Contracts for healthcare referral services: Coordination via outcome-based penalty contracts

Elodie Adida
School of Business, University of California Riverside, elodie.goodman@ucr.edu,

Fernanda Bravo
Anderson School of Management, University of California Los Angeles, fernanda.bravo@anderson.ucla.edu,

This work focuses on the B2B interaction between a service requester and a service provider in a healthcare environment. The requester is the primary caregiver responsible for managing the health of a population of patients. When a patient requires advanced care outside the requester's expertise, the requester refers the patient to a provider and pays for the referral services. Treatment may succeed or fail, and in case of failure the requester incurs further follow-up costs. The requester may exert preventive effort to reduce the volume of referrals. The provider may exert non-reimbursable effort to reduce the chance of treatment failure. We analyze payment contracts between the two firms. We find that fee-for-service (FFS) induces neither system nor social optimum effort outcomes. However, a penalty contract can generally coordinate the effort decisions with either the system optimum or the social optimum. Furthermore, we find that patients may benefit from having a coordinating contract replace FFS. However, the types of procedures that make a coordinating contract most advantageous for the requester and provider are not necessarily the same as those that make the patients better off than under FFS. Yet, in most cases the coordinating contract improves social welfare, as compared to FFS, and brings it close to the social optimum. Hence, the requester-provider coordinating contract can be considered as an improvement over FFS for the entire system.

Key words: Healthcare, B2B, Contracts, Fee-for-service, Pay-for-performance.

History: Submitted September 8, 2016; Revised March 31, 2017, Revised September 8, 2017, Accepted October 18, 2017.

1. Introduction

In today's US healthcare environment, and especially since the recent healthcare reform (Affordable Care Act), a patient's health is often managed by an organization. In many cases, the managing organization is capable of providing routine care but does not have the ability or the expertise to provide complex or highly specialized care. When such advanced care is needed, the organization must outsource delivery of care to an external provider.[1] Cost control initiatives embedded in the

[1] Examples of such partnerships include the deal between Cambridge Health Alliance and Beth Israel Deaconess Medical Center (Hacker et al. 2014), and that between Steward Health Care System and Partners HealthCare Sys-

new law are seeking to make the managing organization responsible not only for managing the patients' health, but also for the cost and the quality of the care that they provide. As a result, the managing organization is held responsible for the cost of referral services (Bravo et al. 2016).

For example, Accountable Care Organizations (ACOs) are groups of caregivers (e.g., hospitals, physicians, and clinics) that come together voluntarily to assume responsibility for managing the health of a patient population. The establishment of ACOs is one of the initiatives promoted by the Affordable Care Act, and the number and size of ACOs has been steadily growing since 2010. For instance, Montefiore ACO, based in New York City, is one of the largest ACOs with 23,000 beneficiaries, 4187 participating physicians and 935 facilities in 2014 (SK&A 2014). Today, over 23 million Americans are being served by an ACO. In 2015, physician-led ACOs (that is, ACOs that can provide primary and/or secondary care, but not tertiary or quaternary care[2]) account for 37% of all the ACOs in the U.S. (Tu et al. 2015, Colla et al. 2016). These ACOs play a major role in the delivery of preventive and low-complexity care. Yet, ACOs are financially responsible for the total cost of care, including inpatient (tertiary or quaternary) care, regardless of who performs the service (Barnes et al. 2014). Robinson and Schaeffer (2015) state that "the ACO's financial responsibility is not limited to the services directly provided by its member physicians, but extends to all the services provided by all the caregivers and resources used by the patient." Thus, if a complication arises that cannot be treated within the ACO, the ACO refers the patient for advanced care (e.g., specialist consultation, inpatient procedure, surgical care) to a third party provider (e.g., medical center). The volume of care that is provided outside the ACO may be significant: McWilliams et al. (2014) report that in 2011, "66.7% of office visits with specialists were provided outside of the assigned ACO". While this figure refers to specialists consultations, it provides evidence that the care volume provided outside the ACO, either as referral or patient leakage, can be large. Regarding specifically inpatient referrals, "one health insurer's 2009 referral data shows that only 35-45% of adult inpatient care, as measured by revenue, goes to the partner hospital", implying that over half of adult referrals may be serviced outside the ACO (Kuraitis 2011).

Similarly, Health Maintenance Organizations (HMOs), introduced in the early 1990s, generally assume full financial responsibility for the cost of care (Malcomson 2004). HMOs negotiate contracts for professional services with physician group practices, or with an Independent Practice Association which in turn contracts with independent physicians (Gold et al. 1995). A well-known

tem (Weisman 2012). Under the latter, the most severely injured patients from emergency rooms at Steward's ten community hospitals are sent to Partners-owned Massachusetts General and Brigham and Women's hospitals in Boston.

[2] Primary and secondary care refer to low-complexity care delivered by a primary care physician or a specialist. Tertiary and quaternary care require highly specialized equipment and expertise, and cannot be provided at a small community hospital (e.g., coronary artery bypass surgery).

example of a fully integrated HMO is Kaiser Permanente in California, serving more than 10.6 million members. More than 80 million people are enrolled in an HMO today nationwide.

Because managing organizations like ACOs and HMOs are responsible for all expenses incurred, even for services rendered by an external provider or hospital, they benefit from shrinking the volume of referrals needed. Hence, the managing organization typically focuses on prevention and on carefully managing patients with serious or chronic conditions, with the aim of reducing the need to outsource care to an external provider. To this end, the organization may invest in preventive measures such as patient education and monitoring (e.g., cholesterol and diabetes programs), wellness activities promotion (e.g., weight loss and tobacco cessation programs), disease screening, nurse care managers hiring, and tools facilitating care coordination including healthcare data sharing, storing and analyzing. While this preventive effort incurs an upfront cost, which grows with the size of the population being served, it also reduces future costs by lowering the volume of external referrals required.

In the event a patient is referred to an external provider for a given procedure, the provider has considerable latitude in the effort it exerts to ensure good health outcomes. Several "best practices" have been shown to improve outcomes, including: proper medication management (ensuring that the prescribing doctor is aware of all medication currently prescribed), discharge management (scheduling remote and in-person follow-up care, ensuring the patient understands discharge instructions), coordinating post-discharge care with the primary care physician (Arbaje et al. 2008). This type of intervention does not constitute a billable medical act that the provider may request reimbursement for. However, these measures carry an immediate cost that is born by the provider. They also offer the potential to reduce the chance of poor patient outcomes and hence to lower future possible complication costs.

The most common way of paying an external provider for outsourced care is fee-for-service (FFS) (Zuvekas and Cohen 2016). Under FFS, the provider is paid a *fixed*, *set-in-advance* price for the procedure provided. Moreover, the payment is the same regardless of the health outcome. An example of this set-in-advance price is the use of Current Procedure Terminology (CPT) codes by doctors for billing medical services. Experts have long recognized that the incentives under FFS are not conducive to providing efficient, high-quality health care (Feder 2013). Indeed, the provider is being paid the same regardless of the patient's eventual health outcome. In particular, should the treatment fail or should complications arise, leading the patient to require further care (e.g., rehabilitation), the provider is not being held financially responsible. Hence, in a FFS contract the provider lacks financial incentives to exert costly, but non-reimbursable, effort to avoid poor patient outcomes. This lack of incentives has contributed to observing a 30-day readmission rate of nearly 20% for Medicare FFS beneficiaries from 2007 through 2011 (Gerhardt et al. 2013).

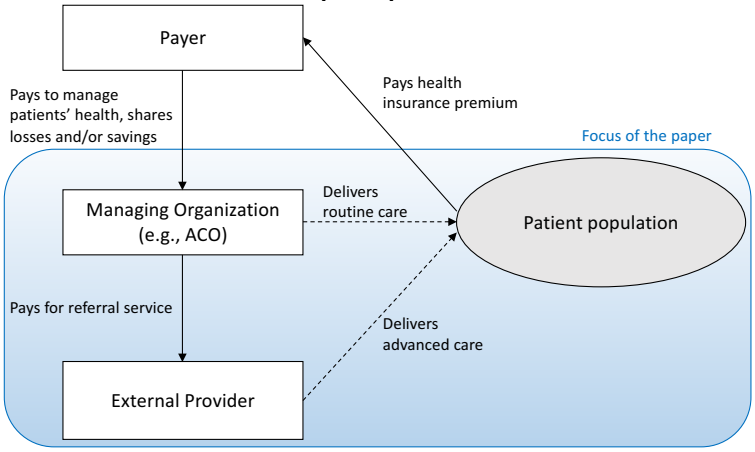
Because the managing organization bears the referral cost, it must carefully negotiate the fee structure used to compensate external providers for their services. Insurers and payers (e.g., Medicare) have investigated and tested several reimbursement contracts with providers to try and better align caregivers' incentives with the goal of obtaining better patient outcomes – including capitation, bundled payments, cost-sharing, shared-savings, and hospital readmission penalties. The aim is generally to move away from a pay-per-transaction contract toward a pay-for-performance contract, where the caregiver is being held, at least partially, responsible for the eventual patient outcome (McCluskey 2015). Some of these contracts have proven useful for aligning incentives in the payer-caregiver setting, where the payer makes no medical decision directly affecting the patient's health status. Yet, it is unclear whether such contracts would align incentives in the requester-provider setting that we focus on, where both parties play a direct role in medical decision-making to ensure a patient's good health.

In this paper, we focus on the business-to-business (B2B) interaction and payment system between the managing organization, or service requester, and the external service provider (see Figure 1). Our model applies best to a physician-led ACO that either does not include an affiliated hospital, or includes a small community hospital unable to offer all highly specialized services and thus must contract with external providers for such procedures.[3] In this context, the provider may exert non-reimbursable effort to lower the chance of poor patient outcomes and thus of the patient requiring further treatment cost. The requester may also exert effort that lowers the patients' need for advanced treatment and thus reduces the volume of referrals sent to the provider. While patients benefit from both types of effort, these efforts incur a cost, but not necessarily a benefit, for the exerting party. In addition, the provider's effort benefits the requester, whereas the requester's effort makes the provider economically worse off. We seek to determine whether a service payment contract may provide incentives to the requester and the provider to exert effort at an optimal level. The setting considered in this paper differs from the study of the interaction between a payer and a medical provider in one fundamental way: contrary to the payer, the service requester makes a medical effort decision that controls the volume of patients treated by the provider. We seek a payment system that aligns incentives both for the provider's effort to enhance treatment outcomes, and for the requester's effort to maintain an adequate volume of referrals.

In our problem, a coordinating payment contract between the requester and the provider maximizes their joint total profits. Such coordination disregards the patient utility, and focuses on reducing operational costs comprising the costs of effort, the treatment cost, and the cost of unsuccessful treatment. Clearly, reducing the need for referrals and the chance of a failed treatment is

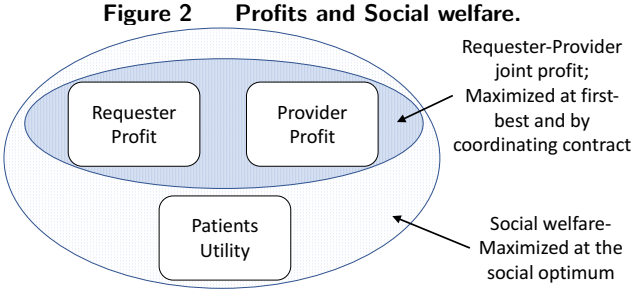
[3] Example of such ACOs include Physician Group Alliances and Expanded Physician Groups (Oss 2016).

Figure 1 Interactions between participants in the healthcare environment.



aligned with patients' interest. Yet, these goals must be balanced with the cost of exerting effort to achieve them, thereby putting a downward pressure on effort in a way that may conflict with the patients' interest. Therefore, in general, coordination between the requester and the provider maximizes neither the patient utility nor the social welfare (which comprises the requester and provider joint profits as well as the patients' utility, see Figure 2). Since patient outcomes are a primary consideration in designing a healthcare payment agreement, we analyze the effect of a coordinating contract on patient utility and on social welfare. Furthermore, we investigate whether there is a contract that may achieve an optimal social welfare outcome (i.e., the social optimum), that is, that may maximize a combination of the firms' profit *and* the patients' utility.

Figure 2 Profits and Social welfare.



This paper makes the following contributions. We introduce a new model of healthcare payment system tailored to capture the interaction between a service requester, such as a physician-led ACO, who makes an effort decision affecting the volume of patients treated, a service provider who makes a non-reimbursable effort decision affecting the patients' health outcomes, and a population of patients. We find the optimal efforts for a system comprising the requester and the provider (i.e., first-best efforts) and compare them to the effort levels under FFS and under a penalty contract. We also obtain efforts at the social optimum for the "society" including patients as well. We find

that FFS can never induce either the (provider-requester) system optimum or the social optimum.[4] Yet, a penalty contract may coordinate the effort decisions of the provider and the requester with either the system optimal or the socially optimal efforts. We find that the coordinating contract is most significantly better for the requester-provider system compared to FFS when the FFS price paid for each service rendered is large (i.e., high-margin procedures), the treatment cost is low, and/or the failure rate is high. Patients may be better off under a coordinating contract than under FFS, especially when the FFS price per procedure is low (i.e., low-margin procedures) and/or the failure rate is high. Generally, the types of procedures for which the requester and provider most benefit from a coordinating contract are not the same as those for which the patients benefit from having a coordinating contract replace FFS. However, in most cases the coordinating contract improves the social welfare compared to FFS and brings it to a level close to the social optimum. Therefore, even though the coordinating contract does not take the patient utility into account, it still represents a meaningful improvement over FFS.

2. Literature Review

The healthcare literature has seen a sustained interest in understanding how a payer (usually the insurer) can align providers' incentives to ensure quality and cost control (see the review by Christianson et al. (2008) and references therein). Agency problems often arise in this context due to *adverse selection*, i.e., when the payer cannot observe the provider's "type" before committing to the agreement (hidden information), and/or *moral hazard*, i.e., when the provider acts in its own interest which might conflict with the payer's objectives (hidden actions) (McGuire 2000).

Even without information asymmetry, performance-based contracting has been investigated in the healthcare operations literature (Selviaridis and Wynstra 2014). Jiang et al. (2012) propose a performance-based contracting scheme for outpatient medical services; the payer designs a contract to minimize costs while achieving a desired outcome, and the provider decides how to allocate capacity among the different patient types. They show that a threshold penalty performance-based contract can coordinate the system. Andritsos and Tang (2015) investigate how bundled payments and pay-for-performance payment systems combined with patient cost-sharing can help prevent readmissions when the care is co-produced by the patient and the provider.

The recent papers by Zhang et al. (2016) and Adida et al. (2017) study how some of the new payment initiatives in the U.S. can better align incentives between a provider or hospital and an insurer. Zhang et al. (2016) focus on the Hospital Readmissions Reduction Program (HRRP), which

[4] We also considered a variety of often-used payment contracts, including capitation, two-part tariff, cost-sharing, shared-savings and find that they can never induce either the system optimum nor the social optimum. The analyses are presented in the online Appendix E.

consists in penalizing hospitals with excess readmissions compared to a risk-adjusted benchmark. They find that, because many hospitals prefer paying penalties, they lack incentives to reduce readmissions, but modifying the benchmarking process would induce less readmissions. In our analysis, the requester plays a role with respect to the provider similar to the role that Medicare plays with respect to a hospital in their paper. Yet there are several key distinctions both in focus and in execution. First, they use a threshold policy on the readmission rate, and they focus on competition among hospitals by determining the (risk-adjusted) threshold according to the performance of all hospitals nationwide. In our research, there is a single service provider and thus no peer-benchmarking defining a threshold. Second, we capture the effect of the service requester's prevention effort on patient volume, while in their setting the payer (Medicare) does not make any medical effort decision. Third, their goal is to reduce the readmission rate, while we consider coordination of care efforts to maximize either the requester and provider joint profit or the social welfare. Fourth, in their paper the penalty for excess readmission is applied to the total revenues across all conditions, not only the condition leading to a too high rate of readmission.

Adida et al. (2017) investigate how a Bundled Payment system may improve performance compared to FFS. They consider a medical provider choosing the "treatment level" among a variety of available treatment routes, where the FFS reimbursement varies according to the intensity of treatment. They find that under FFS the financial incentives lead providers to "treat more, not better" contributing to high costs. Indeed, under reasonable assumptions, they obtain that the provider would systematically select the highest possible treatment intensity to obtain higher payments. Thus they seek contractual incentives via bundled payments to lower the treatment intensity to a socially optimal level. In contrast, in this paper we focus on one given procedure with a single treatment option; the incentives misalignment stems from the provider being paid *the same fixed amount*, regardless of the eventual health outcome of the patient, for this given medical procedure. The effort decision that the provider makes has no bearing on his revenue (hence the lack of effort under FFS) whereas in Adida et al. (2017) the provider's decision is directly linked to his revenue (hence the excessive treatment intensity under FFS). In this manuscript, we analyze the lack of incentives under FFS to exert costly, but non-billable effort to avoid poor patient outcomes. In a sense, we are focusing on the inverse problem of that in Adida et al. (2017): while they seek ways to incentivize providers to implement less measures that are reimbursed but ineffective, we seek ways to incentivize providers to implement more measures that are effective, but non-reimbursable. In addition, we capture the effect of the service requester's effort while in their setting the insurer does not make any medical effort decision. Furthermore, they study bundled payments whereas we focus on a readmission penalty contract.

Our work contributes to this literature by investigating payment contracts that improve cost and quality in referral services, but our focus is on a B2B contract between peer providers, where *both* parties may exert effort influencing patient care, rather than a payer-provider setting where the payer does not make medical effort decisions directly affecting care outcomes.[5] Thus, a distinctive factor in our setting is that we seek to find a contract that aligns both players' efforts simultaneously when the incentives to exert costly effort are in opposition to each other's interests.

Some literature investigates the hospital-physician relationship. Burns and Muller (2008) provide a thorough review of the different types of contractual arrangements between hospital and physician groups and analyze the effectiveness of those arrangements in improving access, quality, and cost in healthcare. The tensions in the hospital-physician relationship often arise due to the misaligned incentives caused by physicians having a large influence in driving hospital costs and service quality. These tensions vary according to the level of integration between the hospital and the physicians (Vlachy et al. 2017). In that context there is no referral between the hospital and the physicians. In addition, the hospital and the physicians are usually paid independently and do not bear the cost of treatment failure. Hence, the incentives in the hospital-physician setting substantially differ to those studied in our setting.

A stream of literature has studied incentives for healthcare "gatekeepers", who control patients' access to specialist services (like the service requester in our paper). Many managed care plans use gatekeepers (primary care physicians, or general practitioners in the UK). Some of this literature focuses on how a payer can incentivize the gatekeeper to effectively refer patients (Mariñoso and Jelovac 2003, Malcomson 2004). Liu et al. (2015) are interested in the optimal policy for two-way referrals between two hospitals and their primary concern is congestion and waste of resources. They design threshold policies to help the hospitals determine which patients to re-direct to the other hospital. The referral decision follows diagnosis and is based on the complexity of the patient's condition, with no effort decision. The authors show that a cost-sharing contract can coordinate the two-way referral system. In our model of a one-way referral system without threshold policy and where both parties exert effort, we find that cost-sharing does not achieve coordination. In our research, the requester does not decide whether or not to make a referral; rather she influences the referral volume by exerting preventive effort. Moreover, we do not focus on capacity and congestion issues. Instead, we are interested in coordinating the efforts exerted by the two parties as a way to quantify the amount of preventive activities by the requester, and the intensity of follow-up care by

[5] The literature on B2B contracts between peer providers in a healthcare setting is sparse. One notable exception is Bravo et al. (2016) who study risk sharing between a service requester and a provider where the demand for the service is uncertain and there are no effort decisions. They show that a two-price piecewise linear contract allows parties to optimally share risks due to demand uncertainty. In contrast, our focus is on coordination of effort decisions in a peer-to-peer setting.

the provider, that improve the chance of treatment success and ensure better health outcomes. In addition, we derive an optimal contract between the requester and the provider, and not between the payer and the requester (gatekeeper).

Our work also relates to the outsourcing and contracting literature in services and product supply chains (Pinker et al. 2010, Kaya and Özer 2011). Ren and Zhou (2008) study the coordination of staffing and effort levels in a call center setting using queuing theory. Cost-sharing contracts can coordinate staffing and quality if the service quality effort is observed. Otherwise, the authors propose a partnership contract. A main difference is that in their setting only the call center exerts effort. Lee et al. (2012) consider outsourcing contracts in a two-level service process and show that first-best contracts exist when only one level of the process is outsourced. Cachon and Lariviere (2005) show that revenue sharing contracts can coordinate price and quantity decisions but not effort decisions when these affect demand. Roels (2014) proposes a framework to study co-productive services between an end-customer and a service provider, where the service value depends on the effort of the two parties, while the firm designs the type of collaboration. In contrast, in our setting the service value, corresponding to the reduction of referrals and follow-up costs, is linked to the requester and provider's efforts, not the patient's, via the volume of referrals and chance of treatment failure. Moreover, the type of collaboration, i.e. how the efforts affect the service value, is not a decision. In addition, we are concerned with designing a contract between the two firms to achieve coordination.

Several papers address the problem of joint cost reduction via appropriate contracting in supply chains (Kim and Netessine 2013). Close to this work, Corbett and DeCroix (2001) focus on efficient shared-savings contracts for jointly reducing consumption of indirect materials. The misalignment of incentives is similar to our setting: the buyer wants to minimize consumption (i.e., cost of indirect material), while the supplier benefits from higher consumption through higher revenue. In a related paper, Corbett et al. (2005) study a similar problem in a more general setting allowing linear cost-of-effort functions and consider also contracts that are nonlinear in the quantity consumed. They find that the first-best can sometimes be achieved by a shared-savings contract. Our model, on the other hand, tailored to a peer-to-peer healthcare contracting problem, captures the variable cost of effort, and also incorporates a quality dimension by capturing the impact of provider's effort on the chances of treatment failure and follow-up costs affecting the requester. These key distinctions result in fundamental differences in the way that efforts affect the outcome and in the incentives driving decisions and efficiency of a variety of contracts. For instance, in Corbett and DeCroix (2001) and Corbett et al. (2005) a shared-savings contract reduces to a two-part tariff contract that incentivizes the supplier to exert some effort. In our setting, the shared-savings contract is not a two-part tariff, and none of these contracts can achieve the first-best.

A stream of research on quality management considers a supplier deciding on the quality of production, and a buyer deciding on the quality inspection policy or quality appraisal mechanism to encourage the supplier to exert higher quality effort. A wide variety of contracts have been proposed, including penalty (Reyniers and Tapiero 1995), cost-sharing (Chao et al. 2009, Ma et al. 2013), revenue sharing (El Ouardighi 2014), warranty (Balachandran and Radhakrishnan 2005), and deferred payment contracts (Babich and Tang 2012). In this paper, the service requester cannot use quality inspection or appraisal like in some manufacturing settings. In addition, the buyer's effort consists in assessing quality without affecting directly the production output. In our setting, the service requester's effort affects the referral volume, and hence negatively impacts the provider's bottom line, which fundamentally changes the problem of finding a coordinating contract.

3. Model

We consider a B2B service chain in the healthcare industry consisting of a service provider (he), and a service requester (she) serving a given population for a given condition.[6] The service requester is in charge of managing the health of a patient population and is also financially responsible for the cost of care that they receive. This situation is common for self-insured healthcare networks, HMOs, and, more recently, for ACOs that have adopted capitation-like contracts with payers (Tu et al. 2016, Lewis et al. 2014).[7] While the majority of the care services can be provided within the requester's network, there are some services (e.g., complex, specialized, or advanced procedures) for which a referral is required. Referrals are serviced by the service provider, but the requester is responsible for the referral service cost.

The requester, in her role of maintaining the patients' health, can exert effort. This effort may include implementing chronic condition management programs, patient education and monitoring programs, promoting prevention and patient health maintenance. The requester's ultimate goal is to avoid more serious health issues that would require referring the patient to the service provider thus generating extra costs. We denote e_R the effort level selected by the service requester. Exerting effort e_R incurs a cost $c_R(e_R)$ for the requester for each patient in the population she is serving. The size of the population is denoted by v_0. When patients require specialized care, the service requester

[6] In practice, both the provider and the requester serve a variety of patients for a variety of conditions. We study coordinating contracts for one specific category of patient and condition, characterized by a treatment cost, treatment failure cost, effort cost function, referral volume function and chance of treatment failure function. The provider and the requester would then have a *portfolio* of contracts applicable to each category. This approach is inspired by the fixed fee per service payment model where the payment to the provider is different for each procedure and where the specific patient type affects the extent of the intervention and thus the price charged to the requester.

[7] Lewis et al. (2014) states that ACO contracts are generally based on one of three payment arrangements: shared-savings, global budgets, and capitated payments. Our model is consistent with two-sided shared-savings contracts (where both savings and losses are shared between the two parties), global budget contracts and capitated payments contracts between payer and requester.

refers them to the service provider. The total volume of patients referred to the service provider is affected by the effort exerted by the requester; we denote it as $v(e_R)$. The procedure provided by the service provider for the condition considered incurs a standard treatment cost T_1 per patient. Once a patient is referred, the provider is in charge of the patient's care, including aspects beyond the standard procedure per se, such as making sure the patient understands discharge instructions (e.g., dietary restrictions, surgical site care to avoid infection) and medication plan, ensuring the patient has appropriate support at home, following up by phone after discharge, providing nurse visits at home if necessary, contacting the primary care physician to coordinate post-discharge care, etc. In this role, the service provider selects effort level e_P, which incurs a cost of effort $c_P(e_P)$ for each patient. We emphasize that the provider's effort consists in optional post-treatment care that is additional to the standard procedure that the provider is paid for. Consistent with Andritsos and Tang (2015), it involves costly, but non-billable interventions. Such interventions are highly beneficial to the patient's health outcome by improving the chance of proper recovery, but the provider cannot bill the requester for implementing them. In particular, the provider's effort selection does not imply any choice of treatment route or intensity that under FFS would lead to different levels of reimbursement. For a study of how the provider's choice among a variety of treatment options is affected by the payment system, the reader may refer to Adida et al. (2017).

Following treatment, the patient returns under the care of the requester. However, there is a chance that the procedure "failed", and that further medical attention is required (e.g., readmission, rehabilitation, stabilization, extensive check-ups, etc.). In such a situation, the requester incurs a cost T_2 per patient. The effort exerted by the provider impacts the probability of treatment failure; we denote $q(e_P)$ the probability that the patient requires further care after receiving referral treatment (i.e., the "failure" probability).[8] Note that the preventive effort exerted by the requester does not affect the chance of treatment failure. This modeling choice is consistent with the definition of the requester's effort as entailing primarily preventive activities that aim at reducing the need for outside referrals. For example, educating a patient in lifestyle changes to avoid a complicated surgery has little impact on the patient's chance of infection of the surgical site, should the surgery end up being required. Essentially, for a given patient the requester's preventive effort either prevents the need of a referral or not. In the latter case, the effect of the requester preventive effort on the outcome of the provider's treatment is negligible. The chance of "success" of the referral procedure may vary mostly based upon proper medication management, steps to avoid infection, understanding discharge instruction, appropriate support at home, follow up nurse

[8] Because we focus on a specific patient type and a specific condition, costs and treatment outcomes tend to be homogeneous and thus we assume that the standard treatment cost T_1 and the complication cost T_2 are the same for all patients in this population, and that the chance of treatment failure is the same for all patients.

visits, etc., which are the responsibility of the provider. In Appendix F, we consider the case where the chance of treatment failure depends on both the provider and the requester's efforts and find that our main coordination results continue to hold. All appendices can be found in the online supplement.

The requester and the provider make their decisions with the goal of maximizing their own individual profit. We make no assumption regarding the order of the decisions. Namely, our results remain valid whether the provider moves first, the requester moves first, or they make simultaneous decisions. For ease of reference we provide a glossary of notation in Appendix A. We make some mild assumptions on the functions defined above.

Assumption 1 (Convexity of the cost, volume, and failure probability functions).

1. Effort costs $c_R(\cdot)$ and $c_P(\cdot)$ are non-negative convex increasing;
2. Patient volume $v(\cdot)$ is non-negative convex decreasing;
3. Probability of referral treatment failure $q(\cdot)$ is non-negative convex decreasing.

Assuming that costs of effort are convex means that the marginal costs are increasing, that is, there are "low-hanging" fruits: some measures may incur a rather low cost while having a positive impact on patients, but further actions become increasingly costlier. The convexity of cost of effort is a common assumption in the literature (e.g., Bhattacharyya and Lafontaine 1995, Xue and Field 2008, Roels 2014, Andritsos and Tang 2015). In a similar fashion, assuming that the patient volume and the failure probability are convex (decreasing) implies that the slope becomes gradually less and less steep, that is, an increment in effort when the current effort level is near zero has a large impact on both measures, but further effort becomes less influential. This assumption is consistent with the common assumption in the literature that efforts yield diminishing returns (Roels 2014).

We also make some technical assumptions:

Assumption 2. $v(\cdot)$, $q(\cdot)$ and $c_P(\cdot)$ satisfy the following inequalities for all effort levels:[9]

$$\frac{v\ v''}{v'^2} \cdot \frac{q\ q''}{q'^2} \geq 1, \qquad \frac{v\ v''}{v'^2} \cdot \frac{c_P\ c_P''}{c_P'^2} \geq 1. \tag{1}$$

These conditions, which are *sufficient* to guarantee that the optimization problems we consider are convex, can be viewed as a stronger version of Assumption 1, by requiring that $v(\cdot)$, $q(\cdot)$ and $c_P(\cdot)$ are *sufficiently* convex. Intuitively, these conditions ensure that there is some opportunity for improvement, that is, that exerting effort can be beneficial from a system perspective. The first expression indicates that the volume and the probability of treatment failure functions have to be, in combination, convex enough. That is, an incremental amount of requester and/or provider effort can

[9] Prime ($'$) and double prime ($''$) respectively denote the first and second derivatives.

rapidly decrease the referral volume and/or probability of treatment failure. Similarly, the second expression indicates that the volume and provider cost of effort functions are, in combination, convex enough. That is, an incremental amount of requester and/or provider effort can rapidly decrease the referral volume without incurring a very large cost of effort (i.e., the change in cost due to an increase in effort is much more gradual). Overall, these conditions imply that there are low-hanging fruits in term of the effectiveness of the requester and provider efforts. From a practitioner's perspective, this assumption is validated by research showing that when implemented, simple steps can have a big impact on prevention (Stampfer et al. 2000) and on avoiding readmissions (Silow-Carroll et al. 2011). Song (2014) observes that ACOs have generated substantial savings to date and that "the low-hanging fruit in Medicare seems to be admissions and readmissions, while that in commercial contracts may be lower prices obtained by changing referral patterns".

Finally, we note that our model focuses on the interaction between requester and provider, and does not explicitly include the contract between the requester and the payer. This approach is consistent with a setting where the requester is financially responsible for the entire healthcare cost of the beneficiaries, and her revenue is independent of the effort decisions made by the requester and the provider. Thus the requester aims at minimizing her operating costs (while maintaining quality of care). This framework matches several types of capitation-like contracts used by ACOs, including the two-sided shared-savings, global budget, and capitated payment contract with the payer (Lewis et al. 2014). Under these contracts, the ACO gets to keep all or a portion of any savings generated compared to a certain benchmark, and bears all or a fraction of the costs incurred beyond the benchmark. Thus, the ACO benefits from expenses lower than the benchmark, and incurs losses if the expenses exceed it. Such payment systems between ACO and payer thus incentivize the ACO to incur operating costs that are as low as possible, as long as quality standards are met, consistent with our modeling approach.[10]

4. Effort Levels Under Different Payment Systems

In this section, we present in detail the centralized system, the FFS payment system, and the penalty contract. Appendix E presents alternative payment systems including capitation, cost-sharing and shared-savings payment contracts.

[10] Having an outcome-based contract between the payer and the requester (Zorc et al. 2017), either as an exogenous penalty for poor outcomes (high failure rate and/or referral volume) or as an exogenous reward for better outcomes, would ultimately increase the incentives for the requester to exert higher levels of effort. However, the incentives for the provider under FFS would not change; hence, the underlying misalignment of incentives in the referral market would be exacerbated. Despite obtaining a different magnitude of the efforts, the coordination results and insights would remain the same.

4.1. Centralized System and the First-Best

In order to evaluate the performance of different payments schemes and understand what would be the "ideal" levels of effort to target, we need to define a benchmark. To this end, we analyze a centralized system comprising both decision-makers: the service provider and the service requester. Forming a single entity in the centralized system, their joint decisions are aimed at maximizing the cumulative total system profits. We note that the centralized system does not take patient utility into account. This is because a benchmark to compare the decentralized decisions must include only the profits of the agents making decisions. However, in Section 6 we define the patient utility and the social welfare, and we obtain the optimal effort levels for a system comprising the service provider, the service requester as well as the patient population.

In a centralized system, the requester and the provider operate as a unified care system. The *first-best* optimizes the total profits for the entire centralized system:

$$\Pi_T(e_P, e_R) = -v_0 c_R(e_R) - v(e_R)[c_P(e_P) + q(e_P)T_2 + T_1].$$

We make some remarks on the above expression. First, as detailed in Section 3, we omit the revenue that the requester receives for taking care of the population of patients[11] (e.g., from the payer in case of an ACO or from patients' premiums in case of an HMO), as well as the operating costs for providing standard care to these patients in-network. These cash flows are independent of the effort levels and volume of referrals, so we treat them as constants and omit them from the profit expression above to be maximized, leaving only the components that depend on the decisions we are analyzing in this model. Second, we observe that the centralized total profit is independent of the payment contract between the requester and the provider since the contract involves cash flows that are internal to the centralized system.

Before analyzing the first-best efforts, we establish a technical result. All the proofs are provided in Appendix D.

LEMMA 1. Π_T is jointly concave in (e_R, e_P).

[11] Our model can capture the situation whereby the payer imposes an exogenous penalty on the requester for poor patients outcomes (e.g., combined capitation and penalty contract). For instance, the requester's revenue from the payer could be contingent on failure probability $q(\cdot)$. Indeed, our current model would capture this situation by increasing the value of parameter T_2 (cost of a failed treatment) accordingly, and so the analysis and findings would continue to hold unchanged. The requester performance could also be measured by the volume of referrals, and the requester could receive a bonus for the patients who are not referred to an external provider. In such a situation, the requester (and the centralized) profit would have an added term $+a(v_0 - v(e_R))$. The first order conditions would be slightly modified; however, all our results remain unchanged under this scenario. The case where the payer endogenously adjusts payments to the requester based on her outcomes would require analyzing a different model incorporating the payers decision-making, and is left as a direction of future research.

Thus, the first-best solution (e_P^*, e_R^*) is obtained from the first-order conditions:

$$c_P'(e_P) = |q'(e_P)|T_2, \tag{2}$$
$$|v'(e_R)|[c_P(e_P) + q(e_P)T_2 + T_1] = v_0 c_R'(e_R).$$

In order to guarantee positive solutions for the first-best efforts (as well as for efforts under other payment systems considered in the paper), a *sufficient* condition is $c_P'(0) = 0$ and $c_R'(0) = 0$ (Bhattacharyya and Lafontaine 1995, Kim and Wang 1998, Corbett and DeCroix 2001). In the rest of the paper, similar to Corbett and DeCroix (2001) and Roels (2014), we focus on non-border equilibria.

We now investigate some comparative statics properties of the first-best solution.

LEMMA 2. The first-best efforts are such that

- e_P^* is invariant in T_1 and increasing in T_2.
- e_R^* is increasing in both T_1 and T_2.

These results are consistent with intuition. The provider exerts more effort at the first-best when the cost of treatment failure T_2 increases, since this cost affects the overall profit. On the other hand, when the treatment cost T_1 increases, there is no reason to increase the provider effort. Indeed, the provider effort only affects the chance of treatment failure and thus costs incurred only in case of treatment failure. Conversely, the centralized system has an incentive to increase the requester effort to either reduce the volume of patients requiring treatment if the treatment cost T_1 increases, or susceptible to the unsuccessful treatment cost if the failure cost T_2 increases.

4.2. Traditional Payment Contract: FFS

The traditional reimbursement practice in the healthcare industry is FFS, that is, a fixed single-payment per service transaction.

While such a payment scheme presents great simplicity and ease of implementation, it has some major drawbacks, which have been studied in the literature (Jiang et al. 2012, Adida et al. 2017). A FFS payment structure rewards the provider for volume (number of patients), not for providing care of better value or improving patient outcomes. The misalignment of incentives under FFS is the main reason why in the Affordable Care Act, several new payment initiatives are being tested to improve quality of care while reducing spending.

Under FFS the requester pays a fixed, set in advance, price w^{FFS} for each patient receiving the treatment from the provider. The fixed price is exogenous and independent of the provider's effort decision. In other words, as detailed in Section 3, we focus on a given treatment procedure subject to a given level of reimbursement, and we model the provider's effort as *additional* care (e.g. follow-up care), which is not billable. This effort does not directly affect the treatment cost,

which corresponds to a standard intensity of care. On the other hand, the provider's effort can increase the chance of a successful treatment, but it is optional (and costly) for the provider. The requester and the provider profits under FFS are given by:

$$\Pi_P(e_P, e_R) = v(e_R)[w^{FFS} - c_P(e_P) - T_1]$$
$$\Pi_R(e_P, e_R) = -v_0 c_R(e_R) - v(e_R)[w^{FFS} + q(e_P)T_2].$$

To ensure participation of the provider, we assume that the price w^{FFS} is high enough to guarantee the provider a non-negative profit at least if he were to exert no effort:

$$w^{FFS} \geq c_P(0) + T_1. \tag{3}$$

We examine the provider and requester decisions. Clearly, Π_P is decreasing in e_P, so consistent with Andritsos and Tang (2015), the optimal decision for the provider is to exert no effort: $e_P^{FFS} = 0$. Indeed, the provider has no incentive to exert effort that incurs a cost for him and only benefits the requester (by lowering the chance of treatment failure) and the patient. Thus, if it is desired that the provider exert some level of effort, the provider must be otherwise compensated to be induced to do so. Furthermore, under Assumption 1, Π_R is concave in e_R. Therefore, the optimal effort level e_R^{FFS} satisfies the first-order condition:

$$|v'(e_R)|[w^{FFS} + q(0)T_2] = v_0 c_R'(e_R). \tag{4}$$

Similarly to the centralized case, we note that $e_R^{FFS} > 0$. The following lemma establishes how the requester effort changes when the contracted price per procedure varies.

LEMMA 3. e_R^{FFS} is increasing in w^{FFS}.

This result illustrates that when the requester must pay more for each referral, she has more incentives to intensify her effort aimed at reducing the volume of referrals required.

4.3. Penalty Contract

The Hospital Readmissions Reduction Program (HRRP) was established under the Affordable Care Act and started being implemented in 2013. Its goal is to create incentives for hospitals to reduce avoidable readmissions, by reducing the amount that the insurer pays when too many patients are readmitted within a pre-defined time window after discharge (Zuckerman et al. 2016, Zhang et al. 2016). The payment penalty has reached up to 3% since 2015.

Inspired by the HRRP, we consider a penalty contract in which the provider receives a fixed fee per patient, and is retrospectively penalized for treatment failures (Lee and Zenios 2012). (For differences between the penalty contract and the HRRP, refer to the literature review in Section 2). The provider receives w^{PEN} for each patient treated successfully. If the treatment fails, the provider only receives a fraction f of payment w^{PEN} (i.e., $1-f$ represents the extent of the penalty

imposed on the provider). Hence, the provider has an incentive to exert effort to reduce the chance of failure, as any unsuccessful treatment lowers the revenue received from the requester. Similar to FFS, the readmission penalty contract incentivizes the requester to exert effort by imposing a payment w^{PEN} for each referral (although the payment is reduced to $w^{PEN}(1-f)$ in case of failure). The provider and the requester profits are given by:

$$\Pi_P(e_P, e_R) = v(e_R)[w^{PEN}(1 - q(e_P)(1-f)) - c_P(e_P) - T_1]$$
$$\Pi_R(e_P, e_R) = -v_0 c_R(e_R) - v(e_R)[w^{PEN}(1 - q(e_P)(1-f)) + q(e_P)T_2].$$

To ensure participation of the provider, we assume that the price w^{PEN} is high enough to guarantee the provider a non-negative profit when he exerts his optimal effort e_P^{PEN}. That is, we assume

$$w^{PEN}(1 - q(e_P^{PEN})(1-f)) \geq c_P(e_P^{PEN}) + T_1. \tag{5}$$

By Assumption 1, the provider's profit is concave in e_P and the requester's profit is concave in e_R. As a result, the optimal effort levels e_P^{PEN} and e_R^{PEN} satisfy the first-order conditions:

$$c_P'(e_P) = |q'(e_P)| w^{PEN}(1-f) \tag{6}$$
$$|v'(e_R)|[w^{PEN}(1 - q(e_P)(1-f)) + q(e_P)T_2] = v_0 c_R'(e_R).$$

We now investigate how the efforts under a given penalty contract are affected by changes in the treatment and failure costs. Lemma 4 and Lemma 5 summarize these findings.

Lemma 4. For fixed penalty contract parameters, f and w^{PEN}, resulting efforts are such that:

- e_P^{PEN} is invariant in T_1 and in T_2.
- e_R^{PEN} is invariant in T_1 and increasing in T_2.

The treatment cost T_1 only affects the provider's profit, so changing it does not affect the effort exerted by the requester. The provider incurs the treatment cost for every patient referred; his level of effort does not change the patient volume, so the provider's effort is not sensitive to the treatment cost T_1. The provider's effort is not influenced by the cost T_2 born by the requester in case of treatment failure, since that cost does not affect the provider's profit. The requester incurs the failure cost T_2 for every patient whose treatment fails. Hence, increasing effort would reduce the volume of patients referred, and thus the volume of failed treatments, therefore a higher failure cost T_2 motivates the requester to increase her effort.

In addition we also investigate how the efforts under the penalty contract are affected by changes in the contract terms.

Lemma 5. The effort levels are such that:

- e_P^{PEN} is increasing in w^{PEN} and decreasing in f;

- e_R^{PEN} is increasing in w^{PEN} for $w^{PEN}(1-f) \geq T_2$, but not necessarily monotonic otherwise; e_R^{PEN} is increasing in f for $w^{PEN}(1-f) \leq T_2$, but not necessarily monotonic otherwise.

The provider has an incentive to exert more effort when the loss of profit per failed treatment experienced due to the penalty, $w^{PEN}(1-f)$, increases (i.e., w^{PEN} increases and/or f decreases), as in that case the provider would exert effort to ensure fewer treatments fail. When the payment per patient keeps increasing past the point where the penalty compensates the failed treatment cost, the requester is incentivized to exert more effort to reduce the volume of referrals. Likewise, if the fraction f received by the provider increases on the domain where the penalty does not compensate the failed treatment cost, the potential cost to the requester incurred by each referred patient increases and thus the requester exerts more effort to lower the number of patients referred.

5. Discussion and Coordination to the First-Best

Consistent with the literature on supply chain management (e.g., Cachon 2003), we use the centralized setting as a benchmark to compare payment contracts to. Thus, we define coordination as matching the first-best decisions. Specifically, a coordinating contract would lead to the maximum possible joint total profits for the system comprising the service requester and the service provider. In Section 6, we consider an alternate coordination goal that aims at aligning decisions with the social optimum, that is, the optimal solution of a system including the patients as well as the requester and provider. In such a case, the social welfare, comprising the joint profits of the service requester and the service provider as well as the patients' utility, would be maximized.

5.1. FFS

Under FFS, the provider exerts zero effort, regardless of the price paid per patient. Hence there is no FFS contract (i.e., there is no price w^{FFS}) that may coordinate the provider's effort. Moreover, the FFS requester effort does not match the first-best requester effort in general, as precised below.

PROPOSITION 1. $e_R^{FFS} > e_R^*$ for all FFS prices w^{FFS} such that (3) holds.

There are two reasons driving the FFS requester effort to exceed the first-best effort. First, under FFS the provider exerts no effort to limit the chance of treatment failure and hence the complications costs that the requester is responsible for. Thus, to curb these costs, the requester must lower the volume of patients more than under the first-best by exerting high effort. Second, under FFS for each referral the requester pays the provider a fee that covers not only the effort and treatment costs incurred, but also a profit margin (due to condition (3)) which is absent at the first-best. Therefore, the requester under FFS has incentives to exert more effort than at the first-best to reduce the volume of referrals and thus these profit margin payments.

It follows from Proposition 1 that there is no price w^{FFS} that may coordinate the requester's effort. As a result, a FFS contract is unable to coordinate either the requester or the provider effort levels with those at the first-best.

5.2. Penalty Contract

Penalty vs. first-best. We compare the efforts under a penalty contract with the first-best efforts for a given contract characterized by w^{PEN} and f.

PROPOSITION 2. The requester effort $e_R^* \leq e_R^{PEN}$. In addition, the provider effort $e_P^* < e_P^{PEN}$ iff $T_2 < w^{PEN}(1-f)$.

Under the penalty contract, for each referral the requester pays the provider a fee that covers not only the effort and treatment cost incurred, but also a profit margin (due to condition (5)) which is absent at the first-best. Thus the requester has incentives to exert more effort than at the first-best to reduce the volume of referrals and these profit margin payments. Also, the failure cost T_2 affects the provider effort at the first-best (Lemma 2) but not under a penalty contract (from (6)). If the failure cost is low, at the first-best the provider exerts less effort as complications have less impact. Moreover, if the revenue loss under the penalty contract is large, the provider's effort is high to curb the loss from the penalty.

Coordination. It appears from Proposition 2 that it is possible to select a price w^{PEN} and a penalty factor f such that the provider effort under the penalty contract aligns with the first-best effort, as long as $w^{PEN}(1-f) = T_2$. Moreover, while in general the requester's effort under the penalty contract is larger than or equal to the first-best effort, it is possible to ensure coordination. We formalize this in the result below.

THEOREM 1. If $c_P(e_P^*) + T_1 > (1 - q(e_P^*))T_2$, a penalty contract with price $w^{PEN} = q(e_P^*)T_2 + c_P(e_P^*) + T_1$ and penalty factor $f = 1 - T_2/w^{PEN}$ coordinates the provider's and requester's efforts.

The above key result shows that the penalty contract can align the decisions of the decentralized system to those at the first-best.[12] We make three observations. First, we believe it is not trivial that a two-parameter penalty contract can coordinate decisions. Indeed, the two-parameter cost-sharing, shared-savings and two-part tariff contracts are unable to achieve coordination (see Appendix E). Second, this result differs from the traditional supply chain coordination result stating that a

[12] The condition in Theorem 1 is not necessary to ensure the coordination result, rather it guarantees that the coordinating fraction f is non-negative. A negative fraction f implies a coordinating penalty $1-f$ that exceeds 100%, that is, the service provider would not be paid anything in case of failed treatment, and would even have to pay the service requester. This would be the case when the treatment cost is low compared to the failure cost ($T_1 \ll T_2$), the provider effort cost is low, and yet at the first-best the probability of success is not sufficiently high. In such a situation, the requester experiences high failure cost, while the provider experiences low treatment and effort cost. Thus the provider would have to compensate the requester for treatment failures, which are encouraged by the too low chance of success due to the too low provider effort.

two-part tariff contract can coordinate the decisions in a supply chain with one upstream agent and one downstream agent (e.g., Cachon and Lariviere 2005). As mentioned in Appendix E, in our framework the penalty contract does not reduce to a two-part tariff contract; furthermore a two-part tariff contract cannot coordinate the effort decisions. Third, we establish the robustness of this result in Appendix F where we show that this finding continues to hold when the chance of treatment failure depends on both the requester's and the provider's efforts.

Comparative statics. We obtain the following comparative statics result on the coordinating contract parameters.

PROPOSITION 3. The coordinating penalty contract parameters are such that w^{PEN} is increasing in both T_1 and T_2, and f is increasing in T_1 and decreasing in T_2.

This result states that, as the treatment cost T_1 increases, the price paid by the requester for treating each patient and the fraction of payment received by the provider in case of failure both increase. This is because a higher treatment cost makes the first-best requester effort increase and leaves the first-best provider effort unchanged (Lemma 2), while the efforts under the penalty contract are both unchanged (Lemma 4). Thus the penalty contract terms must be modified to ensure that (i) the requester adjusts her effort level upwards and (ii) the provider effort remains unchanged. Making the requester pay more for each referral gives the requester incentives to reduce the volume by increasing her effort, achieving goal (i). However, this alone would increase the loss of profit $w^{PEN}(1-f)$ per failed treatment experienced by the provider, which not only would break coordination, but also would incentivize the provider to increase his effort (see Lemma 5). Thus, to ensure that the provider maintains his effort level (goal (ii)), the increase in per-patient payment is accompanied by an increase in the penalty fraction, so that $w^{PEN}(1-f)$ remains constant.

As the cost of unsuccessful treatment T_2 increases, by Lemma 2 the first-best provider and requester efforts increase as well. Moreover, the provider effort under the penalty contract is unchanged, while the requester's increases (Lemma 4). Moreover, because of condition (5), the first-best requester effort increases faster than the penalty contract requester effort. Hence, in order to maintain coordination the penalty contract must provide incentives to (i) increase the provider effort, and (ii) slightly increase the requester effort. Increasing both the per-patient payment w^{PEN} and the penalty fraction $1-f$ achieves both goals. First, the loss of profit $w^{PEN}(1-f)$ per failed treatment experienced by the provider increases, giving him incentives to exert more effort in order to reduce the number of treatment failures. Second, the increase in w^{PEN} tends to incentivize the requester to exert more effort to reduce the volume of referrals, while the increase in the penalty fraction reduces the cost that she bears for each failed treatment, which mitigates the incentive to increase volume and thus ensures that the effort only slightly increases.

6. Patient Utility and Social Welfare

6.1. Patient Utility

So far we have ignored the impact of effort level decisions on patients. Yet, the efforts that the requester and provider exert directly affect the value that patients receive. In our model, effort may reduce the need for advanced treatment by the provider and the chance of complications. Thus we define the patient's utility in terms of a loss that represents the discomfort of undergoing advanced treatment, and, if applicable, the potential distress of facing complications. A referral treatment results in a value loss of $-u_1$. If the patient requires further care due to unsuccessful treatment, the total value loss is $-u_2$, where $u_2 > u_1 > 0$. Thus, the overall patient population's (dis)utility is:

$$\Pi_{PT}(e_P, e_R) = -v(e_R)(u_1 + (u_2 - u_1)q(e_P)).$$

Note that higher effort levels are always preferred from the patients' perspective.

LEMMA 6. $\Pi_{PT}(e_P, e_R)$ is concave.

By coordinating the efforts to the first-best, we avoid a loss of efficiency due to the decentralization of service operations between the requester and the provider. However, the first-best efforts do not generally result in the best outcome for patients. For instance, although FFS does not provide incentives for the provider to exert any effort, it does incentivize the requester to exert an effort larger than at the first-best. This trade-off could potentially result in FFS yielding better outcomes than the first-best for patients. We further investigate this point numerically in Section 7.

6.2. Social Welfare

In this section, we introduce the notion of social welfare. The centralized system studied in Section 4.1 comprises the service requester and the service provider, but not the patients. Optimizing the requester and provider's joint profit yields the first-best. Hence, coordination to the first-best disregards the patients' benefit. When designing a healthcare payment system, it is important to take into consideration the patients' utility in addition to the profits of the firms involved in providing care. To this end, we introduce a broader system that considers the patients, the requester and the provider, and we define the corresponding social welfare value function as

$$\begin{aligned}\Pi_S(e_P, e_R) &= \Pi_R(e_P, e_R) + \Pi_P(e_P, e_R) + \Pi_{PT}(e_P, e_R) = \Pi_T(e_P, e_R) + \Pi_{PT}(e_P, e_R) \\ &= -v_0 c_R(e_R) - v(e_R)[c_P(e_P) + q(e_P)\tilde{T}_2 + \tilde{T}_1],\end{aligned}$$

where $\tilde{T}_2 = T_2 + u_2 - u_1$ includes the treatment failure cost born by the requester and patient, and $\tilde{T}_1 = T_1 + u_1$ includes the treatment cost born by the provider and patient. We observe that the social welfare is identical to the centralized profit $\Pi_T(e_R, e_P)$ after substituting the modified cost

of treatment $\tilde{T_1}$ $(>T_1)$ for T_1, and the modified failure cost $\tilde{T_2}$ $(>T_2)$ for T_2. Therefore, results for socially optimal effort levels follow directly from the results we obtained for the first-best.

COROLLARY 1. By Lemma 2, the socially optimal effort levels (e_P^S, e_R^S) are above the first-best efforts: $e_P^S \geq e_P^*, \quad e_R^S \geq e_R^*$.

Because patients benefit from higher effort levels, adding the patients value function to the centralized profits results in obtaining effort levels that are higher than at the first-best.

Moreover, it follows from Appendix E that capitation, cost-sharing and shared-savings contracts cannot coordinate both types of effort to the socially optimal effort levels. However, a penalty contract can do so. The result below follows from Theorem 1 and Proposition 3.

COROLLARY 2. If $c_P(e_P^S) + \tilde{T_1} > (1 - q(e_P^S))\tilde{T_2}$, a penalty contract can coordinate the efforts to the socially optimal effort levels. Furthermore, the coordinating contract requires a larger price w^{PEN} and penalty $1 - f$ than what is needed to coordinate to the first-best.

A larger payment per patient for the requester, and a larger penalty for the provider are required to coordinate to the socially optimal efforts so as to incentivize higher efforts than at the first-best.

We note that while it is in the requester and the provider's joint best interest to design a payment contract aligning their decisions to those at the first-best, in certain contexts they may choose to align their decisions to the social optimum. For instance, coordination to the social optimum captures the case when the requester and provider are (at least partially) altruistic. This situation may realistically arise in a health care environment where the firms' mission is rooted in public health. Hence the firms may seek a contract that not only increases their profit, but also improves patient heath outcomes, which we incorporate via the patients' utility.

7. Numerical Analysis

In this section, we evaluate numerically the performance of FFS and of the coordinating contract using a variety of performance measures. We have established that a FFS contract cannot coordinate the decisions to the first-best or to the social optimal decisions. Yet, it is the contract most commonly used in practice. Alternatively, the decision-makers could use a coordinating penalty contract to optimize their joint utility, achieving the first-best. However, this contract does not take the patients' utility into consideration, and hence does not maximize the social welfare. Hence, we aim to evaluate how FFS and the coordinating contract affect the patients and the social welfare.

We first give some background on how the inputs are selected in our numerical examples. Then we evaluate how FFS performs compared to a coordinating contract for: (i) the requester-provider system, (ii) the patients, and (iii) the system comprising the requester, provider, and patients (via the social welfare).

7.1. Example of Procedures and Choice of Inputs

Table 1 presents four examples of procedures and their average reimbursement, reimbursement paid for readmission, and readmission rate. For instance, a Coronary Artery Bypass Graft (CABG) is an open heart surgical procedure used to resolve certain blockages in the heart. While it has a rather high readmission rate, the amount paid in case of readmission is only 32% of the initial reimbursement. On the other hand, a Cesarean section (C-section) is a much simpler procedure, less costly, and less likely to require readmission, but the reimbursement for a readmission averages 110% of the initial payment. Albeit imperfect, the relationship between the initial reimbursement and the readmission reimbursement can serve as a reasonable indicator of the relationship between the treatment cost and the failure cost. Thus, CABG can be viewed as an example of procedure for which the treatment cost T_1 is high compared to the failure cost T_2, and the chance of readmission is high, while C-section is an example of procedure where the reverse is true.[13] For the two other procedures presented in the table, the costs ratio and the chance of readmission are intermediate. Our numerical experiments illustrate the role played by these factors.

Table 1 Example of procedures.

Procedure	Number of cases	Mean reimbursement [$/case]	Readmission rate [%]	Mean readmission reimbursement [$/case]
CABG	136,057	40,068	12	13,127
Hip replacement	479,491	18,836	7	14,314
Hysterectomy	252,461	11,604	5	10,953
Cesarean section	1,143,070	6,443	2	7,104

National statistics from 2013 obtained from HCUPnet, Agency for Healthcare Research and Quality (2013).

We consider the following functional forms for the cost of effort, probability of treatment failure, and volume. The cost of effort functions are $c_P(e_P) = c_0^P e^{\lambda e_P}$ and $c_R(e_R) = c_0^R e^{\kappa e_R}$. The probability of treatment failure is modeled as $q(e_P) = q_0 e^{-\gamma e_P}$. The volume is represented as $v(e_R) = v_0 e^{-\delta e_R}$. Parameters λ, κ, γ, and δ are positive. It is easy to check that these functions satisfy Assumptions 1 and 2. The solutions obtained in our numerical experiments result in non-border solutions.

In the examples presented in Table 1, the initial reimbursements, indicative of (but larger than) the treatment costs, are much more heterogeneous than the readmission reimbursements, indicative of the failure costs. Hence, we fix the treatment failure cost at $T_2 = 10,000$, and we consider two possible values of the treatment cost $T_1 \in \{5,000;\ 20,000\}$ to capture the wide possible range for

[13] The requester can exert preventive and educational effort to reduce the risk of requiring CABG (stop smoking, lose weight, consume less alcohol) or a C-section (exercise, limit weight gain, stop smoking, take childbirth classes). If the procedure is needed, the provider can, in addition to delivering the procedure, exert follow up effort to avoid or to identify early some of the complications (e.g., infection) that the patient might experience after treatment.

the ratio of these two costs (while noting that neither scenario is an exact match to the procedures presented in Table 1). The FFS price must cover the treatment cost to ensure participation of the provider, and should also cover expenses like overhead as well as a profit margin. Therefore, we vary the FFS price in the range $[T_1 + c_0^P, 3T_1]$. We consider two possible values for parameter q_0 representing the chance of failure when the provider exerts no effort: $q_0 \in \{5\%;\ 15\%\}$. For the population volume, we use $v_0 = 1000$ (the results are independent of this value).

The impact of the provider's effort on reducing the probability of treatment failure is captured through γ. Jack et al. (2009) estimate that 30-day readmissions can be reduced by 30% by simply educating patients. In the case of CABG, Perk et al. (1990) report that a comprehensive post-operative rehabilitation program reduced readmission rates from 32% to 14%. Thus we consider that the provider effort can reduce failure rates by 30-60% compared to the no-effort failure rate q_0. We choose $\gamma = 0.2$ to ensure that the decrease in the failure rate is within the 30-60% range when the provider exerts the first-best effort.

The impact of the requester's effort on reducing the referral volume is captured through δ. The World Health Organization recommends maintaining C-section rates between 10-15% (World Health Organization 2015); the U.S. C-section rate for low-risk births is 23.9% (Centers for Desease Control and Prevention 2016). To reach recommended rates, the U.S. has to cut down rates by 40-60%.[14] Thus we choose $\delta = 0.2$ so that the reduction in volume due to the requester effort is within the 40-60% range. We tested smaller values of δ and the insights are similar.

For the provider cost of effort function, we use parameters $c_0^P = 10$ and $\lambda = 1$. Jack et al. (2009) tested an educational discharge program for general health services; they estimate that the additional labor cost includes 1.5 hours of nurse time and 0.5 hour of pharmacist time per patient. Considering median hourly salaries of \$27 for a nurse and of \$55 for a pharmacist (PayScale Inc. 2016), the additional cost per patient is \$68. This choice of parameters is such that the cost of effort at the provider's first-best effort matches the additional cost of \$68 per patient.[15] Furthermore, we use $c_0^R = 10$ and $\kappa = 1$ as a way to normalize the cost of effort for both decision-makers.

To estimate the patient disutility parameters u_1 and u_2, we use medical leave time as a proxy.[16] Gehring et al. (1988) report that the patients that returned to work after CABG took 6 months or

[14] Some hospitals in California have recently managed to reduce the C-section rate by 20% in 6 months (California Health Care Foundation 2016), and some hospitals have already reached the recommended rate, so with sufficient effort, this reduction should be achievable.

[15] For a complex procedure like CABG, the cost c_0^P would be much larger because specialists and surgeons might also be involved in the discharge program. However, the insights derived from the numerical analysis do not change significantly by increasing c_0^P.

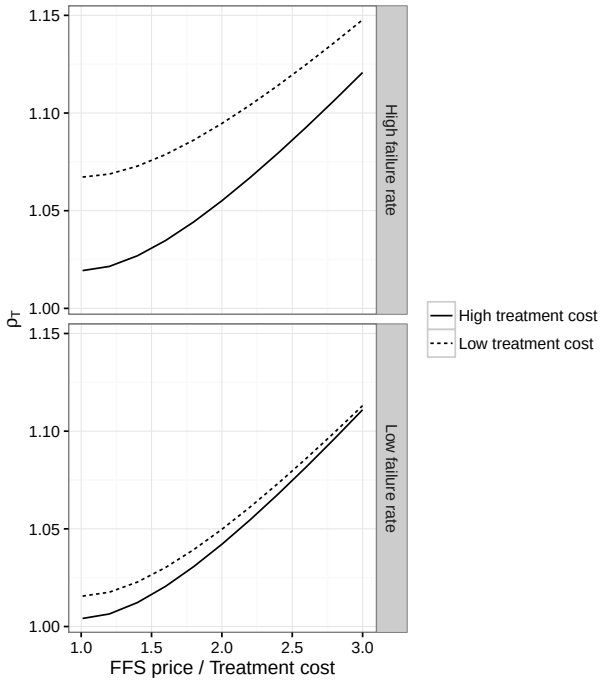
[16] Perk et al. (1990) report that a CABG post-surgical rehabilitation program did not affect the time to return to work. This aligns with our assumption that patient costs are independent of effort levels.

less of medical leave. In the case of a C-section, the California maternity leave disability insurance covers two additional weeks (State of California, Employment Envelopment Department 2016). Since the real median household income is $53,482 (based on the U.S. Census Bureau), we consider $u_1 \in \{2,000;\ 10,000\}$, and set $u_2 = 3u_1$. Table 2 in Appendix B summarizes the value of the parameters that we use in the numerical implementation.

7.2. Requester-Provider System

To measure the performance of the FFS contract compared to a coordinating contract from the perspective of the requester-provider system, we compute the ratio $\rho_T = \Pi_T(e_P^{FFS}, e_R^{FFS})/\Pi_T(e_P^*, e_R^*)$. Notice that the ratio is always above 1 since the first-best efforts maximize the system profit (i.e., minimize the system cost) and $\Pi_T \leq 0$. Thus, the larger the ratio, the more inefficient FFS is. In Figure 3, we illustrate the behavior of ρ_T when the FFS price varies.[17]

Figure 3 Inefficiency of FFS for the Requester-Provider system.



Note. $q_0 = 15\%$ (top plot), $q_0 = 5\%$ (bottom plot), $T_1 = 20,000$ (solid line), $T_1 = 5,000$ (dashed line).

The FFS inefficiency is increasing in the margin (w^{FFS}/T_1) of the FFS contract. This can be explained as follows. The FFS provider effort remains at zero regardless of the FFS price. The FFS requester effort is above the first-best effort (Proposition 1), and it increases with the price paid for the service (Lemma 3). Hence, a higher FFS price results in a larger gap between the requester effort under FFS and first-best, and ultimately worsens the inefficiency of the FFS contract.

[17] Note that the values of u_1 and u_2 have no impact on the inefficiency ratio.

A change in the treatment cost has a pronounced effect on the inefficiency of FFS, viewed by comparing the solid and dashed lines in Figure 3. We note that FFS is more inefficient when the treatment cost is low. We recall from Section 4.2 that the FFS efforts are independent of the treatment cost. However, by Lemma 2, a lower treatment cost induces the requester to exert less effort at the first-best. Hence, a lower treatment cost brings the requester first-best effort further from the FFS effort, and the FFS contract becomes more inefficient.

The effect of the treatment failure probability can be seen by comparing the top and bottom panels in Figure 3. For a fixed treatment cost and margin value (x-axis coordinate), a higher failure rate results in a moderately larger FFS inefficiency. Indeed, the provider first-best effort increases when the failure rate becomes high to directly influence the chance of treatment failure and hence of extra system costs. Since the FFS provider effort remains at zero, the gap between first-best and FFS provider efforts widens, which increases inefficiency.

In summary, the coordinating contract is significantly better than FFS for the requester and the provider especially when the FFS margin is large, the treatment cost is low relative to the failure cost, and/or the treatment failure rate is high (the last factor being less influential).

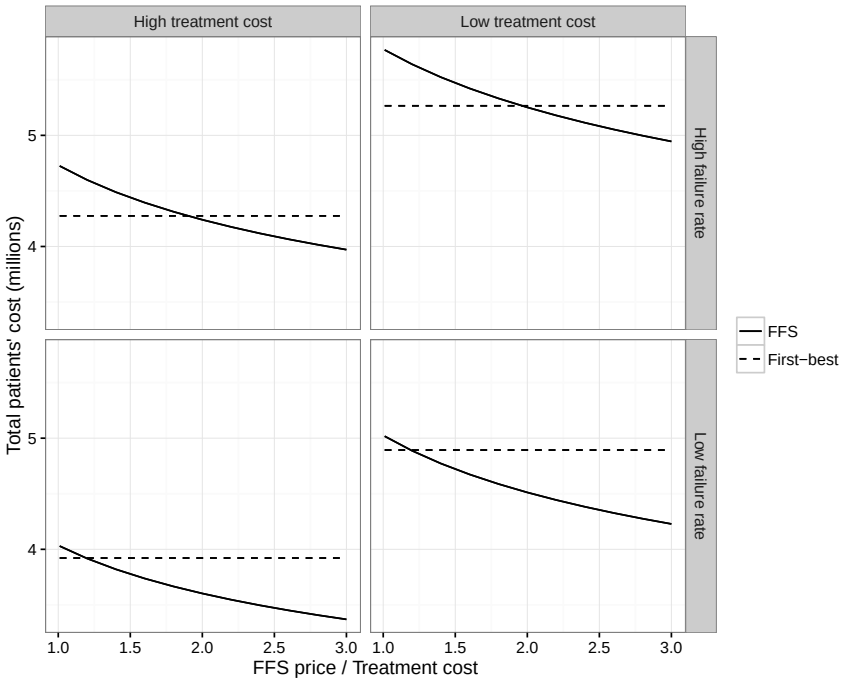
These results indicate that the provider and requester have little to gain from implementing a coordinating contract for a procedure such as CABG due to its high treatment cost relative to the failure cost (despite its less influential high failure rate). However, for a procedure such as a C-section, characterized by a low treatment cost, the provider and the requester have more to gain by replacing FFS with a coordinating contract, despite the less influential low failure rate, especially when the FFS margin is relatively high. This observation is particularly relevant when changing the payment system would incur high fixed costs (communication, staff training, computer system changes, etc.) that must be compensated by significant gains.

7.3. Patient Utility

We now aim to measure the performance of the FFS contract compared to a coordinating contract from the perspective of the patients. We evaluate the patients' cost (i.e., $-\Pi_{PT}$) under both a FFS contract and a coordinating contract. The results are shown in Figure 4.

Interestingly, we observe that, even though the coordinating contract does not take into account the patients' utility, patients may be better off under this contract than a FFS contract in certain situations. Specifically, we observe that when (i) the FFS margin is low and/or (ii) the failure rate is high, the patients' total cost tends to be lower at the first-best than under FFS. We can explain these observations as follows. Recall that patients benefit from higher effort levels from both the requester and the provider, but are not negatively impacted by high prices or costs. (i) A

Figure 4 Total patients' cost: First-best vs. FFS.



Note. $q_0 = 15\%$ (top plots), $q_0 = 5\%$ (bottom plots), $T_1 = 20,000$ (left plots), $T_1 = 5,000$ (right plots), $(u_1, u_2) = (10,000, 30,000)$ (high patient disutility).

low FFS price (or margin) results in low requester effort under FFS (Lemma 3), without affecting the FFS provider effort or the first-best efforts. Hence a low FFS price (or margin) disadvantages patient utility. (ii) A higher failure rate increases the provider first-best effort to directly mitigate the higher chance of incurring failure costs, which benefits the patient utility. Hence, a high FFS price and high failure rate result in overall higher efforts at the first-best than under FFS, and thus lower patients' total cost. Note that the treatment cost does not have a strong impact on whether patients prefer FFS or the coordinating contract. We also analyzed the case where the patient's disutility is low (by assuming $(u_1, u_2) = (2,000, 6,000)$), and recover the same insights and behavior as shown in Figure 4.

For example, for a procedure like CABG, where patients experience high disutility, patients could benefit from the implementation of a coordinating contract compared to the currently common FFS payment scheme, if the FFS margin is not too high (e.g., see top-left panel in Figure 4 when the FFS margin is less than 1.8, corresponding to a FFS price lower than 38,000). However, patients would be slightly worse off if a coordinating contract is used instead of FFS for a procedure like a C-section (where patients experience low disutility, and the patient costs follow a pattern similar to that of the bottom-right panel in Figure 4).

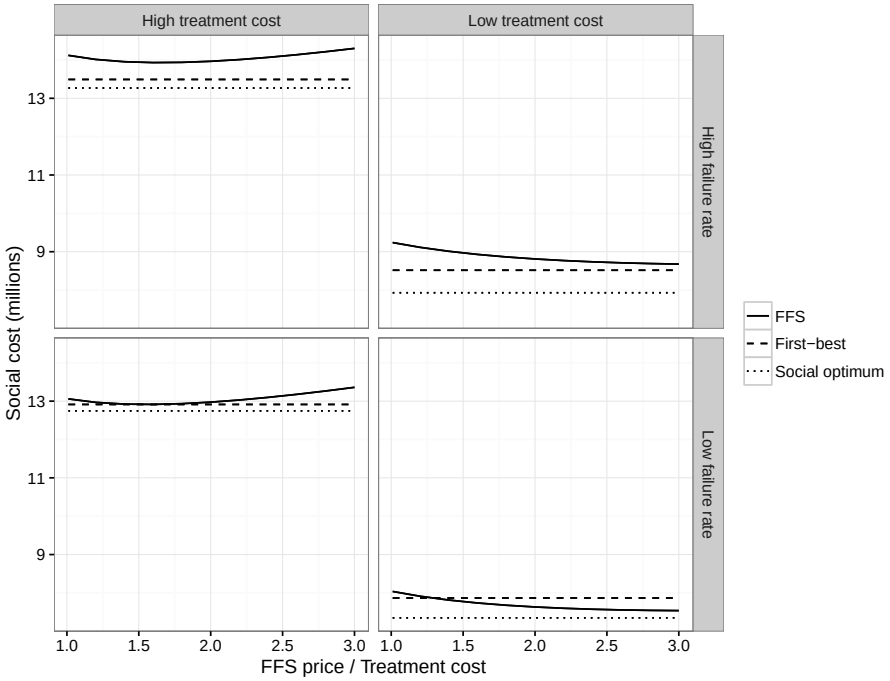
We note that the criteria that make the patients better off under a coordinating contract than under FFS do not agree with those that make a coordinating contract most worthwhile for the

requester and provider (with the exception of the failure rate, which has only a moderate effect on the requester-provider system's contract preference). Hence in general, coordination between requester and provider is more likely to take place in circumstances where the patients' utility would be negatively impacted (e.g., both CABG and C-section lead to patients' interest at odds with the requester-provider system).

7.4. Social Welfare

We now aim to measure the performance of the FFS contract compared to a coordinating contract from the perspective of the social welfare. We determine how the FFS and coordinating contracts affect the social welfare, that is, the system comprising the patients, the requester and the provider. We established above that the interests of the patients are generally at odds with those of the provider and requester regarding when a coordinating contract is most beneficial as compared to FFS. The social welfare is a combination of the patients' utility and the provider and the requester's profits. Hence it is unclear how the FFS and first-best compare to each other in terms of social welfare. Furthermore, we compare the social welfare under FFS and under the first-best coordinating contract to the optimal social welfare (i.e., the social optimum). Figure 5 illustrates the social cost (i.e., $-\Pi_S$) under a FFS contract, a coordinating contract, and at the social optimum.

Figure 5 Social cost: First-best vs. FFS.

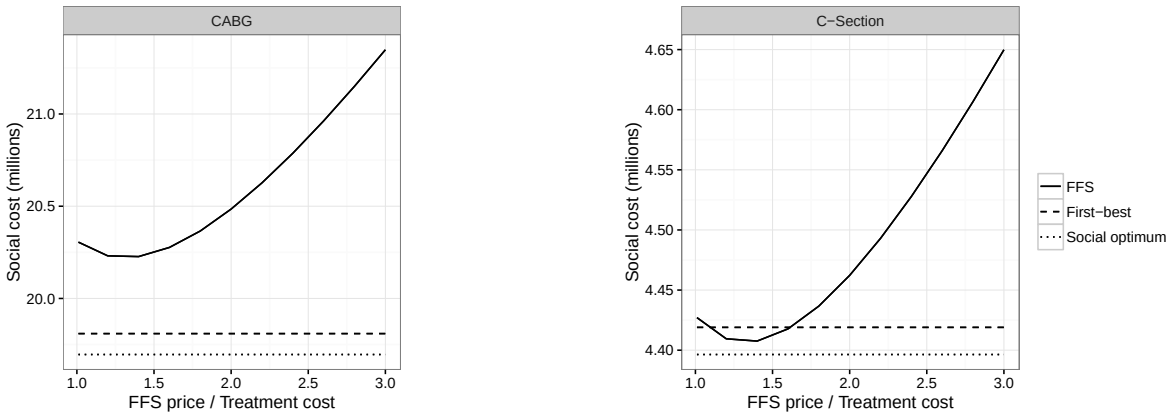


Note. $q_0 = 15\%$ (top plots), $q_0 = 5\%$ (bottom plots), $T_1 = 20,000$ (left plots), $T_1 = 5,000$ (right plots), $(u_1, u_2) = (10,000, 30,000)$ (high patient disutility).

We observe that, while it is possible for the social cost to be higher at the first-best than under FFS (e.g., with high FFS margin, low treatment cost, low failure rate, see bottom-right panel in Figure 5), in general, a FFS contract leads to a higher social cost than the coordinating contract. We found that in the case of low patient utility ($(u_1, u_2) = (2,000, 6,000)$), the first-best contract is better than FFS across all the studied scenarios. In other words, the possible negative impact on patients of implementing a coordinating contract is typically outweighed by the benefit for the provider and requester.

In particular, for procedures like CABG and C-section (corresponding to problem instances close to those represented in Figure 6), the social cost is generally higher under FFS than under a coordinating contract. For a C-Section, however, the FFS contract could be better for society when the FFS margin is low (i.e, when the FFS price is relatively small compared to the treatment cost). Hence, society could benefit from requester-provider coordination for some types of procedures (CABG), but be better off without coordination for others (C-Section under low FFS margins). Admittedly, the social cost is higher under the coordinating contract than at the social optimum,

Figure 6 Social cost for CABG and C-Section procedures.



Note. $T_1 = 40,000$, $T_2 = 13,000$, $q_0 = 12\%$, $(u_1, u_2) = (10,000, 30,000)$.

Note. $T_1 = 6,440$, $T_2 = 7,100$, $q_0 = 2\%$, $(u_1, u_2) = (2,000, 6,000)$.

by definition. However, the requester and provider have no financial incentive to take the patients' utility into account, and thus to coordinate to the social optimum, when they decide on a payment contract. Still, it is reassuring to observe that if the requester and provider decide to adopt a new contract to maximize their joint profits, ignoring the patients' utility, the resulting contract would lead to an outcome that generally benefits the social welfare compared to the currently used FFS contract. Besides, we observe that the social cost at the first-best is generally very close to the optimal social cost. Therefore, little social welfare is lost by implementing the first-best via a coordinating contract selected by the provider and requester, instead of the social optimum.

8. Conclusion

This paper examines a B2B contracting problem for the payment of referral services between healthcare organizations. The way the service provider is paid by the service requester affects the incentives on both parties to exert effort aiming at improving the quality of care while keeping costs under control. There is a consensus that misaligned incentives under FFS result in higher costs for the two organizations and in poor patient outcomes.

A variety of new reimbursement systems between payers and caregivers have been proposed to better align incentives and improve patient outcomes. These proposed payment systems aim at moving away from a fee-for-service payment scheme and toward a fee-for-performance system. However, the findings do not necessarily apply to the interaction between effort-exerting service requester and provider because, contrary to the requester, the payer makes no medical decision directly affecting the patient care. Our analysis sheds some light on how these payment systems may perform within the context of this interaction. Table 3 in Appendix C summarizes our findings.

Our results confirm that FFS lacks incentives to achieve optimal efforts, mainly because the provider does not receive any reward for exerting effort improving chances of good patient outcomes. We also show (in the Appendix) that capitation, cost-sharing and shared-savings cannot coordinate simultaneously the provider's and the requester's efforts. Interestingly, we find that modifying the FFS payment system by incorporating a carefully designed penalty for poor patient outcomes can achieve either the first-best or the social optimum.

We find that the service requester and provider most benefit from implementing a coordinating contract for procedures with a high FFS price, a treatment cost low relative to the failure cost, and/or a high failure rate. Effort decisions made under the coordinating contract do not take patient utility into consideration, as the requester and provider simply aim at maximizing their joint profit and have no incentive to do otherwise. Yet, we find that in some cases, patients may be better off when such a coordinating contract is used than under FFS, especially when the FFS price is low and/or the failure rate is high. In most cases, the social welfare (for a system comprising the requester, provider and patients) improves under the coordinating contract as compared to FFS, and is close to the social optimum.

As the ACO model spreads in the U.S., so does its influence on referrals. Reducing duplication of services, controlling cost, and ensuring continuity and quality of care are some of the benefits that ACOs can obtain by effectively managing patients referrals. Developing a high-quality, low-cost network of external providers is crucial for improving the financial and quality of care performance of any ACO. Our work provides new insights for ACOs (especially for those led by physician groups)

and their interaction with external providers. Specifically, it highlights some of the implications of adopting different contracts on the referral service quality and cost. Furthermore, our work can be used to inform ACOs' and HMOs' decisions on which providers to include as part of the network of preferred providers depending on their cost structures and effectiveness of their effort in reducing the chances of treatment failure.

Acknowledgments

The authors would like to thank Charles Corbett and Victor Tabbush for their comments on an earlier version of the manuscript. The authors are also thankful to the review team for their constructive and helpful suggestions. The discussions during the authors' presentations at the 2016 MEDS seminar at Kellogg School of Management, 2016 INFORMS Annual meeting, 2017 POMS Annual Conference, 2017 MSOM Annual Conference, and 2017 INFORMS Healthcare conference were also beneficial.

References

Adida, E., H. Mamani, S. Nassiri. 2017. Bundled payment vs. fee-for-service: Impact of payment scheme on performance. *Management Science* **63**(5) 1606–1624.

Agency for Healthcare Research and Quality. 2013. H·CUPnet data base. URL `http://hcupnet.ahrq.gov/`.

Andritsos, D.A., C.S. Tang. 2015. Incentive programs for reducing readmissions when patient care is co-produced. Working paper.

Arbaje, A.I., J.L. Wolff, Q. Yu, N.R. Powe, G.F. Anderson, C. Boult. 2008. Postdischarge environmental and socioeconomic factors and the likelihood of early hospital readmission among community-dwelling Medicare beneficiaries. *The Gerontologist* **48**(4) 495–504.

Babich, V., C.S. Tang. 2012. Managing opportunistic supplier product adulteration: Deferred payments, inspection, and combined mechanisms. *Manufacturing & Service Operations Management* **14**(2) 301–314.

Balachandran, K.R., S. Radhakrishnan. 2005. Quality implications of warranties in a supply chain. *Management Science* **51**(8) 1266–1277.

Barnes, A.J., L. Unruh, A. Chukmaitov, E. van Ginneken. 2014. Accountable Care Organizations in the USA: Types, developments and challenges. *Health Policy* **118**(1) 1–7.

Bhattacharyya, S., F. Lafontaine. 1995. Double-sided moral hazard and the nature of share contracts. *The RAND Journal of Economics* 761–781.

Bravo, F., R. Levi, G. Perakis, G. Romero. 2016. A risk-sharing pricing contract in B2B service-based supply chains. Working paper.

Burns, L., R. Muller. 2008. Hospital-physician collaboration: Landscape of economic integration and impact on clinical integration. *Milbank Quarterly* **86**(3) 375–434.

Cachon, G.P. 2003. Supply chain coordination with contracts. A.G. de Kok, S.C. Graves, eds., *Handbooks in Operations Research and Management Science: Volume 11: Supply Chain Management: Design, Coordination and Operation*, chap. 6. Elsevier.

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

California Health Care Foundation. 2016. Reducing unnecessary Cesarean-section deliveries in California. URL `http://www.chcf.org/projects/2015/reducing-cesarean-sections`.

Centers for Desease Control and Prevention. 2016. Births - Method of delivery. URL `http://www.cdc.gov/nchs/fastats/delivery.htm/`.

Chalkley, M., J.M. Malcomson. 2002. Cost sharing in health service provision: An empirical assessment of cost savings. *Journal of Public Economics* **84**(2) 219–249.

Chao, G.H., S.M.R. Iravani, R.C. Savaskan. 2009. Quality improvement incentives and product recall cost sharing contracts. *Management Science* **55**(7) 1122–1138.

Christianson, J.B., S. Leatherman, K. Sutherland. 2008. Lessons from evaluations of purchaser pay-for-performance programs: A review of the evidence. *Medical Care Research and Review* **65**(6 suppl) 5S–35S.

Colla, C.H., V.A. Lewis, E. Tierney, D.B. Muhlestein. 2016. Hospitals participating in ACOs tend to be large and urban, allowing access to capital and data. *Health Affairs* **35**(3) 431–439.

Corbett, C.J., G.A. DeCroix. 2001. Shared-savings contracts for indirect materials in supply chains: Channel profits and environmental impacts. *Management Science* **47**(7) 881–893.

Corbett, C.J., G.A. DeCroix, A.Y. Ha. 2005. Optimal shared-savings contracts in supply chains: Linear contracts and double moral hazard. *European Journal of Operational Research* **163** 653–667.

El Ouardighi, F. 2014. Supply quality management with optimal wholesale price and revenue sharing contracts: A two-stage game approach. *International Journal of Production Economics* **156** 260–268.

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

Frakt, A.B., R. Mayes. 2012. Beyond capitation: How new payment experiments seek to find the 'sweet spot' in amount of risk providers and payers bear. *Health Affairs* **31**(9) 1951–1958.

Gehring, J., W. Koenig, N.W. Rana, P. Mathes. 1988. The influence of the type of occupation on return to work after myocardial infarction, coronary angioplasty and coronary bypass surgery. *European Heart Journal* **9**(Suppl L) 109–114.

Gerhardt, G., A. Yemane, P. Hickman, A. Oelschlaeger, E. Rollins, N. Brennan. 2013. Data shows reduction in Medicare hospital readmission rates during 2012. *Medicare and Medicaid Research Review* **3**(2) E1–E12.

Gold, M.R., R. Hurley, T. Lake, T. Ensor, R. Berenson. 1995. A national survey of the arrangements managed-care plans make with physicians. *The New England Journal of Medicine* **333**(25) 1678–1683.

Hacker, K., R. Mechanic, P. Santos. 2014. Accountable care in the safety net: A case study of the Cambridge Health Alliance. *The Commonwealth Fund* **13**(11756).

Jack, B.W., V.K. Chetty, D. Anthony, J.L. Greenwald, G.M. Sanchez, A.E. Johnson, S.R. Forsythe, J.K. O'Donnell, M.K. Paasche-Orlow, C. Manasseh, S. Martin, L. Culpepper. 2009. A reengineered hospital discharge program to decrease rehospitalization: A randomized trial. *Annals of Internal Medicine* **150**(3) 178–187.

Jack, W. 2005. Purchasing health care services from providers with unknown altruism. *Journal of Health Economics* **24**(1) 73–93.

Jelovac, I. 2001. Physicians' payment contracts, treatment decisions and diagnosis accuracy. *Health Economics* **10**(1) 9–25.

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

Kaya, M., Ö. Özer. 2011. Pricing in business-to-business contracts: Sharing risk, profit and information. Ö. Özer, R. Phillips, eds., *The Oxford Handbook of Pricing Management*, chap. 29. Oxford University Press.

Kim, S.-H., S. Netessine. 2013. Collaborative cost reduction and component procurement under information asymmetry. *Management Science* **59**(1) 189–206.

Kim, S.K., S. Wang. 1998. Linear contracts and the double moral-hazard. *Journal of Economic Theory* **82**(2) 342–378.

Kuraitis, V. 2011. Patient "leakage": Rethinking two field of dreams assumptions about ACOs. URL `https://tinyurl.com/y932rdtc`. Published on June 27.

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

Lee, H.-H., E.J. Pinker, R.A. Shumsky. 2012. Outsourcing a two-level service process. *Management Science* **58**(8) 1569–1584.

Leng, M., M. Parlar. 2010. Game-theoretic analyses of decentralized assembly supply chains: Non-cooperative equilibria vs. coordination with cost-sharing contracts. *European Journal of Operational Research* **204**(1) 96–104.

Lewis, V.A., C.H. Colla, W.L. Schpero, S.M. Shortell, E.S. Fisher. 2014. ACO contracting with private and public payers: A baseline comparative analysis. *The American Journal of Managed Care* **20**(12) 1008–1014.

Liu, X., X. Cai, R. Zhao, Y. Lan. 2015. Mutual referral policy for coordinating health care systems of different scales. *International Journal of Production Research* **53**(24) 7411–7433.

Ma, P., H. Wang, J. Shang. 2013. Contract design for two-stage supply chain coordination: Integrating manufacturer-quality and retailer-marketing efforts. *International Journal of Production Economics* **146**(2) 745–755.

Malcomson, J.M. 2004. Health service gatekeepers. *RAND Journal of Economics* **35**(2) 401–421.

Mariñoso, B.G., I. Jelovac. 2003. GPs' payment contracts and their referral practice. *Journal of Health Economics* **22**(4) 617–635.

McCluskey, P.D. 2015. Blue Cross extends revised contracts for care. *Boston Globe.* Published on October 5, 2015.

McGuire, T.G. 2000. Physician agency. *Handbook of Health Economics* **1** 461–536.

McWilliams, J.M., M.E. Chernew, J.B. Dalton, B.E. Landon. 2014. Outpatient care patterns and organizational accountability in Medicare. *Journal of the American Medical Association – Internal Medicine* **174**(6) 938–945.

Oss, M.E. 2016. ACOs & hospitals – The changing landscape. URL `https://tinyurl.com/yb8wttoq`. Published on May 2.

PayScale Inc. 2016. Pay Scale. URL `http://www.census.gov//`.

Perk, J., B. Hedbäck, J. Engvall. 1990. Effects of cardiac rehabilitation after coronary artery bypass grafting on readmissions, return to work, and physical fitness. A case-control study. *Scandinavian Journal of Public Health* **18**(1) 45–51.

Pinker, E., R. Shumsky, H.-H. Lee, S. Hasija. 2010. Managing the outsourcing of two-level service processes: Literature review and integration. *Proceedings of the Annual Hawaii International Conference on System Sciences* 1–10.

Ren, Z., Y. Zhou. 2008. Call center outsourcing: Coordinating staffing level and service quality. *Management Science* **54**(2) 369–383.

Reyniers, D.J., C.S. Tapiero. 1995. The delivery and control of quality in supplier-producer contracts. *Management Science* **41**(10) 1581–1589.

Robinson, J.C., L.D. Schaeffer. 2015. Referral management and disease management in California's Accountable Care Organizations. Issue Brief 15, Integrated Healthcare Association.

Roels, G. 2014. Optimal design of coproductive services: Interaction and work allocation. *Manufacturing & Service Operations Management* **16**(4) 578–594.

Selviaridis, K., F. Wynstra. 2014. Performance-based contracting: A literature review and future research directions. *International Journal of Production Research* **7543** 37–41.

Silow-Carroll, S., J.N. Edwards, A. Lashbrook. 2011. Reducing hospital readmissions: Lessons from top-performing hospitals. Tech. rep., The Commonwealth Fund.

SK&A. 2014. SK&A Market insight report – Top 30 Accountable Care Organizations. Tech. rep., IMS Health.

Song, Z. 2014. Accountable Care Organizations in the U.S. health care system. *Journal of Clinical Outcomes Management* **21**(8) 364–371.

Song, Z., D.G. Safran, B.E. Landon, Y. He, R.P. Ellis, R.E. Mechanic, M.P. Day, M.E. Chernew. 2011. Health care spending and quality in year 1 of the alternative quality contract. *New England Journal of Medicine* **365**(10) 909–918.

Stampfer, M.J., F.B. Hu, J.E. Manson, E.B. Rimm, W.C. Willett. 2000. Primary prevention of coronary heart disease in women through diet and lifestyle. *The New England Journal of Medicine* **343** 16–22.

State of California, Employment Envelopment Department. 2016. State disability insurance. URL `http://www.edd.ca.gov/disability/`.

Tu, T., W. Caughey, D. Muhlestein. 2016. Accountable Care Organizations and risk-based payment arrangements: Strong preference for upside-only contracts. *Leavitt Partners* URL `http://leavittpartners.com/wp-content/uploads/2016/11/ACO_risks_whitepaper_v1.pdf`.

Tu, T., D. Muhlestein, L. Kocot, R. White. 2015. The impact of accountable care: Origins and future of accountable care organizations. *Leavitt Partners* URL `https://www.brookings.edu/wp-content/uploads/2016/06/Impact-of-Accountable-CareOrigins-052015.pdf`.

Vlachy, J., T. Ayer, M. Ayvaci, S. Raghunathan. 2017. The business of healthcare: Physician integration in bundled payments. Working paper.

Weisman, R. 2012. Steward, Partners hospitals reach deal on trauma care. URL `https://tinyurl.com/ybwo9p5l`. Published on October 19.

World Health Organization. 2015. WHO statement on caesarean section rates. URL `http://www.who.int/reproductivehealth/publications/maternal_perinatal_health/cs-statement/en//`.

Xue, M., J.M. Field. 2008. Service coproduction with information stickiness and incomplete contracts: Implications for consulting services design. *Production and Operations Management* **17**(3) 357–372.

Zhang, D.J., I. Gurvich, J.A. Van Mieghem, E. Park, R.S. Young, M.V. Williams. 2016. Hospital readmissions reduction program: An economic and operational analysis. *Management Science* **62**(11) 3351–3371.

Zorc, S., S.E. Chick, S. Hasija. 2017. Outcomes-based reimbursement policies for chronic care pathways. Working paper.

Zuckerman, R.B., S.H. Sheingold, E.J. Orav, J. Ruhter, A.M. Epstein. 2016. Readmissions, observation, and the Hospital Readmissions Reduction Program. *The New England Journal of Medicine* (374) 1543–1551.

Zuvekas, S.H., J.W. Cohen. 2016. Fee-for-service, while much maligned, remains the dominant payment method for physician visits. *Health Affairs* **35**(3) 411–414.

Online supplement

Appendix A: Notation

v_0	patient population size
e_R	service requester's effort level
e_P	service provider's effort level
$c_R(\cdot)$	service requester's cost of effort function
$c_P(\cdot)$	service provider's cost of effort function
$v(\cdot)$	volume of patient requiring care from the service provider (function of the service requester's effort level)
$q(\cdot)$	probability that the patient requires further care (function of the service provider's effort level)
T_1	treatment cost incurred by the service provider
T_2	cost incurred by the service requester in case the patient requires further care
u_1	utility loss incurred by the patient when undergoing treatment
u_2	utility loss incurred by the patient when undergoing treatment and then requiring further care
$\tilde{T}_1$	$= T_1 + u_1$
$\tilde{T}_2$	$= T_2 + u_2 - u_1$
$\Pi_R(\cdot,\cdot)$	service requester's profit function
$\Pi_P(\cdot,\cdot)$	service provider's profit function
$\Pi_T(\cdot,\cdot)$	service provider and requester's joint profit function
$\Pi_{PT}(\cdot,\cdot)$	patient population utility function
$\Pi_S(\cdot,\cdot)$	social welfare value function
e_R^*, e_P^*	first-best efforts
e_R^S, e_P^S	socially optimal efforts
w^{FFS}	price per patient treated charged by the service provider under FFS
e_R^{FFS}, e_P^{FFS}	efforts selected under FFS
w^{PEN}	price per patient treated charged by the service provider under the penalty contract
f	fraction of the payment kept by the provider in case the patient requires further care under the penalty contract
e_R^{PEN}, e_P^{PEN}	efforts selected under readmission penalty
ρ_T	inefficiency ratio $\Pi_T(e_P^{FFS}, e_R^{FFS})/\Pi_T(e_P^*, e_R^*)$
$c_0^P,\ \lambda$	parameters of the cost of effort function in Section 7: $c_P(e_P) = c_0^P e^{\lambda e_P}$
$c_0^R,\ \kappa$	parameters of the cost of effort function in Section 7: $c_R(e_R) = c_0^R e^{\kappa e_R}$
$q_0,\ \gamma$	parameters of the probability function in Section 7: $q(e_P) = q_0 e^{-\gamma e_P}$
δ	parameter of the patient volume function in Section 7: $v(e_R) = v_0 e^{-\delta e_R}$

Appendix B: Parameter Values in the Numerical Study

Table 2 Summary of the value of parameters.

T_1	$\in \{5,000; 20,000\}$
T_2	10,000
q_0	$\in \{5\%; 15\%\}$
γ	0.2
v_0	1000
δ	0.2
c_0^P, c_0^R	10
λ, κ	1
u_1	$\in \{2,000; 10,000\}$
u_2	$\in \{6,000; 30,000\}$

Appendix C: Summary of the Coordination Results

Table 3 Summary of the coordination results of the various payment systems.

	FFS	Capitation	Cost-sharing	Shared-savings	Penalty
Parameters	w^{FFS}	C	w^{CS}, α	w^{SS}, β	w^{PEN}, f
May coordinate e_R to first-best	No ($e_R^{FFS} > e_R^*$)	No	Yes	Yes	Yes
May coordinate e_P to first-best	No ($e_P^{FFS} = 0$)	No ($e_P^{CAP} = 0$)	No ($e_P^{CS} < e_P^*$)	No ($e_P^{SS} < e_P^*$)	Yes
May coordinate e_R to social optimum	No ($e_R^{FFS} > e_R^S$)	No	Yes	Yes	Yes
May coordinate e_P to social optimum	No ($e_P^{FFS} = 0$)	No ($e_P^{CAP} = 0$)	No ($e_P^{CS} < e_P^S$)	No ($e_P^{SS} < e_P^S$)	Yes

Appendix D: Proofs

Proof of Lemma 1

For the centralized system total profit, we have that

$$\begin{aligned}
\frac{\partial^2 \Pi_T}{\partial e_P^2} &= -v(e_R)(c_P''(e_P) + q''(e_P)T_2) \\
\frac{\partial^2 \Pi_T}{\partial e_R^2} &= -v''(e_R)(c_P(e_P) + q(e_R)T_2 + T_1) - v_0 c_R''(e_R) \\
\frac{\partial^2 \Pi_T}{\partial e_R \partial e_P} &= -v'(e_R)(c_P'(e_P) + q'(e_P)T_2).
\end{aligned}$$

Under Assumption 1, $\frac{\partial^2 \Pi_T}{\partial e_P^2} < 0$, and $\frac{\partial^2 \Pi_T}{\partial e_R^2} < 0$. Thus, for the Hessian to be definite negative we need its determinant to be positive,

$$\underbrace{v(v''T_1 + v_0 c_R'')(q''T_2 + c_P'')}_{>0} + \underbrace{vv''(qT_2 + c_P)(q''T_2 + c_P'') - v'^2(q'T_2 + c_P')^2}_{**} > 0.$$

Conditions (1) in Assumption 2 guarantee $** \geq 0$. Therefore, Π_T is jointly concave. □

Proof of Lemma 2

We recall the first-order conditions for maximizing $\Pi_T(e_R^*, e_P^*)$,

$$c_P'(e_P^*) + q'(e_P^*)T_2 = 0,$$
$$v_0 c_R'(e_R^*) + v'(e_R^*)[c_P(e_P^*) + q(e_P^*)T_2 + T_1] = 0.$$

Note that the first equation uniquely identifies e_P^* and this does not depend on T_1. To see the variation with respect to T_2 we can take total derivatives in both sides and rearrange terms to get

$$[c_P''(e_P^*) + q''(e_P^*)T_2]\frac{\partial e_P^*}{\partial T_2} = -q'(e_P^*).$$

Given the convexity of $c_P(\cdot)$ and $q(\cdot)$ together with $q(\cdot)$ decreasing, we have that $\frac{\partial e_P^*}{\partial T_2} > 0$.

To analyze the requester effort, we use the second first-order condition and take total derivatives with respect to T_1. After rearranging terms we get the following equality

$$-\frac{\partial e_R^*}{\partial T_1}\underbrace{[v''(e_R^*)(c_P(e_P^*) + q(e_P^*)T_2 + T_1) + v_0 c_R''(e_R^*)]}_{\geq 0} = \underbrace{v'(e_R^*)}_{<0}.$$

By the convexity Assumption 1, e_R^* is increasing in T_1.

Similarly, we take total derivatives with respect to T_2, and rearrange terms to obtain

$$-\frac{\partial e_R^*}{\partial T_2}\underbrace{[v''(e_R^*)(c_P(e_P^*) + q(e_P^*)T_2 + T_1) + v_0 c_R''(e_R^*)]}_{\geq 0}$$
$$= v'(e_R^*)[q(e_P^*) + \underbrace{(c_P'(e_P^*) + q'(e_P^*)T_2)}_{=0,\ from\ e_p\ FOC}\frac{\partial e_P^*}{\partial T_2}] = \underbrace{v'(e_R^*)q(e_P^*)}_{<0}.$$

By the convexity Assumption 1, e_R^* is increasing in T_2. □

Proof of Lemma 3

We recall the first-order conditions for finding the requester effort under FFS:

$$|v'(e_R)|[w^{FFS} + q(0)T_2] = v_0 c_R'(e_R).$$

To see the variation with respect to w^{FFS} we take total derivatives in both sides and rearrange terms to get:

$$\frac{\partial e_R^{FFS}}{\partial w^{FFS}}[v_0 c_R''(e_R^{FFS}) + v''(e_R^{FFS})(w^{FFS} + q(0)T_2)] = -v'(e_R).$$

By Assumption 1, we have that $\frac{\partial e_R^{FFS}}{\partial w^{FFS}} > 0$. □

Proof of Lemma 4

The provider's effort under the contract with a fixed f and w^{PEN} is given by the solution of the first-order condition $c_P'(e_P) = |q'(e_P)|w^{PEN}(1-f)$, which is independent of T_1 and T_2, hence e_P^{PEN} is also independent

of T_1 and T_2. The requester's effort under the contract with a fixed f and w^{PEN} is given by the solution of the first-order condition

$$-v'(e_R^{PEN})[w^{PEN}(1-q(e_P^{PEN})(1-f))+q(e_P^{PEN})T_2]=v_0c_R'(e_R^{PEN}),$$

which is independent of T_1, hence e_R^{PEN} is also independent of T_1. In addition, since e_P^{PEN} is independent of T_2, we obtain

$$-\frac{\partial e_R^{PEN}}{\partial T_2}v''(e_R)[w^{PEN}(1-q(e_P^{PEN})(1-f))+q(e_P^{PEN})T_2]-v'(e_R)q(e_P^{PEN})=v_0\frac{\partial e_R^{PEN}}{\partial T_2}c_R''(e_R^{PEN}),$$

thus

$$\frac{\partial e_R^{PEN}}{\partial T_2}=\frac{-v'(e_R^{PEN})q(e_P^{PEN})}{v_0c_R''(e_R^{PEN})+v''(e_R^{PEN})[w^{PEN}(1-q(e_P^{PEN})(1-f))+q(e_P^{PEN})T_2]}>0,$$

where the last inequality follows from Assumption 1. □

Proof of Lemma 5

The provider's effort is given by the solution of the first-order condition $-q'(e_P^{PEN})w^{PEN}(1-f)=c_P'(e_P^{PEN})$. We obtain

$$-\frac{\partial e_P^{PEN}}{\partial f}q''(e_P^{PEN})w^{PEN}(1-f)+q'(e_P^{PEN})w^{PEN}=\frac{\partial e_P^{PEN}}{\partial f}c_P''(e_P^{PEN}),$$

hence

$$\frac{\partial e_P^{PEN}}{\partial f}=\frac{q'(e_P^{PEN})w^{PEN}}{q''(e_P^{PEN})w^{PEN}(1-f)+c_P''(e_P^{PEN})}<0,$$

where the last inequality follows from Assumption 1. Similarly,

$$-\frac{\partial e_P^{PEN}}{\partial w^{PEN}}q''(e_P^{PEN})w^{PEN}(1-f)-q'(e_P^{PEN})(1-f)=\frac{\partial e_P^{PEN}}{\partial w^{PEN}}c_P''(e_P^{PEN}),$$

hence

$$\frac{\partial e_P^{PEN}}{\partial w^{PEN}}=\frac{-q'(e_P^{PEN})(1-f)}{q''(e_P^{PEN})w^{PEN}(1-f)+c_P''(e_P^{PEN})}>0,$$

where the last inequality follows from Assumption 1.

The requester's effort is given by the solution of the first-order condition

$$-v'(e_R^{PEN})[w^{PEN}(1-q(e_P^{PEN})(1-f))+q(e_P^{PEN})T_2]=v_0c_R'(e_R^{PEN}).$$

By taking derivatives, we obtain

$$-\frac{\partial e_R^{PEN}}{\partial f}v''(e_R^{PEN})[w^{PEN}(1-q(e_P^{PEN})(1-f))+q(e_P^{PEN})T_2]$$

$$-v'(e_R^{PEN})[w^{PEN}q(e_P^{PEN})+\frac{\partial e_P^{PEN}}{\partial f}q'(e_P^{PEN})(-w^{PEN}(1-f)+T_2)]=v_0\frac{\partial e_R^{PEN}}{\partial f}c_R''(e_R^{PEN}),$$

thus

$$\frac{\partial e_R^{PEN}}{\partial f}=\frac{-v'(e_R^{PEN})[w^{PEN}q(e_P^{PEN})+\frac{\partial e_P^{PEN}}{\partial f}q'(e_P^{PEN})(-w^{PEN}(1-f)+T_2)]}{v_0c_R''(e_R^{PEN})+v''(e_R^{PEN})[w^{PEN}(1-q(e_P^{PEN})(1-f))+q(e_P^{PEN})T_2]}.$$

By Assumption 1 and on the domain $T_2\geq w^{PEN}(1-f)$, we have that $\frac{\partial e_R^{PEN}}{\partial f}>0$.

We also obtain

$$-\frac{\partial e_R^{PEN}}{\partial w^{PEN}} v''(e_R^{PEN})[w^{PEN}(1-q(e_P^{PEN})(1-f))+q(e_P^{PEN})T_2]$$
$$-v'(e_R^{PEN})\left[1-q(e_P^{PEN})(1-f)+\frac{\partial e_P^{PEN}}{\partial w^{PEN}} q'(e_P^{PEN})(T_2-w^{PEN}(1-f))\right] = v_0 \frac{\partial e_R^{PEN}}{\partial w^{PEN}} c_R''(e_R^{PEN}),$$

thus

$$\frac{\partial e_R^{PEN}}{\partial w^{PEN}} = \frac{-v'(e_R^{PEN})\left[1-q(e_P^{PEN})(1-f)+\frac{\partial e_P^{PEN}}{\partial w^{PEN}} q'(e_P^{PEN})(T_2-w^{PEN}(1-f))\right]}{v_0 c_R''(e_R^{PEN})+v''(e_R^{PEN})[w^{PEN}(1-q(e_P^{PEN})(1-f))+q(e_P^{PEN})T_2]}.$$

By Assumption 1 and on the domain $T_2 \le w^{PEN}(1-f)$, we have that $\frac{\partial e_R^{PEN}}{\partial w^{PEN}} > 0$. □

Proof of Proposition 1

Let us denote the first order conditions for e_R of the first-best and FFS as

$$\phi^T(e_P,e_R) = |v'(e_R)|[c_P(e_P)+q(e_P)T_2+T_1]-v_0 c_R'(e_R)$$
$$\phi^{FFS}(e_P,e_R) = |v'(e_R)|[w^{FFS}+q(e_P)T_2]-v_0 c_R'(e_R).$$

The first-best effort levels satisfy $\phi^T(e_P^*,e_R^*)=0$. The FFS effort levels satisfy $\phi^{FFS}(e_P^{FFS},e_R^{FFS}) = \phi^{FFS}(0,e_R^{FFS})=0$. In addition, since Π_T is concave in e_R (Lemma 1), we have that $\phi^T(e_P,e_R)$ is decreasing in e_R. As a result, if $\phi^T(e_P^*,e_R^{FFS})\le 0$, then $e_R^* \le e_R^{FFS}$.

$$\phi^T(e_P^*,e_R^{FFS}) = \phi^T(e_P^*,e_R^{FFS})-\phi^{FFS}(0,e_R^{FFS}) = |v'(e_R^{FFS})|[c_P(e_P^*)+T_1-w^{FFS}+(q(e_P^*)-q(0))T_2].$$

Furthermore, by definition e_P^* maximizes the first-best total profit over e_P and by concavity (Lemma 1) it is the unique maximizer. That is, e_P^* is the unique minimizer of $c_P(e_P)+q(e_P)T_2$. Thus,

$$c_P(e_P^*)+q(e_P^*)T_2 \le c_P(0)+q(0)T_2, \tag{7}$$

and the inequality is strict if $e_P^*>0$. It follows that

$$\phi^T(e_P^*,e_R^{FFS}) \le |v'(e_R^{FFS})|[c_P(0)+T_1-w^{FFS}] \le 0,$$

where the second inequality follows from (3) and the first inequality is strict if $e_P^*>0$. □

Proof of Proposition 2

The requester first order condition under the penalty contract corresponds to

$$\phi^{PEN}(e_P,e_R) = |v'(e_R)|[w^{PEN}(1-q(e_P)(1-f))+q(e_P)T_2]-v_0 c_R'(e_R)$$

and $\phi^{PEN}(e_P^{PEN},e_R^{PEN})=0$. Thus,

$$\phi^{PEN}(e_P^{PEN},e_R^*) = \phi^{PEN}(e_P^{PEN},e_R^*)-\phi^T(e_P^*,e_R^*)$$
$$= |v'(e_R^*)|[w^{PEN}(1-q(e_P^{PEN})(1-f))+q(e_P^{PEN})T_2-q(e_P^*)T_2-c_P(e_P^*)-T_1].$$

Because e_P^* is the unique minimizer of $c_P(e_P)+q(e_P)T_2$, we have

$$c_P(e_P^*)+q(e_P^*)T_2 \leq c_P(e_P^{PEN})+q(e_P^{PEN})T_2.$$

Therefore,

$$\phi^{PEN}(e_P^{PEN},e_R^*) \geq |v'(e_R^*)|[w^{PEN}(1-q(e_P^{PEN})(1-f))-c_P(e_P^{PEN})-T_1] \geq 0.$$

where the last inequality follows from condition (5). Since $\phi^{PEN}(e_P,e_R)$ is decreasing in e_R, this means that $e_R^* \leq e_R^{PEN}$.

The optimal effort for the provider is e_P^{PEN} such that $c_P'(e_P^{PEN})=|q'(e_P^{PEN})|(1-f)w^{PEN}$. We recall that e_P^* is defined by $c_P'(e_P^*)=|q'(e_P^*)|T_2$. It is clear that $e_P^* < e_P^{PEN}$ iff $T_2 < w^{PEN}(1-f)$. □

Proof of Theorem 1

It is clear from Proposition 2 that $f=1-T_2/w^{PEN}$ ensures $e_P^{PEN}=e_P^*$. From the proof of Proposition 2, it follows that if $f=1-T_2/w^{PEN}$, and $w^{PEN}=q(e_P^*)T_2+c_P(e_P^*)+T_1$, then $e_R^{PEN}=e_R^*$. The condition in the proposition guarantees $f>0$. □

Proof of Proposition 3

The coordinating penalty contract $f=1-T_2/w^{PEN}$, and $w^{PEN}=q(e_P^*)T_2+c_P(e_P^*)+T_1$ induces first-best effort levels (e_P^*,e_R^*). Taking total derivatives, $\frac{\partial w^{PEN}}{\partial T_2}=(q'(e_P^*)T_2+c_P'(e_P^*))\frac{\partial e_P^*}{\partial T_2}+q(e_P^*)=q(e_P^*)>0$, where the last equality follows from the first-best first-order conditions. By Lemma 2, e_P^* is invariant in T_1, hence, $\frac{\partial w^{PEN}}{\partial T_1}=1>0$.

Similarly, we use the definition of the penalty and the result above to take total derivatives:

$$\frac{\partial f}{\partial T_2}=-\frac{1}{w^{PEN}}+\frac{T_2}{(w^{PEN})^2}\frac{\partial w^{PEN}}{\partial T_2}=\frac{-w^{PEN}+q(e_P^*)T_2}{(w^{PEN})^2}=\frac{-c_P(e_P^*)-T_1}{(w^{PEN})^2}<0.$$

Finally, $\frac{\partial f}{\partial T_1}=\frac{T_2}{(w^{PEN})^2}\frac{\partial w^{PEN}}{\partial T_1}=\frac{T_2}{(w^{PEN})^2}>0$. □

Proof of Lemma 6

We check that the Hessian of $\Pi_{PT}(e_P,e_R)$ is definite negative. Note that $\Pi_{PT}(e_P,e_R)$ is concave in the individual effort levels. The determinant of the Hessian is

$$det(H(\Pi_{PT}))=(u_2-u_1)[u_1v(e_R)v''(e_R)q''(e_P)+(u_2-u_1)[v(e_R)v''(e_R)q(e_P)q''(e_P)-v'(e_R)^2q'(e_P)^2]].$$

Assumption 2 ensures that $v(e_R)v''(e_R)q(e_P)q''(e_P)-v'(e_R)^2q'(e_P)^2>0$, and hence by convexity of $v(\cdot)$ and of $q(\cdot)$, we have $det(H(\Pi_{PT}))>0$. □

Appendix E: Alternative Contracts

We now consider different pricing contracts that have been suggested in the literature and implemented in limited practical settings as alternatives to FFS in a variety of contexts. We first present these contracts and obtain the resulting effort levels. Because FFS is the most common payment system, we compare the outcome under these alternative payment systems to that under FFS. We also compare these effort levels to the first-best efforts to show that they cannot coordinate the efforts to the first-best levels.

E.1. Effort levels under different payment systems

E.1.1. Capitation Under a capitation contract, the requester pays the provider a fixed amount C for each patient in the population independently of the actual volume of patients the provider treats. Capitation was adopted by managed care organizations in the mid-to-late 1990s to control rising health care spending (Frakt and Mayes 2012). By the end of the decade, about one third of physicians had capitation contracts. A capitation system is often criticized because it submits caregivers to a high level of financial risk, and does not give the caregiver any incentive to deliver care of high quality, since the payment is fixed and disconnected from patient outcomes. Original capitation models proved unsustainable because payment rates were fast outpaced by medical spending, causing severe financial losses for caregivers. Contemporaneous payment systems, however, often use capitation-like contracts, but with tight spending budgets and incentives for quality performance (Song et al. 2011). For instance, ACOs are sometimes paid via partial capitation by the payer (e.g., Medicare), implemented through a combination of a pre-set budget with fee-for-service payments, while being held to quality targets.

We analyze the efforts under capitation. The requester and the provider profits are given by:

$$\Pi_P(e_P, e_R) = v_0 C - v(e_R)[c_P(e_P) + T_1]$$
$$\Pi_R(e_P, e_R) = -v_0 C - v_0 c_R(e_R) - v(e_R) q(e_P) T_2.$$

We observe that Π_P is decreasing in e_P, so the optimal decision for the provider is to exert no effort as intuitively explained, that is, $e_P^{CAP} = 0$. By Assumption 1, the requester's profit is concave in the requester's effort, thus the requester effort under capitation, e_R^{CAP}, can be obtained by solving the first-order condition:

$$|v'(e_R)| q(0) T_2 = v_0 c_R'(e_R). \tag{8}$$

E.1.2. Cost-sharing We now consider a payment system where the requester and the provider share observable costs, i.e., the cost of treatment T_1 and of treatment failure T_2. The rationale for not sharing costs of effort is that these costs are not easily contractible since they are internal and hard to observe and verify by a third-party. The idea behind a cost-sharing contract is that when the provider bears some of the cost of treatment failure, he has incentives to exert some effort to reduce the chance that treatment fails. Similarly, the requester must take into account the treatment cost when setting her effort level, as she does at the first-best, since she bears a fraction of that cost. In supply chain management research, cost-sharing contracts between firms have been shown to enable coordination in certain settings (e.g., Leng and Parlar 2010). Cost sharing is used in a healthcare setting when consumers have to pay a portion of their health care costs, via deductibles, co-payments or co-insurance. Medical cost-sharing with patients incentivizes them to be more efficient users of the healthcare system. Cost-sharing between payer and provider has also been studied in a variety of healthcare settings (Jelovac 2001, Chalkley and Malcomson 2002, Mariñoso and Jelovac 2003, Jack 2005, Liu et al. 2015).

We consider a cost-sharing contract in which the requester pays a fixed fee w^{CS} per service transaction, and the cost of treatment and treatment failure is shared. Namely, the requester is responsible for a fraction

$\alpha \in (0,1)$ of these costs, and the provider is responsible for the remaining $1-\alpha$. The requester and the provider profits are given by:

$$\Pi_P(e_P, e_R) = v(e_R)[w^{CS} - c_P(e_P) - (1-\alpha)(q(e_P)T_2 + T_1)]$$
$$\Pi_R(e_P, e_R) = -v_0 c_R(e_R) - v(e_R)[w^{CS} + \alpha(q(e_P)T_2 + T_1)].$$

By Assumption 1, the provider's profit is concave in e_P and the requester's profit is concave in e_R. As a result, the optimal effort levels e_P^{CS} and e_R^{CS} satisfy the first-order conditions:

$$c_P'(e_P) = (1-\alpha)|q'(e_P)|T_2 \tag{9}$$
$$|v'(e_R)|[w^{CS} + \alpha(q(e_P)T_2 + T_1)] = v_0 c_R'(e_R).$$

E.1.3. Shared-savings As previously described, the service requester needs to keep the cost of referring patients to an external provider down. The referral cost covers the direct cost of paying the provider and possible treatment failure costs (depending on who incurs those costs). It does not cover the cost of effort which comprises prevention programs rather than treatment, and which, as mentioned earlier, is not contractible as it is hard to observe and verify by an external party. If the requester is able to reduce her referral expenses, the reduction constitutes savings for the requester. Clearly the requester has every incentive to make savings as high as possible. To incentivize the provider to help the requester increase these savings by lowering the fraction of failed treatments, the requester can share a fraction with the provider.

Shared-savings contracts have been studied in a supply chain setting (Corbett and DeCroix 2001). In a healthcare setting, Medicare implements a shared-savings program with ACOs as a reward for spending less than a benchmark while satisfying quality performance standards.

We consider a shared-savings contract in which the requester pays an amount w^{SS} for each patient referred, and in addition the requester keeps a fraction $\beta \in (0,1)$ of any savings from a budget M dedicated to the direct cost of referrals (e.g., the amount spent in the previous year), while fraction $1-\beta$ is granted to the provider. The requester's savings are equal to $M - v(e_R)(w^{SS} + q(e_P)T_2)$. The major difference between a shared-savings contract and the cost-sharing contract analyzed above regards who bears the treatment cost, T_1. Under cost-sharing, both agents bear a fraction of this cost, while under shared-savings, the provider is responsible for its entirety. The requester and the provider profits are given by:

$$\Pi_P(e_P, e_R) = (1-\beta)M + v(e_R)[\beta w^{SS} - c_P(e_P) - (1-\beta)q(e_P)T_2 - T_1]$$
$$\Pi_R(e_P, e_R) = \beta M - \beta v(e_R)[w^{SS} + q(e_P)T_2] - v_0 c_R(e_R).$$

By Assumption 1, the provider's profit is concave in e_P and the requester's profit is concave in e_R. As a result, the optimal effort levels e_P^{SS} and e_R^{SS} satisfy the first-order conditions:

$$c_P'(e_P) = (1-\beta)|q'(e_P)|T_2 \tag{10}$$
$$\beta|v'(e_R)|[w^{SS} + q(e_P)T_2] = v_0 c_R'(e_R).$$

E.1.4. Two-part Tariff Under a two-part tariff contract, the requester pays the provider a fixed fee as well as a marginal payment for each referral. In particular, a two-part tariff contract, such as the one

considered in Corbett and DeCroix (2001) and in Corbett et al. (2005) as a generalization of a management fee, leasing or shared-savings contract, can be viewed as a hybrid between a FFS and capitation payment. The two-part tariff contract leads to the effort levels similar to FFS in our model, that is, the provider exerts no effort, and the requester exerts a positive effort. Furthermore, we note that contrary to Corbett and DeCroix (2001) and Corbett et al. (2005), in our setting, a shared-savings contract does not reduce to a two-part tariff contract, mainly because in our model the cost of effort is proportional to the volume of referral (which depends on requester's effort) and the requester post-treatment operational cost does not apply to each patient referred, but only to a fraction (patient whose treatment does not succeed) that depends on provider's effort.

E.1.5. Bundled Payments Under bundled payments, the requester pays the provider a single fee to cover treatment for a single episode of care. The episode of care is defined within a certain time windows (e.g., pre-operative care and 30 days after treatment). Hence, under bundled payments, the provider is responsible not only for the services directly related to treating the patient, but also for possible complications within the pre-defined time window. Adida et al. (2017) show that bundled payments can reduce providers' incentives to provide unnecessary services.

E.2. Comparison and Lack of Coordination

E.2.1. Capitation

Capitation vs. FFS. From (8), we notice that the requester's effort is independent of the capitation payment C, and that it corresponds to e_R^{FFS} when $w^{FFS}=0$. It follows that the requester's effort under capitation is lower than the effort under FFS if $w^{FFS}>0$, that is, $e_R^{CAP}<e_R^{FFS}$. Under capitation, the requester is not directly sensitive to the volume of patients referred, so she has less incentives to exert effort than under FFS. In addition, $e_P^{CAP}=e_P^{FFS}=0$.

Capitation vs. first-best. The provider effort under capitation cannot be coordinated since it equals zero regardless of the capitation payment. Moreover, depending on the parameters of the problem, the requester's effort under capitation may be larger or smaller than the first-best effort. Indeed, the first-best effort depends on the treatment cost T_1 (Lemma 2), while under capitation the treatment cost has no effect on the requester's effort decision. Hence, a higher cost of treatment leads to a higher first-best requester effort, but does not impact the capitation effort.

PROPOSITION 4. The requester effort $e_R^{CAP}<e_R^*$ iff $T_1>(q(0)-q(e_P^*))T_2-c_P(e_P^*)$.

Proof: Let us denote the first order condition of the requester profit as

$$\phi^{CAP}(e_P,e_R)=|v'(e_R)|q(e_P)T_2-v_0c_R'(e_R).$$

The capitation efforts satisfy $\phi^{CAP}(e_P^{CAP},e_R^{CAP})=\phi^{CAP}(0,e_R^{CAP})=0$. In addition, since Π_R is concave in the requester effort, we have that $\phi^{CAP}(0,e_R)$ decreases in e_R, so if $\phi^{CAP}(0,e_R^*)<0$, then $e_R^{CAP}<e_R^*$.

$$\begin{aligned}\phi^{CAP}(0,e_R^*)&=\phi^{CAP}(0,e_R^*)-\phi^T(e_P^*,e_R^*)\\&=|v'(e_R^*)|[q(0)T_2-c_P(e_P^*)-q(e_P^*)T_2-T_1].\end{aligned}$$

By (7), we have $\phi^{CAP}(0,e_R^*) \geq |v'(e_R^*)|[-c_P(0)-T_1]$, which is negative, but $\phi^{CAP}(0,e_R^*)$, being greater than a negative number, may be negative or positive. Namely, if $T_1 > q(0)T_2 - c_P(e_P^*) - q(e_P^*)T_2$, then $e_R^{CAP} < e_R^*$. Alternatively, if $T_1 \leq q(0)T_2 - c_P(e_P^*) - q(e_P^*)T_2$, then $e_R^{CAP} \geq e_R^*$. □

Because the capitation payment C plays no role in how e_R^{CAP} compares to e_R^*, it follows that there is no capitation contract that can coordinate the efforts to those at the first-best.

E.2.2. Cost-sharing

Cost-sharing vs. FFS. Contrary to FFS, the cost-sharing contract gives rise to a positive provider effort. Furthermore, under cost-sharing the requester effort is subject to opposing forces: the requester bears only a fraction $\alpha \in (0,1)$ of the failure costs, which diminishes incentives to exert effort, but if the price w^{CS} is large enough, it should exert effort to keep the volume of referrals low.

PROPOSITION 5. The provider effort $e_P^{CS} > e_P^{FFS} = 0$. In addition, there exists $\alpha_1 > 0$, such that for any $\alpha < \alpha_1$, the requester effort $e_R^{CS} < e_R^{FFS}$.

Proof: It is clear that $e_P^{CS} > e_P^{FFS} = 0$. We denote the first order condition for the requester effort under cost-sharing as

$$\phi^{CS}(e_P,e_R) = |v'(e_R)|[w^{CS} + \alpha(q(e_P)T_2 + T_1)] - v_0 c_R'(e_R).$$

The optimal efforts under cost-sharing satisfy $\phi^{CS}(e_P^{CS},e_R^{CS}) = 0$.

$$\begin{aligned}\phi^{CS}(e_P^{CS},e_R^{FFS}) &= \phi^{CS}(e_P^{CS},e_R^{FFS}) - \phi^{FFS}(e_P^{FFS},e_R^{FFS}) \\ &= |v'(e_R^{FFS})|[w^{CS} + \alpha(q(e_P^{CS})T_2 + T_1) - w^{FFS} - q(0)T_2].\end{aligned}$$

Let us consider the case $w^{CS} < w^{FFS} + q(0)T_2$. Then, $e_R^{CS} < e_R^{FFS}$ iff $\phi^{CS}(e_P^{CS},e_R^{FFS}) < 0$, that is,

$$\alpha(q(e_P^{CS})T_2 + T_1) < w^{FFS} + q(0)T_2 - w^{CS}. \tag{11}$$

We notice that the left-hand-side of (11) is increasing in α. To see this, we take the derivative with respect to α which is given by $q(e_P^{CS})T_2 + T_1 + \alpha q'(e_P^{CS})T_2 \frac{\partial e_P^{CS}}{\partial \alpha}$. Now, in order to conclude that this is positive, we take the derivative with respect to α of the first-order condition for e_P in equation (9), to find

$$\frac{\partial e_P^{CS}}{\partial \alpha} c_P''(e_P^{CS}) = -\frac{\partial e_P^{CS}}{\partial \alpha}(1-\alpha)q''(e_P^{CS})T_2 + q'(e_P^{CS})T_2,$$

and hence

$$\frac{\partial e_P^{CS}}{\partial \alpha} = \frac{q'(e_P^{CS})T_2}{c_P''(e_P^{CS}) + (1-\alpha)q''(e_P^{CS})T_2} < 0.$$

Because $q'(e_P^{CS}) < 0$, it follows that the left-hand-side of (11) has a positive derivative with respect to α, and thus it is monotonically increasing in α. Since (11) trivially holds when $\alpha = 0$, we have that there exists $\alpha_1 > 0$, such that $e_R^{CS} < e_R^{FFS}$ for any $\alpha < \alpha_1$. Finally, if $\alpha_1 \geq 1$ (that is, if $q(e_P^{CS})T_2 + T_1 < w^{FFS} + q(0)T_2 - w^{CS}$), then $e_R^{CS} < e_R^{FFS}$ for any value of $\alpha \in (0,1)$.

On the other hand, if the price of the cost-sharing contract satisfies $w^{CS} \geq w^{FFS} + q(0)T_2$, then $\phi^{CS}(e_P^{CS},e_R^{FFS}) \geq 0$, and $e_R^{CS} \geq e_R^{FFS}$ for any value of $\alpha \in (0,1)$. □

Our goal is to coordinate the decisions to those at the first-best. By Proposition 1 the FFS contract leads to a too high requester effort. This result indicates that going from FFS to cost-sharing is a step in the right direction for both types of effort, provided that the price w^{CS} is not too high.[18]

Cost-sharing vs. first-best. We compare the cost-sharing efforts with the first-best efforts for a given contract, and then determine whether there exists a coordinating cost-sharing contract, i.e., a price w^{CS} and fraction $\alpha \in (0,1)$ that induce first-best effort levels.

PROPOSITION 6. The provider effort $e_P^{CS} < e_P^*$. In addition, there exists $\alpha_2 > 0$ such that the requester effort $e_R^{CS} < e_R^*$ iff $\alpha < \alpha_2$.

Proof: Comparing the first-order condition for e_P in (2) and (9), and by Assumption 1, it is clear that $e_P^{CS} < e_P^*$ when $\alpha > 0$.

Because $\phi^T(e_P^*, e_R^*) = 0$, we have that

$$\begin{aligned}\phi^{CS}(e_P^{CS}, e_R^*) &= \phi^{CS}(e_P^{CS}, e_R^*) - \phi^T(e_P^*, e_R^*) \\ &= |v'(e_R^*)|[w^{CS} + \alpha(q(e_P^{CS})T_2 + T_1) - c_P(e_P^*) - q(e_P^*)T_2 - T_1)].\end{aligned}$$

We have $e_R^{CS} < e_R^*$ iff $\phi^{CS}(e_P^{CS}, e_R^*) < 0$. Let us consider the case $w^{CS} < c_P(e_P^*) + q(e_P^*)T_2 + T_1$. Then, $e_R^{CS} < e_R^*$ iff

$$\alpha(q(e_P^{CS})T_2 + T_1) < c_P(e_P^*) + q(e_P^*)T_2 + T_1 - w^{CS}. \tag{12}$$

The left-hand-side of this inequality is identical to that of (11) in the proof of Proposition 5. Hence, it is also monotonically increasing in α. Since (12) trivially holds when $\alpha = 0$, we have that there exists $\alpha_2 > 0$, such that $e_R^{CS} < e_R^*$ for any $\alpha < \alpha_2$. Finally, if $\alpha_2 \geq 1$ (that is, if $q(e_P^{CS})T_2 + T_1 < c_P(e_P^*) + q(e_P^*)T_2 + T_1 - w^{CS}$), then $e_R^{CS} < e_R^*$ for any value of $\alpha \in (0,1)$.

In the case where $w^{CS} \geq c_P(e_P^*) + q(e_P^*)T_2 + T_1$, we have that $e_R^{CS} \geq e_R^*$, for any value of $\alpha \in (0,1)$. □

Since the provider bears only a fraction of the failure cost under cost-sharing, he has less incentives than the centralized system to exert effort. Similarly, if the fee per patient and the fraction of costs that the requester is responsible for are low enough, she exerts less effort than at the first-best.

This result shows that while a cost-sharing contract cannot coordinate the provider's effort, it may be possible to find a cost share $\alpha \in (0,1)$ that coordinates the requester's effort as long as the price per patient w^{CS} is not too high.[19]

E.2.3. Shared-savings

Shared-savings vs. FFS. Similarly to cost-sharing, the shared-savings payment system gives the provider incentives to exert a positive effort to incur less treatment failure costs as he receives a portion of the unused budget. Moreover, the requester only bears a fraction $\beta \in (0,1)$ of the payment for each patient requiring treatment (and of the treatment failure cost), which can decrease the requester effort compared to the FFS contract. These observations lead to the following Proposition.

[18] It can be shown in a very similar fashion that if, in addition to the cost-sharing component, instead of paying a fee per patient w^{CS}, the requester transfers a fixed lump sum to the provider to take care of all referred volume, then $e_R^{CS} < e_R^{FFS}$, for any lump sum and cost share α.

[19] We find similar structural results if instead of paying a fee per patient w^{CS}, the requester transfers a fixed lump sum to the provider to take care of all the referred volume.

PROPOSITION 7. The provider effort $e_P^{SS} > e_P^{FFS} = 0$. In addition, there exists $\beta_1 > 0$, such that for any $\beta < \beta_1$, the requester effort $e_R^{SS} < e_R^{FFS}$.

Proof: It is clear that $e_P^{SS} > e_P^{FFS} = 0$. We denote the first order condition for e_R^{SS} as

$$\phi^{SS}(e_P, e_R) = |v'(e_R)|[\beta w^{SS} + \beta q(e_P)T_2] - v_0 c_R'(e_R).$$

The optimal effort level satisfies $\phi^{SS}(e_P^{SS}, e_R^{SS}) = 0$. We have $e_R^{SS} < e_R^{FFS}$ iff $\phi^{SS}(e_P^{SS}, e_R^{FFS}) < 0$.

$$\begin{aligned}\phi^{SS}(e_P^{SS}, e_R^{FFS}) &= \phi^{SS}(e_P^{SS}, e_R^{FFS}) - \phi^{FFS}(e_P^{FFS}, e_R^{FFS}) \\ &= |v'(e_R^{FFS})|[\beta(w^{SS} + q(e_P^{SS})T_2) - w^{FFS} - q(0)T_2].\end{aligned}$$

Thus $e_R^{SS} < e_R^{FFS}$ iff

$$\beta(w^{SS} + q(e_P^{SS})T_2) < w^{FFS} + q(0)T_2. \tag{13}$$

We notice that the left-hand-side of (13) is increasing in β. To see this, we take the derivative with respect to β which is given by $w^{SS} + q(e_P^{SS})T_2 + \beta q'(e_P^{SS})T_2 \frac{\partial e_P^{SS}}{\partial \beta}$. Now, in order to conclude that this is positive, we take the derivative with respect to β of the first-order condition for e_P in equation (10), to find

$$\frac{\partial e_P^{SS}}{\partial \beta} c_P''(e_P^{SS}) = -\frac{\partial e_P^{SS}}{\partial \beta}(1-\beta)q''(e_P^{SS})T_2 + q'(e_P^{SS})T_2,$$

and hence

$$\frac{\partial e_P^{SS}}{\partial \beta} = \frac{q'(e_P^{SS})T_2}{c_P''(e_P^{SS}) + (1-\beta)q''(e_P^{SS})T_2} < 0.$$

Because $q'(e_P^{CS}) < 0$, it follows that the left-hand-side of (13) has a positive derivative with respect to β, and thus it is monotonically increasing in β. Since (13) trivially holds when $\beta = 0$, we have that there exists $\beta_1 > 0$, such that $e_R^{SS} < e_R^{FFS}$ for any $\beta < \beta_1$. Finally, if $\beta_1 \geq 1$ (that is, if $w^{SS} + q(e_P^{SS})T_2 < w^{FFS} + q(0)T_2$), then $e_R^{SS} < e_R^{FFS}$ for any value of $\beta \in (0,1)$. □

This result indicates that going from FFS to shared-savings may be a step in the right direction toward coordination to the first-best for both types of effort (at least when β is not too high).

Shared-savings vs. first-best. We compare the shared-savings efforts with the first-best efforts for a given contract, and then determine whether there exists a coordinating shared-savings contract, i.e., a price w^{SS} and fraction $\beta \in (0,1)$ that induce first-best effort levels.

PROPOSITION 8. The provider effort $e_P^{SS} < e_P^*$. In addition, there exists $\beta_2 > 0$ such that the requester effort $e_R^{SS} < e_R^*$ iff $\beta < \beta_2$.

Proof: Comparing the first-order condition for e_P in (2) and (10), and by Assumption 1, it is clear that $e_P^{SS} < e_P^*$ when $\beta > 0$. In addition, we have

$$\begin{aligned}\phi^{SS}(e_P^{SS}, e_R^*) &= \phi^{SS}(e_P^{SS}, e_R^*) - \phi^{T}(e_P^*, e_R^*) \\ &= |v'(e_R^*)|[\beta w^{SS} + \beta q(e_P^{SS})T_2 - c_P(e_P^*) - q(e_P^*)T_2 - T_1].\end{aligned}$$

Thus, $e_R^{SS} < e_R^*$ iff $\beta(w^{SS} + q(e_P^{SS})T_2) < c_P(e_P^*) + q(e_P^*)T_2 + T_1$. The left-hand-side of this inequality is identical to (13) in the proof of Proposition 7. Hence, it is also monotonically increasing in β. Since the

inequality trivially holds when $\beta = 0$, we have that there exists $\beta_2 > 0$ such that $e_R^{SS} < e_R^*$ iff $\beta < \beta_2$. Finally, if $\beta_2 \geq 1$ (that is, if $w^{SS} + q(e_P^{SS})T_2 < c_P(e_P^*) + q(e_P^*)T_2 + T_1$), then $e_R^{SS} < e_R^*$ for any value of $\beta \in (0,1)$. □

Since the provider bears only a fraction of the failure cost under shared-savings, he has less incentives than the centralized system to exert effort. Similarly, if the fraction of the savings that the requester keeps is too low, she exerts less effort than at the first-best.

This result demonstrates that while a shared-savings contract cannot coordinate the provider's effort, it may be possible to find a share β that coordinates the requester's effort as long as $\beta_2 < 1$ (e.g., when T_1 or T_2 are not too large).

E.2.4. Two-Part Tariff The supply chain management literature has shown that in many cases a two-part tariff contract can coordinate decisions (e.g., order quantity) between two firms (e.g., Cachon and Lariviere 2005). However, in our framework specific to the interaction between a service requester and a service provider making effort decisions that relate to quality outcomes in a peer-to-peer healthcare setting, a two-part tariff does not achieve coordination. It is possible to adjust the marginal payment to coordinate the requester effort to that at the first-best. Yet, a two-part tariff is unable to incentivize the provider to exert effort, and hence to achieve full coordination.

E.2.5. Bundled Payments While a Bundled Payment system does aim at incentivizing high quality by making the provider responsible for treatment outcomes, the bundle includes activities and associated outcomes only *within a certain time window* (e.g., 30 days). However, there may be additional costs associated with treatment failure, that would *not* be part of such a bundle, but that the ACO would still have to cover. For instance, after an inpatient treatment, under a bundle the provider may be responsible for readmission costs within 30 days; however, because of the readmission complication, the patient may require additional monitoring, access to medications, and other ancillary services even after the episode of care is completed. For this reason, although a bundled payment system can help move effort decisions in the right direction, it may not be sufficient to induce first-best efforts.

Appendix F: Model Extensions

F.1. Probability of treatment failure depends on both, provider and requester's efforts.

In this model extension, we are concerned with the impact of considering $q = q(e_P, e_R)$, as opposed to $q = q(e_P)$, on the main paper results. Let us assume $q = q(e_P, e_R)$, under this formulation the requester's effort has two desirable effects: it decreases the volume of referrals and also decreases the probability of treatment failure. Intuitively, this may better represent the case of chronic conditions (e.g., Diabetes). In this case, the preventive care delivered by the requester not only reduces the need for advance treatment, but it can also improve the general health status of the patients which may increase the likelihood of provider's treatment success. In the following analysis, we denote $\frac{\partial q(e_P,e_R)}{\partial e_P} = q'_P(e_P, e_R)$, $\frac{\partial^2 q(e_P,e_R)}{\partial e_P^2} = q''_P(e_P, e_R)$, and $\frac{\partial^2 q(e_P,e_R)}{\partial e_R \partial e_P} = q''_{PR}(e_P, e_R)$, and we would omit the dependency on e_P and e_R where clear within the context.

Assumption 3. The probability of treatment failure $q(e_P, e_R)$ is non-negative convex decreasing in e_P and e_R, and $q''_{PR} \geq 0$. Further, we also consider the following technical conditions

$$\frac{v\ v''}{{v'}^2} \cdot \frac{c_P\ c''_P}{{c'_P}^2} \geq 1, \quad \frac{v\ v''}{{v'}^2} \cdot \frac{q\ q''_P}{{q'_P}^2} \geq 1, \quad \frac{q''_P}{|q'_P|} \geq \frac{q''_{PR}}{|q'_R|}, \quad q''_P\ q''_R \geq {q''_{PR}}^2. \tag{14}$$

The probability of treatment failure decreases as the requester and provider exert more effort, and we assume that there are decreasing marginal returns on efforts. The second order partial derivative of q being positive means that e_P and e_R are complements. As previously mentioned, we can interpret this as if an additional unit of preventive effort (requester) makes patients 'healthier' in some way. Thus provider's post-treatment effort delivered on 'healthier' patients is more effective in the sense that it can decrease the likelihood of treatment failure further. Regarding the technical conditions, note that the first two conditions in (14) are the same as those in Assumption 2 equation (1) in the main section of the paper. The last two conditions in (14) state that the cross-effect between the requester and provider efforts is small (i.e., q''_{PR} is small). For instance, if $q(e_P, e_R) = \tilde{q}_P(e_P) + \tilde{q}_R(e_R)$, the last two conditions in (14) are trivially satisfied.

LEMMA 7. Π_T is jointly concave in (e_P, e_R).

Proof: For the centralized system total profit, we have that

$$\begin{aligned}
\frac{\partial^2 \Pi_T}{\partial e_P^2} &= -v(c''_P v + q''_P T_2) \\
\frac{\partial^2 \Pi_T}{\partial e_R^2} &= -v''(c_P + qT_2 + T_1) - 2v' q'_R T_2 - v q''_R T_2 - v_0 c''_R \\
\frac{\partial^2 \Pi_T}{\partial e_P \partial e_R} &= -v'(c'_P + q'_P T_2) - v q''_{PR} T_2.
\end{aligned}$$

Under the convexity Assumption 1 and 3, $\frac{\partial^2 \Pi_T}{\partial e_R^2} < 0$, and $\frac{\partial^2 \Pi_T}{\partial e_P^2} < 0$. Thus, for the Hessian to be definite negative we need its determinant to be positive; the following sufficient conditions ensure this.

$$\begin{aligned}
vv''(c''_P + q''_P T_2)(c_P + qT_2) - v'^2(c'_P + q'_P T_2)^2 \geq 0, \\
2vv'[(c''_P + q''_P T_2)q'_R - (c'_P + q'_P T_2)q''_{PR}]T_2 \geq 0, \text{ and} \\
v^2[(c''_P + q''_P T_2)q''_R - q''^2_{PR} T_2]T_2 \geq 0.
\end{aligned}$$

The conditions in Assumption 3 equation (14) guarantee that the three above inequalities are satisfied. Therefore, Π_T is jointly concave. □

THEOREM 2. The FFS, Capitation, Cost-sharing, and Shared-saving contracts cannot **simultaneously** coordinate the Provider and Requester's effort decisions to the first-best. Alternatively, a Penalty contract can coordinate both the Provider and Requester's effort decisions to the first-best and to the socially optimum effort levels.

Proof: In this Theorem we are concerned with the impact of considering $q = q(e_P, e_R)$, as opposed to $q = q(e_P)$, on the main coordination result of the paper. The first-order conditions of the centralized profit are given by

$$c'_P(e_P) = |q'_P(e_P, e_R)|T_2, \tag{15}$$

$$|v'(e_R)|[c_P(e_P) + q(e_P, e_R)T_2 + T_1] + \underbrace{v(e_R)|q'_R(e_P, e_R)|T_2}_{**} = v_0 c'_R(e_R). \tag{16}$$

Note that the additional term $**$ in the first-order condition is due to the dependence of the failure probability on the requester's effort. Under Assumption 1, the first-best effort levels are $(\tilde{e}^*_P, \tilde{e}^*_R) > (0, 0)$ (we use $\sim$ to differentiate from the first-best solution in the case $q = q(e_P)$). We proceed by analyzing the first-order

conditions of each contract and comparing them to (15) and (16) to show that at least one of the effort decisions cannot be coordinated under the FFS, Capitation, Cost-sharing, and Shared-saving contracts. For the Penalty contract we show that both effort decisions can be coordinated to the first-best and socially optimum efforts.

- FFS: The provider's profit is given by $\Pi_P(e_P, e_R) = v(e_R)[w^{FFS} - c_P(e_P) - T_1]$ which does not depend on the probability of treatment failure. Thus following the same reasoning as in Section 4.2, $e_P^{FFS} = 0$ for all payment w^{FFS}. Hence, FFS cannot coordinate the provider's effort decision.

- Capitation: Similar to FFS, we note that the provider's profit does not depend on the failure probability, $\Pi_P(e_P, e_R) = v_0 C - v(e_R)[c_P(e_P) + T_1]$, thus following the same reasoning as in Section E.2.1, $e_P^C = 0$ for all capitation payment C. Hence, Capitation cannot coordinate the provider's effort decision.

- Cost-sharing: The requester profit is given by $\Pi_R(e_P, e_R) = -v_0 c_R(e_R) - v(e_R)[w^{CS} + \alpha(q(e_P, e_R)T_2 + T_1)]$. Let us assume that we can coordinate the requester's effort using a Cost-sharing contract. This means that there exists contract parameters α and w^{CS} such that the requester effort $e_R^{CS} = \tilde{e}_R^*$. The provider chooses his effort level to maximize $\Pi_P(e_P, e_R) = v(e_R)(w^{CS} - c_P(e_P) - (1-\alpha)[q(e_P, e_R)T_2 + T_1])$. Thus, assuming coordination of the requester's effort, the provider's optimal effort decision e_P^{CS} satisfies

$$c'_P(e_P) = (1-\alpha)|q'_P(e_P, \tilde{e}_R^*)|T_2.$$

However, first-best efforts satisfy (15), thus for any value of $\alpha \in (0,1)$, $e_P^{CS} < \tilde{e}_P^*$. Therefore, it is not possible to simultaneously coordinate the requester and provider's effort decisions under Cost-sharing.

- Shared-saving: The requester profit is given by $\Pi_R(e_P, e_R) = \beta M - \beta v(e_R)[w^{SS} + q(e_P, e_R)T_2] - v_0 c_R(e_R)$. Let us assume that we can coordinate the requester's effort using Shared-savings contract. This means that we can choose contract parameters β and w^{SS} such that the requester effort $e_R^{SS} = \tilde{e}_R^*$. The provider chooses his effort level to maximize $\Pi_P(e_P, e_R) = (1-\beta)M + v(e_R)(\beta w^{SS} - c_P(e_P) - (1-\beta)q(e_P, e_R)T_2 + T_1)$. Thus, assuming coordination of the requester's effort, the provider's optimal effort decision e_P^{SS} satisfies

$$c'_P(e_P) = (1-\beta)|q'_P(e_P, \tilde{e}_R^*)|T_2.$$

However, first-best efforts satisfy (15), thus for any value of $\beta \in (0,1)$, $e_P^{SS} < \tilde{e}_P^*$. Therefore, it is not possible to simultaneously coordinate the requester and provider's effort decisions under the Shared-saving contract.

- Penalty: The provider's profit is given by $\Pi_P(e_P, e_R) = v(e_R)[w^{PEN}(1 - q(e_P, e_R)(1-f)) - c_P(e_P) - T_1]$ and the Requester's profit is $\Pi_R(e_P, e_R) = -v_0 c_R(e_R) - v(e_R)[w^{PEN}(1 - q(e_P, e_R)(1-f)) + q(e_P, e_R)T_2]$. The first-order conditions for both profit functions are

$$\begin{aligned} &c'_P(e_P) = |q'_P(e_P, e_R)|w^{PEN}(1-f), \\ &|v'(e_R)|[w^{PEN}(1 - q(e_P, e_R)(1-f)) + q(e_P, e_R)T_2] \\ &+ v(e_R)|q'_R(e_P, e_R)|(T_2 - w^{PEN}(1-f)) = v_0 c'_R(e_R). \end{aligned}$$

By choosing $f = 1 - \frac{T_2}{w^{PEN}}$ and $w^{PEN} = c_P(\tilde{e}_P^*) + q(\tilde{e}_P^*, \tilde{e}_R^*)T_2 + T_1 + \frac{v(\tilde{e}_R^*)}{|v'(\tilde{e}_R^*)|}|q'_R(\tilde{e}_P^*, \tilde{e}_R^*)|T_2$, the provider and requester effort decisions can be coordinated **simultaneously**. To show this we just need to plug f and

w^{PEN} back into the above first-order conditions and after some algebraic manipulations, we recover the first-order conditions (15) and (16), and by the concavity of Π_T, the unique solution is $(e_P^{PEN}, e_R^{PEN}) = (\tilde{e}_P^*, \tilde{e}_R^*)$. Therefore, the penalty contract can achieve coordination of effort decisions for both players when $q = q(e_P, e_R)$.

Coordination to the social optimum efforts: We first note that under Assumption 3 the patients utility function $\Pi_{PT}(e_P, e_R) = -v(e_R)(u_1 + (u_2 - u_1)q(e_P, e_R))$ is concave and so is the social welfare function $\Pi_S = \Pi_P + \Pi_R + \Pi_{PT}$. Given this, we can do a similar analysis as we did in Section 6, hence coordination to the social optimum efforts follows directly from realizing that the social welfare function is the same as the centralized profit function with modified costs of treatment $(T_1 + u_1)$ and treatment failure $(T_2 + (u_2 - u_1))$. □

In the following Proposition we compare the first-best efforts and coordinating contract parameters under both models $q = q(e_P)$ and $q = q(e_P, e_R)$.

PROPOSITION 9. Let us assume $q(e_P) = q(e_P, \tilde{e}_R^*)$. The first-best efforts are such that $e_P^* = \tilde{e}_P^*$ and $e_R^* < \tilde{e}_R^*$. Furthermore, under $q = q(e_P, e_R)$ the coordinating contract fee w^{PEN} is larger and the fraction f is smaller than under $q = q(e_P)$.

Proof: The first-best first-order conditions for both models are summarized in Table 4. Considering that

Table 4 First-best first-order conditions

	$\frac{\partial \Pi_T}{\partial e_P}$	$\frac{\partial \Pi_T}{\partial e_R}$
$q = q(e_P)$	$c_P'(e_P) = \lvert q'(e_P)\rvert T_2$	$\lvert v'(e_R)\rvert[c_P(e_P) + q(e_P)T_2 + T_1] = v_0 c_R'(e_R)$
$q = q(e_P, e_R)$	$c_P'(e_P) = \lvert q_P'(e_P, e_R)\rvert T_2$	$\lvert v'(e_R)\rvert[c_P(e_P) + q(e_P, e_R)T_2 + T_1] + v(e_R)\lvert q_R'(e_P, e_R)\rvert T_2 = v_0 c_R'(e_R)$

the impact of provider's effort on the probability of treatment failure is the same in both models, that is, $q'(e_P) = q_P'(e_P, \tilde{e}_R^*)$, and since the first-order condition with respect to e_P under both models is the same, the provider first-best effort must be the same under both models. From the second first-order condition (with respect to e_R), we note that given the convexity of the requester cost of effort and the probability of treatment failure the additional term $v(e_R)|q_R'(e_P, e_R)|T_2 > 0$ results in larger requester first-best effort under $q = q(e_P, e_R)$.

Now we look at the coordinating contract parameters under the two modeling assumptions (see Table 5). Given the volume and probability of treatment failure convexity assumptions, the additional term

Table 5 Coordinating Penalty contract

	f	w^{PEN}
$q = q(e_P)$	$1 - \frac{T_2}{w^{PEN}}$	$c_P(e_P^*) + q(e_P^*)T_2 + T_1$
$q = q(e_P, e_R)$	$1 - \frac{T_2}{w^{PEN}}$	$c_P(\tilde{e}_P^*) + q(\tilde{e}_P^*, \tilde{e}_R^*)T_2 + T_1 + \frac{v(\tilde{e}_R^*)}{\lvert v'(\tilde{e}_R^*)\rvert}\lvert q_R'(\tilde{e}_P^*, \tilde{e}_R^*)\rvert T_2$

$\frac{v(\tilde{e}_R^*)}{|v'(\tilde{e}_R^*)|}|q_R'(\tilde{e}_P^*, \tilde{e}_R^*)|T_2 > 0$ under $q = q(e_P, e_R)$ results in a larger fee w^{PEN}. The higher fee provides more incentives for the requester so she exerts the higher first-best effort $e_R^* < \tilde{e}_R^*$. The behavior of the fraction f follows directly from the previous observation. In addition, we note that the net provider's profit loss for treatment failure $w^{PEN}(1 - f) = T_2$ is the same under the two modeling assumptions as the provider exerts the same first-best effort.