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 $\Pi_T(e_P^{FFS}, e_R^{FFS})/\Pi_T(e_P^*, e_R^*)$
$c_0^P, \quad \lambda$	parameters of the cost of effort function in Section 7: $c_P(e_P) = c_0^P e^{\lambda e_P}$
$c_0^R, \quad \kappa$	parameters of the cost of effort function in Section 7: $c_R(e_R) = c_0^R e^{\kappa e_R}$
$q_0, \quad \gamma$	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 A.1 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 A.2 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

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

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

$$\phi^T(e_P,e_R)=|v'(e_R)|[c_P(e_P)+q(e_P)T_2+T_1]-v_0c_R'(e_R)$$
$$\phi^{FFS}(e_P,e_R)=|v'(e_R)|[w^{FFS}+q(e_P)T_2]-v_0c_R'(e_R).$$

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

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

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

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

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

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

Therefore,

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

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

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

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

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

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

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.

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

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

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

$$|v'(e_R)|[c_P(e_P)+q(e_P,e_R)T_2+T_1]+\underbrace{v(e_R)|q'_R(e_P,e_R)|T_2}_{**}=v_0c'_R(e_R). \tag{16}$$

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, $\Pi_P(e_P,e_R)=v_0C-v(e_R)[c_P(e_P)+T_1]$, 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 $\Pi_R(e_P,e_R)=-v_0c_R(e_R)-v(e_R)[w^{CS}+\alpha(q(e_P,e_R)T_2+T_1)]$. 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 $\Pi_R(e_P,e_R)=\beta M-\beta v(e_R)[w^{SS}+q(e_P,e_R)T_2]-v_0c_R(e_R)$. 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 $\Pi_R(e_P,e_R)=-v_0c_R(e_R)-v(e_R)[w^{PEN}(1-q(e_P,e_R)(1-f))+q(e_P,e_R)T_2]$. The first-order conditions for both profit functions are

$$\begin{aligned}
&c_P'(e_P)=|q_P'(e_P,e_R)|w^{PEN}(1-f),\\
&|v'(e_R)|[w^{PEN}(1-q(e_P,e_R)(1-f))+q(e_P,e_R)T_2]\\
&+v(e_R)|q_R'(e_P,e_R)|(T_2-w^{PEN}(1-f))=v_0c_R'(e_R).
\end{aligned}$$

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 A.3. Considering

Table A.3 First-best first-order conditions

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

that 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 A.4). Given the volume and probability of treatment failure convexity assumptions, the additional term

Table A.4 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.

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.

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

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.

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.

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.

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.

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.

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.

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

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.