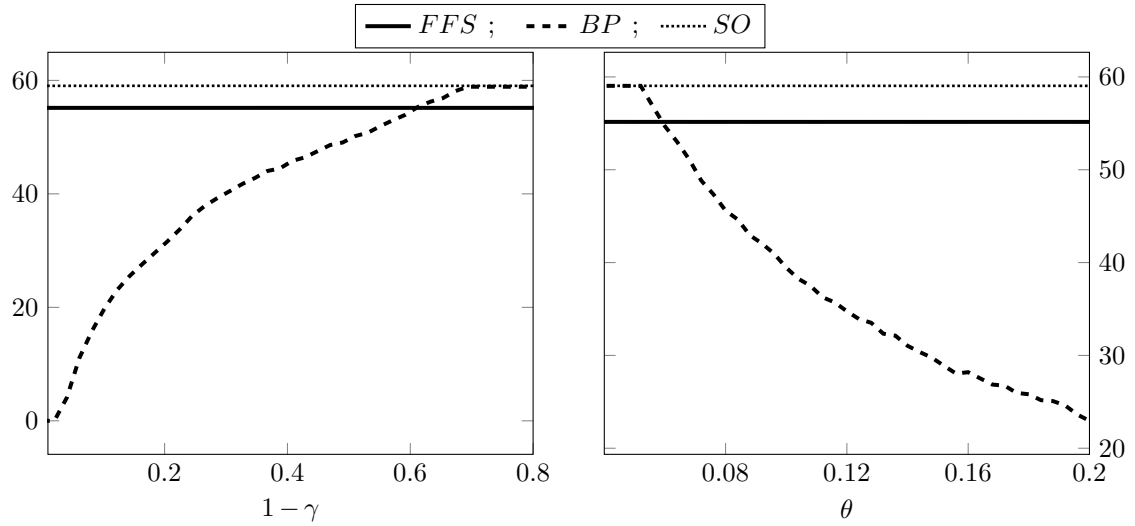
Figure 4(b) plots the average system payoff for increasing values of the provider's risk aversion. From Propositions 1 and 8 and Assumption 4, it is clear that the system-optimal and the FFS treatment levels are independent of θ. Hence, the total system payoff is also independent of θ under SO and FFS, as can be observed in Figure 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 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 (15) and (16). Clearly, under BP the provider's utility increases with the value of B as illustrated in Figure 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 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.

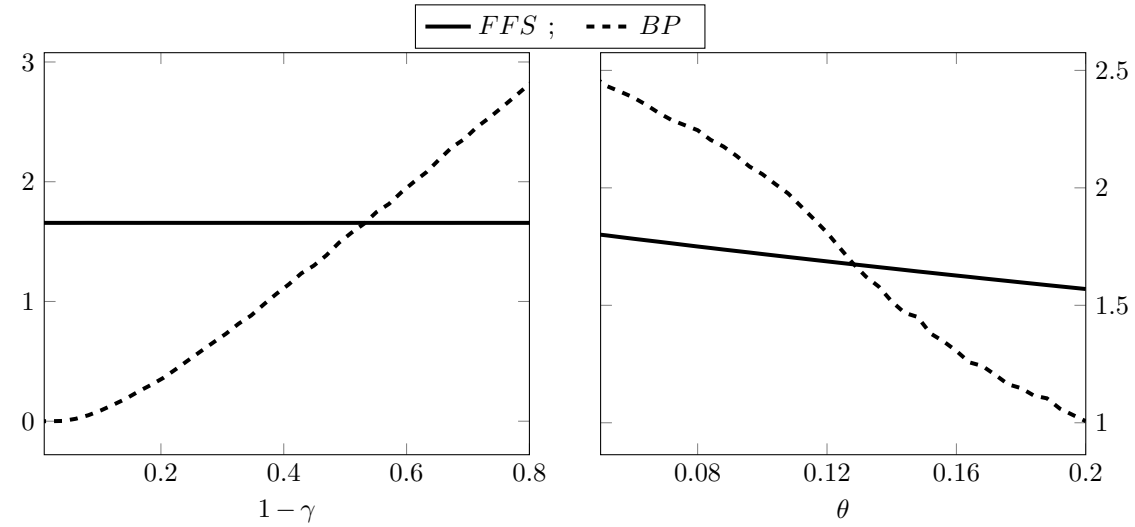
It is interesting to observe that when providers are not too risk averse and if B is sufficiently large, both the system and providers themselves benefit from BP as compared to FFS. This is because they no longer treat at the highest possible level, the gain due to the difference between treatment costs 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 4 and 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 4, the whole population has been selected under the system optimum due to our choice of parameters.



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

Figure 4 Average system payoff in \$1000s for $V = \$60,000$



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

Figure 5 Average provider utility in \$1000s for $V = \$60,000$

D.3. Coordinating mechanisms

We start this section by investigating coordination under the proposed mechanisms, namely hybrid payment and stop-loss. Figure 6 shows the effect of the risk-aversion parameter, θ, 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, which states that when provider's risk aversion is small enough, $t^{BP}(\mu) \leq t^*(\mu)$ and therefore the HP mechanism coordinates (Proposition 17). For large enough values of

θ, $t^{BP}(\mu) \geq t^*(\mu)$ and condition (14) is satisfied so the stop-loss mechanism coordinates, consistent with Proposition 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 (14) is violated, no hybrid or stop-loss coordinating mechanism can be found.

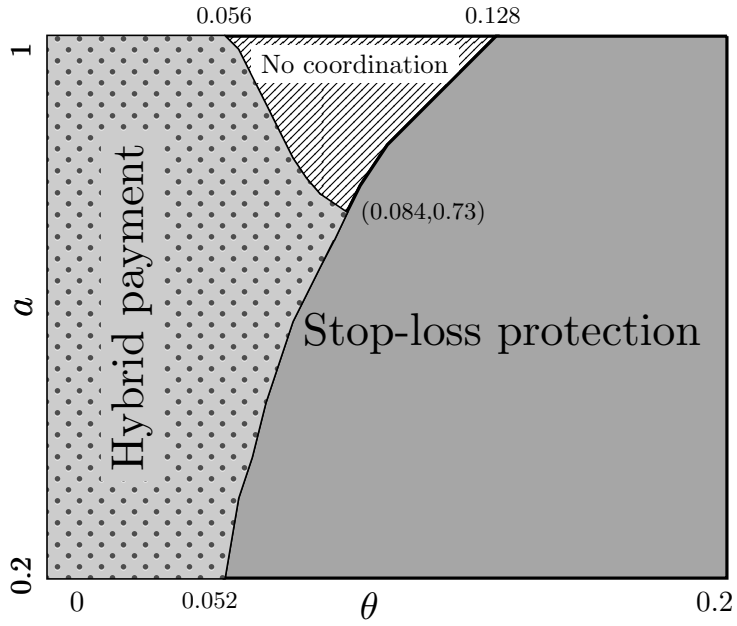


Figure 6 **Coordination by stop-loss mechanism or hybrid payment scheme or none for** $V = \$60,000$, $T^P = 0.02V$ **and** $\mu = 8,000$

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 4.4.3. 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 (10) 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 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 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 (11)) to reach a certain level of risk exposure under each payment mechanisms. As demonstrated in Figure 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 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.

	$1-\gamma$					
	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 3 Patient population size N needed for provider risk exposure of $\rho = 0.9(B-\delta)$

In Table 3(b), we perform a similar comparison with the stop-loss (SL) payment mechanism. We notice from Figure 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 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 terms of provider risk exposure, and especially for smaller values of the BP bundle payment.

Appendix E: Proofs

Proof of Lemma 1 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, we apply Lemma 4 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. □

Proof of Proposition 1. 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 $\overline{t}$. 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 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 $\overline{t}$, whichever leads to the highest value of the provider's objective function. We have

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

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

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

By Lemma 1,

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

where μ_1 is such that

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

□

Proof of Proposition 2. By Lemma 1, 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, it is clear that $E_{c_2|\mu}[\pi^P(t)|\mu]$ is continuous at the breakpoint $\mu = \mu_1$. The result thus follows. □

Proof of Proposition 3. 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

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

□

Proof of Lemma 2 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, we apply Lemma 4 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. □

Proof of Proposition 4. 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

$$\frac{d\, E_{c_2|\mu}\left[\pi^P(t)|\mu\right]}{d\, t} = e^{-\theta(B-c_1(t))}\left(-c'_1(t)(q(t) + (1 - q(t))I_\mu) - \frac{1}{\theta}(1 - I_\mu)q'(t)\right) \tag{20}$$

$$\begin{aligned}\frac{d^2\, E_{c_2|\mu}\left[\pi^P(t)|\mu\right]}{d\, t^2} &= e^{-\theta(B-c_1(t))}\left(-\theta c'_1(t)^2(q(t) + (1 - q(t))I_\mu) - (1 - I_\mu)q'(t)c'_1(t)\right. \\ &\quad \left. - c''_1(t)(q(t) + (1 - q(t))I_\mu) - c'_1(t)q'(t)(1 - I_\mu) - \frac{1}{\theta}(1 - I_\mu)q''(t)\right) \\ &= -e^{-\theta(B-c_1(t))}\left[(1 - I_\mu)\left(\frac{1}{\theta}q''(t) + 2c'_1(t)q'(t)\right) + \theta c'_1(t)^2(q(t) + (1 - q(t))I_\mu) + c''_1(t)(q(t) + (1 - q(t))I_\mu)\right].\end{aligned}$$

The second bracketed term above is clearly non-negative. by Assumption 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 (4). □

Proof of Proposition 5. If there exists a $\underline{\mu} \le \tilde{\mu} \le \overline{\mu}$ for which $t^{BP}(\tilde{\mu}) = \overline{t}$, we show that for all $\tilde{\mu} \le \mu \le \overline{\mu}$, $t^{BP}(\mu) = \overline{t}$. To see this, note that if $t^{BP}(\tilde{\mu}) = \overline{t}$, then

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

On the other hand, because $I_\mu - 1 > 0$ and increasing in μ by Lemma 2, it follows that

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

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

For all other values of $\mu \in [\underline{\mu}, \overline{\mu}]$, t^{BP} is the solution to (4). Taking the derivative of (4), we have

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

where by Lemma 2, $I_\mu' \equiv \partial I_\mu / \partial \mu \ge 0$. By assumption 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 2, $I_\mu - 1 > 0$ and $I_\mu' \ge 0$. By Assumption 2, $c_1''\left(t^{BP}(\mu)\right) \ge 0$. It thus follows that $d\,t^{BP}(\mu)/d\,\mu \ge 0$ for all $\mu \in [\underline{\mu}, \overline{\mu}]$. □

Proof of Proposition 6. 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 $\underline{t} \le t_0 \le \overline{t}$ 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 (4) we get

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

where the second equality follows from (4), and the inequality follows from Lemma 2. 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. □

Proof of Proposition 7 This proof is adapted from Gan et al. (2004). 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:

$$\pi^P(S) = E[U^P(w^P(S))]$$
$$\pi^I(S) = E[w^I(S)]$$
$$\pi^B(S) = E[U^B(w^B(S))].$$

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.

Proof of Proposition 8. Clearly, for a given μ, by Assumptions 2 and 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. □

Proof of Proposition 9. Clearly, if $t_1 < \underline{t}$ or $t_1 > \bar{t}$, we have $dt^*(\mu)/d\mu = 0$ and the result holds. If $\underline{t} \le t_1 \le \bar{t}$, taking the derivative of (6), we have

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

By Assumptions 2 and 3, and because q is increasing in t, we find that $dt^*(\mu)/d\mu \geq 0$. □

Proof of Proposition 10. Clearly, if $t_1 < \underline{t}$ or $t_1 > \overline{t}$, we have $dE_{c_2|\mu}[\pi^P(t)|\mu]/d\mu \leq 0$. If $\underline{t} \leq t_1 \leq \overline{t}$, taking the derivative of the expected total system payoff for a given μ, with respect to μ, and using (6) we get

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

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. □

Proof of Proposition 11. 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

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

where t_0^N is the unique solution of the equation

$$c_1'(t) = (T^P+\mu)q'(t). \tag{21}$$

If $t_0^N < \underline{t}$, then by definition $t^*(\mu) \geq t^{BP}(\mu) = \underline{t}$ and the left inequality holds. If $t_0^N > \overline{t}$ then by concavity of the hospital's expected payoff we have $c_1'(t) \leq (T^P+\mu)q'(t)$ for all $t \in [\underline{t}, \overline{t}]$. Therefore, clearly $c_1'(t) \leq (T^P+\mu+T^B)q'(t)$ and $t^* = \overline{t}$ and the left inequality is tight. Finally if $\underline{t} \leq t_0^N \leq \overline{t}$ 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 $t_1 \leq \overline{t}$, then $t_1 = t^*(\mu)$, and the first inequality follows. If $t_1 > \overline{t}$, then $t^*(\mu) = \overline{t} \geq t^{BP}(\mu)$. So the left inequality holds for all cases. By Assumption 4, $t^{FFS}(\mu) = \overline{t}$ hence the second inequality is clear. □

Proof of Corollary 1. Since the first order Taylor expansion of I_μ as $\theta \to 0$ is

$$\int_{\underline{c}}^{\overline{c}} (1+\theta(c_2+T^P))g_\mu(c_2)dc_2 = 1 + \theta(\mu+T^P),$$

the limit of the left hand side of (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 (4) approaches the (risk-neutral) solution of (21).

Hence, the result follows from Proposition 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. □

Proof of Proposition 12. We want to show that there exists a coordinating bundled payment $B_C\geq 0$ such that $\mu^*=\mu^{BP}$. From (8) and (5), 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 (3) leads to the result.

□

Proof of Proposition 13. Replacing the term for $\overline{w^P}$ from (10) 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 (Hoeffding, 1963, 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'.

□

Proof of Lemma 3 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, we apply Lemma 4 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. □

Proof of Proposition 14. 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

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

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 2. The first one is positive by Lemma 3 and Assumption 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 (12).

□

Proof of Proposition 16. Similarly to the proof of Proposition 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

$$B' - (1-\beta)c_1(t^{HP}(\mu^{HP})) - (T^P + (1-\beta)\mu^{HP})(1 - q(t^{HP}(\mu^{HP}))) = 0. \tag{22}$$

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 (8) and (22), 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

$$\frac{V - c_1(t^*(\mu^*))}{1 - q(t^*(\mu^*))} - T^P - T^B = \frac{B'/(1-\beta) - c_1(t^{HP}(\mu^{HP}))}{1 - q(t^{HP}(\mu^{HP}))} - \frac{T^P}{1-\beta}. \tag{23}$$

In this case, when $\beta = T^B/(T^B + T^P)$ we also have $t^{HP}(\mu^{HP}) = t^*(\mu^*)$, hence equation (23) can be rewritten

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

□

Proof of Proposition 17. As shown in the proof of Proposition 14, the objective of the provider under HP is a concave function with the following first derivative.

$$\frac{d\, E_{c_2|\mu}[\pi^P(t)|\mu]}{dt} = \frac{e^{-\theta(B' - (1-\beta)c_1(t))}}{\theta}\left(q'(t) - \theta(1-\beta)c_1'(t)\left[1 - q(t) + \frac{1}{L_\mu - 1}\right]\right), \tag{24}$$

where

$$L_\mu = \int_{\underline{c}}^{\overline{c}} e^{\theta((1-\beta)c_2 + T^P)} g_\mu(c_2) dc_2 = e^{\theta T^P} \int_{\underline{c}}^{\overline{c}} e^{\theta(1-\beta)c_2} g_\mu(c_2) dc_2.$$

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) $t^*(\mu) = \overline{t}$, (2) $t^*(\mu) = \underline{t}$, and (3) $\underline{t} < t^*(\mu) < \overline{t}$.

Case 1: $t^*(\mu) = \overline{t}$. If $\beta \to 1$ then (24) implies that the derivative of $E_{c_2|\mu}[\pi^P(t)|\mu]$ is non-negative for all $t \in [\underline{t}, \overline{t}]$ since $c_1'(t), q(t)$, and L_μ are all finite parameters. Therefore, $t^{HP}(\mu)|_{\beta\to 1} = \overline{t} = t^*(\mu)$.

Case 2: $t^*(\mu) = \underline{t}$. If $\beta \to 0$ then (24) is the same as the derivative of the hospital's utility under BP with bundled payment B'. Note that we assumed $t^{BP}(\mu) \le t^*(\mu) = \underline{t}$. Since $t^*(\mu) = \underline{t}$ it follows $t^{HP}(\mu)|_{\beta\to 0} = t^{BP}(\mu) = \underline{t} = t^*(\mu)$.

Case 3: $\underline{t} < t^*(\mu) < \overline{t}$. 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 (24), where $t^*(\mu)$ solves the following at the system optimum:

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

Plugging $t^*(\mu)$ into (24) and using the property of the system optimum solution, we need to show the existence of β for which:

$$\frac{1}{T^P+\mu+T^B}-\theta(1-\beta)\left[1-q(t^*(\mu))+\frac{1}{L_\mu-1}\right]=0. \tag{25}$$

Thus, a hybrid system coordinates the treatment level decision iff there exists a $\beta\in[0,1]$ that satisfies (25) (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 (25) *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 (25) equals $1/(T^P+\mu+T^B)>0$.

When $\beta=0$, $L_\mu=I_\mu$ and thus the left-hand-side of (25) 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 (25) 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 (25) 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

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

□

Proof of Proposition 18. 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:

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

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

Note that the first term in the bracket is negative due to Assumption 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)

$$\left.\frac{\partial\, E_{c_2|\mu}\left[\pi^P(t)|\mu\right]}{\partial\, t}\right|_{t=t^*, S=\underline{c}} = e^{-\theta(B-c_1(t^*))}\left[-c_1'(t^*)(q(t^*) + (1-q(t^*))e^{\theta(\underline{c}+T^P)}) - \frac{1}{\theta}q'(t^*)(1-e^{\theta(\underline{c}+T^P)})\right] \quad (26)$$

and

$$\left.\frac{\partial\, E_{c_2|\mu}\left[\pi^P(t)|\mu\right]}{\partial\, t}\right|_{t=t^*, S=\overline{c}} = e^{-\theta(B-c_1(t^*))}\left[-c_1'(t^*)(q(t^*) + (1-q(t^*))I_\mu) - \frac{1}{\theta}q'(t^*)(1-I_\mu)\right]. \quad (27)$$

Note that (26) is negative due to (14). Furthermore, (27) is identical to the derivative with respect to t of the expected provider utility under BP (20) evaluated at $t = t^*$. As shown in the proof of Proposition 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

$$\mu^* = \frac{V - c_1(t^*(\mu^*))}{1 - q(t^*(\mu^*))} - T^P - T^B. \quad (28)$$

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

$$B=\frac{1}{\theta}\ ln\left(q(t^*(\mu^{SL}))+\left(1-q(t^*(\mu^{SL}))\right)M_{\mu^{SL}}(S^*)\right)+c_1(t^*(\mu^{SL})). \tag{29}$$

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 (29), it is clear that B is non-negative ($M_{\mu^{SL}}>1$). μ^* as appeared in Equation (28), can be replaced with μ^{SL} in Equation (29). □

Appendix F: 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

$$\pi^P(t)=(\kappa-1)c_1(t)+(1-q(t,\mu))(-T^P+(\kappa-1)c_2), \qquad \text{(provider's payoff under FFS).} \tag{30}$$

$$\pi^P(t)=B-c_1(t)+(1-q(t,\mu))(-T^P-c_2), \qquad \text{(provider's payoff under BP).} \tag{31}$$

$$\pi^P(t)=B'+(\beta-1)c_1(t)+(1-q(t,\mu))(-T^P+(\beta-1)c_2), \qquad \text{(provider's payoff under HP).} \tag{32}$$

$$\pi^S(t)=-c_1(t)+(1-q(t,\mu))(-T^P-c_2-T^B)+V, \qquad \text{(system's payoff).} \tag{33}$$

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:

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

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 11 and 16, respectively. In the remainder of the this section we state and derive the results required for the proof of these statements.

F.1. Proofs

Proof of Proposition AC1. Using Lemmas AC1, AC2, and AC3 below, the proof is similar to the proof of Proposition 11. □

Proof of Proposition AC2. Using Lemmas AC3 and AC4 below, the proof is similar to the proof of Proposition 16. □

LEMMA AC1. Under the FFS mechanism defined in (30), the optimal treatment intensity is $t^{FFS} = \bar{t}$.

Proof. Using Assumption 4, the proof is similar to the proof of Proposition 1 when $\theta \to 0$. □

LEMMA AC2. Under the BP mechanism defined in (31),

(a) the optimal treatment intensity is given by

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

(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 4 when $\theta \to 0$.

Part (b): The proof is similar to the proof of Proposition 5 when $\theta \to 0$ with the qualification that taking derivative of (31) with respect to μ results in

$$\frac{\partial q(t,\mu)}{\partial t} + \left(T^P + \mu\right) \frac{\partial^2 q(t,\mu)}{\partial t\, \partial \mu} = \frac{d\ t^{BP}(\mu)}{d\ \mu} \underbrace{\left(c_1''(t) - (T^P + \mu) \frac{\partial^2 q(t,\mu)}{\partial t^2} \right)}_{\ge 0}.$$

The necessary and sufficient condition for the claim to hold true is then, $\dfrac{\partial q(t,\mu)}{\partial t} + \left(T^P + \mu\right) \dfrac{\partial^2 q(t,\mu)}{\partial t\, \partial \mu} \ge 0$. This condition is satisfied considering Assumption AC1.

Part (c): The proof is similar to the proof of Proposition 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

$$B-c_1(t^{BP}(\overline{\mu}))-\left(T^P+\overline{\mu}\right)(1-q(t^{BP}(\overline{\mu}),\overline{\mu}))<0.$$

□

Lemma AC3. Under the system optimum case defined in (33),

(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 (32),

(a) the optimal treatment intensity is given by

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

(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. □