INSEAD The Business School for the World®

Working Paper

2023/05/TOM

(Revised version of 2021/38/ TOM)

Outcomes-Based Reimbursement Policies for Chronic Care Pathways

Sasa Zorc
University of Virginia, zorcs@darden.virginia.edu

Stephen E. Chick
INSEAD, stephen.chick@insead.edu

Sameer Hasija
INSEAD, sameer.hasija@insead.edu

Outcomes-based reimbursement is one active area of theory and practice that supports roader efforts to improve health value by rewarding provider services with better health outcomes with higher payments. The design of outcomes-based reimbursement policies involves several choices - the type of contract to use (e.g., capitation or fee-for-service), on which level to measure outcomes (e.g., population- or provider-level), and whether to contract with individual providers or with a group of providers. Such choices may involve economic challenges (including collusion and free-riding) and health-specific challenges (e.g., referrals from practitioners who support chronic disease management to specialists that treat complications). We present a parsimonious game-theoretic model that identifies differences in the impact on health, costs, and system efficiency as a function of potential collusion and free-riding under different reimbursement policies. We give illustrative numerical experiments calibrated to data from UK pathways for diabetes. Our results indicate a strong performance of outcomes-adjusted capitation contracts with individual providers using population-level data. We also provide theory to interpret the performance of contracts in use in the USA and UK.

Keywords: Health Care; Contracting; Moral Hazard; Queue; Outcomes; Value Based; Collusion; Free riding.

History : 08/02/2023. Earlier versions: 27/07/2021. 24/05/2017.

Electronic copy available at: http://ssrn.com/abstract=2973048

Outcomes-Based Reimbursement Policies for Chronic Care Pathways

Saša Zorc
Darden School of Business, University of Virginia, 100 Darden Blvd, Charlottesville, VA 22903
zorcs@darden.virginia.edu

Stephen E. Chick
INSEAD, Boulevard de Constance, 77300 Fontainebleau, France; stephen.chick@insead.edu

Sameer Hasija
INSEAD, 1 Ayer Rajah Avenue, Singapore 138676, Singapore; sameer.hasija@insead.edu

Outcomes-based reimbursement is one active area of theory and practice that supports broader efforts to improve health value by rewarding provider services with better health outcomes with higher payments. The design of outcomes-based reimbursement policies involves several choices—the type of contract to use (e.g., capitation or fee-for-service), on which level to measure outcomes (e.g., population- or provider-level), and whether to contract with individual providers or with a group of providers. Such choices may involve economic challenges (including collusion and free-riding) and health-specific challenges (e.g., referrals from practitioners who support chronic disease management to specialists that treat complications). We present a parsimonious game-theoretic model that identifies differences in the impact on health, costs, and system efficiency as a function of potential collusion and free-riding under different reimbursement policies. We give illustrative numerical experiments calibrated to data from UK pathways for diabetes. Our results indicate a strong performance of outcomes-adjusted capitation contracts with individual providers using population-level data. We also provide theory to interpret the performance of contracts in use in the USA and UK.

History: 08/02/2023. Earlier versions: 27/07/2021. 24/05/2017. https://ssrn.com/abstract=2973048.

Economic pressure, public accountability, political will, and new technology that eases the gathering of health outcomes data are some of the reasons behind recent attention to outcomes-based contracts in health care. Such contracts have been a central part of the recent shift in healthcare payment models from ones focusing on the volume of care to ones focusing on the value provided by care (Porter 2010, WEF 2018, ICHOM 2021). The idea behind outcomes-based contracting (also called pay-for-performance, or P4P) is simple: compensation to care providers should be based on the health benefits they create. The USA's Centers for Medicare & Medicaid Services (CMS 2020), the UK's National Health Service (UK NHS 2020), and private sector firms (Zhu et al. 2020), among others, are exploring such contracts.

Designing an outcomes-based system is a daunting task. The institutions implementing them have used very different designs (e.g., Eijkenaar et al. 2013, Bastani et al. 2016, Hsieh et al. 2017, Koff and Lyons 2020), and there is mixed evidence with no consensus as to the best reimbursement policy (Christianson et al. 2008, Eijkenaar et al. 2013, Burns and Pauly 2018). A patient's care pathway often involves several providers, resulting in challenges in designing reimbursement policies that can align the interests of the providers along the pathway (Cebul et al. 2013).

The presence of multiple providers along the pathway creates a danger of *collusion*, where providers jointly make decisions to increase their profit (Rodwin 1995). Collusion often manifests as *demand inducement*, also known as kickbacks (McGuire 2000), where one colluding provider uses referrals to generate additional demand for another (Pauly 1979). Collusion in the form of demand inducement and kickbacks is widespread in developing markets (Gadre 2015). Despite explicit laws which deem such practice illegal (e.g., by the Stark Law and Anti-Kickback Statute in the USA),[1] collusion *does* happen in developed markets. Cases of demand inducement collusion have been the subject of multiple high-profile lawsuits in the US (Woolhandler and Himmelstein 2004, Mannava et al. 2013, Thornton et al. 2013, Dyer 2015), which resulted in more than $153 million in penalties and settlements in 2021 (US DoJ 2021). Moreover, the laws protecting against collusion in the USA's Medicare were updated in 2021, exempting outcomes-based contracts from those laws (Fanburg et al. 2022). Thus, there exist concerns that the introduction of outcomes-based contracts could open doors to new opportunities for collusion (US Senate 2016, WEF 2018).

Group contracts[2] have been suggested as one way to coordinate care and address possible collusion issues. However, they have well-known limitations of their own. *Free-riding* is a phenomenon occurring when the payouts from the group's joint effort are shared amongst the group (Hölmstrom 1982), resulting in individuals who form the group putting in less effort than optimal (effectively free-riding on the work of others). Unfortunately, it has also been documented in existing group-based incentives in healthcare where it has been an impediment to their success (Pauly and Redisch 1973, Christianson et al. 2008, Redding 2022). Thus, free-riding creates an argument against the use of group contracts. This tension between collusion and free-riding contributes to making the reimbursement policy design a challenging endeavor.

This context motivates the following two core questions. Should we be worried about collusion in outcomes-adjusted contracts? If so, can we do something about it by our choice of reimbursement

[1] See the United States Code, https://uscode.house.gov/, Title 42, U.S.C. § 1395nn and § 1320a-7b. Regulators primarily refer to the demand-inducing collusion we study as "kickbacks." The Anti-Kickback Statute forbids not only explicit payment in return for referrals, but also any kind of profit sharing agreement, such as profit-sharing or co-ownership, which would enable one provider to profit from generating demand for another.

[2] A single payment to the group is shared by the group's providers, as in an alliance contract in the UK (Sanderson et al. 2018), a bundled pay contract in the USA (NEJM Catalyst 2018), or contracts between Medicare and Accountable Care Organizations which render individual contracts redundant.

policy? In particular, can these issues be attenuated, if not eliminated, through different combinations of important choices for the design of outcomes-based *reimbursement policies* (Rosenthal and Dudley 2006, 2007), including the choice of *reimbursement contract* (e.g., an outcomes-adjusted capitation payment for the care of a given population), *outcome measurements* (e.g., population-level data or provider-level data) used to adjust the amount reimbursed, and *reimbursement structure* (e.g., contracts with providers individually, or with a group of providers).

We formulate a parsimonious game-theoretic model to answer these questions and analyze the model to identify if or when optimal outcomes can be achieved by a variety of reimbursement policy choices. We then use diabetes care data from the UK to calibrate the model, address some important implementation issues, and explore how different classes of contracts currently used in practice compare to the theoretical optima.

For the first core question, we find below that the answer depends critically on whether specialist compensation scales with volume. In systems where it does, like the US, we have every reason to worry as they are vulnerable to collusion, which can make both outcomes and costs significantly worse. On the other hand, systems where the outcomes adjustment is purely on the level of primary care, while the specialist compensation is fixed (as in the UK), are virtually immune to collusion.

For the second question, we find below that individual outcomes-adjusted capitation contracts – a small modification of contracts already used in practice – have the potential to resolve any issues with collusion, as they can be shown to be both optimal and collusion-proof in the vast majority of the cases we study.

We also provide theory and discuss implications for the choice of reimbursement policies and reimbursement structures for both population-level and individual provider-level outcome metrics.

Overview: To improve our understanding of the performance of P4P[3] reimbursement policies, we develop a contracting model for care services in the principal-agent framework. A government acts as principal and contracts with two health care providers who are responsible for the care of a population of patients who manage a chronic condition. The government's objective is to maximize population health minus the government's health care expenditure.

In Section 2, we give our parsimonious model of a pathway for the care of patients with a chronic condition, in which there are two providers: a general practitioner (GP) and a specialist (SP), each of which chooses their quality of care, with a higher quality of care also carrying a higher cost. Increased quality of GP care can improve health and reduce the rate of complications that require SP care. Longer queues for the SP may result in worse health outcomes due to delays in treatment.

[3] The term P4P has been used for a variety of variable compensation systems. We focus here on P4P schemes based on health outcomes rather than other metrics such as adherence to specified processes (Petersen et al. 2006).

In Section 3, we consider the simplest version of this contracting problem, which we term the *naïve* problem, as it ignores the possibility of collusion and free-riding occurring. We show that, in the absence of free-riding and collusion, two existing contract types (capitation and per-patient) can achieve the first-best, if the reimbursement amounts are adjusted to take outcomes into account. Both individual and group contracts can achieve the optimum. Section 4 studies the consequences of collusion and free-riding if those issues are ignored in contract design. Once these considerations are taken into account, the first-best outcome can not be attained.

Section 5 gives our main model, where all of the components introduced in previous sections are united. Collusion and free-riding are endogenous (agents in the model free-ride and collude if it is in their interest to do so), and the principal designs the contracts with these issues in mind. Here, we show that (a) for any group contract there are individual contracts that are better, (b) within all individual contracts based on population health, a relatively simple linear contract is optimal and collusion-proof. Section 6 shows the size of the effects of reimbursement policies, free-riding, and collusion on health, costs, and system efficiency. We do so with illustrative computations calibrated to data about diabetes care in the UK NHS (OBH 2014, Diabetes UK 2019).

The theory in Section 5 and numerical evidence in Section 6 show that disregarding these two issues in contract design can lead to considerable negative consequences; yet, these consequences can be mitigated through the choice of reimbursement policy. Our results favor individual outcomes-adjusted capitation contracts, as a) if free-riding and collusion occur under those contracts, the negative effects will be smaller than with any other contract type we consider, b) collusion is highly unlikely to occur: in our numerical results, these contracts are collusion-proof for 100% of plausible parameter sets. In the optimal contracts we identify, population-level health measurement is required for capitation contracts, and provider-level measurement is required for per-patient ones.

To the best of our knowledge, our paper is the first in the healthcare operations literature to examine how the impact of collusion or free-riding depends on the contracts held by the providers.

We also respond to two secondary questions in the outcomes-based incentives debate. One, all of the contracts we show to be optimal include both penalties and bonuses, not just one or the other. Two, the optimal contracts we identify use rather large incentives. This is consistent with the UK NHS's Quality and Outcomes Framework (QOF, the world's largest P4P incentive scheme), which awards outcomes-based bonuses to GPs that are a large part of their total compensation (UK NHS 2020). It is less consistent with USA Medicare's use of smaller ($\leq 7\%$) outcomes-based incentives.

Section 7 interprets our results relative to some reimbursement policies used in the UK and USA. Our analytical results suggest that collusion may be of greater concern in applications where the specialist may hold volume-scaling contracts, as in the USA, while settings like the UK's QOF, where only the GP's contract is outcomes-adjusted, are more resistant to collusion.

As do all stylized models, our model makes simplifying assumptions. Section 8 shows that our assumption that provider costs are a known function of their effort can be addressed in practice by using realized costs with multi-period contracting or yardstick contracts. Another limit is the structure of our assumed care pathway with a GP for prevention or chronic disease management and a single specialist. Broader care networks are not yet modeled.

An online companion provides notation, proofs of mathematical claims, a discussion of model calibration, and model extensions that assess the robustness of our conclusions to some of our assumptions (allowing for multiple specialists, broader queue regimes, endogenous compliance, the GP directly affecting health outcomes, biased measures, and a different model of collusion).

1. Related Literature

Our work on outcomes-based reimbursement on care pathways links to literature on contract theory for multiple service providers, the healthcare operations management literature related to incentives, and the health economics literature on pay for performance.

We draw upon methods from the principal-agent framework (e.g., Bolton and Dewatripont 2005). The problem considered is a type of a *moral hazard* problem (Hölmstrom 1982), which may lead to *free-riding* in teams. A challenge of moral hazard in multi-agent situations is the inseparability of efforts: based on the final results, the principal cannot distinguish which agent exerted effort and which did not. While the same holds true in our setting, the fact that some patients are treated by just the GP and not the SP, and vice versa, provides some means to derive information about provider-specific decisions and thereby design individual rewards.

The use of individual rewards instead of group ones can alleviate free-riding, but introduces another source of inefficiency: potential collusion between agents (Tirole 1986). Several economics papers study whether it is beneficial to introduce individual rewards in light of collusion, but do so in settings with fundamental differences from ours. Laffont and Martimort (1998) study a principal who hires two agents separately, each producing one component, where one of each component is needed to form a final product. Collusion between agents is modeled via side contracts which are proposed and enforced by a third party. Baliga and Sjöström (1998) study a similar setting, but focus on private information (where one of the agents is better informed about actions of the other than the principal) and non-transferable utility between the agents limits their ability to use side-contracts. Rather than enforcing collusion through contracts, agents in our setting may rely on informal bargaining to split the gains from collusion.

The literature on contracting in multi-tier service systems (Shumsky and Pinker 2003, Lee et al. 2012, Freeman et al. 2017), like our paper, studies contracting when two agents are responsible for providing a service, but focuses on the first agent's role as the gatekeeper: the agent either provides

the service himself (herself) or redirects to the second agent. Only one agent in this model provides the service. In our paper, patient care results from the joint efforts of both agents.

Our paper fits well with the health care operations management literature that studies incentives (e.g., Lee and Zenios 2012, Ata et al. 2013, Gupta and Mehrotra 2015, Bastani et al. 2015, 2016, Zhang et al. 2016, Dai et al. 2017, Andritsos and Tang 2018, Guo et al. 2019, Aswani et al. 2019, Zhang et al. 2020, So and Tang 2000, Arifoglu et al. 2021, de Vericourt et al. 2021, She et al. 2022, Suen et al. 2022).

We comment on some papers in this stream to position our work. Jiang et al. (2012) assess performance-based contracts in access to care, considering how to incentivize providers to provide an appropriate capacity for outpatient treatment. As we do, they use a principal-agent model with queueing dynamics, yet there are two distinctions from our work: Jiang et al. (2012) consider a sole care provider (whereas in this paper, the involvement of multiple providers is a key source of complexity) and the contracts they study are based on observed waiting times for patients, rather than health outcomes. Adida et al. (2017) examine the efficiency of Medicare's bundled payment and its impact on health compared to the usual fee-for-service. They also consider a two-tier healthcare provision model, where low quality in primary care results in a higher probability of complications, requiring costly secondary care. There are differences. Only the primary care provider is strategic in their model (removing the complexity of a multi-agent system), and queueing dynamics for referrals are not considered. Unlike us, they consider the providers' ability to cherry-pick which patient they will take under their care. In a similar two-agent model to ours, Adida and Bravo (2019) model a system consisting of a payer who pays an Accountable Care Organization (ACO) for the treatment of patients, but the ACO outsources advanced treatment of patients to an external provider, paying them per referral. Ghamat et al. (2021) study a modified model for gain-sharing agreements between a payer and hospital and care provider, focusing on billing reduction issues in a study of targets for price and quality. Rajagopalan and Tong (2022) also consider a GP that provides referrals to an SP. They also consider congestion for the SP and outcomes-based reimbursement, and they consider partial attribution of efforts to providers. None of the cited papers in this paragraph and the previous consider free-riding or collusion.

Lastly, the health economics literature on the performance of outcomes-based contracts in practice is relevant to our work (e.g., Christianson et al. 2008, Eijkenaar et al. 2013, Burns and Pauly 2018). The evidence on the performance of these incentive schemes is mixed, with some success stories (Hsieh et al. 2017), but also programs that did not result in noticeable improvements, such as the US Hospital Value-Based Purchasing program (Figueroa et al. 2016, Ryan et al. 2017).

2. Model of Care Pathway and Effects of Services on Health

Before delving into the contracting problem in Section 3, we give our model of a chronic care pathway in Section 2.1 and give our model of how operational issues related to care provision and queueing delays influence health outcomes in Section 2.2.

2.1. Care Pathway Model

The system consists of two health care providers: a general practitioner (GP) and a specialist (SP). The GP serves as the primary care provider for a population of n people with a chronic condition and provides periodic ongoing care for the primary condition (e.g., diabetes). Thus, the GP provides routine checkups, manages symptoms, and prevents complications. The SP provides *acute* care for any arising complications (e.g., diabetic neuropathy). If the SP is not immediately available, patients queue for the first available appointment. After being treated by the SP, patients continue to receive care from the GP. The primary decision variables we consider are investments in improving the quality of care made by the GP and the SP, over a time period during which contracts are to hold (whose duration is selected to be one year for expository purposes).

A fraction ϕ of patients (such that ϕn is integer-valued) take advantage of the GP's services and benefit from his or her treatment. Patients who so visit the GP are called *adherent* patients (clinical adherence in the medical literature, McNabb 1997). The GP's primary decision variable is the quality of care for their patients, which we denote by $d_G \in [0, \infty)$. This variable abstracts a number of smaller decisions, including selecting a frequency of patient visits, the amount and type of prevention activities, investments in equipment, training, and so forth. Depending on the quality of GP care, each adherent patient has a probability λ of developing a complication during the period, which will then require SP care. Of course, higher quality of care will lead to fewer complications, which we model as $\lambda(d_G) \doteq \overline{\lambda}/(1+d_G)$, where $0 < \overline{\lambda} \leq 1$ is the probability that complications will occur without the GP's treatment. Thus, assuming that these events are independent,[4] the number of people who develop complications is a random variable $\mathcal{S}(d_G) \sim \text{Bin}(\phi n, \lambda(d_G)) + \text{Bin}\left((1-\phi)n, \overline{\lambda}\right)$.

The GP's costs are given by $k_G(d_G)$, a twice continuously differentiable, convex, and increasing function. To ensure an interior solution, we assume that the marginal cost of providing at least a small amount of treatment is very low, i.e., $k_G'(0) = 0$.

Figure 1 illustrates this pathway. Table EC.1 in the online companion summarizes the notation.

Patients who develop complications require acute care from the SP. If the SP is not immediately available, patients experience a queueing delay (which may influence outcomes). We model this

[4] The independence assumption is not entirely innocuous, as environmental conditions could, in principle, lead to correlated complications between patients.

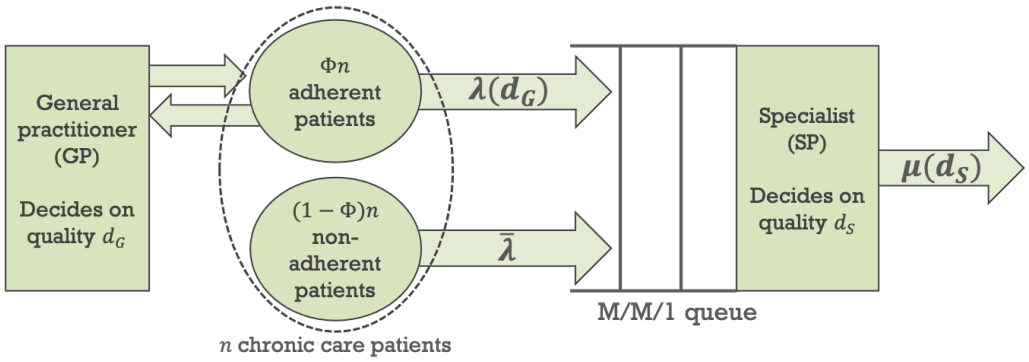


Figure 1 Model of primary care and specialist care service delivery choices and resulting queueing dynamics that may affect the progression of a chronic condition.

delay by embedding a queueing model and approximating the queueing dynamics. The arrival process of the $\mathcal{S}(d_G)$ patients with complications is approximated as a Poisson process, with arrival rate $n\phi\lambda(d_G) + n(1-\phi)\overline{\lambda}$. We denote the average complication rate across all patients by

$$\Lambda(d_G) \doteq \phi\lambda(d_G) + (1-\phi)\bar{\lambda}. \tag{1}$$

The SP makes numerous decisions: the number of appointments per week, nurses staffed, equipment, and so forth. We model these decisions by $d_S \in [0, \infty)$, the SP's quality of care. As higher quality of care will lead to the condition being stabilized and possibly resolved more quickly, we assume that the service rate is proportional to the quality of care: $\mu(d_S) \doteq \underline{\mu} + \theta d_S$, where $\underline{\mu}$ is the basic acceptable service rate and θ is a positive constant. Providing higher quality care is costly to the SP, with the cost $k_{SF}(d_S)$ being a twice continuously differentiable, increasing, and convex function with $k'_{SF}(0) = 0$. The function $k_{SF}(d_S)$ can also be interpreted as the cost of providing capacity for care. In either case, $k_{SF}(d_S)$ includes all fixed costs incurred by the SP, independent of the number of patients treated. The SP also bears a variable cost k_{SV} for every patient treated (disposable materials, diagnostic tests, or medication per patient). The GP can have such variable costs as well. Because the number of patients under the GP's care is deterministic in our model, those costs are included in the cost function $k_G(d_G)$ without loss of generality.

Assuming the treatment times are exponential and the arrival rate of patients with complications is homogeneous throughout the year, the queueing dynamics can be approximated by an M/M/1 queue with arrival rate $n\Lambda(d_G)$ (a result of the GP's decisions) and service rate $\mu(d_S)$ (a result of the SP's decision). Thus, we explicitly model congestion (and queueing) for the SP but not for the GP. The reason for this is that the GP provides periodic care throughout the year (consisting of periodic checkups), where the outcomes are insensitive to the timing of these checkups. In contrast, the SP provides *acute* care for complications, the treatment of which can be time-sensitive. This

formulation implicitly assumes that a patient with complications can access the SP directly or that the time needed to obtain a referral is negligible compared to the waiting time to access the SP.

The providers' quality of care decisions may also affect things beyond the rates of complications and service. In Appendix C.1, we generalize the model, allowing those decisions to directly impact the health of patients and patients' decisions on whether to adhere to treatment. Our main results are robust to these generalizations.

2.2. Impact of Care on Health

We now describe our model of how this care pathway influences the health of patients. The influence of the care pathway on health outcomes is modeled with respect to quality-adjusted life years (QALY) as a measure of health. We compare QALYs and financial results through a willingness to pay conversion factor (e.g., £20k-£30k per QALY generated in the UK NHS, Claxton et al. 2015).

Patient i's health state at the beginning of the period of care is modeled by q_i^0, the monetary value of their health capital (Grossman 1972), for $i = 1, 2, \ldots, n$. We denote by $q_P^0 = \sum_{i=1}^{n} q_i^0$ the total (sum) of initial health capital in the population, and by $\bar{q}_P^0 = q_P^0/n$ the average. (When discussing the health of other patient groups, we will also follow this notational convention, with a bar denoting the average.) We assume that the average initial health capital of adherent and non-adherent patients is equal. The health of a patient after a period of care is stochastic: if the patient develops a complication, their health will experience an immediate decline, this decline will further increase with time spent waiting for care, and finally, after treatment is completed, some amount of the lost health will be restored.

Let $Q_i(d_G, d_S)$ be the random variable representing patient i's health capital at the end of the period, and let q_i be its realization. A patient develops a complication with probability $\lambda_i(d_G)$, where $\lambda_i(d_G) = \lambda(d_G)$ if the patient was treated by the GP or $\lambda_i(d_G) = \overline{\lambda}$ otherwise. If a patient does not develop a complication, the health state remains constant ($q_i = q_i^0$). If there is a complication, health declines as a function of the time spent waiting for treatment, which is captured by increasing function $w(t)$. Once SP treatment is received, a fraction ζ of lost health is restored by the treatment. We consider ζ to be exogenous here. (We allow for endogenous ζ in Appendix C.1.) Thus,

$$Q_i(d_G, d_P) = \begin{cases} q_i^0(1 - w(W(d_G, d_P))(1 - \zeta)) & \text{with probability } \lambda_i, \\ q_i^0 & \text{otherwise,} \end{cases}$$

where $W(d_G, d_P)$ is a random variable with the same distribution as the sojourn times in the system (which depends on the providers' decisions, d_G and d_P).

Formally, our queue is over a finite time period. For tractability, we use steady-state queueing metrics to approximate queueing outcomes, and assume the queue has a steady state even with a minimal quality of care ($\overline{\lambda}\underline{\mu} - n > 0$). With this steady-state approximation and the results of

Taylor and Karlin (1998, Ch. IX) for the density of steady-state sojourn times in M/M/1 queues, the expected health of patient i at the end of the period of care is

$$\mathbb{E}Q_i(d_G, d_S) = q_i^0 \left(1 - \lambda_i(d_G) \int_0^\infty (\mu(d_S) - n\Lambda(d_G)) e^{-(\mu(d_S) - n\Lambda(d_G))t} w(t)(1-\zeta) dt\right). \tag{2}$$

For now, we assume the health decline is linear, so $w(t) = a + bt$, for some $b > 0$ and $a \in [0, 1)$.[5] Thus, a is the health loss from developing a complication (even if treated immediately), whereas b is the marginal decline in health over time. Using this form for $w(t)$ yields

$$\mathbb{E}Q_i(d_G, d_S) = q_i^0 \left(1 - a\lambda_i(d_G)(1-\zeta) - \frac{b\lambda_i(d_G)(1-\zeta)}{\mu(d_S) - n\Lambda(d_G)}\right). \tag{3}$$

Thus, the expected sum of the health capital in the population ($Q_P = \sum_{i=1}^n Q_i$) is

$$\mathbb{E}Q_P(d_G, d_S) = q_P^0 \left(1 - a\Lambda(d_G)(1-\zeta) - \frac{b\Lambda(d_G)(1-\zeta)}{\mu(d_S) - n\Lambda(d_G)}\right). \tag{4}$$

We assume the average health capital of the population is such that $\bar{q}_P^0 \geq k_{SV}$. (If this were not the case, then it would be better not to have any SP treatment at all.)

The *expected system efficiency* function u (the expected value of the whole population's health capital minus the GP's costs and the SP's fixed and variable costs) is, therefore:

$$u(d_G, d_S) = \mathbb{E}Q_P(d_G, d_S) - k_G(d_G) - k_{SF}(d_S) - n\Lambda(d_G)k_{SV}. \tag{5}$$

It will be useful below to discuss the health of patients in different sets. Let G be the set of all adherent patients (those seen by the GP). Note that $|G| = \phi n$. The sum of health capital of the GP's patients is then $Q_G(d_G, d_S) = \sum_{i \in G} Q_i(d_G, d_S)$. Applying (3) yields

$$\mathbb{E}Q_G(d_G, d_S) = q_P^0 \phi \left(1 - a\lambda(d_G)(1-\zeta) - \frac{b\lambda(d_G)(1-\zeta)}{\mu(d_S) - n\Lambda(d_G)}\right). \tag{6}$$

Similarly, the expected sum of the health of the SP's patients ($\mathbb{E}Q_S$) can be derived by separating the patient pool into the (random) set of those who develop complications (S) and the set of those who do not (S'). Noting that the SP treats all of the patients with complications but none of the others, we have $Q_S = Q_P - Q_{S'}$. Taking an expectation and applying (4) gives

$$\mathbb{E}Q_S(d_G, d_S) = q_P^0 \Lambda(d_G) \left(1 - a(1-\zeta) - \frac{b(1-\zeta)}{\mu(d_S) - n\Lambda(d_G)}\right). \tag{7}$$

[5] Appendix C.5 shows that our model remains tractable using alternative specifications for $w(t)$ where the integral in (2) admits a closed-form solution, including the threshold function $w(t) = a + \mathbb{1}(t \geq T)b$, where T is a critical response time, or an exponential function $w(t) = a + \exp(bt)$ (where the steady state condition becomes $\bar{\lambda}\underline{\mu} - n > b$).

3. Naïve Contracting Problem and the First Best Benchmark

Now we turn our attention to the contracting problem faced by the principal (the government), whose objective is to maximize the health benefits generated by the system minus its own spending. We start with the simplest version of this problem in a classical contract theory setup, modeling all of the challenges discussed in the introduction *except* collusion and free-riding. Due to the neglect of these two issues, we will refer to the problem here as the *naïve contracting problem*. This model and its analysis are used in Section 4 to explore the consequences of ignoring collusion and free-riding, and again used to analyze our main model in Section 5, where the principal designs contracts being mindful that these two issues may occur.

Individual contracts. Consider a principal who contracts with each provider individually. The principal chooses contracts to offer to the GP and the SP ($v_1()$ and $v_2()$, respectively) with the goal of maximizing total health in the system minus its own costs.

The primary variable of the principal's interest (q_P, a realization of Q_P) is unobservable, and so not contractible. The principal may, however, observe a signal $\mathcal{Q}_P(d_G, d_S)$, which consists of the actual population health and independent, zero-mean noise ($\mathcal{Q}_P(d_G, d_S) = Q_P(d_G, d_S) + \varepsilon$, where $\mathbb{E}[\varepsilon] = 0$).[6] We assume the distribution of ε is common knowledge.

We may also be interested in estimating the total health of patients seen by the GP, denoted Q_G, with related signal $\mathcal{Q}_G$ and realization q_G. Analogously, the total health of patients treated by the SP (Q_S) may be estimated by an unbiased but noisy signal $\mathcal{Q}_S$ with realization q_S. Each agent's contract can be based on either the signal of patient health for patients under that agent's care ($\mathcal{Q}_G$ and $\mathcal{Q}_S$, respectively) or the signal of the whole population's health ($\mathcal{Q}_P$).

After the principal offers contracts v_1 and v_2, the agents evaluate whether they should take the contract by looking at how much they will make if they accept the contract, and both agents choose Nash equilibrium quality of care decisions $(\tilde{d_G}, \tilde{d_S})$. For agents to accept the contract, their profit under equilibrium decisions needs to be at least the same as the value of their outside option ($V_i, i \in \{G, S\}$), i.e., of the amount they could earn (perhaps through other employment) if they reject the principal's offer. This gives us the *naïve individual contracting problem* (NICP):

$$\max_{\mathcal{Q}_1 \in \{\mathcal{Q}_P, \mathcal{Q}_G\}, \mathcal{Q}_2 \in \{\mathcal{Q}_P, \mathcal{Q}_S\}, v_1(\mathcal{Q}_1), v_2(\mathcal{Q}_2)} \quad \mathbb{E}\left[Q_P(\tilde{d_G}, \tilde{d_S}) - v_1(\mathcal{Q}_1(\tilde{d_G}, \tilde{d_S})) - v_2(\mathcal{Q}_2(\tilde{d_G}, \tilde{d_S}))\right], \tag{8}$$

$$\text{subject to} \quad \tilde{d_G} \in \arg\max_{d_G \in [0,\infty)} \mathbb{E}\left[v_1(\mathcal{Q}_1(d_G, \tilde{d_S})) - k_G(d_G)\right], \tag{9}$$

$$\tilde{d_S} \in \arg\max_{d_S \in [0,\infty)} \mathbb{E}\left[v_2(\mathcal{Q}_2(\tilde{d_G}, d_S)) - k_{SF}(d_S) - \mathcal{S}(\tilde{d_G}) k_{SV}\right], \tag{10}$$

$$V_G \leq \mathbb{E}\left[v_1(\mathcal{Q}_1(\tilde{d_G}, \tilde{d_S})) - k_G(\tilde{d_G})\right], \tag{11}$$

$$V_S \leq \mathbb{E}\left[v_2(\mathcal{Q}_2(\tilde{d_G}, \tilde{d_S})) - k_{SF}(\tilde{d_S}) - \mathcal{S}(\tilde{d_G}) k_{SV}\right]. \tag{12}$$

[6] Appendix C.2 identifies the degradation in the contract's performance if the noise has unknown or uncorrected bias.

In summary, the principal first offers contracts v_1, v_2, which the agents accept if the individual rationality (IR) constraints (11)–(12) hold. After accepting the contracts, the agents make equilibrium decisions as given by the incentive compatibility (IC) constraints (9)–(10). During the contract, patients arrive for treatment, as illustrated in Figure 1. At the end of the period, the number of people who developed complications is realized (s), as are the health signals (q_P, q_G, q_S). The agents are then paid the stipulated amounts, $v_1(q_1)$ to the GP and $v_2(q_2)$ to the SP.

The best outcome the principal can hope to achieve in this problem (the *first-best*) is the one where agents make decisions that maximize the value generated by the system, and all of this value is appropriated by the principal. Formally, a decision (d_G^*, d_S^*) which solves the optimization problem $\arg\max u(d_G, d_S)$ is called first-best, where $u(d_G, d_S)$ is the system efficiency function in (5). The contracts (v_1, v_2) are said to attain the first-best if the decisions induced by those contracts $(\tilde{d_G}, \tilde{d_S})$ equal the first-best decisions (d_G^*, d_S^*) and the IR constraints (11)–(12) bind.

Prop. 1 shows that the principal has several options available that can achieve the first-best.

Proposition 1 (Optimal individual contracts) *There exists an optimal pair of* outcomes-adjusted capitation contracts (v_{GC}, v_{SC}) *that achieves the first-best in the NICP. Under these contracts, provider $i \in \{G, S\}$ is paid a capitation fee c_i, which is paid for every patient in the population (treated or not), and which is outcomes-adjusted according to the measured population health:*

$$c_i(q_P) = f_{iC} + r_{iC}(q_P - t_{iC}). \tag{13}$$

The first-best can also be achieved by outcomes-adjusted per-patient contracts (v_{GP}, v_{SP})*, which award providers a fee p_i per patient they treat. This fee is outcomes-adjusted based on the health of patients treated by that provider: $p_G(q_G) = f_{GP} + r_{GP}(q_G - t_{GP})$, $p_S(q_S) = f_{SP} + r_{SP}(q_S - t_{SP})$.*

In the capitation expression (13), $f_{iC} > 0$ is the fixed capitation rate, and $r_{iC}(q_P - t_{iC})$ is the outcomes-adjustment reimbursement component where the per-capita rate is increased at rate r_{iC} for any improvements in the population health above the target health threshold t_{iC} (or decreased when falling short of the health target). Hence, outcomes-adjusted variants of commonly used contracts are capable of achieving the first-best here. As these contracts achieve the first-best, they are optimal over the space of *all* contracts, without restricting the principal's choice of contracts to any parametric form. Denoting the optimal capitation (per-patient) contracts by additional subscript G (P), the payments to providers under the contracts given in Prop. 1 are

$$\begin{aligned} v_{iC}(q_P) &= nc_i(q_P), \text{ for } i \in \{G, S\}, \quad \text{(capitation contracts)} \\ v_{GP}(q_G) &= \phi np_G(q_G), \quad v_{SP}(q_S) = sp_S(q_S), \quad \text{(per-patient contracts).} \end{aligned} \tag{14}$$

Remark 1 *The proof of Prop. 1 shows that under the optimal contracts, target health thresholds can be chosen so that* $t_{GC}, t_{GP}, t_{SC}, t_{SP} < \mathbb{E}Q_P(d_G^*, d_S^*)$. *(For expressions for these thresholds, see (EC.13),(EC.17),(EC.19), and (EC.21) in Appendix A.1.). A consequence is that equilibrium outcomes adjustment typically results in bonuses rather than penalties (compared to the base rate).*

Group contracts. The principal may also create a single group contract for the two providers. To examine this, we introduce some additional notation. Let Q_A be the total health of patients treated by at least one of the GP or the SP (A for alliance/group), let $\mathcal{Q}_A$ be the related unbiased signal, and let q_A be its realization. (This notation is analogous to Q_G and Q_S in Section 3.)

Consider a principal who contracts with one agent who chooses both decision variables (d_G, d_S). Instead of the NICP in (8)–(12), this principal solves the *naïve group contracting problem* (NGCP), whose solution is characterized by Prop. 2 that follows:

$$\max_{\mathcal{Q}\in\{\mathcal{Q}_P,\mathcal{Q}_A\},v(\mathcal{Q})} \quad \mathbb{E}\left[Q_P(\tilde{d_G},\tilde{d_S}) - v(\mathcal{Q}(\tilde{d_G},\tilde{d_S}))\right], \tag{15}$$

$$\text{subject to} \quad (\tilde{d_G},\tilde{d_S}) \in \underset{(d_G,d_S)\in[0,\infty)^2}{\arg\max} \mathbb{E}\left[v(\mathcal{Q}(d_G,d_S)) - k_G(d_G) - k_{SF}(d_S) - \mathcal{S}(d_G)k_{SV}\right], \tag{16}$$

$$\mathbb{E}\left[v(\mathcal{Q}(\tilde{d_G},\tilde{d_S})) - k_G(\tilde{d_G}) - k_{SF}(\tilde{d_S}) - \mathcal{S}(\tilde{d_G})k_{SV}\right] \geq V_G + V_S. \tag{17}$$

Proposition 2 (Optimal group contracts) *Optimal group contracts are an outcomes-adjusted capitation contract* (v_{AC}) *with per-capita fee* $c_A(q_P) = f_{AC} + r_{AC}(q_P - t_{AC})$, *and an outcomes-adjusted per-patient contract* (v_{AP}) *with per-patient fee* $p_A(q_A) = f_{AP} + r_{AP}(q_A - t_{AP})$. *Both contracts achieve the first-best for the NGCP.*

The optimal individual contracts in Prop. 1 and the optimal group contracts in Prop. 2 are not identical. They have the same form but have different parameter values.

4. Consequences of Naïvité

Section 3 presented a naïve contract design that did not account for two potential concerns, free-riding and collusion. The purpose of this section is twofold. First, it establishes what unintended consequences can occur as a result of ignoring these concerns if the contracts from Section 3 were implemented. Second, the models of free-riding in Section 4.1 and collusion in Section 4.2 provide insights for how to mitigate those concerns, as discussed in Section 5.

4.1. Free-riding in Group Contracts

A group contract reimburses the GP and SP jointly, and the GP and SP share rewards. This type of contract runs the risk of free-riding. We follow the approach of Hölmstrom (1982) to study free-riding within a group in our model as follows. Agents in the group split the final payout of their contract, which they do according to the Nash bargaining solution. Each agent will receive the value

of their outside option (V_G for the GP, V_S for the SP) plus an equal split of any reimbursement in excess of the sum of the values of those options, i.e., $V_G + V_S$. Instead of making decisions jointly (by solving the incentive compatibility constraint (16) in the NGCP), each agent chooses their own decision individually, and bears the associated cost. Thus, after receiving a contract $v(\mathcal{Q})$ from the principal, instead of making cooperative decisions $\tilde{d}_G$, $\tilde{d}_S$ as given by (16), free-riding agents will choose $\tilde{d}_G^{FR}$ and $\tilde{d}_S^{FR}$ that maximize their own benefit under this reimbursement policy. That is, they solve the *free-riding problem* (FRP):

$$\tilde{d}_G^{FR} \in \underset{d_G \in [0,\infty)}{\arg\max}\, \mathbb{E}\left[\frac{1}{2}\left(v(\mathcal{Q}(d_G, \tilde{d}_S^{FR})) - V_G - V_S\right) + V_G - k_G(d_G)\right], \tag{18}$$

$$\tilde{d}_S^{FR} \in \underset{d_S \in [0,\infty))}{\arg\max}\, \mathbb{E}\left[\frac{1}{2}\left(v(\mathcal{Q}(\tilde{d}_G^{FR}, d_S)) - V_G - V_S\right) + V_S - k_{SF}(d_S) - \mathcal{S}(\tilde{d}_G^{FR})k_{SV}\right]. \tag{19}$$

Let $\tilde{d}_{G,A}^{FR}$ and $\tilde{d}_{S,A}^{FR}$ denote equilibrium decisions of free-riding agents under one of the group contracts of Proposition 2. When we consider different kinds of organizational issues, we will use this notation, with superscript standing for the type of organizational issue (in this case, FR for free-riding and C for collusion) and the additional subscript standing for the contract which the agents hold (A for alliance/group contract). If a superscript is omitted, that means neither free-riding nor collusion and present, and thus the decisions are made according to the model in Section 3. If the additional subscript is omitted, we are talking about decisions under an unspecified contract. We also let $\pi_j(v, d_G, d_S)$ denote the expected profit of agent $j \in \{G, S, A\}$ when holding contract $v(\mathcal{Q})$, for $\mathcal{Q} \in \{\mathcal{Q}_P, \mathcal{Q}_G, \mathcal{Q}_S, \mathcal{Q}_A\}$ and making decisions (d_G, d_S):

$$\pi_G(v, d_G, d_S) = \mathbb{E}\left[v(\mathcal{Q}(d_G, d_S)) - k_G(d_G)\right], \tag{20}$$

$$\pi_S(v, d_G, d_S) = \mathbb{E}\left[v(\mathcal{Q}(d_G, d_S)) - k_{SF}(d_S) - \mathcal{S}(d_G)k_{SV}\right], \tag{21}$$

$$\pi_A(v, d_G, d_S) = \mathbb{E}\left[v(\mathcal{Q}(d_G, d_S)) - k_G(d_G) - k_{SF}(d_S) - \mathcal{S}(d_G)k_{SV}\right]. \tag{22}$$

We formalize the effects of free-riding on the key metrics of interest in Theorem 1.

Theorem 1 (Effects of free-riding) *Suppose the GP and SP are given one of the optimal group contracts from Prop. 2, but act in individual interest by solving the FRP in (18)-(19). The GP and SP will fail to achieve the first-best. Compared to the first-best:*

(i) *System efficiency is lower:* $u(\tilde{d}_{G,A}^{FR}, \tilde{d}_{S,A}^{FR}) < u(d_G^*, d_S^*)$, *with* u *given by (5).*

(ii) *Complication rate is higher:* $\Lambda(\tilde{d}_{G,A}^{FR}) > \Lambda(d_G^*)$.

(iii) *Expected health of patients is lower:* $\mathbb{E}Q_P(\tilde{d}_{G,A}^{FR}, \tilde{d}_{S,A}^{FR}) < \mathbb{E}Q_P(d_G^*, d_S^*)$, *with* $\mathbb{E}Q_P$ *given by (4).*

(iv) *Expected profit of the group is lower:* $\pi_A(v_A, \tilde{d}_{G,A}^{FR}, \tilde{d}_{S,A}^{FR}) < \pi_A(v_A, d_G^*, d_S^*)$, *with* π_A *given by (22).*

(v) *Expected government expenditure is lower:* $\mathbb{E}v_A(\mathcal{Q}_P(\tilde{d}_{G,A}^{FR}, \tilde{d}_{S,A}^{FR})) < \mathbb{E}v_A(\mathcal{Q}_P(d_G^*, d_S^*))$.

Free-riding incentivizes each agent to decrease their quality of care, compared to the interests of the group. However, the expected profit of both agents is submodular in (d_G, d_S) (verified by taking $\partial^2/\partial d_G \partial d_S$ of (22)), so one agent decreasing treatment quality incentivizes the other one to increase theirs. Yet, the effects on the system are consistent with the expectations based on simpler free-riding models. As can be seen from parts (ii)-(iii) of the theorem, the net effect of free-riding on health and complication rates is negative. A consequence of that is also lower government expenditure, as the expenditure is proportional to health benefits created, as shown in part (v). Any cost savings by the agents due to free-riding appear insufficient to compensate for the lower income, resulting in lower overall profits for the agents, as shown in part (iv).

4.2. Collusion in Individual Contracts

The concern with individual contracts is that the agents might collude in order to make both of them better off (Rodwin 1995). This section shows that a type of collusion that can emerge in this setting is *demand inducement*, where colluding agents act by generating demand for each other.

A common economic approach for modeling collusion would be to model colluding agents as a single entity, with the entity making both agents' decisions to maximize its profits. We study that approach, an informative "corner case," in Appendix C.4, but we steer away from it here, because it would be inconsistent for us to assume that colluding agents are able to perfectly coordinate (acting as a single entity), whereas formal groups and alliances are not. Moreover, we wish to compare the negative effects of collusion with those of free-riding. To make such comparisons, we require a set of assumptions that allows for both behaviors simultaneously.

As our main model of collusion, therefore, we adopt the same assumptions about the behavior of agents as in Section 4.1. Thus, under collusion, colluding agents share their joint revenue according to the Nash bargaining solution but are responsible for their individual decisions and costs. Agents are said to be engaged in collusion if, after being given individual contracts $v_1(q_1)$, $v_2(q_2)$, they do not make the non-cooperative decisions $\tilde{d}_G$, $\tilde{d}_S$ as given by (9)-(10). Instead, they choose $\tilde{d}_G^C$, $\tilde{d}_S^C$ to maximize individual profits, knowing that excess revenues will be split via Nash bargaining, as in the following *collusion problem* (CP):

$$\tilde{d}_G^C \in \underset{d_G \in [0,\infty)}{\arg\max}\, \mathbb{E}\left[\frac{1}{2}\sum_{i\in\{1,2\}}\Big(v_i(\mathcal{Q}_i(d_G, \tilde{d}_S^C)) - \pi_i(v_i, \tilde{d}_G, \tilde{d}_S)\Big) + \pi_G(v_1, \tilde{d}_G, \tilde{d}_S) - k_G(d_G)\right], \tag{23}$$

$$\tilde{d}_S^C \in \underset{d_S \in [0,\infty)}{\arg\max}\, \mathbb{E}\left[\frac{1}{2}\sum_{i\in\{1,2\}}\Big(v_i(\mathcal{Q}_i(\tilde{d}_G^C, d_S)) - \pi_i(v_i, \tilde{d}_G, \tilde{d}_S)\Big) + \pi_S(v_2, \tilde{d}_G, \tilde{d}_S) - k_{SF}(d_S) - \mathcal{S}(\tilde{d}_G^C)k_{SV}\right]. \tag{24}$$

In the absence of collusion, the proof of Prop. 1 shows that, with optimal contracts, both agents receive the full value of marginal benefits created by the system. If agents act in their own interest, such compensation is beneficial, serving the purpose of aligning the agents' interests. The presence of collusion, however, creates an incentive to over-treat patients because colluding parties can be

compensated at a higher marginal rate than the value they create. While the resulting increase in quality and patient health looks appealing, it comes paired with inefficiently high costs.

If the SP holds a per-patient contract, there is an additional perverse incentive. An SP with such a contract earns more the more patients he or she has, creating an incentive for a colluding GP to decrease service quality so as to increase complication rates and generate demand for the SP. Also, there is an incentive to decrease service quality because of a lack of coordination, as above with free-riding agents. Lastly, the colluding agents make their decisions in a way that disregards the SP's variable costs because those costs depend only on the GP's decision, yet are borne by the SP.

In the presence of such countervailing effects, not all comparative statics we are interested in are conclusive. Therefore we adopt a threefold approach. One, conclusive results are presented in Theorem 2. Two, questions that remain inconclusive are addressed by numerical analysis in Section 6. Three, Prop. 3 below identifies special cases of the parameter space in which we can establish results that do not hold in general. Following the notation used in Theorem 1, in the theorem below, $\tilde{d}^C_{G,k,j}$ ($\tilde{d}^C_{S,k,j}$) stands for the equilibrium decision of a colluding GP (SP) when the providers hold contracts $(k,j) \in \{C,P\}^2$.

Theorem 2 (Effects of collusion) *Suppose the GP and SP are given individual contracts as defined in Prop. 1, but engage in collusion by solving (23)-(24) instead of (9)-(10). Then, the first-best is not achieved. For every GP contract v_{Gk} for $k \in \{C,P\}$, we can compare outcomes when the SP holds v_{SC} to ones where the SP holds v_{SP}:*

(i) *The complication rate is lower under v_{SC}* ($\Lambda(\tilde{d}^C_{G,k,C}) < \Lambda(\tilde{d}^C_{G,k,P})$).

(ii) *The specialist's quality of service is lower under v_{SC}* ($\tilde{d}^C_{S,k,C} < \tilde{d}^C_{S,k,P}$).

(iii) *If $k = C$, the pop. health is higher under v_{SC}* ($\mathbb{E}Q_P(\tilde{d}^C_{G,k,C}, \tilde{d}^C_{S,k,C}) > \mathbb{E}Q_P(\tilde{d}^C_{G,k,P}, \tilde{d}^C_{G,k,P})$).

The effect is more complex when the GP holds a per-patient contract: GP decisions have a second-order impact on patients not under the GP's care (due to the GP's influence on congestion for the SP), but the GP's compensation does not account for the outcomes of those patients.

Prop. 3 gives additional results for special cases. In summary, if the SP's variable costs are low enough and the adherence rate is high enough, the second-order effects of the GP's decision become negligible, and some contracts are able to eliminate the adverse effects of collusion altogether.

Proposition 3 (Special cases of collusion) *If the variable costs of the SP are negligible ($k_{SV} = 0$), then first-best is achieved under collusion when both agents hold optimal capitation contracts (v_{GC} for the GP, v_{SC} for the SP in Prop. 1). Moreover, if $k_{SV} = 0$ and there is full adherence ($\phi = 1$), then first-best is achieved under collusion when the GP holds a per-patient contract v_{GP} and the SP holds a capitation contract v_{SC}.*

A variable cost k_{SV} close to zero is more realistic in settings where most costs (such as labor, equipment, facilities) are fixed, and marginal treatment costs (consumables, etc.) are very inexpensive. Full adherence ($\phi = 1$) may hold in some settings, such as for hospitalized patients. But full adherence might not be realistic in other settings, such as when managing to improve control of diabetes, especially among young patients (Borus and Laffel 2010, García-Pérez et al. 2013).

5. Main model: Endogenous Collusion and Free-riding

In Section 3, the principal designed contracts oblivious of the potential free-riding and collusion issues. If those issues do occur (as in Section 4), it is unsurprising that a contract designed ignoring them may have degraded performance. Here, we turn our attention to the mindful principal, who designs contracts with these potential issues in mind, and show potential gains from doing so.

5.1. Mindful Group Contracts

We model free-riding with group contracts using the same approach as in Section 4. The principal designs the contract $v(\mathcal{Q})$ with the goal of maximizing population health minus the health care expenditure, knowing that agents under a group contract will experience free-riding. Thus, the principal's problem is given by the *main (mindful) group contracting problem* (MGCP):

$$\max_{\mathcal{Q}\in\{\mathcal{Q}_P,\mathcal{Q}_A\},v(\mathcal{Q})} \quad \mathbb{E}\left[Q_P(\tilde{d}_G^{FR},\tilde{d}_S^{FR}) - v(\mathcal{Q}(\tilde{d}_G^{FR},\tilde{d}_S^{FR}))\right], \tag{25}$$

$$\text{subject to} \quad \tilde{d}_G^{FR} \in \arg\max_{d_G\in[0,\infty)} \mathbb{E}\left[\frac{1}{2}\left(v(\mathcal{Q}(d_G,\tilde{d}_S^{FR})) - V_G - V_S\right) + V_G - k_G(d_G)\right], \tag{26}$$

$$\tilde{d}_S^{FR} \in \arg\max_{d_S\in[0,\infty)} \mathbb{E}\left[\frac{1}{2}\left(v(\mathcal{Q}(\tilde{d}_G^{FR},d_S)) - V_G - V_S\right) + V_S - k_{SF}(d_S) - \mathcal{S}(\tilde{d}_G^{FR})k_{SV}\right], \tag{27}$$

$$\mathbb{E}\left[\frac{1}{2}\left(v(\mathcal{Q}(\tilde{d}_G^{FR},\tilde{d}_S^{FR})) - V_G - V_S\right) + V_G - k_G(\tilde{d}_G^{FR})\right] \geq V_G, \tag{28}$$

$$\mathbb{E}\left[\frac{1}{2}\left(v(\mathcal{Q}(\tilde{d}_G^{FR},\tilde{d}_S^{FR})) - V_G - V_S\right) + V_S - k_{SF}(\tilde{d}_S^{FR}) - \mathcal{S}(\tilde{d}_G^{FR})k_{SV}\right] \geq V_S. \tag{29}$$

Compared to the NICP in Section 3, the principal's objective (25) is the same. The incentive compatibility constraints in (26)-(27) are exactly those of the FRP of Section 4.1, as given in (18)-(19). There are now two participation constraints (28)-(29), as the profit of two agents from the contract might be uneven, and it needs to be in the interest of both to accept the contract.

For the remainder of the section, we adopt the following technical assumption.

Assumption 1 *The SP's variable costs are strictly positive, $k_{SV} > 0$. Contracts $v()$ are differentiable and allow for interchanging the order of differentiation and expectation (e.g., $v()$ satisfies the hypothesis of the dominated convergence theorem).*

Theorem 3 (Linear mindful group contracts) *Consider the MGCP with group contract $v(\boldsymbol{q})$ satisfying Assumption 1 and let $\tilde{d}_G, \tilde{d}_S$ be the interior ($\tilde{d}_G, \tilde{d}_S > 0$) decisions that the contract*

induces. Then, there exists a linear outcomes-adjusted capitation contract $v^\dagger(q_P)$ *that induces the same decisions* $(\tilde{d}_G, \tilde{d}_S)$*, at an equal or lower cost to the principal.*

As a simple consequence of Theorem 3, a linear outcomes-adjusted capitation contract is optimal, but not necessarily uniquely so. There are also two interesting properties which hold for all group contracts, not just the optimal ones:

Proposition 4 (Properties of all group contracts)

1. *By choosing a group contract, the principal can induce any expected population health in* $[\mathbb{E}Q_P(0,0), q_P^0)$*, but not all* d_G, d_S *pairs can be induced. Specifically, any inducible* d_G, d_S *such that* $d_G, d_S \neq 0$ *satisfy* $\frac{\partial \mathbb{E}Q_P(d_G,d_S)}{\partial d_S} / \frac{\partial \mathbb{E}Q_P(d_G,d_S)}{\partial d_G} = k'_{SF}(d_S)/k'_G(d_G)$.
2. *For any* d_G, d_S *inducible by a group contract, there exist* $d_G^\dagger, d_S^\dagger$ *which yield the same expected population health, at a lower treatment cost, so that* $\mathbb{E}Q_P(d_G, d_S) = \mathbb{E}Q_P(d_G^\dagger, d_S^\dagger)$ *and* $\mathbb{E}(k_G(d_G) + k_{SF}(d_S) + \mathcal{S}(d_G)k_{SV}) > \mathbb{E}(k_G(d_G^\dagger) + k_{SF}(d_S^\dagger) + \mathcal{S}(d_G^\dagger)k_{SV})$.

Informally, under any group contract, the providers choose inefficiently expensive decisions. The main issue with all group contracts under free-riding is that the agents will not be inherently mindful of the externalities they impose on each other. For example, the GP's decision affects the expected health of the patients (which can be accounted and incentivized for), but it also affects the SP's cost structure: higher demand for the SP increases the SP's costs. Another way to put it is that under free-riding, both agents prefer for the work to be executed by the other one even when they can achieve the same results at a lower cost themselves. A consequence of such behavior (as demonstrated in Prop. 4) is that any decisions induced by group contracts will be cost-inefficient ones (and therefore not achieve the first-best).

Proposition 5 (Properties of optimal mindful group contracts)

1. *While outcomes-adjusted capitation contracts are optimal both in this model and the naïve one, those contracts are* not *identical as they can have different per-capita reimbursement rates.*
2. *The optimal contract (one that solves the MGCP) does* not *achieve first-best.*
3. *If an agent has strictly lower equilibrium costs than the other, he or she extracts positive rent.*

A notable similarity between optimal contracts in this setting and ones in the naïve model is that the same contract type (linear outcomes-adjusted capitation contract) is guaranteed to be optimal. There is also a notable difference. Despite being the same contract type, these contracts can have large differences in reimbursement rates. In simulations of Section 6, we see that optimal group contracts under free-riding typically reward improved health at a rate (r_{AC} or r_{AP} in Prop. 2) which is roughly twice as high as the naïve ones. This poses a challenge for a principal who is worried that

free-riding might be an issue, but is not sure how widespread it is. Giving contracts that assume all will free-ride will also result in distorted incentives if this behavior is not ubiquitous.

The strictly positive rent property (Prop. 5 part 3) adds another layer of additional costs to the principal, in addition to the inherent cost inefficiency.

5.2. Mindful Individual Contracts

In this section, the idea is to a) endogenize collusion, so that the agents collude if and only if it is profitable for both of them, and b) have the principal design individual contracts with the idea of addressing potential collusion in mind. Recall that when colluding, a GP under contract $v_1(\mathcal{Q}_1)$ and an SP under contract $v_2(\mathcal{Q}_2)$ will choose decisions $\tilde{d}_G^C, \tilde{d}_S^C$ which solve the collusion problem (23)-(24), and if not colluding, they will choose $\tilde{d}_G^{NC}, \tilde{d}_S^{NC}$, which solve the NICP incentive compatibility constraints (9)-(10). Denote by π_G^C and π_S^C expected profit of the two colluding agents, i.e.,

$$\pi_G^C = \mathbb{E}\Big[\frac{1}{2}\sum_{i\in\{1,2\}}\big(v_i(\mathcal{Q}_i(\tilde{d}_G^C,\tilde{d}_S^C)) - v_i(\mathcal{Q}_i(\tilde{d}_G^{NC},\tilde{d}_S^{NC}))\big) + v_1(\mathcal{Q}_1(\tilde{d}_G^{NC},\tilde{d}_S^{NC})) - k_G(\tilde{d}_G^C)\Big],$$

$$\pi_S^C = \mathbb{E}\Big[\frac{1}{2}\sum_{i\in\{1,2\}}\big(v_i(\mathcal{Q}_i(\tilde{d}_G^C,\tilde{d}_S^C)) - v_i(\mathcal{Q}_i(\tilde{d}_G^{NC},\tilde{d}_S^{NC}))\big) + v_2(\mathcal{Q}_2(\tilde{d}_G^{NC},\tilde{d}_S^{NC})) - k_{SF}(\tilde{d}_S^C) - \mathcal{S}(\tilde{d}_G^C)k_{SV}\Big].$$

Analogously, we can denote the profit of the two agents when not colluding with π_1^{NC} and π_2^{NC}. We also use the logic notation $\wedge$ and $\vee$ to denote "and" and "or," respectively. With this notation, the main *(mindful) individual contracting problem* (or MICP) is given by:

$$\max_{\substack{\mathcal{Q}_1,\mathcal{Q}_2\in\{\mathcal{Q}_P,\mathcal{Q}_G,\mathcal{Q}_S\}\\ v_1(\mathcal{Q}_1),v_2(\mathcal{Q}_2)}} \quad \mathbb{E}\left[Q_P(\tilde{d}_G,\tilde{d}_S) - v_1(\mathcal{Q}_1(\tilde{d}_G,\tilde{d}_S)) - v_2(\mathcal{Q}_2(\tilde{d}_G,\tilde{d}_S))\right] \tag{30}$$

$$\text{subject to} \quad \tilde{d}_G^{NC} \in \arg\max_{d_G\in[0,\infty)} \mathbb{E}\left[v_1(\mathcal{Q}_1(d_G,\tilde{d}_S^{NC})) - k_G(d_G)\right] \tag{31}$$

$$\tilde{d}_S^{NC} \in \arg\max_{d_S\in[0,\infty)} \mathbb{E}\left[v_2(\mathcal{Q}_2(\tilde{d}_G^{NC},d_S)) - k_{SF}(d_S) - \mathcal{S}(\tilde{d}_G^{NC})k_{SV}\right] \tag{32}$$

$$\tilde{d}_G^C \in \arg\max_{d_G\in[0,\infty)} \mathbb{E}\Big[\frac{1}{2}\sum_{i\in\{1,2\}}\Big(v_i(\mathcal{Q}_i(d_G,\tilde{d}_S^C)) - v_i(\mathcal{Q}_i(\tilde{d}_G^{NC},\tilde{d}_S^{NC}))\Big) + v_1(\mathcal{Q}_1(\tilde{d}_G^{NC},\tilde{d}_S^{NC})) - k_G(d_G)\Big] \tag{33}$$

$$\tilde{d}_S^C \in \arg\max_{d_S\in[0,\infty)} \mathbb{E}\Big[\frac{1}{2}\sum_{i\in\{1,2\}}\Big(v_i(\mathcal{Q}_i(\tilde{d}_G^C,d_S)) - v_i(\mathcal{Q}_i(\tilde{d}_G^{NC},\tilde{d}_S^{NC}))\Big) + v_2(\mathcal{Q}_2(\tilde{d}_G^{NC},\tilde{d}_S^{NC})) - k_{SF}(d_S) - \mathcal{S}(\tilde{d}_G^C)k_{SV}\Big] \tag{34}$$

$$(\tilde{d}_G,\tilde{d}_S) = \begin{cases} (\tilde{d}_G^C,\tilde{d}_S^C) & \mid \pi_i^C > \pi_i^{NC}, \forall i\in\{G,S\} \\ (\tilde{d}_G^{NC},\tilde{d}_S^{NC}) & \mid \text{otherwise} \end{cases} \tag{35}$$

$$((\tilde{d}_G,\tilde{d}_S) = (\tilde{d}_G^C,\tilde{d}_S^C) \wedge V_G \le \pi_G^C \wedge V_S \le \pi_S^C) \vee (V_G \le \pi_G^{NC} \wedge V_S \le \pi_S^{NC}) \tag{36}$$

Here, (30) is the same objective function as in the other models. (31)-(32) gives the decisions that the agents will make if they do not collude, which are the same as incentive compatibility constraints (9)-(10) in the naïve model. (33)-(34) give the decisions that the agents will make if

they do decide to collude, which correspond to (23)-(24) in the collusion model of Section 4. The new constraint in (35) states that agents will collude only if it is strictly profitable for them to do so. We refer to contracts v_1 and v_2 as collusion-proof if under them $(\tilde{d}_G, \tilde{d}_S) = (\tilde{d}_G^{NC}, \tilde{d}_S^{NC})$, i.e., it is not profitable for the agents to collude. Lastly, (36) is the new participation constraint, which states that whatever decisions the agents end up making (collusion or no collusion), the resulting profit has to be equal to or better than the outside option.

Theorem 4 (Mindful individual contracts for the MICP)

1. *The health outcome of optimal group contracts can be replicated at a lower cost by using two individual linear outcomes-adjusted capitation contracts.*
2. *Amongst the class of contracts using $\mathcal{Q}_P$ as the signal, linear outcomes-adjusted capitation contracts are optimal, and they are collusion-proof.*
3. *If the naïve individual contracts (solutions of the NICP in Section 3) are collusion-proof, they are also optimal and achieve first-best in this model.*

Part 1 of the theorem creates a strong argument against the use of group contracts, because anything they can accomplish can also be accomplished using individual contracts at a lower cost. For the intuition behind this result, consider any group contract and the following process of obtaining a better set of individual contracts. First, find an equivalent outcomes-adjusted capitation group contract, as in Theorem 3. This capitation contract can be "split in half," giving each provider one-half of the group contract's reimbursement rate for health benefits created. At this point, we have two individual contracts that jointly induce the same outcome at the same cost as the group contract we wanted to replicate. However, individual contracts can go one step further, as they have an additional degree of freedom: the fixed capitation rates can be individually adjusted for each provider, thereby removing the rent issue of group contracts.

Part 2 of the theorem is an argument for the use of individual linear outcomes-adjusted capitation contracts, as they are optimal within the class of all contracts using the population health measure $\mathcal{Q}_P$ as the signal. There remains, however, a possibility that collusion-inducing contracts that use different signals for different providers could perform better.

Part 3 of the theorem considers the most attractive scenario, where the contracts which solve the naïve problem are also collusion-proof. If this is true, it is very convenient for the principal, as it is then sufficient to solve the comparatively simple naïve problem (NICP), which has closed-form solutions available. It also resolves the dilemma of a principal who is unsure how often do free-riding and collusion arise, as in this scenario, the same contract (with identical parameters), is optimal irrespective of whether collusion and free-riding are present or not.

One key question we will address in the numerical analysis in Section 6 is: how often can we expect this attractive scenario to arise in practice? We will demonstrate that this scenario occurs for almost all realistic parameter values, but *only* if using outcomes-adjusted capitation contracts.

6. Model Calibration and Numerical Analysis

In this section, we numerically explore outcomes-based contracts and the consequences of naïvité (i.e., using the contracts of Section 3 rather than of Section 5), when the models are calibrated to data on type 2 diabetes treatment in the UK. We focus on the specialist being an ophthalmologist who treats diabetic retinopathy. Diabetic retinopathy is one of the most common complications of diabetes (30% prevalence amongst 2 diabetes patients, Mathur et al. 2017) and the most common cause of blindness in the working-age population. Our analysis is illustrative over a 'reasonable range' of parameters, and is not intended for detailed policy recommendations.

The goal of this analysis is three-fold. First, to identify the *magnitude* of deviations that result from naïvité, as the structural results of Section 4 are only directional. Second, to explore outcomes in situations where definitive structural results were unavailable in Section 4 – e.g., how per-patient contracts affect health. Third, to identify the sensitivity of our conclusions to parameter values, as some results do not hold for all parameter values (notably Theorem 4, part 3).

6.1. Data Modeling

An important characteristic of type 2 diabetes treatment is that the model parameters vary widely for different settings. E.g., the number of patients per GP (and thus access to care) varies depending on the geographical region (UK NHS 2016), and adherence rates have high variability over different geographies and different population segments (Taddeo et al. 2008, Borus and Laffel 2010).

To account for these differences, we focus not only on point estimates of parameters but also a distribution of parameters on a *plausible* range. Tables 1–2 report the considered range of parameters, sources used to estimate them, and comments about the assumed distribution for sampling parameters. Unless otherwise specified, parameters were sampled uniformly over the considered range.

Some of the model's parameters are functions (the costs k_G and k_{SF}), which we cannot observe in data directly, but we can observe their realizations. When considering those parameters, we rely on *acceptance sampling* to ensure their values are realistic. That is, we start by considering a wide range of possible values for those parameters. After drawing each parameter set, we test whether the parameter values are plausible along three dimensions: (a) The cost realizations under first-best decisions need to be within $\frac{1}{4}$x to 4x of the observed costs reported in the literature (Mitchell et al. 2012, Diabetes UK 2014), (b) the first-best complication rate cannot be higher than the complication rates observed in practice (Khalid et al. 2014), (c) the first-best waiting times need

to be between one day and one year (literature estimates for waiting times range from a few days to almost a year in different settings). Parameter sets that did not pass these tests were dropped (6157 parameter sets of 20,000 generated were plausible). Appendix B provides more details.

For each plausible parameter set, optimal naïve contracts were calculated as in Section 3, as was their performance under collusion and free-riding (as in Section 4), and an attempt was made to identify the optimal contract in the setting of our main model (Section 5). The results of the analysis presented use a monetary value of a QALY equal to $£30,000$. Conclusions based on other considered values of the monetary values of QALY are similar to the conclusions drawn below.

Table 1 Parameters Related to Initial Population Health and Complications

Parameter	Range Considered	Estimation Methods and Sources
Population Size	Empirical Distribution	Full distribution of patients per full-time equivalent GP from the NHS census of GPs (UK NHS 2019). We sample the number of patients per GP (m) from this distribution (truncating both 1% tails), then sample $n \sim \text{Bin}(m, 6.4\%)$, where 6.4% is the incidence rate of type 2 diabetes (Diabetes UK 2019).
Initial Avg. Health of Population	$\bar{q}_0^P = 7.61\ QALY$	Method of Cutler and Richardson (1998), life expectancy data (Leal et al. 2009, Khalid et al. 2014), quality of life estimates (Clarke et al. 2002).
Monetary Value of One QALY	QALY$\in [£20k, £100k]$.	NHS uses $£20k$ to $£30k$ (Claxton et al. 2013), US regulators use \$50k to \$100k (Neumann et al. 2014). Some estimates in health economics are even higher (Neumann et al. 2014). We do not randomly sample this value, as it is subject to the regulator's choice. Rather, we conduct the analysis for four different values: $£20k$, $£30k$, $£70k$, and $£100k$.
Adherence Rate	$\phi \in [0.1, 1]$	Currie et al. (2012) find that 61% of patients are adherent. Sub-populations can have almost full adherence (hospitalized), or low (.1) adherence (adolescents) (Taddeo et al. 2008, Borus and Laffel 2010). We use a triangular distribution with support $[0.1, 1]$ and mode 0.61.
Non-adherent Patient Complication Rate	$\overline{\lambda} \in [3.54\%, 5.08\%]$	Base incidence of diabetic retinopathy (3.22%, Mathur et al. 2017), adjusted for non-adherence (García-Pérez et al. 2013, Currie et al. 2012).
Health Deterioration Due to Untreated Complications	$b \in [0.106, 0.2]$	Based on untreated retinopathy progressing to legal blindness in 3.2 years (Ferris 1993), and the range of estimates for QALY effects of vision loss (Javitt and Aiello 1996, Rein et al. 2007).
Health Cost of a Complication	$a = b/2$	Based on early-stage retinopathy being asymptomatic, but the expected time to a diagnosis being 6 months after onset.
Health Improvement from SP Treatment	$\zeta = 0$	The vision loss from retinopathy is irreversible, but successful treatment can stop further deterioration of vision (NIH 2019).

Table 2 Cost Functions

Parameter	Functional form	Estimation Methods and Sources
GP's Cost Function	$k_G(d_G) = \gamma_1 (d_G)^{\gamma_2}$	We consider a wide range for the hyper-parameters, specifically $\gamma_1 \in [1, 10^6]$ and $\gamma_2 \in [1, 4]$, which we filter for plausible parameters using acceptance sampling and realized cost data of Diabetes UK (2014).
SP's Fixed Costs	$k_{SF}(d_S) = \delta_1 (d_S)^{\delta_2}$	We consider a wide range for the hyper-parameters, specifically $\delta_1 \in [1, 10^6]$ and $\delta_2 \in [1, 4]$, which we filter for plausible parameters using acceptance sampling and realized cost data of Mitchell et al. (2012).
SP's Variable Costs	$k_{SV} \in [£823, £9356]$	Variable cost range of Mitchell et al. (2012).

Table 3 Mean of deviation from first-best (with standard error) in a numerical analysis of naïve contracts

	Individual Contract (Collusion)				Group Contract (Free-riding)
Metric	C,C	P,C	C,P	P,P	C or P
Total Population Health (QALY)	0 (0)	0 (0)	-1.71 (0.02)	-1.69 (0.02)	-0.46 (0.00)
Government Expenditure (k£)	0 (0)	0 (0)	71.03 (1.28)	67.00 (1.24)	-13.89 (0.11)
Provider Profit (k£)	0 (0)	0 (0)	78.25 (1.31)	74.71 (1.27)	-4.54 (0.04)
System Efficiency (k£)	0 (0)	0 (0)	-44.10 (0.50)	-43.18 (0.49)	-4.54 (0.04)

6.2. Numerical Comparison of Contracts

This section summarizes numerical results on the consequences of naïvité, and reports mean values and standard errors of the main performance metrics under all naïve contracts, as derived in Section 3, when those contracts are held by agents of Section 5 (agents which *can* free ride or collude). These metrics include total population health (summed over all patients, measured in QALYs), governmental expenditure, provider profit, and system efficiency. Estimates of expected values represent the difference in outcome for the given contract scenario from the first-best outcome.

The first and most salient result that can be observed in Table 3 is that there is no deviation from the first-best whenever the specialist holds a naïve outcomes-adjusted capitation contract. This can be observed by the 0s in the columns for individual contracts where the second entry is a 'C' as compared to a 'P.' This is driven by such contracts being collusion-proof for 100% of plausible parameter sets; thus, they are also solutions to the MICP (Theorem 4). [If we also consider non-plausible parameter sets, dropped as part of acceptance sampling, collusion can possibly be stable under outcomes-adjusted capitation contracts (it occurred in 1.3% of cases, data not shown[7]).] Even if we focus solely on the (unrealistic) parameter sets which lead to collusion under capitation contracts, the deviation from the first-best under such contracts is drastically smaller than in any other scenario considered. The average impact on Total Population Health in the parameter sets which lead to collusion when both providers hold naïve outcomes-adjusted capitation contracts is -0.06 QALY, while the increase in Government Expenditure is £485.

The second result which can be observed from the table is that per-patient SP contracts severely under-perform relative to all other contracts we consider, no matter which of the metrics we use as the criteria. This can be observed by the numbers in the columns for individual contracts, where the second entry is a 'P' rather than a 'C.' Naïve per-patient contracts were collusion-inducing in 83.3% of cases (data not shown). If collusion happens under such contracts, the negative effects it causes on both health and finances are drastically worse than in any other scenario considered.

[7] Such cases required a particular set of circumstances: the GP faced very high costs of improving the complication rate, to the point of not putting any effort into the quality of care under the first-best ($d_G^* \approx 0$) *and* the SP could improve the health of patients at a much lower cost. Thus, the collusion worked by the SP increasing the quality of care (but not the GP), with the resulting increase in GP compensation effectively being shared by the providers.

This is not particularly surprising if we consider how collusion manifests under such contracts: the income of colluding agents scales with the number of patients who develop complications, creating an incentive for a colluding GP to *reduce* the quality of care, which in turn decreases patient health and increases provider profit and government spending.

Group contracts were shown in Theorem 1 to be inefficient, which is primarily caused by the providers under them making decisions that are not cost-efficient (see Prop. 4). The column for Group Contract in Table 3 shows that there is inefficiency, but that the inefficiency is not as bad, on average, as compared to the inefficiency due to the SP holding a per-patient individual contract (for example, -4.54 is a better system inefficiency than -44.01 or -43.11). The result is the same whether the group contract is a C or a P. Thus, if a per-patient contract is required for some reason for the SP, then it is better to use a group per-patient contract.

Thus, combining this insight with the analytical results of Section 5, all results point in the direction of individual outcomes-adjusted capitation contracts as the best contract type. They outperform all group contracts (Theorem 4, part 1), and can be expected to be collusion-proof in the vast majority of cases, in which case they also solve the MICP (Theorem 4, part 3), which implies they cannot be outperformed by any other contract. Furthermore, even if the rare situation arises when collusion can exist under such contracts, its effect is small.

7. Implications for practice

The contracts that emerge as optimal in our model thus far are close variations of ones already used in practice, but they do not perfectly correspond to any extant medical system. By looking at properties of entire contract classes (which *do* include the contracts currently in place) rather than only optimal contracts, we can identify strengths, weaknesses, and potential focus areas for the future development of those systems. Section 7.1 examines a setting where outcomes-adjustment is made for the GP and not the SP, motivated by the UK NHS's QOF. Section 7.2 analyzes the case of volume-scaling contracts that are volume adjusted, particularly for an SP, motivated by the US.

7.1. The NHS and Other Systems with Outcomes-adjustment Only for GPs

The UK NHS's QOF is the world's largest outcomes-based reimbursement scheme. It defines outcomes-based bonuses for GPs that comprise a considerable part of GP compensation (UK NHS 2020). Let $\mathscr{G}$ be the class of all population-health-based contracts for the GP that satisfy Assumption 1. This class allows for a wide variety of outcomes-based contracts for the GP, does not make any linearity assumptions, and includes all of the capitation contracts we study in the paper.

SPs (consultants) within the NHS have a fixed salary which increases in seniority. Bonuses they receive are *not* outcomes-adjusted, but are rewards for additional duties such as teaching or hospital management.[8] Let $\mathscr{F}$ be the class of all fixed-compensation contracts for the SP.

[8] See also https://www.healthcareers.nhs.uk/explore-roles/doctors/pay-doctors, Accessed Nov 25, 2022.

We consider the usual individual contracting problems studied above in both their naïve and mindful forms, but now restrict the choice of contracts for the GP and SP to $\mathscr{G}$ and $\mathscr{F}$, respectively.

Theorem 5 (Outcomes-adjustment only for the GP) *Consider the NICP in (8)-(12) where the choice of* (v_1, v_2) *is restricted to* $\mathscr{G} \times \mathscr{F}$*. Then,*

(i) *A linear optimal (second-best) outcomes-adjusted capitation contract for the GP exists.*[9]

Consider the MICP in (30)-(36) where the choice of (v_1, v_2) *is restricted to* $\mathscr{G} \times \mathscr{F}$*. Then,*

(ii) *The performance, in terms of the principal's objective function, of any collusion-proof* $(v_1, v_2) \in \mathscr{G} \times \mathscr{F}$ *is bounded from the above by the performance of the optimal NICP contract in* (i)*.*

(iii) *The performance of a collusion-inducing* $(v_1, v_2) \in \mathscr{G} \times \mathscr{F}$ *is bounded from the above by the solution of a modified MICP with no* $\mathscr{G} \times \mathscr{F}$ *restriction, and where the providers* have to *collude.*

(iv) *Any collusion-inducing* $(v_1, v_2) \in \mathscr{G} \times \mathscr{F}$ *can only induce decisions* (d_G, d_S) *that can also be induced by group contracts and have been characterized as inefficient by Prop. 4.*

(v) *For any decision pair* (d_G, d_S) *that is inducible by collusion-inducing contracts in* $\mathscr{G} \times \mathscr{F}$*, the participation constraint (36) can be made to bind for both providers.*

Part (i) of Theorem 5 solves the NICP. This auxiliary result supports parts (ii) and (iii), which identify the performance bounds of contracts in $\mathscr{G} \times \mathscr{F}$, where the optimal NICP contracts form an upper bound for the collusion-proof contracts.

Parts (iv) and (v) of Theorem 5 speak of the connection between group contracts of the MGCP and the collusion-inducing contracts in the MICP. Those two contract types share the issue of only being able to induce actions along an inefficient frontier. However, collusion-inducing contracts have one advantage over group contracts: the principal can alter each individual's outside option (what happens if they *do not* collude), which cannot be done with group contracts. This gives the principal an added degree of freedom and allows the principal to prevent the rent that the more cost-efficient agent is able to extract in group contracts (recall Prop. 5, part 3).

Table 4 extends the experiments in Section 6 to explore the results in Theorem 5. We see that the bounds of collusion-proof contracts can, in all studied cases, be attained by linear contracts, and that those contracts exhibit performance that is remarkably close to the first-best (notice the gap to the first-best is smaller than all cases with deviations reported in Table 3).

Collusion-inducing contracts can perform even better. To see why, consider the form that collusion takes when providers hold contracts in $\mathscr{G} \times \mathscr{F}$, as they do in the UK. Here, the GP, whose compensation is outcomes-adjusted, may share some reward with the SP, in return for which the SP helps improve outcomes. Collusion may even be a misnomer here, it is a case of the system not

[9] A second-best contract is a contract that is optimal in the sense that there does not exist any other contract which can outperform it, yet it does not eliminate all agency frictions like a first-best contract would.

incentivizing cooperation between providers, so they (inefficiently) create those incentives themselves. The inefficient attempts at coordination can be better than no coordination at all.

The bottom line is that it appears that systems like the UK, where outcomes adjustment is for GPs and not for SPs, need not worry about collusion. It is not hard to make such a system collusion-proof, but even if collusion occurs, it will take a benign form. Perhaps a more relevant question for these systems is "can we reap the full benefit of outcomes-based contracting without incentives for the specialists?" When looking at the diabetes-calibrated model, the gap between the true optimum and results achievable by a system with fixed SP compensation is remarkably small.

7.2. The USA and Other Systems with Volume-scaling Contracts

In Sections 4 and 6, the property that drives the undesirable performance of individual outcomes-adjusted per-patient contracts is that SP compensation scales with the volume of patients treated. This volume-scaling property is hardly unique to reimbursement schemes studied in this paper. *All* fee-for-service and per-patient contracts, outcomes-adjusted or not, have this property. Volume-scaling compensation is prevalent in the US, and is also found in Canada and Australia. As of 2022, 93.3% of US specialists have compensation that is volume-scaling, and the volume-scaling component makes up the majority (73.7%) of their income (Reid et al. 2022). Medicare's Merit-based Incentive Payment System (MIPS) introduced in 2017 is one the largest outcomes-adjusted systems in the US; yet, under it, the provider compensation is still volume-scaling (Rathi and McWilliams 2019, Bond et al. 2022).

If compensation scales with demand, one provider can conceivably generate demand for another, and they can share the spoils. Demand-inducing collusion happens: the U.S. Department of Justice recovered more than $153 million in settlements and judgments for violations of the Stark Law and Federal Anti-Kickback Statute (US DoJ 2021).

We now explore the vulnerability to collusion via demand inducement under volume-scaling contracts, particularly for specialists. A contract is volume-scaling if, keeping the average health of patients constant, the variable amount it pays is proportional to the number of patients treated.

Table 4 **Mean of deviation from first-best (with standard error) in a numerical analysis of the performance bounds for contracts in $\mathcal{G} \times \mathcal{F}$ established in Theorem 5.**

Metric	Collusion-Proof Contracts	Collusion-Inducing Contracts
Total Population Health (QALY)	-0.14 (0.01)	0.00 (0.00)
Government Expenditure (k£)	-0.37 (0.04)	0.01 (0.00)
Provider Profit (k£)	0 (0)	0 (0)
System Efficiency (k£)	-3.94 (0.21)	-0.01 (0.00)
Bound attainable by linear contracts (% cases)	100%	83%

Let $\mathcal{V}$ be the set of volume-scaling contracts, with $v(\boldsymbol{q}_S, s) \in \mathcal{V}$ if and only if $v(\boldsymbol{q}_S, s)$ satisfies Assumption 1 and there exist constants $h \in \mathbb{R}, r \in \mathbb{R}_+$ and a weakly increasing non-zero function $p(\bar{\boldsymbol{q}}_S)$ such that $v(\boldsymbol{q}_S, s) = h + rsp(\bar{\boldsymbol{q}}_S)$ (s is the number of patients treated by the SP).

It is useful to define the following equivalence relation: $v_1 \sim v_2$ if there exists a function $p(\bar{\boldsymbol{q}}_S)$ and constants $h_1, h_2 \in \mathbb{R}, r_1, r_2 \in \mathbb{R}_{++}$ such that $v_1(\boldsymbol{q}_S, s) = h_1 + sr_1 p(\bar{\boldsymbol{q}}_S)$ and $v_2(\boldsymbol{q}_S, s) = h_2 + sr_2 p(\bar{\boldsymbol{q}}_S)$. The relation $\sim$ partitions $\mathcal{V}$ into equivalence classes. Each member of an equivalence class shares a common way in which reimbursement is outcomes adjusted ($p(\bar{\boldsymbol{q}}_S)$), but members may differ in the scale of that adjustment (r) and the constant payment (h).

Two of these equivalence classes are of special interest to us. Let $\mathcal{V}_L = \{v \in \mathcal{V} \,|\, v \sim h + sr\bar{\boldsymbol{q}}_S\}$; this is the class of all volume-scaling contracts in which outcomes adjustment is linear in health ($p(\bar{\boldsymbol{q}}_S) = \bar{\boldsymbol{q}}_S$). The optimal naïve per-patient contracts studied earlier in the paper belong to this class. Similarly, let $\mathcal{V}_C = \{v \in \mathcal{V} \,|\, v \sim h + sr\}$; this is the class of all volume-scaling contracts which are *not* outcomes adjusted ($p(\bar{\boldsymbol{q}}_S)$ is a constant).

Theorem 6 (Collusion under volume-scaling SP contracts) *Consider the MICP where the SP contract is chosen from $\mathcal{V}$ and let the GP hold $v_1 \in \mathcal{G}$ such that $0 < v_1'(\boldsymbol{q}_S) \leq \Omega$, where $\Omega \in \mathbb{R}$.*

(i) *There exists $r^* \in \mathbb{R}$ such that for every $v_2 = h + sr\bar{\boldsymbol{q}}_S \in \mathcal{V}_C$, (v_1, v_2) is collusion-inducing and $d_C^G = 0$ if $r \geq r^*$.*

(ii) *If $\mathbb{E}[\mathcal{S}(d_G)\bar{\mathcal{Q}}_S(d_G, d_S)]$ is decreasing in d_G, then for every $v_2 = h + srp(\bar{\boldsymbol{q}}_S) \in \mathcal{V}_L$, (v_1, v_2) is collusion-inducing and $d_G^C = 0$ if $\mathbb{E}Q_S(\tilde{d}_G^C, \tilde{d}_S^C) - \mathbb{E}Q_S(\tilde{d}_G^{NC}, \tilde{d}_S^{NC}) > 0$ and r is sufficiently high.*

(iii) *If $(v_1, v_2) \in \mathcal{G} \times \mathcal{V}$ are collusion-proof and induce $(\tilde{d}_G^*, \tilde{d}_S^*)$, there exists $v_2^* \in \mathcal{V}_L$ (unique up to the constant h) such that under (v_1, v_2^*) we have $(\tilde{d}_G^{NC}, \tilde{d}_S^{NC}) = (\tilde{d}_G^*, \tilde{d}_S^*)$. (v_1, v_2^*) need not be collusion-proof.*

Part (i) addresses the well-known vulnerability to collusion that non-outcomes-adjusted but volume-scaling SP contracts possess. This vulnerability depends not on how GP compensation is outcomes-adjusted. If the volume-scaling component of the SP is large, it will induce collusion.

Part (ii) gives conditions under which collusion occurs under all linear volume-scaling contracts. The dependency on the size of the incentive exists here, as in (i), under two additional conditions. First, $\mathbb{E}[\mathcal{S}(d_G)\bar{\mathcal{Q}}_S(d_G, d_S)]$ needs to be decreasing in d_G (meaning the SP needs to be able to benefit from increased demand). This is universally true if there is no outcomes-adjustment as more volume leads to more income. With linear outcomes-adjustment, this condition is independent of which contracts are held, and b being small enough (the condition progressing relatively slowly over time) is sufficient for it to hold. There is also an opposite effect if the SP compensation is outcomes-adjusted: the more patients the SP treats, the more expensive it is to maintain the same level of average outcomes, or conversely, with the same level of care at the SP's level, outcomes will be

worse if the number of patients increases (which will be reflected in the income). Second, colluding introduces inefficiency into the system due to the free-riding effect, so the benefits from colluding need to outweigh free-riding. This is captured by the $\mathbb{E}Q_S(\tilde{d}_G^C, \tilde{d}_S^C) - \mathbb{E}Q_S(\tilde{d}_G^{NC}, \tilde{d}_S^{NC}) > 0$ condition.

Part (iii) explores whether we can do better if we resort to non-linear contracts. In Propositions 1-2 and Theorems 3-4, we saw that the restriction to linear contracts can be made without loss of optimality in many settings. The same is not the case here. While linear contracts can replicate what other contracts are doing in the absence of collusion, non-linearity offers an additional tool to disincentivize collusion by disproportionally penalizing the bad outcomes.

The results on systems with volume-scaling specialist stand in contrast with those of Section 7.1. Here, a very different form of collusion arises: the SP, whose compensation benefits from increased demand, shares some compensation with the GP, who lowers the quality of care. This generates additional demand for the SP. This form of collusion is *not* benign and typically results in increased costs and worse outcomes. This is a known issue with volume-scaling contracts without outcomes-adjustment (which is regulated through laws rather than incentives). The same issue appears to plague many outcomes-adjusted volume-scaling contracts as well.

For an example of how outcomes-adjustment under volume-scaling contracts can possibly result in worse outcomes than its absence, notice that even volume-scaling non-outcomes-adjusted contracts (those in $\mathscr{V}_C$) can approximate the performance of contracts in $\mathscr{F}$ (as r approaches zero, contracts in $\mathscr{V}$ converge to elements of $\mathscr{F}$). Thus, the performance seen of such contracts in the left columns of Table 4, can be approximated even by conventional volume-scaling contracts by making r very small and using larger outcomes adjustment purely on the GP level. Yet, Section 6 gives an example of how collusion can make a well-intended volume-scaling outcomes-adjusted contract perform much worse (see the third and fourth column of Table 3). This occurs because maximizing the benefit of outcomes-adjustment in the absence of collusion required making the incentives relatively large, but large incentives lead to more vulnerability to collusion.

8. Implementation challenges and yardstick compatibility

The majority of our analysis above points towards reimbursing providers individually with outcomes-adjusted capitation contracts. While knowing which contract type to use has value by itself, implementing the contract in practice also requires optimally setting the contract's parameters. Section 6 and Appendix B demonstrate that most of the required inputs can be retrieved from publicly available data. Notable exceptions are the shapes of the cost functions k_G and k_{SF}. These functions are required to compute the first-best decisions d_G^* and d_S^* and expected rewards.

One way to deal with that challenge is to estimate these functions, as in Section 6. But estimation may degrade the performance of the contracts due to estimation errors. In Appendix C.6, we

conduct a sensitivity analysis of how the performance of our contracts will be impacted by systematic misestimation of the cost functions, and show that the consequences will primarily manifest as a wealth transfer between the principal and the providers, while having little effect on health.

We can avoid these errors and estimation altogether by relying on data that we *do* have. What is typically accessible in the data are ex-post *cost realizations*. Denote the cost realizations by $\varrho_G \doteq k_G(d_G), \varrho_{SF} \doteq k_{SF}(d_S), \varrho_{SV} \doteq s k_{SV}$. Thus we can work under the alternative assumption, which is also common in the Yardstick literature (Shleifer 1985, Savva et al. 2019, Arifoglu et al. 2021):

Assumption 2 *The principal does not know k_G, k_{SF}, k_{SV} (thus also d_G^* and d_S^*) but can observe and contract on the realized costs.*

Most of our results on outcomes-adjusted capitation contracts are replicable under this assumption.

Proposition 6 (Contracts using realized costs) *Let Assumption 2 hold and denote by v_G (v_S) an outcomes-adjusted capitation contract for the GP (SP) which pays a per-capita fee c_G (c_S) for every patient in the population (treated or not), where*

$$c_G = \bar{\boldsymbol{q}}_P + f_G - \varrho_{SV}/n \quad \text{and} \quad c_S = \bar{\boldsymbol{q}}_P + f_S,$$

where f_G and f_S are any constants that are sufficiently high not to violate the participation constraints (11)-(12). Then, these contracts induce first-best decisions in the naïve problem (NICP) and, if they are collusion-proof, in the main problem (MICP).

Inducing the desired decisions in the providers is not an issue in this framework. On the contrary, the alignment of incentives between the providers can be more easily accomplished if the realized costs can be contracted on, as it allows providers to be directly held accountable for the costs that they are generating. However, the main issue above, and the missing piece needed to attain the first-best, is that lacking knowledge of the cost functions, the principal is not able to ensure that the participation constraint is binding (or even not violated).

Realistically, finding values of the fixed pay components (f_G and f_S) that are sufficiently high so that the providers participate should should be easy to identify through negotiation with providers. That process can ensure that participation constraints are not violated, but they are unlikely to be binding and may result in positive rent. Thankfully, identifying binding values for these constants also appears to be resolvable. Prop. 7 identifies two ways to do so.

Proposition 7 *The two results below hold for variants of the NICP always, but also for the MICP if the contracts given in Prop. 6 are collusion-proof.*

1. ***Repeated contracting.*** *Let there be $N \geq 2$ periods, such that an identical contracting game is repeated every period, and there is a common discount factor $1 > \delta > 0$. Consider the principal who offers the contracts v_G and v_S as given by Prop. 6 in every period, but adjusts the fixed pay components f_G and f_S based on the results of the previous period as follows (the period i is denoted in the superscript)*

$$f_G^i = (V_G + \varrho_G^{i-1} + \varrho_{SV}^{i-1})/n - \bar{\boldsymbol{q}}_P^{i-1}, \quad f_S^i = (V_S + \varrho_{SF}^{i-1} + \varrho_{SV}^{i-1})/n - \bar{\boldsymbol{q}}_P^{i-1}.$$

   Then, the providers choosing the first-best decisions in every period will be the sole subgame perfect equilibrium, and the participation constraint will be binding (thus, the first-best will be attained) in every period except *the first.*

2. ***Yardstick competition.*** *Let there be $M \geq 2$ identical GP-SP pairs, each with a separate population of patients. Denote by $\hat{\varrho}_G^{-j}$ the average cost realization of all GPs but the one in the jth GP-SP pair. Define $\hat{\varrho}_{SV}^{-j}, \hat{\varrho}_{SF}^{-j}$, and $\hat{\bar{\boldsymbol{q}}}_P^{-j}$ analogously. Then the principal can attain the first-best by giving all providers a variant of the contracts v_G and v_S as given by Prop. 6, such that for every $j \in \{1, ..., M\}$, f_G and f_S depend on the average performance of other providers:*

$$f_G^j = (V_G + \hat{\varrho}_G^{-j} + \hat{\varrho}_{SV}^{-j})/n - \hat{\bar{\boldsymbol{q}}}_P^{-j}, \quad f_S^j = (V_S + \hat{\varrho}_{SF}^{-j} + \hat{\varrho}_{SV}^{-j})/n - \hat{\bar{\boldsymbol{q}}}_P^{-j}.$$

One way to handle repeated contracting, as in the first part of this proposition, is by making the constants f_G and f_S in every period such that the participation constraint will be binding if the providers once again experience costs equal to the historic ones. There exists an opportunity to game this system: a provider could choose inefficiently high treatment quality, which will increase their profit in the following period as they will be reimbursed based on the previous period's costs. Yet, this opportunity is never worth it (as evident from the equilibrium): this future gain is insufficiently high to offset the profit loss due to inefficiently high costs in the current period.

Another way to accomplish the same is using yardstick competition within the same contracts. Rearranging the terms in these contracts above shows that they effectively consist of cost reimbursement (based on the cost of other similar providers, not the provider's own cost) and a variable component, where the provider j is rewarded at per-capita rate $\bar{\boldsymbol{q}}_P - \hat{\bar{\boldsymbol{q}}}_P^{-j}$ for the health benefits created *in excess* of what other providers are creating.

9. Concluding Remarks

We studied two core questions related to the choice of reimbursement policy. Should we be worried about collusion in outcomes-adjusted contracts? If so, can we do something about it by our choice of the type of reimbursement contract (capitation or per patient), structure (contract with providers

individually or with a group of providers), or health outcome data granularity (population or per-provider)? To answer these questions, we proposed a parsimonious model of a care pathway to derive both general results and results relevant to systems like that found in the UK, for which we also provide illustrative numerical results, and in the USA, where these questions are particularly relevant due to recent exemptions to anti-kickback laws for outcomes-based contracts.

Our analysis suggests that there is reason to worry about collusion when compensation is volume scaling to those getting referrals (as may be found in the USA). If a volume-scaling system is required, our analysis suggests two ways in which collusion could still be averted: either by making the volume-scaling part of compensation a relatively small part of SP income (as may be typical in the UK), or by penalizing bad outcomes in a non-linear way to the point that the compensation of SPs is as sensitive to outcomes as it is to volume.

We observed, in an illustrative example based on UK diabetes data, that outcomes-adjusted capitation contracts with individual providers were almost immune to collusion and free-riding issues, to the point that a mechanism designer could restrict their attention to just such contracts. Consequently, they were the only contract type that can be expected to perform optimally in our application, irrespective of whether collusion and free-riding issues are present or not.

We also found that the choices of contract type and granularity of health outcome measurement are closely linked. Capitation contracts perform the best when used in conjunction with population-level health measures, and per-patient contracts perform the best with provider-level measures.

Although most results were derived using the assumption that the regulator knows the provider's cost function, we showed that in absence of this knowledge, the contracts we propose can still be implemented by using Yardstick competition contracts or with realized cost data over time.

Our analysis relied on a number of assumptions, but also suggested potential areas for further research. First, we did not model competition between medical providers. As such, our conclusions more readily apply in settings where local monopolies are present. Second, it may be useful to extend the model to account for more general pathways or networks of care. Third, our analysis of the impact of noise is limited due to assumptions that the available signal is unbiased and the providers are risk-neutral. Fourth, through our use of the classical Nash bargaining solution, we implicitly assumed that any differences in the bargaining power of providers are reflected in the value of their outside option. If this assumption is relaxed by allowing for arbitrary bargaining power, most results hold with mild adaptation, but some of our results no longer hold (notably Theorem 4). Our model provides a framework for considering these issues.

References

Adida E, Bravo F (2019) Contracts for healthcare referral services: Coordination via outcome-based penalty contracts. *Management Science* 65(3):1322–1341.

Adida E, Mamani H, Nassiri S (2017) Bundled payment vs. fee-for-service: Impact of payment scheme on performance. *Management Science* 63(5):1271–1656.

Andritsos DA, Tang CS (2018) Incentive programs for reducing readmissions when patient care is co-produced. *Production and Operations Management* 27(6):999–1020.

Arifoglu K, Ren H, Tezcan T (2021) Hospital readmissions reduction program does not provide the right incentives: Issues and remedies. *Management Science* 67(4):1993–2656.

Aswani A, Shen Z, Siddiq A (2019) Data-driven incentive design in the Medicare shared savings program. *Operations Research* 67(4):1002–1026.

Ata B, Killaly BL, Olsen TL, Parker RP (2013) On hospice operations under medicare reimbursement policies. *Management Science* 59(5):1027–1044.

Baliga S, Sjöström T (1998) Decentralization and collusion. *Journal Economic Theory* 83(2):196–232.

Bastani H, Bayati M, Braverman M, Gummadi R, Johari R (2016) Analysis of Medicare pay-for-performance contracts. https://ssrn.com/abstract=2839143, Accessed 1/12/2023.

Bastani H, Goh J, Bayati M (2015) Evidence of upcoding in pay-for-performance programs. *Management Science* 65(3):1042–1060.

Bolton P, Dewatripont M (2005) *Contract theory* (MIT Press).

Bond AM, Schpero WL, Casalino LP, Zhang M, Khullar D (2022) Association between individual primary care physician merit-based incentive payment system score and measures of process and patient outcomes. *JAMA* 328(21):2136–2146.

Borus JS, Laffel L (2010) Adherence challenges in the management of type 1 diabetes in adolescents: prevention and intervention. *Current Opinion in Pediatrics* 22(4):405.

Burns LR, Pauly MV (2018) Transformation of the health care industry: curb your enthusiasm? *The Milbank Quarterly* 96(1):57–109.

Cebul R, et al. (2013) Organizational fragmentation and care quality in the US health care system. Elgauge E, ed., *The Fragmentation of US Health Care: Causes and Solutions*, 37–61 (Oxford University Press).

Christianson JB, Leatherman S, Sutherland K (2008) Lessons from evaluations of purchaser pay-for-performance programs. *Medical Care Research and Review* 65(S6):S5–S35.

Clarke P, Gray A, Holman R (2002) Estimating utility values for health states of type 2 diabetic patients using the EQ-5D (UKPDS 62). *Medical Decision Making* 22(4):340–349.

Claxton K, Martin S, Soares M, Rice N, Spackman E, Hinde S, Devlin N, Smith PC, Sculpher M, et al. (2013) *Methods for the estimation of the NICE cost effectiveness threshold* (University of York).

Claxton K, et al. (2015) Methods for the estimation of the National Institute for Health and Care Excellence cost-effectiveness threshold. *Health Technology Assessment* 19(14):1–503.

CMS (2020) Medicare value based programs. Accessed 26/06/2021, https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/Value-Based-Programs/Value-Based-Programs.

Currie CJ, Peyrot M, Morgan CL, et al. (2012) The impact of treatment noncompliance on mortality in people with type 2 diabetes. *Diabetes care* 35(6):1279–1284.

Cutler DM, Richardson E (1998) The value of health: 1970-1990. *American Economic Review* 88(2):97–100.

Dai T, Akan M, Tayur S (2017) Imaging room and beyond: The underlying economics behind physicians' test-ordering behavior in outpatient services. *MSOM* 19(1):99–113.

de Vericourt F, Gurkan H, Wang S (2021) Informing the public about a pandemic. *Management Science* 67(10):6350–6357.

Diabetes UK (2014) The cost of diabetes report. Diabetes UK, https://www.diabetes.org.uk/resources-s3/2017-11/diabetes%20uk%20cost%20of%20diabetes%20report.pdf, Accessed 12/05/2021.

Diabetes UK (2019) Diabetes: Facts and stats. https://www.diabetes.org.uk/professionals/position-statements-reports/statistics, Accessed 1 May 2021.

Dyer O (2015) US crackdown on cash for referrals brings more arrests. *BMJ* 351:h6269.

Eijkenaar F, Emmert M, Scheppach M, Schöffski O (2013) Effects of pay for performance in health care: a systematic review of systematic reviews. *Health Policy* 110(2-3):115–130.

Fanburg JD, et al. (2022) Federal stark law and anti-kickback statute changes for 2022 and beyond. Brach Eichler LLC, Accessed 06/01/2023, https://www.bracheichler.com/insights/webinar-federal-stark-law-and-anti-kickback-statute-changes-for-2022-and-beyond-2/.

Ferris FL (1993) How effective are treatments for diabetic retinopathy? *JAMA* 269(10):1290–1291.

Figueroa JF, et al. (2016) Association between the value-based purchasing pay for performance program and patient mortality in US hospitals: observational study. *BMJ* 353:i2214, URL http://dx.doi.org/10.1136/bmj.i2214.

Freeman M, Savva N, Scholtes S (2017) Gatekeepers at work: An empirical analysis of a maternity unit. *Management Science* 63(10):3147–3167.

Gadre A (2015) India's private healthcare sector treats patients as revenue generators. *BMJ* 350:h826.

García-Pérez LE, Álvarez M, Dilla T, Gil-Guillén V, Orozco-Beltrán D (2013) Adherence to therapies in patients with type 2 diabetes. *Diabetes Therapy* 4(2):175–194.

Ghamat S, Zaric GS, Pun H (2021) Care-coordination: Gain-sharing agreements in bundled payment models. *Production and Operations Management* 30(5):1457–1478.

Grossman M (1972) On the concept of health capital and the demand for health. *J. Political Economy* 80(2):223–255.

Guo P, Tang CS, Wang Y, Zhao M (2019) The impact of reimbursement policy on social welfare, revisit rate, and waiting time in a public healthcare system: Fee-for-service versus bundled payment. *MSOM* 21(1):154–170.

Gupta D, Mehrotra M (2015) Bundled payments for healthcare services: Proposer selection and information sharing. *Operations Research* 63(4):772–788.

Hölmstrom B (1982) Moral hazard in teams. *The Bell Journal of Economics* 13(2):324–340.

Hsieh HM, He JS, Shin SJ, Chiu HC, Lee TC (2017) A diabetes pay-for-performance program and risks of cancer incidence and death in patients with type 2 diabetes in Taiwan. *Preventing chronic disease* 14.

ICHOM (2021) International consortium for outcomes measurement. https://www.ichom.org/.

Javitt JC, Aiello LP (1996) Cost-effectiveness of detecting and treating diabetic retinopathy. *Annals of Internal Medicine* 125(11):939.

Jiang H, Pang Z, Savin S (2012) Performance-based contracts for outpatient medical services. *MSOM* 14(4):654–669.

Khalid JM, et al. (2014) Rates and risk of hospitalisation among patients with type 2 diabetes: retrospective cohort study using the UK general practice research database linked to English hospital episode statistics. *International Journal of Clinical Practice* 68(1):40–48.

Koff E, Lyons N (2020) Implementing value-based health care at scale: the NSW experience. *Medical Journal of Australia* 212(3):104–106.

Laffont JJ, Martimort D (1998) Collusion and delegation. *RAND Journal of Economics* 29(2):280–305.

Leal J, Gray AM, Clarke PM (2009) Development of life-expectancy tables for people with type 2 diabetes. *European Heart Journal* 30(7):834–839.

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

Lee HH, Pinker EJ, Shumsky RA (2012) Outsourcing a two-level service process. *Management Science* 58(8):1569–1584.

Mannava KA, Bercovitch L, Grant-Kels JM (2013) Kickbacks, Stark violations, client billing, and joint ventures: Facts and controversies. *Clinics in Dermatology* 31(6):764–768.

Mathur R, et al. (2017) Population trends in the 10-year incidence and prevalence of diabetic retinopathy in the uk: a cohort study in the clinical practice research datalink 2004–2014. *BMJ Open* 7(2).

McGuire TG (2000) Physician agency. *Handbook of Health Economics* 1:461–536.

McNabb WL (1997) Adherence in diabetes: can we define it and can we measure it? *Diabetes Care* 20(2):215.

Mitchell P, et al. (2012) Cost-effectiveness of ranibizumab in treatment of diabetic macular oedema (dme) causing visual impairment: evidence from the RESTORE trial. *Brit. Jour. Ophthalmol.* 96(5):688–693.

NEJM Catalyst (2018) What are bundled payments? https://catalyst.nejm.org/doi/full/10.1056/CAT.18.0247.

Neumann PJ, Cohen JT, Weinstein MC (2014) Updating cost-effectiveness: the curious resilience of the $50,000-per-qaly threshold. *New England Journal of Medicine* 371(9):796–797.

NIH (2019) Diabetic retinopathy facts. National Eye Institute: http://www.nei.nih.gov/learn-about-eye-health/eye-conditions-and-diseases/diabetic-retinopathy, Accessed 12/05/2021.

OBH (2014) Contracting for outcomes: A value-based approach. Outcomes Based Healthcare. Accessed 1/12/2023, https://outcomesbasedhealthcare.com/wp-content/uploads/2018/12/Contracting_for_Outcomes-1.pdf.

Pauly M, Redisch M (1973) The not-for-profit hospital as a physicians' cooperative. *Amer Econ Rev* 63(1):87–99.

Pauly MV (1979) The ethics and economics of kickbacks and fee splitting. *The Bell Journal of Economics* 344–352.

Petersen LA, Woodard LD, Urech T, Daw C, Sookanan S (2006) Does pay-for-performance improve the quality of health care? *Annals of internal medicine* 145(4):265–272.

Porter ME (2010) What is value in health care? *New England Journal of Medicine* 363(26):2477–2481.

Rajagopalan S, Tong C (2022) Payment models to coordinate healthcare providers with partial attribution of outcome costs. *MSOM* 24(1):600–616.

Rathi VK, McWilliams JM (2019) First-year report cards from the merit-based incentive payment system (mips): what will be learned and what next? *JAMA* 321(12):1157–1158.

Redding K (2022) Participation and performance in accountable care organizations. Working Paper, https://kolereddig.github.io/reddig_acos.pdf, Accessed 25/01/2023.

Reid RO, Tom AK, Ross RM, Duffy EL, Damberg CL (2022) Physician compensation arrangements and financial performance incentives in us health systems. *JAMA Health Forum* 3(1):e214634–e214634.

Rein DB, Wirth KE, Johnson CA, Lee PP (2007) Estimating quality-adjusted life year losses associated with visual field deficits using methodological approaches. *Ophthalmic epidemiology* 14(4):258–264.

Rodwin MA (1995) *Medicine, money, and morals: physicians' conflicts of interest* (Oxford Univ. Press on Demand).

Rosenthal MB, Dudley RA (2006) Pay for performance: a decision guide for purchasers. *Agency for Healthcare Research and Quality (AHRQ)* 6:1–47.

Rosenthal MB, Dudley RA (2007) Pay-for-performance: will the latest payment trend improve care? *JAMA* 297(7):740–744.

Ryan AM, Krinsky S, Maurer KA, Dimick JB (2017) Changes in hospital quality associated with hospital value-based purchasing. *New England Journal of Medicine* 376(24):2358–2366.

Sanderson M, et al. (2018) New models of contracting in the public sector: A review of alliance contracting, prime contracting and outcome-based contracting literature. *Social Policy & Administration* 52(5):1060–1083.

Savva N, Tezcan T, Yıldız Ö (2019) Can yardstick competition reduce waiting times? *Management Science* 65(7):3196–3215.

She Z, Ayer T, Montanera D (2022) Can big data cure risk selection in healthcare capitation programs? a game theoretical analysis. *MSOM* 24(6):3117–3134.

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

Shumsky RA, Pinker EJ (2003) Gatekeepers and referrals in services. *Management Science* 49(7):839–856.

So KC, Tang CS (2000) Modeling the impact of an outcome-oriented reimbursement policy on clinic, patients, and pharmaceutical firms. *Management Science* 46(7):875–892.

Suen Sc, Negoescu D, Goh J (2022) Design of incentive programs for optimal medication adherence in the presence of observable consumption. *Operations Research* 70(3):1691–1716.

Taddeo D, Egedy M, Frappier J (2008) Adherence to treatment in adolescents. *Paediatrics & Child Health* 13(1):19.

Taylor HM, Karlin S (1998) *An Introduction to Stochastic Modeling* (Academic press).

Thornton D, et al. (2013) Predicting healthcare fraud in Medicaid: a multidimensional data model and analysis techniques for fraud detection. *Procedia Technology* 9:1252–1264.

Tirole J (1986) Hierarchies and bureaucracies: On the role of collusion in organizations. *Journal of Law, Economics, & Organization* 2(2):181–214.

UK NHS (2016) General and personal medical services, England September 2015-March 2016, Provisional experimental statistics. NIHR, http://content.digital.nhs.uk/catalogue/PUB21772, Accessed 8/06/2021.

UK NHS (2019) General Practice Workforce, Official Statistics, 30 September 2019. NIHR, https://digital.nhs.uk/data-and-information/publications/statistical/general-and-personal-medical-services/final-30-september-2019, Accessed 10/05/2021.

UK NHS (2020) Quality and outcomes framework. United Kingdom National Health Service, http://content.digital.nhs.uk/qof, Accessed 7/5/2021.

US DoJ (2021) Justice department recovers over $2.2 billion from false claims act cases in fiscal year 2020. U.S. Department of Justice Public Release, available online at https://www.justice.gov/opa/pr/justice-department-recovers-over-22-billion-false-claims-act-cases-fiscal-year-2020, Accessed 20/12/2022.

US Senate (2016) Examining the stark law: current issues and opportunities. U.S. Senate Hearing 114-668, https://www.congress.gov/event/114th-congress/senate-event/LC52008/text, Accessed 27/12/2022.

WEF (2018) Value in healthcare: mobilizing cooperation for health system transformation. World Economic Forum, Accessed 18/04/2021, https://www.weforum.org/reports/value-in-healthcare-mobilizing-cooperation-for-health-system-transformation.

Woolhandler S, Himmelstein DU (2004) The high costs of for-profit care. *Canadian Medical Association Journal* 170(12):1814–1815.

Zhang C, Atasu A, Ayer T, Toktay LB (2020) Truthful mechanisms for medical surplus product allocation. *MSOM* 22(4):735–753.

Zhang DJ, Gurvich I, Van Mieghem JA, Park E, Young RS, Williams MV (2016) Hospital readmissions reduction program: An economic and operational analysis. *Management Science* 62(11):3351–3371.

Zhu J, et al. (2020) New payment models in medtech. Deloitte Insights, Accessed 26/01/2023, https://www2.deloitte.com/us/en/insights/industry/life-sciences/medical-device-business-model-payments.html.

Online Companion: Appendices

Table EC.1 summarizes the principal notation used in the main paper. Appendix A formally justifies the mathematical claims in the main paper. Appendix B describes how literature and analysis were used to select 'reasonable' ranges and combinations of parameters for numerical experiments. Appendix C gives several extensions to the model to assess the robustness of the model to several (but not all) assumptions in the main paper, including allowing for multiple specialists, broader queue regimes, the effect of the GP on care, and a different model of collusion.

Appendix A: Proofs of Mathematical Claims

We first establish a few preliminaries. There are several substitutions that will get reused in the proofs:

$$\tau \doteq (1+d_G)/\bar{\lambda} \tag{EC.1}$$

$$\mu \doteq \underline{\mu} + \theta d_S. \tag{EC.2}$$

Table EC.1 Summary of notation for patient flows, costs, and health outcomes.

Parameters	
n	Number of patients with a chronic condition in the target population
q_i^0	Initial (beginning of period) health capital of patient i
q_P^0	Sum of initial health capitals of the whole population
$\bar{q}_P^0$	Average initial health capital per person in the whole population
ϕ	Proportion of adherent patients
$\bar{\lambda}$	Probability that an untreated patient develops a complication
$k_G(\cdot)$	GP's costs of providing treatment
$k_{SF}(\cdot)$	Specialist's (fixed) costs of providing treatment
k_{SV}	Specialist's variable (per-patient) cost
a	Health impact of developing a complication
b	Rate of health degradation with an untreated complication per unit of time
ζ	Fraction of lost health restored by the SP's treatment
Decisions	
d_G	Quality of care of the general practitioner (or GP)
d_S	Quality of care of the specialist (or SP)
Composites	which depend on both parameters and decisions
λ	Probability that a patient treated by the GP develops a complication
Λ	Average complication probability across all patients
μ	Specialist's service rate (with lower bound of $\underline{\mu}$)
$\mathcal{S}$	Number of people who develop complications (RV), or $\mathcal{S}(d_g)$ to stress the distribution's dependence on d_g
s	Realization of $\mathcal{S}$
Q_i	Health capital of patient i at the end of period (RV)
q_i	Realization of Q_i
$\bar{Q}_P, \bar{Q}_G, \bar{Q}_S$	Average end-of-period health capital of all patients ($\bar{Q}_P$), GP's patients ($\bar{Q}_G$), SP's patients ($\bar{Q}_S$)
$\bar{q}_P, \bar{q}_G, \bar{q}_S$	Realizations of $\bar{Q}_P, \bar{Q}_G, \bar{Q}_S$, respectively
Q_P, Q_G, Q_S	Sum of end-of-period health capitals of all patients (Q_P), GP's patients (Q_G), SP's patients (Q_S)
q_P, q_G, q_S	Realizations of Q_P, Q_G, Q_S, respectively
$\mathcal{Q}_P, \mathcal{Q}_G, \mathcal{Q}_S$	Unbiased noisy signals of q_P, q_G, q_S, respectively ($\mathbb{E}[\mathcal{Q}_P \mid Q_P] = Q_P, \mathbb{E}[\mathcal{Q}_G \mid Q_G] = Q_G$, $\mathbb{E}[\mathcal{Q}_S \mid Q_S] = Q_S$)
$\boldsymbol{q}_P, \boldsymbol{q}_G, \boldsymbol{q}_S$	Realizations of $\mathcal{Q}_P, \mathcal{Q}_G, \mathcal{Q}_S$ respectively
$\pi_j(v, d_G, d_S)$	Expected profit of the GP (if $j = G$), the SP (if $j = S$), or group (if $j = A$) when holding contract $v()$ and agent decisions are d_G and d_S

Intuitively, these are the inverse of the complication rate and the service rate, which are induced by provider decisions d_G and d_S. Notice that (EC.1) and (EC.2) depend only on provider decisions and constants and establish bijections between possible values of τ and d_G (respectively μ and d_S). Another useful substitution is $A \doteq a(1-\zeta)$, $B \doteq b(1-\zeta)$. Using these substitutions, we can more concisely restate the elementary functions of the naïve model as functions of τ, μ instead of d_G, d_S:

$$\Lambda(\tau) = \phi/\tau + (1-\phi)\bar{\lambda}, \tag{EC.3}$$

$$\mathbb{E}Q_P(\tau,\mu) = q_P^0\left(1 - A\Lambda(\tau) - \frac{B\Lambda(\tau)}{\mu - n\Lambda(\tau)}\right), \tag{EC.4}$$

$$\mathbb{E}Q_G(\tau,\mu) = q_P^0\phi\left(1 - \frac{A}{\tau} - \frac{B/\tau}{\mu - n\Lambda(\tau)}\right), \tag{EC.5}$$

$$\mathbb{E}Q_S(\tau,\mu) = q_P^0\Lambda(\tau)\left(1 - A - \frac{B}{\mu - n\Lambda(\tau)}\right), \tag{EC.6}$$

$$\mathcal{S}(\tau) \sim \text{Poisson}(n\Lambda(\tau)). \tag{EC.7}$$

We also introduce one notational convention: for any function $f(\cdot)$ defined on a closed set $[\underline{f}, \overline{f}]$, the function has only one-handed derivatives at $\underline{f}$ and $\overline{f}$. Throughout, we use the normal derivative notation, $f'(\underline{f})$ and $f'(\overline{f})$, to denote those one-handed derivatives.

For any function $f(\cdot)$, we use $f(\cdot)\uparrow x$ ($f(\cdot)\downarrow x$) to denote that the function is increasing (decreasing) in x.

Denoting $\underline{\tau} \doteq 1/\bar{\lambda}$ and using the substitution (EC.1)-(EC.2) for the cost functions, we have that $k_G(\tau), k_{SF}(\mu)$ are increasing and convex with $k_G'(\underline{\tau}) = k_{SF}'(\underline{\mu}) = 0$. Thus, for the analysis of claims in Sections 3–4, the problem of providers deciding on d_G and d_S is equivalent to them deciding on τ and μ directly. These substitutions will be of limited use for the main model in Section 5 and the extensions in Appendix C, due to the more complicated role of d_G, d_S, and ζ there.

Lemma 1 establishes the sufficiency of first-order conditions (FOCs) for several functions used to prove the claims from Sections 3–4 below.

Lemma 1 *Let assumptions of the care model stated in Section 2 hold. Then, for $\alpha_0, \alpha_1, \alpha_2, \alpha_3, \alpha_4 \geq 0$, the function $f : [0,\infty)^2 \to \mathbb{R}$, with mapping rule*

$$f(d_G, d_S) = \alpha_0\mathbb{E}Q_P(d_G,d_S) + \alpha_1\mathbb{E}Q_G(d_G,d_S) - \alpha_2 k_G(d_G) - \alpha_3 k_{SF}(d_S) - \alpha_4 k_{SV} n\Lambda(d_G),$$

is jointly concave. If $\max\{\alpha_0,\alpha_1\} > 0$ and $\max\{\alpha_2,\alpha_3\} > 0$ then $f(d_G,d_S)$ has an interior maximum.

Proof of Lemma 1. Let us first demonstrate the joint concavity of $\mathbb{E}Q_P$ as a function of τ and μ, using the substitution given in (EC.1)-(EC.2). Denote $T(\tau) = \tau(1-\phi) + \underline{\tau}\phi$. Then, from (EC.4), we have $\frac{\partial}{\partial\tau}\mathbb{E}Q_P(\tau,\mu) = q_P^0\phi\,(A/\tau^2 + B\mu\underline{\tau}^2/(\mu\tau\underline{\tau} - nT(\tau))^2)$, $\frac{\partial}{\partial\mu}\mathbb{E}Q_P(\tau,\mu) = q_P^0 B\tau\underline{\tau}T(\tau)/(\mu\tau\underline{\tau} - nT(\tau))^2$, and the Hessian of $\mathbb{E}Q_P$ is

$$\mathbf{H}(\mathbb{E}Q_P(\tau,\mu)) = \begin{bmatrix} -\dfrac{2q_P^0\phi\,(A(\mu\tau\underline{\tau} - nT(\tau))^3 + B\mu\tau^3\underline{\tau}^2(\mu\underline{\tau} + n(\phi-1)))}{\tau^3(\mu\tau\underline{\tau} - nT(\tau))^3} & -\dfrac{Bq_P^0\underline{\tau}^2\phi(\mu\tau\underline{\tau} + nT(\tau))}{(\mu\tau\underline{\tau} - nT(\tau))^3} \\ -\dfrac{Bq_P^0\underline{\tau}^2\phi(\mu\tau\underline{\tau} + nT(\tau))}{(\mu\tau\underline{\tau} - nT(\tau))^3} & -\dfrac{2Bq_P^0\tau^2\underline{\tau}^2T(\tau)}{(\mu\tau\underline{\tau} - nT(\tau))^3} \end{bmatrix}.$$

Observe that $T(\tau) \geq \underline{\tau} > 1$ as $\phi \in (0,1]$ and $\tau \geq \underline{\tau} > 1$. Also observe that the steady state condition $\underline{\mu} > \bar{\lambda}n$ implies both $\mu\tau\underline{\tau} - nT(\tau) > 0$ and $\mu\underline{\tau} + n(\phi-1) > 0$. It follows from these observations that the leading principal minor is negative ($\mathbf{H}_{1,1}(\mathbb{E}Q_P(\tau,\mu)) < 0$) and $|\,\mathbf{H}(\mathbb{E}Q_P(\tau,\mu))\,| =$

$[B(q_P^0)^2\underline{\tau}^2\phi\,(B\tau\underline{\tau}^2(\mu\tau(3\underline{\tau}\phi+4\tau(1-\phi))+n\phi T(\tau))+4AT(\tau)(\mu\tau\underline{\tau}-nT(\tau))^2)]/[\tau(\mu\tau\underline{\tau}-nT(\tau))^5]>0$. Consequently, $\mathbf{H}(\mathbb{E}Q_P(\tau,\mu))$ is negative definite and thus $\mathbb{E}Q_P(\tau,\mu)$ is jointly concave.

The joint concavity of $\mathbb{E}Q_G(\tau,\mu)$ can be shown analogously. Because $f(\tau,\mu)$ is a weighted sum of jointly concave functions and univariate concave functions, it is jointly concave as well. The existence of an interior maximum follows from $k_G'(\underline{\tau})=k_{SF}'(\underline{\mu})=0$. The same holds for $f(d_G,d_S)$ as (EC.1)-(EC.2) are affine and composition with affine functions preserves (joint) concavity. □

A.1. Proofs for Claims About the Naïve Model in Section 3.

Proof of Prop. 1. The structure of the proof is as follows. We identify a set of necessary conditions that any differentiable contract which achieves first-best must satisfy. We characterize all linear contracts that satisfy those conditions in four parts, and then show that those indeed do achieve first-best.

From (5), the first-best decisions (d_G^*,d_S^*) are solutions of the optimization problem

$$\underset{(d_G,d_S)\in[0,\infty)^2}{\arg\max}\ \mathbb{E}Q_P(d_G,d_S)-k_G(d_G)-k_{SF}(d_S)-k_{SV}n\Lambda(d_G). \tag{EC.8}$$

Because this objective function is jointly concave and has an interior maximum (from Lemma 1), the first-best decisions are unique solutions to FOCs

$$k_G'(d_G) = \frac{\partial\mathbb{E}Q_P(d_G,d_S)}{\partial d_G}-k_{SV}n\Lambda'(d_G), \tag{EC.9}$$

$$k_{SF}'(d_S) = \frac{\partial\mathbb{E}Q_P(d_G,d_S)}{\partial d_S}. \tag{EC.10}$$

Part 1: deriving the optimal capitation contract for the GP ($\boldsymbol{v_{GC}}$). Assume there exists a linear contract $v_{GC}(\boldsymbol{q}_P)=a_{GC}+b_{GC}\boldsymbol{q}_P$ for the GP, under which first-best can be achieved. Then, when holding this contract, the GP's best response to the SP choosing d_S^* needs to be d_G^* such that $d_G^*\in\arg\max_{d_G\in[0,\infty)}\mathbb{E}\left[a_{GC}+b_{GC}\mathcal{Q}_P(d_G,d_S^*)-k_G(d_G)\right]$. Using Lemma 1, this problem can be replaced by its FOC:

$$k_G'(d_G)=b_{GC}\frac{\partial\mathbb{E}Q_P(d_G,d_S^*)}{\partial d_G}. \tag{EC.11}$$

Thus, $(d_G,d_S)=(d_G^*,d_S^*)$ needs to simultaneously solve (EC.9) and (EC.11). Equating the right-hand sides (RHS) of those two equations and solving for b_{GC} yields

$$b_{GC}=1-\frac{k_{SV}n\Lambda'(d_G^*)}{\frac{\partial}{\partial d_G}\mathbb{E}Q_P(d_G^*,d_S^*)} \tag{EC.12}$$

as the unique solution. For the contract to achieve FB, it is also necessary that the individual rationality constraint (11) is binding. Therefore, solving $\mathbb{E}\left[a_{GC}+b_{GC}\mathcal{Q}_P(d_G^*,d_S^*)-k_G(d_G^*)\right]=V_G$ for a_{GC} yields $a_{GC}=V_G+k_G(d_G^*)-b_{GC}\mathbb{E}Q(d_G^*,d_S^*)$ as the unique solution. Setting

$$r_{GC}=\frac{1}{n}+\frac{k_{SV}(\mu(d_S^*)-n\Lambda(d_G^*))^2}{q_P^0\left(A(\mu(d_S^*)-n\Lambda(d_G^*))^2+B\mu(d_S^*)\right)},\quad f_{GC}=k_G(d_G^*)/n,\quad t_{GC}=\mathbb{E}Q_P(d_G^*,d_S^*)-\frac{V_G}{nr_{GC}}, \tag{EC.13}$$

ensures that $v_{GC}(\boldsymbol{q}_P)=a_{GC}+b_{GC}\boldsymbol{q}_P=n\left(f_{GC}+r_{GC}(\boldsymbol{q}_P-t_{GC})\right)$.[10]

[10] The optimal contract has an additional degree of freedom between t_{GC} and f_{GC} in identifying the intercept $f_{iC}-r_{iC}t_{iC}$ of the unique solution among linear contracts in (13). So, one can increase the fixed compensation (but then also increase the target health threshold), while keeping the expected payout to the providers and, thus, contract performance the same. These focal values for f_{GC} and t_{GC} were chosen to be more in line with existing policies of a fixed reimbursement component being designed to cover costs. However, if the mechanism designers have any additional objectives we did not consider, they could potentially be accomplished by optimizing over this degree of freedom. The other optimal contracts in this proposition have the same property.

Part 2: deriving the optimal per-patient contract for the GP ($\boldsymbol{v_{GP}}$). Assume there exists a linear contract $v_{GP}(\boldsymbol{q}_G) = a_{GP} + b_{GP}\boldsymbol{q}_G$ for the GP, under which first-best can be achieved. Then, when holding the aforementioned contract, $d_G^* \in \arg\max_{d_G\in[0,\infty)} \mathbb{E}\left[a_{GP} + b_{GP}\mathcal{Q}_G(d_G, d_S^*) - k_G(d_G)\right]$, a maximization problem that can be reduced (using Lemma 1) to its FOC:

$$k_G'(d_G) = b_{GP}\frac{\partial \mathbb{E}Q_G(d_G, d_S^*)}{\partial d_G}. \tag{EC.14}$$

Thus, $(d_G, d_S) = (d_G^*, d_S^*)$ needs to simultaneously solve (EC.9) and (EC.14). Denote by $\mathbb{E}\hat{Q}_G(d_G, d_S)$ the sum of expected health of *non-adherent* patients so that $\mathbb{E}Q_P(d_G, d_S) = \mathbb{E}Q_G(d_G, d_S) + \mathbb{E}\hat{Q}_G(d_G, d_S)$. Equating the RHS of those two equations and solving for b_{GP} yields

$$b_{GP} = 1 + \frac{\frac{\partial}{\partial d_G}\mathbb{E}\hat{Q}_G(d_G^*, d_S^*) - k_{SV}n\Lambda'(d_G^*)}{\frac{\partial}{\partial d_G}\mathbb{E}Q_G(d_G^*, d_S^*)} \tag{EC.15}$$

as the unique solution. For the contract to achieve FB, it is also necessary that the individual rationality constraint (11) is binding, and therefore, $a_{GP} = V_G - b_{GP}\mathbb{E}Q_G(d_G^*, d_S^*) + k_G(d_G^*)$. Then, set

$$r_{GP} = \frac{\frac{\partial}{\partial d_G}\mathbb{E}Q_P(d_G^*, d_S^*) - k_{SV}n\Lambda'(d_G^*)}{\phi n \frac{\partial}{\partial d_G}\mathbb{E}Q_G(d_G^*, d_S^*)}, \tag{EC.16}$$

$$f_{GP} = k_G(d_G^*)/(\phi n), \quad t_{GP} = \mathbb{E}Q_G(d_G^*, d_S^*) - \frac{V_G}{n\phi r_{GP}}, \tag{EC.17}$$

to ensure that $v_{GP}(\boldsymbol{q}_G) = a_{GP} + b_{GP}\boldsymbol{q}_G = n\phi\left(f_{GP} + r_{GP}(\boldsymbol{q}_G - t_{GP})\right)$.

Part 3: deriving the optimal capitation contract for the SP ($\boldsymbol{v_{SC}}$). Assume there exists a linear contract $v_{SC}(\boldsymbol{q}_P) = a_{SC} + b_{SC}\boldsymbol{q}_P$ for the SP, under which first-best can be achieved. Then, when holding the aforementioned contract, $d_S^* \in \arg\max_{d_S\in[0,\infty)} \mathbb{E}\left[a_{SC} + b_{SC}\mathcal{Q}_P(d_G^*, d_S) - k_{SF}(d_S) - \mathcal{S}(d_G^*)k_{SV}\right]$, a maximization problem that can be reduced (using Lemma 1) to its FOC:

$$k_{SF}'(d_S) = b_{SC}\frac{\partial \mathbb{E}Q_P(d_G^*, d_S)}{\partial d_S}. \tag{EC.18}$$

Thus, $(d_G, d_S) = (d_G^*, d_S^*)$ needs to simultaneously solve (EC.10) and (EC.18). Equating the RHS of those two equations and solving for b_{SC} yields $b_{SC} = 1$ as the unique solution. For the contract to achieve FB, it is also necessary that the individual rationality constraint (12) is binding, which yields, $a_{SC} = V_S + k_{SV}n\Lambda(d_G^*) + k_{SF}(d_S^*) - b_{SC}\mathbb{E}Q_P(d_G^*, d_S^*)$. Setting

$$r_{SC} = \frac{1}{n}, \quad f_{SC} = k_{SV}\Lambda(d_G^*) + k_{SF}(d_S^*)/n, \quad t_{SC} = \mathbb{E}Q_P(d_G^*, d_S^*) - V_S \tag{EC.19}$$

ensures that $v_{SC}(\boldsymbol{q}_P) = a_{SC} + b_{SC}\boldsymbol{q}_P = n\left(f_{SC} + r_{SC}(\boldsymbol{q}_P - t_{SC})\right)$.

Part 4: deriving the optimal per-patient contract for the SP ($\boldsymbol{v_{SP}}$). Note that, unlike the other expected health functions, $\mathbb{E}Q_S(d_G, d_S)$ is concave in d_S but not in d_G. Assume there exists a linear contract $v_{SP}(\boldsymbol{q}_S) = a_{SP} + b_{SP}\boldsymbol{q}_S$ for the SP, under which first-best can be achieved. Then, when holding the aforementioned contract, $d_S^* \in \arg\max_{d_S\in[0,\infty)} \mathbb{E}\left[a_{SP} + b_{SP}\mathcal{Q}_S(d_G^*, d_S) - k_{SF}(d_S) - \mathcal{S}(d_G^*)k_{SV}\right]$, a maximization problem that can be reduced (using Lemma 1) to its FOC:

$$k_{SF}'(d_S) = b_{SP}\frac{\partial \mathbb{E}Q_S(d_G^*, d_S)}{\partial d_S}. \tag{EC.20}$$

Thus, $(d_G, d_S) = (d_G^*, d_S^*)$ needs to simultaneously solve (EC.10) and (EC.20). Equating the RHS of those two equations and solving for b_{SP} yields $b_{SP} = 1$ as the unique solution. For the contract to achieve FB, it is also necessary that the individual rationality constraint (12) is binding, using which yields $a_{SP} = V_S + k_{SV} n\Lambda(d_G^*) + k_{SF}(d_S^*) - b_{SP}\mathbb{E}Q_S(d_G^*, d_S^*)$ as the unique solution for a_{SP}. Setting

$$r_{SP} = \frac{1}{n\Lambda(d_G^*)}, \quad f_{SP} = k_{SV} + \frac{k_{SF}(d_S^*)}{n\Lambda(d_G^*)}, \quad t_{SP} = \mathbb{E}Q_S(d_G^*, d_S^*) - V_S, \tag{EC.21}$$

ensures that $v_{SP}(q_S) = a_{SP} + b_{SP} q_S = n\Lambda(d_G^*)\,(f_{SP} + r_{SP}(q_S - t_{SP}))$.

Part 5: sufficiency. Finally, giving the GP either of the contracts v_{GC} or v_{GP} as given in parts 1 and 2 of this proof, while giving the SP either of the contracts v_{SC} or v_{SP} as defined in parts 3 and 4 of this proof will result in the following. The GP's best response to the SP choosing d_S^* will be to choose d_G^*, as shown in parts 1 and 2 of this proof, while the SP's best response to the GP choosing d_G^* will be to choose d_S^*, as shown in parts 3 and 4 of this proof. Thus, (d_G^*, d_S^*) is a Nash Equilibrium and solves the individual compatibility constraints (9)–(10). The system efficiency function u, as given by (5), is maximized for (d_G^*, d_S^*), from the definition of d_G^*, d_S^*. The individual rationality constraints (11)–(12) are binding, as shown in parts 1-4, thus agents will accept the contracts and all value generated will be appropriated by the principal. Hence all four of such contract pairs achieve first-best and thus solve the NICP in (8)–(12). □

Proof of Prop. 2. We first derive an expression for $\mathbb{E}Q_A$. Denote by M the (random) number of patients who are treated by neither provider. M is distributed according to $\text{Bin}(n(1-\phi), 1-\overline{\lambda})$ as M is equal to the number of non-adherent patients $(n(1-\phi))$ who do not develop complications (each has a $1-\overline{\lambda}$ probability of not developing a complication). Wald's equation gives $\mathbb{E}M = n(1-\overline{\lambda})(1-\phi)$, applying which to (4) yields

$$\mathbb{E}Q_A(d_G, d_S) = \mathbb{E}Q_P(d_G, d_S) - q_P^0(1-\overline{\lambda})(1-\phi) \tag{EC.22}$$

From (5), the FB decisions (d_G^*, d_S^*) are solutions to $\arg\max_{(d_G, d_S)\in[0,\infty)^2} u(d_G, d_S)$, or equivalently (from Lemma 1) to FOCs (EC.9) and (EC.10). Equilibrium decisions of a unified provider under group contract v are given by the incentive compatibility constraint (16). Specifically, for the linear group contracts

$$v_{AC}(q_P) = a_{AC} + q_P \text{ and } v_{AP}(q_A) = a_{AP} + q_A, \tag{EC.23}$$

the objective function in the incentive compatibility constraint (16) differs from the system efficiency function (5) only by a constant; thus, the set of maximizers is the same. Consequently, a group under such a contract will make the first-best decisions. The optimal a_{AC} and a_{AP} are then derived from (17) to ensure individual rationality is binding, which yields $a_{AC} = V_G + V_S + k_{SF}(d_S^*) + k_G(d_G^*) + k_{SV} n\Lambda(d_G^*) - \mathbb{E}Q_P(d_G^*, d_S^*)$ and $a_{AP} = V_G + V_S + k_{SF}(d_S^*) + k_G(d_G^*) + k_{SV} n\Lambda(d_G^*) - \mathbb{E}Q_A(d_G^*, d_S^*)$. Then, set

$$r_{AC} = \frac{1}{n}, \quad f_{AC} = \frac{k_{SF}(d_S^*) + k_G(d_G^*)}{n} + k_{SV}\Lambda(d_G^*), \quad t_{AC} = EQ_P(d_G^*, d_S^*) - V_G - V_S,$$

$$r_{AP} = \frac{1}{n(1-(1-\overline{\lambda})(1-\phi))}, \quad f_{AP} = \frac{k_{SF}(d_S^*) + k_G(d_G^*) + k_{SV} n\Lambda(d_G^*)}{n(1-(1-\overline{\lambda})(1-\phi))}, \quad t_{AP} = EQ_A(d_G^*, d_S^*) - V_G - V_S,$$

to ensure that $v_{AC}(q_P) = a_{AC} + q_P = n(t_{AC} + r_{AC}(q_P - t_{AC}))$ and $v_{AP}(q_A) = a_{AP} + q_A = n(1-(1-\overline{\lambda})(1-\phi))(t_{AP} + r_{AP}(q_P - t_{AP}))$. □

A.2. Proofs for Claims About Free-riding and Collusion in Section 4.

Lemma 2 serves as the "machinery" behind Theorem 1, using the implicit function theorem to derive how equilibrium decisions change under free-riding. For the proofs of claims in Section 4, including Lemma 2, we find it useful to express all functions using the τ, μ substitution given in (EC.1)-(EC.2). The function $f(\tau, \mu, \alpha)$ defined in Lemma 2 is constructed so that for $\alpha = 0$, it is maximized by the first-best decisions, while for $\alpha = 1$, it is maximized by the decisions of free-riding agents under contract $v_A \in \{v_{AC}, v_{AP}\}$ in Prop. 2.

Lemma 2 *Let assumptions of Section 2 hold and define $f : [\underline{\tau}, \infty) \times [\underline{\mu}, \infty) \times [0,1] \to \mathbb{R}$ with mapping rule*

$$f(\tau, \mu; \alpha) = \left(1 - \frac{1}{2}\alpha\right) \mathbb{E}Q_P(\tau, \mu) - k_G(\tau) - k_{SF}(\mu) - (1-\alpha) n \Lambda(\tau) k_{SV}. \tag{EC.24}$$

Furthermore, let $\tilde{\tau}(\alpha)$ and $\tilde{\mu}(\alpha)$ be functions such that $(\tilde{\tau}(\alpha), \tilde{\mu}(\alpha)) \in \arg\max_{\tau,\mu} f(\tau, \mu; \alpha)$. Then, $\tilde{\tau}'(\alpha) < 0$ and $\frac{\partial}{\partial \alpha}\left(\mathbb{E}Q_P(\tilde{\tau}(\alpha), \tilde{\mu}(\alpha))\right) < 0$.

Proof of Lemma 2. Because f is jointly concave in (τ, μ) and has an interior maximum (by Lemma 1), for any $\alpha \in [0,1]$, the sole maximizer of f is the unique solution of FOCs. Thus, by implicit function theorem $\tilde{\tau}(\alpha), \tilde{\mu}(\alpha)$ are well-defined and unique. Denote $F_1(\tau, \mu; \alpha) \doteq \frac{\partial}{\partial \tau} f(\tau, \mu; \alpha)$ and $F_2(\tau, \mu; \alpha) \doteq \frac{\partial}{\partial \mu} f(\tau, \mu; \alpha)$. By application of implicit function theorem, we can find derivatives of $\tilde{\tau}(\alpha), \tilde{\mu}(\alpha)$ by solving

$$\mathbf{J} \begin{bmatrix} \frac{\partial \tilde{\tau}}{\partial \alpha} & \frac{\partial \tilde{\mu}}{\partial \alpha} \end{bmatrix}^T = -\mathbf{F}, \qquad \text{where} \qquad \mathbf{J} = \begin{bmatrix} \frac{\partial F_1}{\partial \tau} & \frac{\partial F_1}{\partial \mu} \\ \frac{\partial F_2}{\partial \tau} & \frac{\partial F_2}{\partial \mu} \end{bmatrix} \qquad \text{and} \qquad \mathbf{F} = \begin{bmatrix} \frac{\partial F_1}{\partial \alpha} \\ \frac{\partial F_2}{\partial \alpha} \end{bmatrix},$$

with all of the partial derivatives evaluated at $(\tilde{\tau}(\alpha), \tilde{\mu}(\alpha); \alpha)$. Cramer's rule yields $\tilde{\tau}'(\alpha) = -\text{Det}(\mathbf{J}_1)/\text{Det}(\mathbf{J})$ and $\tilde{\mu}'(\alpha) = -\text{Det}(\mathbf{J}_2)/\text{Det}(\mathbf{J})$ with $\mathbf{J}_i$ standing for $\mathbf{J}$ in which the i-th column is replaced by $\mathbf{F}$. Denote $T = (\tau(1-\phi) + \underline{\tau}\phi)$ and $\Psi = \mu\tau\underline{\tau} - n(\tau(1-\phi) + \underline{\tau}\phi)$ and notice that $\Psi > 0$ (from steady state condition and $\tau \geq \underline{\tau}$), then $\text{Det}(\mathbf{J}_1) = \phi[2k''_{SF}(\mu)\Psi^2 \left(q_P^0 \left(A\Psi^2 + B\mu\tau^2\underline{\tau}^2\right) + 2k_{SV} n \Psi^2\right) + (2-\alpha) B q_P^0 \tau^2 \underline{\tau}^2 T (q_P^0 (2A\Psi + B\tau\underline{\tau}) + 4k_{SV} n \Psi)]/[4\tau^2\Psi^4] > 0$. We also have $\text{Det}(\mathbf{J}) > 0$ (from Lemma 1), thus $\tilde{\tau}'(\alpha) < 0$.

It remains to show that $\frac{\partial}{\partial \alpha}\left(\mathbb{E}Q_P(\tilde{\tau}(\alpha), \tilde{\mu}(\alpha))\right) < 0$. Using (4), we have $\mathbb{E}Q_P(\tilde{\tau}(\alpha), \tilde{\mu}(\alpha)) = q_P^0 \left(1 - A\Lambda(\tilde{\tau}(\alpha)) - B/((\tilde{\mu}(\alpha)/\Lambda(\tilde{\tau}(\alpha)) - n)\right)$, from which it can be seen that showing $\Lambda(\tilde{\tau}(\alpha)) \uparrow \alpha$ and $\tilde{\mu}(\alpha)/\Lambda(\tilde{\tau}(\alpha)) \downarrow \alpha$ is sufficient to demonstrate that $\mathbb{E}Q_P(\tilde{\tau}(\alpha), \tilde{\mu}(\alpha)) \downarrow \alpha$. From $\Lambda(\tilde{\tau}(\alpha)) = (1 - \phi)/\underline{\tau} + \phi/\tilde{\tau}(\alpha)$ we have $\frac{\partial}{\partial \alpha}\Lambda(\tilde{\tau}(\alpha)) = -\phi\tilde{\tau}'(\alpha)/\tilde{\tau}(\alpha)^2 > 0$ and $\frac{\partial}{\partial \alpha}\left(\tilde{\mu}(\alpha)/\Lambda(\tilde{\tau}(\alpha))\right) = \underline{\tau}\left[\underline{\tau}\phi\tilde{\tau}(\alpha)\tilde{\mu}'(\alpha) + (1-\phi)\tilde{\tau}(\alpha)^2\tilde{\mu}'(\alpha) + \underline{\tau}\phi\tilde{\mu}(\alpha)\tilde{\tau}'(\alpha)\right]/\left[-\phi\tilde{\tau}(\alpha) + \tilde{\tau}(\alpha) + \underline{\tau}\phi\right]^2$. Thus,

$$\begin{aligned}
\text{sgn}\left(\frac{\partial}{\partial \alpha}\left(\frac{\tilde{\mu}(\alpha)}{\Lambda(\tilde{\tau}(\alpha))}\right)\right) &= \text{sgn}\left((\underline{\tau}\phi\tilde{\tau}(\alpha) + (1-\phi)\tilde{\tau}(\alpha)^2)\tilde{\mu}'(\alpha) + \underline{\tau}\phi\tilde{\mu}(\alpha)\tilde{\tau}'(\alpha)\right) \\
&= \text{sgn}\left((\underline{\tau}\phi\tau + (1-\phi)\tau^2)\frac{-\text{Det}\mathbf{J}_2}{\text{Det}\mathbf{J}} + \underline{\tau}\phi\mu\frac{-\text{Det}\mathbf{J}_1}{\text{Det}\mathbf{J}}\right)\Big|_{(\tau,\mu)=(\tilde{\tau}(\alpha),\tilde{\mu}(\alpha))} \\
&= \text{sgn}\Big(-\frac{\underline{\tau}\phi}{4\tau^2\Psi^4}\Big(2\mu\phi k''_{SF}(\mu)\Psi^2\left(q_P^0\left(a\Psi^2 + b\mu\tau^2\underline{\tau}^2\right) + 2k_{SV}n\Psi^2\right) + 2\phi^{-1}bq_P^0\tau^4 k''_G(\tau)T^2\Psi^2 \\
&\quad + b(q_P^0)^2\tau T(2-\alpha)\left(a(2T + \underline{\tau}\phi)\Psi^2 + 2b\mu\tau^2\underline{\tau}^2 T\right) \\
&\quad + 2bq_P^0\tau T k_{SV} n(2(\alpha-1)\tau(\phi-1) + (4-3\alpha)\underline{\tau}\phi)\Psi^2\Big)\Big)\Big|_{(\tau,\mu)=(\tilde{\tau}(\alpha),\tilde{\mu}(\alpha))}.
\end{aligned}$$

which is negative because $\Psi > 0$, $T > 0$, $\alpha \in [0,1]$ and $\phi \in (0,1]$, thus $\frac{\partial}{\partial \alpha}\left(\tilde{\mu}(\alpha)/\Lambda(\tilde{\tau}(\alpha))\right) < 0$ and hence $\frac{\partial}{\partial \alpha}\mathbb{E}Q_P(\tilde{\tau}(\alpha), \tilde{\mu}(\alpha)) < 0$. □

Proof of Theorem 1. Observe in the proof of Prop. 2 that the payout of the group will not depend on which contract type (out of the two given in Prop. 2) the group holds. Thus for the purposes of this theorem, which of the two contracts does the group hold bears no impact. Using the τ, μ substitution given by (EC.1)-(EC.2), from (18)–(19), it follows that free-riding agents will choose $\tilde{\tau}_A^{FR}$, $\tilde{\mu}_A^{FR}$ which solve the system $\tilde{\tau}_A^{FR} \in \arg\max_{\tau \in [\underline{\tau}, \infty)} \mathbb{E}\left[(v_A(\mathcal{Q}_P(\tau, \tilde{\mu}_A^{FR}))/2 - V_G - V_S) + V_G - k_G(\tau)\right]$, $\tilde{\mu}_A^{FR} \in \arg\max_{\mu \in [\underline{\mu}, \infty)} \mathbb{E}\left[(v_A(\mathcal{Q}_P(\tilde{\tau}_A^{FR}, \mu))/2 - V_G - V_S) + V_S - k_{SF}(\mu) - \mathcal{S}(\tilde{\tau}_A^{FR}) k_{SV}\right]$. Using Lemma 1 and the definition of $v_A \in \{v_{AC}, v_{AP}\}$ given by Prop. 2: both of these objective functions are concave, twice differentiable and have an interior maximum. Moreover, v_A is a linear function, so $\mathbb{E}v_A(\mathcal{Q}_P) = v_A(\mathbb{E}\mathcal{Q}_P) = v_A(\mathbb{E}Q_P)$.

Thus, the system above can be replaced by its FOCs, given by $\mathbb{E}Q_P(\tilde{\tau}_A^{FR}, \tilde{\mu}_A^{FR})/2 - k_G(\tilde{\tau}_A^{FR}) = 0$ and $\mathbb{E}Q_P(\tilde{\tau}_A^{FR}, \tilde{\mu}_A^{FR})/2 - k_{SF}(\tilde{\mu}_A^{FR}) = 0$. By Lemma 1, there is a unique solution of this system: the sole maximizer of $\mathbb{E}Q_P(\tau, \mu)/2 - k_G(\tau) - k_{SF}(\mu) = f(\tau, \mu, 1)$, where f is the function defined in Lemma 2. Thus, the first-best decisions are maximizers of $f(\tau, \mu, 0)$ whereas the decisions under free-riding are the maximizers of $f(\tau, \mu, 1)$. Applying Lemma 2 and using its functions $\tilde{\tau}(\alpha)$ and $\tilde{\mu}(\alpha)$ gives $\tilde{\tau}_A^{FR} = \tilde{\tau}(1)$, $\tilde{\mu}_A^{FR} = \tilde{\mu}(1)$, $\tau^* = \tilde{\tau}(0)$, $\mu^* = \tilde{\mu}(0)$, thus $\tilde{\tau}_A^{FR} < \tau^*$ and $\mathbb{E}Q_P(\tilde{\tau}_A^{FR}, \tilde{\mu}_A^{FR}) < \mathbb{E}Q_P(\tau^*, \mu^*)$ showing parts (ii) and (iii) of the Theorem. Because (τ^*, μ^*) is the sole maximizer of $u(\tau, \mu)$ and $(\tau_A^{FR}, \mu_A^{FR}) \neq (\tau^*, \mu^*)$, system efficiency decreases under free-riding, which shows part (i). Analogously, part (iv) follows from (τ^*, μ^*) being the sole maximizer of the group profit function $\pi_A(v_A, \tau, \mu)$.

Lastly, using a_{AC} derived in the proof of Prop. 2, we have $\mathbb{E}v_A(\mathcal{Q}_P(\tau^*, \mu^*)) = \mathbb{E}Q_P(\tau^*, \mu^*) + a_{AC} > \mathbb{E}Q_P(\tilde{\lambda}_A^{FR}, \tilde{\mu}_A^{FR}) + a_{AC} = \mathbb{E}v_A(\mathcal{Q}_P(\tilde{\lambda}_A^{FR}, \tilde{\mu}_A^{FR}))$, which shows part (v). □

Lemma 3 serves the same role for Theorem 2 as Lemma 2 does for Theorem 1, allowing us to see how decisions made by the agents and population health depend on contract types in the presence of collusion. The function $g(\tau, \mu; \alpha, \beta)$ is constructed so that several functions of interest in the theorem correspond to specific values of parameters α and β.

Lemma 3 *Let assumptions of the care model stated in Section 2 hold and define a function* $g : [\underline{\tau}, \infty) \times [\underline{\mu}, \infty) \times [1/2, \infty) \times [0, \infty) \to \mathbb{R}$ *with mapping rule*

$$g(\tau, \mu; \alpha, \beta) = \alpha \mathbb{E}Q_P(\tau, \mu) + \beta \mathbb{E}Q_G(\tau, \mu) - k_G(\tau) - k_{SF}(\mu) - n\Lambda(\tau) k_{SV}. \tag{EC.25}$$

Let $\tilde{\tau}(\alpha, \beta)$ *and* $\tilde{\mu}(\alpha, \beta)$ *be functions such that* $(\tilde{\tau}(\alpha, \beta), \tilde{\mu}(\alpha, \beta)) \in \arg\max_{\tau, \mu} g(\tau, \mu; \alpha, \beta)$. *Then,* $\frac{\partial}{\partial \beta} \tilde{\tau}(\alpha, \beta) > 0$ *and* $\frac{\partial}{\partial \beta} \left(\mathbb{E}Q_P(\tilde{\tau}(\alpha, \beta), \tilde{\mu}(\alpha, \beta))\right) > 0$. *Also, let* $\hat{\mu}(\tau, \alpha, \beta)$ *be a function such that* $\hat{\mu}(\tau, \alpha, \beta) \in \arg\max_{\mu} g(\tau, \mu; \alpha, \beta)$. *Then,* $\frac{\partial}{\partial \tau} \hat{\mu}(\tau, \alpha, \beta) < 0$. *Furthermore, if* $B = 0$ *or if*

$$\alpha \geq \beta \frac{\tau(1 - \phi) - \underline{\tau}\phi}{\tau(1 - \phi) + \underline{\tau}\phi}, \tag{EC.26}$$

then $\frac{\partial}{\partial \alpha} \tilde{\tau}(\alpha, \beta) > 0$, $\frac{\partial}{\partial \alpha} \left(\mathbb{E}Q_P(\tilde{\tau}(\alpha, \beta), \tilde{\mu}(\alpha, \beta))\right) > 0$ *and* $\frac{\partial}{\partial \tau} \mathbb{E}Q_P(\tau, \hat{\mu}(\tau, \alpha, \beta)) > 0$.

Proof of Lemma 3. As g is jointly concave in (τ, μ) and has an interior maximum (by Lemma 1), for any α and β, the sole maximizer of g is the unique solution of FOCs. Denote $G_1(\tau, \mu; \alpha, \beta) \doteq \frac{\partial}{\partial \tau} g(\tau, \mu; \alpha, \beta)$

and $G_2(\tau,\mu;\alpha,\beta)\doteq\frac{\partial}{\partial\mu}g(\tau,\mu;\alpha,\beta)$. By the implicit function theorem: $\tilde{\tau}(\alpha,\beta)$ and $\tilde{\mu}(\alpha,\beta)$ are well defined and unique for every $y\in\{\alpha,\beta\}$ and we can find the partial derivatives $\frac{\partial}{\partial y}\tilde{\tau}(\alpha,\beta)$, $\frac{\partial}{\partial y}\tilde{\mu}(\alpha,\beta)$ by solving the equation

$$\mathbf{J}\begin{bmatrix}\frac{\partial\tilde{\tau}}{\partial y} & \frac{\partial\tilde{\mu}}{\partial y}\end{bmatrix}^T=-\mathbf{G}_y,\qquad\text{where}\qquad\mathbf{J}=\begin{bmatrix}\frac{\partial G_1}{\partial\tau} & \frac{\partial G_1}{\partial\mu}\\ \frac{\partial G_2}{\partial\tau} & \frac{\partial G_2}{\partial\mu}\end{bmatrix}\qquad\text{and}\qquad\mathbf{G}_y=\begin{bmatrix}\frac{\partial G_1}{\partial y}\\ \frac{\partial G_2}{\partial y}\end{bmatrix},$$

with all of the partial derivatives evaluated at $(\tilde{\tau}(\alpha,\beta),\tilde{\mu}(\alpha,\beta);\alpha,\beta)$.

Part 1: derivatives w.r.t. α,β. We show the derivatives w.r.t. β, as stated in the Lemma; the derivatives w.r.t α can be obtained analogously. Denote by $\mathbf{J}_i(\mathbf{G}_y)$ the matrix $\mathbf{J}$ in which the i-th column is replaced by the vector $\mathbf{G}_y$. Cramer's rule gives us $\frac{\partial}{\partial\beta}\tilde{\tau}(\alpha,\beta)=-\text{Det}(\mathbf{J}_1(\mathbf{G}_\beta))/\text{Det}(\mathbf{J})$ and $\frac{\partial}{\partial\beta}\tilde{\mu}(\alpha,\beta)=-\text{Det}(\mathbf{J}_2(\mathbf{G}_\beta))/\text{Det}(\mathbf{J})$. Denote $T=(\tau(1-\phi)+\underline{\tau}\phi)$ and $\Psi=\mu\tau\underline{\tau}-n(\tau(1-\phi)+\underline{\tau}\phi)$ and notice that $\Psi>0$ (from steady state condition and $\tau\geq\underline{\tau}$), then $\text{Det}(\mathbf{J}_1(\mathbf{G}_\beta))=-\phi q_P^0[Bnq_P^0\tau^2\underline{\tau}^2(2A(\alpha T+\beta\underline{\tau}\phi)\Psi+B\tau\underline{\tau}(\alpha(2T-\underline{\tau}\phi)+\beta\underline{\tau}\phi))+k''_{SF}(\mu)\Psi^2\left(A\Psi^2+B\tau^2\underline{\tau}(\mu\underline{\tau}+n(\phi-1))\right)]/[\tau^2\Psi^4]<0$. We also have $\text{Det}(\mathbf{J})>0$ because g is jointly concave (Lemma 1). Thus $\frac{\partial}{\partial\beta}\tilde{\tau}(\alpha,\beta)>0$. Now we wish to show that $\frac{\partial}{\partial\beta}\left(\mathbb{E}Q_P(\tilde{\tau}(\alpha,\beta),\tilde{\mu}(\alpha,\beta))\right)>0$. From (4), we have

$$\mathbb{E}Q_P(\tilde{\tau}(\alpha,\beta),\tilde{\mu}(\alpha,\beta))=q_P^0\left(1-A\Lambda(\tilde{\tau}(\alpha,\beta))-\frac{B}{\tilde{\mu}(\alpha,\beta)/(\Lambda(\tilde{\tau}(\alpha,\beta))-n}\right).\tag{EC.27}$$

From (EC.27) it can be seen that showing $\Lambda(\tilde{\tau}(\alpha,\beta))\downarrow\beta$ and $\tilde{\mu}(\alpha,\beta)/\Lambda(\tilde{\tau}(\alpha,\beta))\uparrow\beta$ is sufficient to demonstrate that $\mathbb{E}Q_P(\tilde{\tau}(\alpha,\beta),\tilde{\mu}(\alpha,\beta))\uparrow\beta$. From $\Lambda(\tilde{\tau}(\alpha,\beta))=(1-\phi)/\underline{\tau}+\phi/\tilde{\tau}(\alpha,\beta)$, we have $\frac{\partial}{\partial\beta}\Lambda(\tilde{\tau}(\alpha,\beta))=-\phi\tilde{\tau}(\alpha,\beta)^{-2}\frac{\partial}{\partial\beta}\tilde{\tau}(\alpha,\beta)<0$ and $\frac{\partial}{\partial\beta}\left(\tilde{\mu}(\alpha,\beta)/\Lambda(\tilde{\tau}(\alpha,\beta))\right)=\underline{\tau}\left[\underline{\tau}\phi\tilde{\tau}(\alpha,\beta)\frac{\partial}{\partial\beta}\tilde{\mu}(\alpha,\beta)+(1-\phi)\tilde{\tau}(\alpha,\beta)^2\frac{\partial}{\partial\beta}\tilde{\mu}(\alpha,\beta)+\underline{\tau}\phi\tilde{\mu}(\alpha,\beta)\frac{\partial}{\partial\beta}\tilde{\tau}(\alpha,\beta)\right]/[-\phi\tilde{\tau}(\alpha,\beta)+\tilde{\tau}(\alpha,\beta)+\underline{\tau}\phi]^2$, thus

$$\begin{aligned}
\text{sgn}\left(\frac{\partial}{\partial\beta}\left(\frac{\tilde{\mu}(\alpha,\beta)}{\Lambda(\tilde{\tau}(\alpha,\beta)}\right)\right)&=\text{sgn}\left((\underline{\tau}\phi\tilde{\tau}(\alpha,\beta)+(1-\phi)\tilde{\tau}(\alpha,\beta)^2)\frac{\partial}{\partial\beta}\tilde{\mu}(\alpha,\beta)+\underline{\tau}\phi\tilde{\mu}(\alpha,\beta)\frac{\partial}{\partial\beta}\tilde{\tau}(\alpha,\beta)\right),\\
&=\text{sgn}\left((\underline{\tau}\phi\tau+(1-\phi)\tau^2)\frac{-\text{Det}(\mathbf{J}_2(\mathbf{G}_\beta))}{\text{Det}(\mathbf{J})}+\underline{\tau}\phi\mu\frac{-\text{Det}(\mathbf{J}_1(\mathbf{G}_\beta))}{\text{Det}(\mathbf{J})}\right)\Big|_{(\tau,\mu)=(\tilde{\tau}(\alpha,\beta),\tilde{\mu}(\alpha,\beta))},\\
&=\text{sgn}\Big(\frac{q_P^0\underline{\tau}\phi}{\tau^2\Psi^4}\Big(B\tau^4\underline{\tau}k''_G(\tau)T\Psi^2+\phi\Big(\mu k''_{SF}(\mu)\Psi^2\left(A\Psi^2+B\tau^2\underline{\tau}(\mu\underline{\tau}+n(\phi-1))\right)\\
&\quad+B\tau\underline{\tau}\Big(q_P^0\Big(A(3\alpha T+\beta(\tau(1-\phi)+3\underline{\tau}\phi))\Psi^2+B\tau^2\underline{\tau}\big(T(\alpha+\beta)(\mu\underline{\tau}-n(1-\phi))\\
&\quad+\mu\underline{\tau}\big(\alpha(\tau(1-\phi)+T)+\beta\underline{\tau}\phi\big)\big)\Big)+2k_{SV}nT\Psi^2\Big)\Big)\Big)\Big|_{(\tau,\mu)=(\tilde{\tau}(\alpha,\beta),\tilde{\mu}(\alpha,\beta))}.
\end{aligned}$$

Because $\phi\in(0,1]$ gives us $(1-\phi)\geq0$ and $\mu\underline{\tau}-n(1-\phi)\geq\mu\underline{\tau}-n>0$, the partial derivative above is strictly positive, and consequently $\mathbb{E}Q_P(\tilde{\tau}(\alpha,\beta),\tilde{\mu}(\alpha,\beta))\uparrow\beta$.

Part 2: properties of $\hat{\mu}(\tau,\alpha,\beta)$. As g is concave in μ and has an interior maximum (Using Lemma 1), for any given α, β, and τ, the sole maximizer of g is the unique solution of its FOC, thus by implicit function theorem: $\hat{\mu}(\tau,\alpha,\beta)$ is well defined and unique and its partial derivative w.r.t. τ is given by

$$\frac{\partial}{\partial\tau}\hat{\mu}(\tau,\alpha,\beta)=-\frac{\frac{\partial^2}{\partial\mu\partial\tau}g(\tau,\mu;\alpha,\beta)}{\frac{\partial^2}{\partial\mu^2}g(\tau,\mu;\alpha,\beta)}\Big|_{\mu=\hat{\mu}(\tau,\alpha,\beta)}\tag{EC.28}$$

The denominator above is negative, as g is concave in μ (by Lemma 1). The numerator is given by $\frac{\partial^2}{\partial\mu\partial\tau}g(\tau,\mu;\alpha,\beta)=-\Psi^{-3}Bq_P^0\underline{\tau}^2\phi(\mu\tau\underline{\tau}+n(\tau(\phi-1)+\underline{\tau}\phi))$, which is negative as $(\mu\tau\underline{\tau}+n(\tau(\phi-1)+$

$\underline{\tau}\phi)) > (\mu\tau\underline{\tau} + n(\tau(\phi-1) - \underline{\tau}\phi)) = \Psi > 0$, hence (EC.28) is negative. It remains only to examine if $\frac{\partial}{\partial\tau}\mathbb{E}Q_P(\tau,\hat{\mu}(\tau,\alpha,\beta)) > 0$. From (4), we have

$$\mathbb{E}Q_P(\tau,\hat{\mu}(\tau,\alpha,\beta)) = q_P^0\left(1 - A\Lambda(\tau) - \frac{B}{\hat{\mu}(\tau,\alpha,\beta)/\Lambda(\tau) - n}\right),$$

$$\frac{\partial}{\partial\tau}\mathbb{E}Q_P(\tau,\hat{\mu}(\tau,\alpha,\beta)) = \frac{q_P^0\left(An^2\phi T^2 + B\tau^3\underline{\tau}T\frac{\partial}{\partial\tau}\hat{\mu}(\tau,\alpha,\beta) + \tau\underline{\tau}\phi\mu(-2AnT + B\tau\underline{\tau}) + A\tau^2\underline{\tau}^2\phi\mu^2\right)}{\tau^2(\tau\underline{\tau}\mu - nT)^2}\Bigg|_{\mu=\hat{\mu}(\tau,\alpha,\beta)},$$

$$\operatorname{sgn}\left(\frac{\partial}{\partial\tau}\mathbb{E}Q_P(\tau,\hat{\mu}(\tau,\alpha,\beta))\right) = \operatorname{sgn}\left(\phi\Psi\left(A\Psi + \frac{B\tau^2\underline{\tau}^2\left(Bq_P^0\tau\underline{\tau}(\alpha T + \beta(\tau(\phi-1)+\underline{\tau}\phi)) + \mu k''_{SF}(\mu)\Psi^2\right)}{2Bq_P^0\tau^2\underline{\tau}^2(\alpha T + \beta\underline{\tau}\phi) + k''_{SF}(\mu)\Psi^3}\right)\right)\Bigg|_{\mu=\hat{\mu}(\tau,\alpha,\beta)},$$

which is strictly positive if $B=0$ or if (EC.26) holds as (EC.26) guarantees that $(\alpha T + \beta(\tau(\phi-1)+\underline{\tau}\phi))) \geq 0$. Hence, (EC.26) implies $\frac{\partial}{\partial\tau}\mathbb{E}Q_P(\tau,\hat{\mu}(\tau,\alpha,\beta)) > 0$, as does $B=0$. □

Proof of Theorem 2. We show all parts of the theorem for the case when the GP holds a capitation contract ($k=C$). The proof for the case when the GP holds a per-patient contract ($k=P$) is analogous.

From (23)–(24): if the SP holds v_{SC}, colluding agents will make decisions $\tilde{\tau}^{UC}_{C,C}$,$\tilde{\mu}^{UC}_{C,C}$ which solve $\tilde{\tau}^{C}_{C,C} \in \arg\max_{\tau\in[\underline{\tau},\infty)} \mathbb{E}\left[\frac{1}{2}\sum_{i\in\{1,2\}}\left(v_{iC}(\mathcal{Q}_P(\tau,\tilde{\mu}^C_{C,C})) - V_i\right) + V_G - k_G(\tau)\right]$ and $\tilde{\mu}^{C}_{C,C} \in \arg\max_{\mu\in[\underline{\mu},\infty)} \mathbb{E}\left[\frac{1}{2}\sum_{i\in\{1,2\}}\left(v_{iC}(\mathcal{Q}_P(\tilde{\tau}^C_{C,C},\mu)) - V_i\right) + V_S - k_{SF}(\mu) - \mathcal{S}(\tilde{\tau}^C_{C,C})k_{SV}\right]$. Using expressions for parameters of v_{GC}, v_{SC} in (EC.13) and (EC.19), the system above is equivalent to

$$(\tilde{\tau}^C_{C,C},\tilde{\mu}^C_{C,C}) \in \underset{(\tau,\mu)\in[\underline{\tau},\infty)\times[\underline{\mu},\infty)}{\arg\max} \frac{1}{2}(1+nr_{GC})\mathbb{E}Q_P(\tau,\mu) - k_G(\tau) - k_{SF}(\mu). \tag{EC.29}$$

Here, r_{GC} is the rate which GP reimbursement is adjusted to outcomes, under the naïve capitation contract, as given by (13); explicit expression for it is given by (EC.13). Analogously, if the SP holds v_{SP}, colluding agents will make decisions $\tilde{\tau}^C_{C,P}$,$\tilde{\mu}^C_{C,P}$ given by

$$(\tilde{\tau}^C_{C,P},\tilde{\mu}^C_{C,P}) \in \underset{(\tau,\mu)\in[\underline{\tau},\infty)\times[\underline{\mu},\infty)}{\arg\max} \frac{r_{GC}n}{2}\mathbb{E}Q_P(\tau,\mu) + \frac{1}{2}\mathbb{E}Q_S(\tau,\mu) - k_G(\tau) - k_{SF}(\mu). \tag{EC.30}$$

Applying (EC.6), we have $\mathbb{E}Q_P(\tau,\mu) = (1-\Lambda(\tau))q_P^0 + \mathbb{E}Q_S(\tau,\mu)$, so (EC.30) is equivalent to $(\tilde{\tau}^C_{C,P},\tilde{\mu}^C_{C,P}) \in \arg\max_{(\tau,\mu)\in[\underline{\tau},\infty)\times[\underline{\mu},\infty)} ((1+nr_{GC})\mathbb{E}Q_P(\tau,\mu)/2 - k_G(\tau) - k_{SF}(\mu)) + \Lambda(\tau)q_P^0/2$. Here, the term in brackets is the objective function of (EC.29): a jointly concave function (by Lemma 1) with a sole interior maximum at $(\tilde{\tau}^C_{C,C},\tilde{\mu}^C_{C,C})$. The term $\Lambda(\tau)q_P^0/2$ is a univariate, decreasing, and convex function of τ. Consequently, as the objective function in (EC.29) is a sum of jointly concave and univariate decreasing convex, it either has an interior maximum with $\tilde{\tau}^C_{C,P} \in (\underline{\tau},\tilde{\tau}^C_{C,C})$ or a corner maximum in which case $\tilde{\tau}^C_{C,P} = \underline{\tau}$. In either case $\tilde{\tau}^C_{C,P} < \tilde{\tau}^C_{C,C}$, showing part (i) of the theorem. Because the objective function of (EC.29) is submodular and $\Lambda(\tau)$ is independent of μ, by Topkis (1978), we have $\tilde{\mu}^C_{C,P} > \tilde{\mu}^C_{C,C}$, showing part (ii). Using the function $\hat{\mu}$ as defined in Lemma 3 when $k_{SV}=0$ yields $\tilde{\mu}^C_{C,P} = \hat{\mu}(\tilde{\tau}^C_{C,P},(1+nr_{GC})/2,0)$, $\tilde{\mu}^C_{C,C} = \hat{\mu}(\tilde{\tau}^C_{C,C},(1+nr_{GC})/2,0)$. Furthermore, because $\tilde{\tau}^C_{C,C} > \tilde{\tau}^C_{C,P}$ and $\mathbb{E}Q_P(\tau,\hat{\mu}(\tau,(1+nr_{GC})/2,0)) \uparrow \tau$ (by Lemma 3), part (iii) follows. □

Proof of Prop. 3. Let $k_{SV}=0$ and let both agents hold capitation contracts (v_{GC} for the GP and v_{SC} for the SP). From (EC.13) we see that $k_{SV}=0$ implies $r_{GC}=1/n$. From (23)–(24) and the expressions for v_{GC}, v_{SC} in (EC.13), the decisions under collusion are the solutions of the system:

$$\tilde{\tau}^C_{C,C} \in \underset{\tau\in[\underline{\tau},\infty)}{\arg\max}\, \mathbb{E}Q_P(\tau,\tilde{\mu}^C_{C,C}) - k_G(\tau), \qquad \tilde{\mu}^C_{C,C} \in \underset{\mu\in[\underline{\mu},\infty)}{\arg\max}\, \mathbb{E}Q_P(\tilde{\tau}^C_{C,C},\mu) - k_{SF}(\mu). \tag{EC.31}$$

Applying Lemma 1, this system is equivalent to $(\tilde{\tau}^C_{C,C}, \tilde{\mu}^C_{C,C}) \in \arg\max_{(\tau,\mu)\in[\underline{\tau},\infty)\times[\underline{\mu},\infty)} \mathbb{E}Q_P(\tau,\mu) - k_G(\tau) - k_{SF}(\mu)$, which is just the system efficiency function as given by (5). Thus by definition of first-best decisions, we have $(\tilde{\tau}^C_{C,C}, \tilde{\mu}^C_{C,C}) = (\tau^*, \mu^*)$.

Now, consider the situation when the GP holds v_{GP} instead of v_{GC}. If we also have $\phi = 1$, then from (EC.16), it follows that $r_{GP} = 1/n$, and from (1), (4), (6) that $\mathbb{E}Q_G(\tau,\mu) = \mathbb{E}Q_P(\tau,\mu)$. Therefore, applying the expressions for v_{GP}, v_{SC} given in Prop. 1 and (EC.16)-(EC.19), we get that $(\tau^C_{P,C}, \mu^C_{P,C})$ also solve (EC.31) and thus are equal to (τ^*, μ^*). □

A.3. Proofs for Claims in Section 5 with Endogenous Free-riding and Collusion.

Proof of Theorem 3. Denote the index of the contract's metric of choice as $i \in \{P, A\}$. The induced GP decision $\tilde{d}_G$ satisfies the FOC of the first incentive compatibility constraint (26): $\frac{\partial}{\partial d_G}\mathbb{E}v(\mathcal{Q}_i(\tilde{d}_G, \tilde{d}_S)) = 2k'_G(\tilde{d}_G)$. As the error term is independent of provider decisions, we can express the signal as $\mathcal{Q}_i(\tilde{d}_G, \tilde{d}_S) = \mathbb{E}Q_i(\tilde{d}_G, \tilde{d}_S) + \varepsilon_i$, where ε_i is the stochastic error term. Using this decomposition, we can rewrite the FOC as $\mathbb{E}\left[v'(\mathcal{Q}_i(\tilde{d}_G, \tilde{d}_S))\frac{\partial}{\partial d_G}\left(\mathbb{E}Q_i(\tilde{d}_G, \tilde{d}_S) + \varepsilon_i\right)\right] = 2k'_G(\tilde{d}_G)$, or

$$\mathbb{E}\left[v'(\mathcal{Q}_i(\tilde{d}_G, \tilde{d}_S))\right] = \frac{2k'_G(\tilde{d}_G)}{\frac{\partial}{\partial d_G}\mathbb{E}Q_i(\tilde{d}_G, \tilde{d}_S)}. \tag{EC.32}$$

Doing the same for the second incentive compatibility constraint (27) yields

$$\mathbb{E}\left[v'(\mathcal{Q}_i(\tilde{d}_G, \tilde{d}_S))\right] = \frac{2k'_{SF}(\tilde{d}_S)}{\frac{\partial}{\partial d_S}\mathbb{E}Q_i(\tilde{d}_G, \tilde{d}_S)}. \tag{EC.33}$$

If $i = P$ the performance of this contract can be replicated by a linear contract $v^\dagger(q_P) \doteq a + b q_P$ by setting $b \doteq 2k'_G(\tilde{d}_G)/\frac{\partial}{\partial d_G}\mathbb{E}Q_P(\tilde{d}_G, \tilde{d}_S)$. Using Lemma 1, such a $v^\dagger$ induces the unique solution of the very same FOCs, and is thus guaranteed to also induce decisions $\tilde{d}_G, \tilde{d}_S$. From (EC.22), we have $\frac{\partial}{\partial d_G}\mathbb{E}Q_A(d_G, d_S) = \left(\frac{\partial}{\partial d_G}\mathbb{E}Q_P(d_G, d_S)\right)$; thus, an identical contract works if $i = A$.

We can also ensure that this is done at the lowest possible cost to the principal by lowering the fixed pay a until one of the participation constraints (28)-(29) is binding. Analogously to proof of Prop. 1, $v^\dagger$ can also be expressed as outcomes-adjusted capitation contract so that $v^\dagger(q_P) = n(f + r(q_P - t))$, which is then jointly concave. □

Proof of Prop. 4. Equating the RHS of (EC.32) and (EC.33), it follows that

$$\frac{\partial \mathbb{E}Q_P(d_G, d_S)}{\partial d_S} \Big/ \frac{\partial \mathbb{E}Q_P(d_G, d_S)}{\partial d_G} = k'_{SF}(d_S)/k'_G(d_G). \tag{EC.34}$$

Recall that the range of possible expected health outcomes is $[\mathbb{E}Q_P(0,0), q^0_P)$. To show that any expected health outcome is inducible, we first observe that from incentive compatibility constraints

(26)-(27) and Lemma 1, it follows that an outcomes-adjusted capitation contract with reimbursement rate r will induce decisions that solve

$$\max_{(d_G,d_S)\in[0,\infty)^2} rn\mathbb{E}Q_p(d_G,d_S) - k_G(d_G) - k_{SF}(d_S). \tag{EC.35}$$

Denote by $d_G(r)$ and $d_S(r)$ decisions induced by a contract with reimbursement rate r. Trivially, we have $d_G(0) = d_S(0) = 0$. To see what happens when we increase r, we can use the function g defined in Lemma 3, which equals the objective function of (EC.35) when $\alpha = rn$, $\beta = 0$ and $k_{SV} = 0$. From Lemma 3, we have $\frac{d}{dr}\mathbb{E}Q_P(d_G(r),d_S(r)) > 0$ and $\lim_{r\to\infty}\mathbb{E}Q_P(d_G(r),d_S(r)) = q_P^0$; thus, by intermediate value theorem, all possible values of $\mathbb{E}Q_P(d_G,d_S)$ can be attained by adjusting the value of r.

The cost-efficient way of attaining $\mathbb{E}Q_P(\tilde{d}_G,\tilde{d}_S)$ is the solution of the constrained problem $(d_G^\dagger,d_S^\dagger) \in \arg\min_{(d_G,d_S)\in[0,\infty)^2} k_G(d_G) + k_{SF}(d_S) + k_{SV}n\Lambda(d_G)$, s.t. $\mathbb{E}Q(d_G,d_S) = \mathbb{E}Q(\tilde{d}_G,\tilde{d}_S)$. Cost inefficiency then follows from a contradiction between the KKT conditions of this problem and (EC.34). □

Proof of Prop. 5. From (EC.32) and (EC.33) and Prop. 4, it follows that the reimbursement rate of the optimal outcomes-adjusted capitation contract satisfies $r = 2k_G'(\tilde{d}_G)/\left(n\frac{\partial}{\partial d_G}\mathbb{E}Q_P(\tilde{d}_G,\tilde{d}_S)\right) = 2k_{SF}'(\tilde{d}_S)/\left(n\frac{\partial}{\partial d_S}\mathbb{E}Q_P(\tilde{d}_G,\tilde{d}_S)\right)$, while from the proof of Prop. 2, we have that the optimal reimbursement rates of the naïve contracts satisfy $r_{AC} = 1/n$, $r_{AP} = 1/(n(1-(1-\bar{\lambda})(1-\phi)))$. Equality between the two does not hold except in special cases. The non-achievement of first-best then follows directly from the cost inefficiency shown in Prop. 4, part 2.

To show the rent, denote by $v^\dagger(q_P) = a + bq_P$ the optimal group contract (following the notation of Theorem 3), and by $\tilde{d}_G,\tilde{d}_S$ the decisions induced by it. From the proof of Theorem 3 (the step when a is set so that at least one of the participation constraints bind), it follows that $a = V_G + V_S - b\mathbb{E}Q_P(\tilde{d}_G,\tilde{d}_S) + 2\max\{k_G(\tilde{d}_G), k_{SF}(\tilde{d}_S) + k_{SV}n\Lambda(\tilde{d}_G)\}$. The statement of the proposition follows by inserting this expression and $v^\dagger(q_P) = a + bq_P$ into the constraints (28)-(29). □

Proof of Theorem 4. *Part 1.* Let $v(q_P)$ be the optimal group contract. Note that we can assume without loss of generality that its argument is q_P, as any group contract which is a function of q_A can be expressed using (EC.22) as a function of q_P instead. Denote by $v^\dagger(q_P) = a + bq_P$ the linear capitation contract which replicates the performance of $v(q_P)$ (as introduced in the proof of Theorem 3).

Consider now the situation when the following two individual contracts are given to the agents: $v_1(q_P) \doteq a_G + \frac{b}{2}q_P$ for the GP and $v_2(q_P) \doteq a_S + \frac{b}{2}q_P$ for the SP (for the moment, a_G and a_S are undefined constants). Under these contracts, the incentive compatibility constraints without collusion (31)-(32) and under collusion (33)-(34) are equivalent, thus the same provider decisions

will be induced no matter if the agents collude. They are also equivalent to incentive compatibility constraints in the group problem (26)-(27), when the group holds $v^{\dagger}(\boldsymbol{q}_P)$, so this set of individual contracts will induce the same decisions as $v^{\dagger}(\boldsymbol{q}_P)$ (thus also $v(\boldsymbol{q}_P)$). Because the induced decisions are the same irrespective of collusion occurring, then so will be the sum of the agent's incomes. Consequently, these contracts will be collusion-proof as there is no way that the same joint income can be split in such a way to make both of the agents better off when colluding. The constants a_G and a_S can be set to make participation constraint (36) binding, which we can do with $a_G \doteq V_G + k_G(\tilde{d}_G^{NC}) - \frac{b}{2}\mathbb{E}Q_P(\tilde{d}_G^{NC}, \tilde{d}_S^{NC})$ and $a_S \doteq V_S + k_{SF}(\tilde{d}_S^{NC}) + n\Lambda(\tilde{d}_G^{NC})k_{SV} - \frac{b}{2}\mathbb{E}Q_P(\tilde{d}_G^{NC}, \tilde{d}_S^{NC})$. Finally, with these a_G, a_S, the contracts $v_1(\boldsymbol{q}_P)$ and $v_2(\boldsymbol{q}_P)$ induce the same decisions and outcomes as $v(\boldsymbol{q}_P)$, but unlike $v(\boldsymbol{q}_P)$ do so in a cost-efficient way, completing part 1 of the theorem.

Part 2. Let $v_1(\boldsymbol{q}_P)$ and $v_2(\boldsymbol{q}_P)$ be the optimal contracts. We first show that collusion-proof contracts are optimal. Assume $v_1(\boldsymbol{q}_P)$ and $v_2(\boldsymbol{q}_P)$ are such that the agents will collude under them. Then, taking partial derivatives of the collusive incentive compatibility constraints (33)-(34) yields FOCs $\mathbb{E}[v_G'(\mathcal{Q}_P(\tilde{d}_G^C, \tilde{d}_S^C)) + v_S'(\mathcal{Q}_P(\tilde{d}_G^C, \tilde{d}_S^C))] = 2k_G'(\tilde{d}_G^C)/\frac{\partial}{\partial d_G}\mathbb{E}Q_P(\tilde{d}_G^C, \tilde{d}_S^C)$, and $\mathbb{E}[v_G'(\mathcal{Q}_P(\tilde{d}_G^C, \tilde{d}_S^C)) + v_S'(\mathcal{Q}_P(\tilde{d}_G^C, \tilde{d}_S^C))] = 2k_{SF}'(\tilde{d}_S^C)/\frac{\partial}{\partial d_S}\mathbb{E}Q_P(\tilde{d}_G^C, \tilde{d}_S^C)$. Define individual outcomes-adjusted capitation contracts $\bar{v}_1(\boldsymbol{q}_P) \doteq \bar{a}_G + \bar{b}\boldsymbol{q}_P$ and $\bar{v}_2(\boldsymbol{q}_P) \doteq \bar{a}_S + \bar{b}\boldsymbol{q}_P$ where $\bar{b} \doteq \mathbb{E}\left[v_G'(\mathcal{Q}_P(\tilde{d}_G^C, \tilde{d}_S^C)) + v_S'(\mathcal{Q}_P(\tilde{d}_G^C, \tilde{d}_S^C))\right]/2$, $\bar{a}_G \doteq -\bar{b}\mathbb{E}Q_P(\tilde{d}_G^C, \tilde{d}_S^C) + V_G + k_G(\tilde{d}_G^C)$, and $\bar{a}_S \doteq V_S + k_{SF}(\tilde{d}_S^C) + n\Lambda(\tilde{d}_G^C)k_{SV} - \bar{b}\mathbb{E}Q_P(\tilde{d}_G^C, \tilde{d}_S^C)$. In this notation, $\tilde{d}_G^C, \tilde{d}_S^C$ are still the (collusive) decisions induced by the optimal contracts $v_1(\boldsymbol{q}_P)$ and $v_2(\boldsymbol{q}_P)$. It is easily verifiable that such defined $\bar{v}_1(\boldsymbol{q}_P)$ and $\bar{v}_2(\boldsymbol{q}_P)$ will also induce decisions $\tilde{d}_G^C, \tilde{d}_S^C$, but will be collusion-proof and cost-efficient. Thus, not only are collusion-proof contracts optimal, any outcome of collusion-inducing contracts can also be replicated with collusion-proof capitation ones at the same or lower cost. Completely analogously, it can be shown that the performance of any optimal collusion-proof contracts can also be replicated by collusion-proof linear capitation contracts. Thus, collusion-proof linear capitation contracts are optimal in all cases.

Part 3. The naïve contracts are guaranteed to solve the non-collusive incentive compatibility constraints (31)-(32), as those are equivalent (9)-(10) in the naïve problem. Naïve participation constraints (9)-(10) imply the MICP participation constraint (36) irrespective of whether collusion occurs. Additionally, if the contracts are collusion proof, then if (9)-(10) bind, then so does (36). Finally, constraints (33)-(35) are redundant for collusion-proof contracts. Thus, collusion-proof solutions of the NICP also solve the MICP, and the same incentive compatibility constraints induce the same decisions (the first-best ones). Because the participation constraints are binding, first-best is achieved in the MICP as well. □

A.4. Proofs for Claims in Section 7 about Implications for Practice.

Proof of Theorem 5. *Part* (i). Consider the NICP with the restriction $(v_1, v_2) \in \mathcal{G} \times \mathcal{F}$. Note that $v_2 \in \mathcal{F}$ implies that the objective function (10) is decreasing in d_S (compensation is constant, but increasing quality of care is costly), so $\tilde{d_S} = 0$. Because $v_2 \in \mathcal{F}$, it is a constant function; we commit a slight abuse of notation by denoting that constant by v_2 also.

Note that (8) is decreasing in v_2. From (12), it follows that $v_2 = V_S + k_{SF}(0) + k_{SV}\Lambda(\tilde{d_G})n$, where V_S is the outside option for the SP, as above. Thus we can simplify (8)–(12) to

$$\max_{v_1 \in \mathcal{G}} \quad \mathbb{E}\left[Q_P(\tilde{d_G}, 0) - v_1(\mathcal{Q}_P(\tilde{d_G}, 0)) - k_{SV}\Lambda(\tilde{d_G})n\right] \tag{EC.36}$$

$$\text{s. t.} \quad \tilde{d_G} \in \arg\max_{d_G \in [0,\infty)} \mathbb{E}\left[v_1(\mathcal{Q}_P(d_G, 0)) - k_G(d_G)\right] \tag{EC.37}$$

$$V_G \le \mathbb{E}\left[v_1(\mathcal{Q}_P(\tilde{d_G}, 0)) - k_G(\tilde{d_G})\right]. \tag{EC.38}$$

Let $v_1^\dagger \in \mathcal{G}$ and denote by $d_G^\dagger$ the decision it induces. Using the assumption that $\mathcal{Q}_P(d_G, d_S) = \mathbb{E}Q_P(d_G, d_S) + \epsilon$, where ϵ is a zero-mean random variable, yields that if $d_G^\dagger > 0$, it solves the FOC $\mathbb{E}[(v_1^\dagger)'(\mathcal{Q}_P(d_G^\dagger, 0))](\mathbb{E}Q_P)'(d_G^\dagger, 0) - k_G'(d_G^\dagger) = 0$.

For any $a_C \in \mathbb{R}$, $b_C = \mathbb{E}[(v_1^\dagger)'(\mathcal{Q}_P(d_G^\dagger, 0))]$, the linear contract $v_1^{\dagger\dagger}(q_P) \doteq a_C + b_C q_P$ induces the same decision. Making the constant part of the contract $a_C \doteq V_G + k_G(d_G^\dagger) - \mathbb{E}[(v_1^\dagger)'(\mathcal{Q}_P(d_G^\dagger, 0))]\mathbb{E}Q_P(d_G^\dagger, 0)$ ensures this linear contract has the lowest possible cost to the principal. It will be useful to express $v_1^{\dagger\dagger}$ in a form like that in (13) (as an outcome-adjusted capitation contract). This can be done by setting

$$f_{GC} = V_G + k_G(d_G^\dagger), \quad r_{GC} = \mathbb{E}[(v_1^\dagger)'(\mathcal{Q}_P(d_G^\dagger, 0))], \quad t_{GC} = \mathbb{E}Q_P(d_G^\dagger, 0). \tag{EC.39}$$

The decisions induced by the optimal contract can also be induced by the linear contract in (EC.39). This reduces the principal's problem to $\max_{d_G^\dagger} \mathbb{E}Q_P(d_G^\dagger, 0) - V_G - k_G(d_G^\dagger) - n\Lambda(d_G^\dagger)k_{SV}$. From Lemma 1, this problem has a concave objective function with an interior solution. Thus, the optimal $d_G^\dagger$ is uniquely characterized by the FOC

$$\frac{\partial}{\partial d_G}\left(\mathbb{E}Q_P(d_G^\dagger, 0)\right) - k_G'(d_G^\dagger) - n\Lambda'(d_G^\dagger)k_{SV} = 0. \tag{EC.40}$$

Part (ii). From (30)-(36), the best collusion-proof contracts solve

$$\begin{aligned}
\max_{v_1 \in \mathcal{G}, v_2 \in \mathbb{R}} \quad & \mathbb{E}\left[Q_P(\tilde{d_G}, 0) - v_1(\mathcal{Q}_P(\tilde{d_G}, 0)) - v_2\right] \qquad \text{(EC.41)}\\
\text{subject to} \quad & \tilde{d_G^{NC}} \in \arg\max_{d_G \in [0,\infty)} \mathbb{E}\left[v_1(\mathcal{Q}_P(d_G, 0) - k_G(d_G)\right]\\
& \tilde{d_G^C} \in \arg\max_{d_G \in [0,\infty)} \mathbb{E}\left[\frac{1}{2}\left(v_1(\mathcal{Q}_P(d_G, \tilde{d_S^C})) + v_1(\mathcal{Q}_P(\tilde{d_G^{NC}}, 0))\right) - k_G(d_G)\right]\\
& \tilde{d_S^C} \in \arg\max_{d_S \in [0,\infty)} \mathbb{E}\left[\frac{1}{2}\left(v_1(\mathcal{Q}_P(\tilde{d_G^C}, d_S)) - v_1(\mathcal{Q}_P(\tilde{d_G^{NC}}, 0))\right) + v_2 - k_{SF}(d_S) - \mathcal{S}(\tilde{d_G^C})k_{SV}\right]\\
& \pi_G^{NC} \ge \pi_G^C \vee \pi_S^{NC} \ge \pi_S^C\\
& V_G \le \pi_G^{NC} \wedge V_S \le \pi_S^{NC}
\end{aligned}$$

We now relax the problem by dropping the constraints $\pi_G^{NC} \geq \pi_G^C \vee \pi_S^{NC} \geq \pi_S^C$. We observe that in this relaxed problem, the principal's objective function is decreasing in v_2, while the only constraint affected by v_2 is $V_S \leq \pi_S^{NC}$ where π_S^{NC} is increasing in v_2. Thus, we can assume without loss of generality that $V_S = \pi_S^{NC}$. This observation reduces the problem in (EC.41) to (EC.36)-(EC.38), which was solved in part (i) of this proof.

Part (iii). Denote by mMICP the modified MICP in the statement of part (iii). The statement follows from making two observations. First, if a contract that solves the mMICP is collusion-inducing, then it also solves the MICP with the $(v_1, v_2) \in \mathcal{G} \times \mathcal{F}$ restriction. Second, if a contract v_1 that solves the mMICP is *not* collusion-inducing, then there cannot exist a collusion-inducing contract that outperforms it (otherwise, the other contract would solve the mMICP).

Part (iv) follows from observing that if $(v_1, v_2) \in \mathcal{G} \times \mathcal{F}$ are collusion-inducing and would result in inducing (d_G, d_S), then $v_G(\boldsymbol{q}_P) \doteq v_1(\boldsymbol{q}_P) + v_2(\boldsymbol{q}_P)$ satisfies the constraints of the MGCP and also results in inducing (d_G, d_S) in that problem. In Prop. 4, all decisions that are inducible in the MGCP are on the same inefficient frontier (that is characterized by the differential equation in Prop. 4, part 1).

Part (v) follows by noting that adding or subtracting constant payments to either contract in the MICP changes not whether the contracts are collusion-proof or which decisions they induce. □

Proof of Theorem 6. We proceed directly, and assume the setup in the hypothesis of the theorem.

Part (i). From the objective function of (32), we have that the SP's income does not depend on d_S, but the SP's costs are increasing in d_S; thus, providing care of minimal quality is optimal for the SP ($\tilde{d}_S^{NC} = 0$). Inserting $\tilde{d}_S^{NC} = 0$ into (31) we notice that $\tilde{d}_G^{NC}$ does not depend on r.

We continue by demonstrating that if r is sufficiently high, collusion will lead to the GP making corner decision $\tilde{d}_G^C = 0$. Dropping the constant terms from (33) yields

$$\tilde{d}_G^C \in \arg\max_{d_G \in [0,\infty)} \frac{1}{2}\mathbb{E}\big[v_1(\mathcal{Q}_P(d_G, \tilde{d}_S^C) + rn\Lambda(d_G)\big] - k_G(d_G). \tag{EC.42}$$

Using $0 < v_1' \leq \Omega$ and the joint concavity of $\mathbb{E}Q_P$ established in Lemma 1, the slope of the objective function in (EC.42) is no greater than $\Omega(\partial \mathbb{E}Q_P(d_G, 0))/(\partial d_G) + rn\Lambda'(d_G) = \Lambda'(d_G)\big(rn - \Omega q_P^0\big(A + B\mu(0)/(\mu(0) - n\Lambda(d_G))^2\big)\big)$. Thus, for $r \geq r^\dagger \doteq \Omega q_0^P(A + B\mu(0)/(\mu(0) - n\Lambda(0))^2)/n$, the objective of (EC.42) is decreasing, so $\tilde{d}_G^C = 0$. This immediately implies that $\tilde{d}_S^C$ is independent of r for all $r \geq r^\dagger$.

To show that a contract pair (v_1, v_2) is collusion-inducing, we also need to demonstrate that both the GP and the SP are better off under collusion. From (34), when $r \geq r^\dagger$, colluding changes the SP's profit by

$$\begin{aligned}\frac{1}{2}\mathbb{E}\big[v_1(\mathcal{Q}_P(0, \tilde{d}_S^C)) - v_1(\mathcal{Q}_P(\tilde{d}_G^{NC}, 0)) + rn(\Lambda(0) - \Lambda(\tilde{d}_G^{NC}))\big] \\ + k_{SF}(\tilde{d}_S^C) - k_{SF}(0) + k_{SV}n(\Lambda(0) - \Lambda(\tilde{d}_G^{NC})).\end{aligned}$$

Here, the only term that depends on r is $rn(\Lambda(0)-\Lambda(\tilde{d}_G^{NC}))$, so the function is linearly increasing in r and thus unbounded. It follows that the SP is better off colluding for all sufficiently high r. Similarly, from (33), it follows that colluding changes the GP's profit by

$$\frac{1}{2}\mathbb{E}\big[v_1(\mathcal{Q}_P(0,\tilde{d}_S^C))-v_1(\mathcal{Q}_P(\tilde{d}_G^{NC},0))+rn(\Lambda(0)-\Lambda(\tilde{d}_G^{NC}))\big]+k_G(0)-k_G(\tilde{d}_G^{NC}),$$

which is also linearly increasing in r, so the same conclusion applies.

Part (ii). As in part (i) of this proof, we examine the GP's choice under collusion as given by (33). Taking an additional step that uses the property that $\mathbb{E}Q_P(d_G,d_S)=\mathbb{E}Q_S(d_G,d_S)+q_P^0(1-\Lambda(d_G))$ allows us to obtain that $d_C^G=0$ if $r\geq r^{\dagger\dagger}\doteq[\Omega(A+B\mu(\tilde{d}_S^C)/(\mu(\tilde{d}_S^C)-n\Lambda(0))^2]/[q_P^0-A-B/\mu(\tilde{d}_S^C)]$. Then, from (34), when $r\geq r^{\dagger\dagger}$, colluding changes the SP's profit by

$$\begin{aligned}\frac{1}{2}\mathbb{E}\big[v_1(\mathcal{Q}_P(0,\tilde{d}_S^C))-v_1(\mathcal{Q}_P(\tilde{d}_G^{NC},0))+r(\mathcal{Q}_S(0,\tilde{d}_S^C)-\mathcal{Q}_S(\tilde{d}_G^{NC},\tilde{d}_S^{NC}))\big]\\ +k_{SF}(\tilde{d}_S^C)-k_{SF}(\tilde{d}_S^{NC})+k_{SV}n(\Lambda(0)-\Lambda(\tilde{d}_G^{NC})).\end{aligned}$$

If $\mathbb{E}Q_S(\tilde{d}_G^C,\tilde{d}_S^C)-\mathbb{E}Q_S(\tilde{d}_G^{NC},\tilde{d}_S^{NC})>0$ and r is sufficiently high, this expression will be positive, thus collusion profitable for the SP. The profitability of collusion for the GP under these two conditions follows analogously.

Part (iii). Recall (32) and consider substituting v_2 with $v_2^*(\boldsymbol{q}_S)\doteq h_2^*+r_2^*\boldsymbol{q}_S$, where $r_2^*\doteq(\partial\mathbb{E}[v_2(\mathcal{Q}_S(\tilde{d}_G^*,\tilde{d}_S^*))])/(\partial d_S)$. This results in the decision $\tilde{d}_S^{NC}=\tilde{d}_S^*$.

Uniqueness up to the constant h_2^* follows from interior decisions that linear contracts induce being uniquely characterized by the FOCs of the objective functions. Consequently, (31) is the same under the two different contracts. This implies $\tilde{d}_G^{NC}=\tilde{d}_G^*$. Notice, however, that $\tilde{d}_G^C$ and $\tilde{d}_S^C$ may change with this change of contracts, and thus, so can the property of being collusion-proof. □

A.5. Proofs for Claims in Section 8 with Realized Costs or Yardstick Competition.

Proof of Prop. 6. Decisions made by agents holding a contract are determined by the incentive compatibility (IC) constraints (9)-(10). Inserting the contracts v_G and v_S into these constraints yields and taking partial derivatives yields FOCs that are equivalent to the FOCs (EC.9)-(EC.10), which determine the first-best decisions. Applying Lemma 1, the decisions induced by the contract are uniquely determined by the FOCs; thus, the contract achieves the first-best. The same holding in MICP if the contracts are collusion-proof then follows from equality between non-collusive IC constraints in the MICP and the IC constraints in the NICP. □

Proof of Prop. 7. *Part 1.* We show this part by backward recursion. Consider the last period N. As f_G^N and f_S^N are independent of provider decisions in this period, Prop. 6 applies and the providers will choose d_G^*, d_S^*. In the next to last period $(N-1)$, the providers' decisions affect not only the current period's payout but also the next one, thus they will choose decisions that solve

$$
\begin{aligned}
d_G^{N-1} \in \underset{d_G\in[0,\infty)}{\arg\max}\quad & f_G^{N-1} + \mathbb{E}Q_P(d_G, d_S^{N-1}) - n\Lambda(d_G)k_{SV} - k_G(d_G) \\
& + \delta\left(-\mathbb{E}Q_P(d_G, d_S^{N-1}) + V_G + n\Lambda(d_G)k_{SV} + k_G(d_G) + \mathbb{E}Q_P(d_G^*, d_S^*) - n\Lambda(d_G^*)k_{SV} - k_G(d_G^*)\right), \\
d_S^{N-1} \in \underset{d_S\in[0,\infty)}{\arg\max}\quad & f_S^{N-1} + \mathbb{E}Q_P(d_G^{N-1}, d_S) - n\Lambda(d_G^{N-1})k_{SV} - k_S(d_S) \\
& + \delta\left(-\mathbb{E}Q_P(d_G^{N-1}, d_S) + V_S + n\Lambda(d_G^{N-1})k_{SV} + k_S(d_S) + \mathbb{E}Q_P(d_G^*, d_S^*) - n\Lambda(d_G^*)k_{SV} - k_S(d_S^*)\right).
\end{aligned}
$$

In both of the expressions above, the first line is the current $(N-1)$ period payout, whereas the second line is for period N. Rearranging terms and dropping constants (which have no bearing on $\arg\max$) yields $d_G^{N-1} \in \arg\max_{d_G\in[0,\infty)}(1-\delta)\left(\mathbb{E}Q_P(d_G, d_S^{N-1}) - n\Lambda(d_G)k_{SV} - k_G(d_G)\right)$ and $d_S^{N-1} \in \arg\max_{d_S\in[0,\infty)}(1-\delta)\left(\mathbb{E}Q_P(d_G^{N-1}, d_S) - n\Lambda(d_G^{N-1})k_{SV} - k_S(d_S)\right)$, which is by Lemma 1 equivalent to the first-best system. Thus, the providers will choose the first-best decisions in period $N-1$. Inserting $d_G^{N-1} = d_G^*$ and $d_S^{N-1} = d_S^*$ into the profit calculation for period N also yields that both participation constraints are binding in period N. The statement of the proposition follows by repeating the same argument recursively for each preceding period.

Part 2. For $j \in \{1, .., M\}$, denote by $\bar{q}_P^j$ the realized population health of the population of patients served by the j-th GP-SP pair, and denote by ϱ_G^j, ϱ_{SV}^j, and ϱ_{SF}^j the cost realizations within that pair. Suppose now all providers are given outcomes-adjusted capitation contracts with per-capita rates

$$
\bar{q}_P^j + f_G^j - \varrho_{SV}^j/n \quad \text{(for GP } j\text{)}, \qquad \bar{q}_P^j + f_S^j \quad \text{(for SP } j\text{)}. \tag{EC.43}
$$

If f_G^j and f_S^j are independent of decisions made by provider pair j (d_G^j and d_S^j), then Prop. 6 applies and all providers making first-best decisions d_G^*, d_S^* will be in equilibrium under those contracts. The remaining step needed to achieve the first-best is to construct f_G^j and f_S^j so that the participation constraints bind, while preserving the property that f_G^j and f_S^j are independent of decisions of provider pair j. Taking an expectation of (EC.43) conditional on the providers making first-best decisions yields the expected equilibrium profits of GP j and SP j respectively: $\pi_G^j = \mathbb{E}\left[Q_P(d_G^*, d_S^*) + nf_G^j - \mathcal{S}(d_G^*)k_{SV}\right] - k_G(d_G^*)$, $\pi_S^j = \mathbb{E}\left[Q_P(d_G^*, d_S^*) + nf_S^j - \mathcal{S}(d_G^*)k_{SV}\right] - k_{SF}(d_S^*)$. Setting these equal to the value of the providers' outside option and solving for f_G^j and f_S^j yields

$$
\mathbb{E}f_G^j = V_G/n + \Lambda(d_G^*)k_{SV} + k_G(d_G^*)/n - \mathbb{E}Q_P(d_G^*, d_S^*)/n, \tag{EC.44}
$$

$$
\mathbb{E}f_S^j = V_S/n + \Lambda(d_G^*)k_{SV} + k_{SF}(d_S^*)/n - \mathbb{E}Q_P(d_G^*, d_S^*)/n. \tag{EC.45}
$$

Here, the principal cannot simply set f_G^j and f_S^j to be equal to the RHS of these two equations, as the principal lacks knowledge of the cost functions as well as d_G^* and d_S^*. However, following the idea of yardstick competition, the principal can exploit the property that for all provider

pairs other than the j-th, d_G^* and d_S^* are also equilibrium decisions. Thus, in equilibrium, we have $\mathbb{E}\hat{\varrho}_G^{-j} = k_G(d_G^*)$, $\mathbb{E}\hat{\varrho}_{SF}^{-j} = k_{SF}(d_S^*)$, $\mathbb{E}\hat{\varrho}_{SV}^{-j} = n\Lambda(d_G^*)k_{SV}$, and $\mathbb{E}\hat{\bar{\boldsymbol{q}}}_P^{-j} = \mathbb{E}Q_P(d_G^*, d_S^*)/n$.

Consequently, for every provider in pair j, we can define f_G^j and f_S^j by

$$f_G^j \doteq (V_G + \hat{\varrho}_G^{-j} + \hat{\varrho}_{SV}^{-j})/n - \hat{\bar{\boldsymbol{q}}}_P^{-j}, \quad f_S^j \doteq (V_S + \hat{\varrho}_{SF}^{-j} + \hat{\varrho}_{SV}^{-j})/n - \hat{\bar{\boldsymbol{q}}}_P^{-j},$$

which then ensures that (EC.44) and (EC.45) hold while preserving the property that f_G^j and f_S^j do not depend on the decisions of provider pair j. Thus, the participation constraints are binding, and the contract achieves first-best. □

Appendix B: Methodology of the Numerical Analysis

This section provides additional detail on how the plausible range was derived for each of the model's parameters. We first discuss how ranges of certain observable parameters were assessed for an example related to type 2 diabetes in the UK. We then discuss challenges associated with assessing cost functions. All numeric optimization used Mathematica's native NArgMax routine.

Note that parameters that may have a reasonable value when considered individually might not be reasonable when considered in combination with other parameters. For this reason, we then describe and use an acceptance-sampling algorithm designed to assess whether parameters, taken together, are reasonably consistent with observed data.

Number of patients. The complete distribution of the number of patients per full-time equivalent (FTE) GP is available from the NHS census of GPs in the UK (full data available from UK NHS 2019). This distribution has a mean of 2218, a median of 1866, and a standard deviation of 2127. However, it includes several outliers, most notably practices that have 0 patients and a few practices with an extremely high number of patients per FTE GP (more than 60,000). In order to eliminate these outliers, we use a truncated version of this distribution where the top 1% and bottom 1% have been removed. The full histogram of the resulting distribution is given in Figure EC.1.

We draw the total number of patients under the GP's care, including the ones without diabetes, from this distribution and denote that variable as m. The national rate of type 2 diabetes incidence is 6.4% (Diabetes UK 2019). To find the number of diabetes patients under a GP's care, we use a draw from $n \sim \text{Bin}(m, 6.4\%)$.

Health capital. One of our primary metrics of interest is the initial population health capital ($\bar{q}_P^0$), which measures the "present monetary value of a person's health," averaged over all members of the population, as per Grossman (1972). For each individual person, we measure health capital using the method of Cutler and Richardson (1998) as the expected discounted sum (with discount rate r) of their remaining years of life weighted by the quality of life in those years, and multiplied

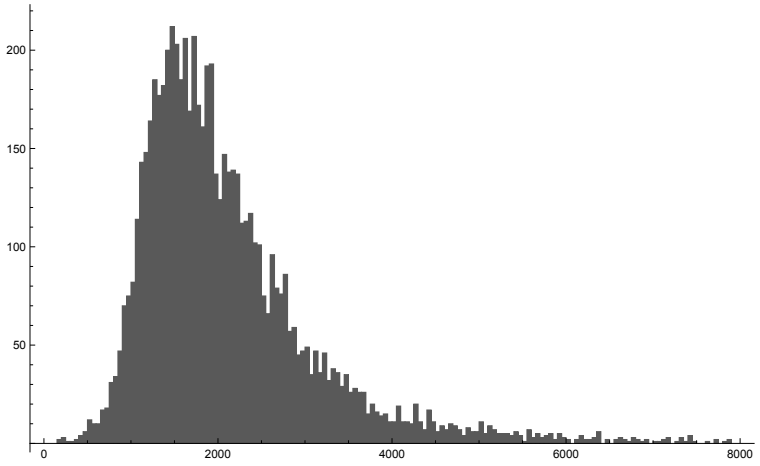


Figure EC.1 Histogram of the number of patients per FTE GP (including non-diabetic patients)

by the monetary value of one QALY. From Khalid et al. (2014), the average age of type 2 diabetes patients is 64.9 years. Based on Leal et al. (2009), the residual life expectancy of type 2 diabetes patients with moderate risk factors at 65 is 11.5 years. From Clarke et al. (2002), the mean quality of life of type 2 diabetes patients is 0.77. We use the discount rate of $r = 0.03$, as in many related papers, including Cutler and Richardson (1998). This gives us $q_0^P = \left(\sum_{i=0}^{10} \frac{0.77}{1.03^i} + \frac{0.77}{2\times 1.03^{11}}\right)$ QALY $\approx$ 7.61 QALY as an estimate of an average type 2 diabetes patient's health capital. Note that, as in Clarke et al. (2002), this is a (possibly biased) approximation; a better estimate could be acquired if the distribution of residual life was known rather than just the expectation.

Monetary Value of One QALY. The UK has used cost-effectiveness thresholds of £20,000 to £20,000 per QALY. US regulators typically use a valuation between \$50,000 to \$100,000 for one QALY (Neumann et al. 2014). The empirical study of Lee et al. (2009) estimates the implied value of one QALY at approximately \$129,000, based on the current treatment practices for end-stage renal patients in the US.

We do not randomly sample this parameter, as it is subject to the regulator's choice. We primarily use the £30,000 value (it is used for all the figures in the main text). However, to check robustness, we also conducted the numerical analysis for four different values of this choice: £20,000, £30,000, £70,000 ($\approx$ \$84,000 at January 2023 exchange rates), and £100,000 ($\approx$ \$120,000). Our results are qualitatively consistent across these choices, which is due to the relatively simple role of this parameter: it is a linear scaling parameter for health in the objective function. Whenever possible, we report the health-related results in terms of QALY instead of monetary terms to reduce the reliance on this measure.

Adherence rate. Non-adherence rates vary greatly in the literature depending exactly on how non-adherence is defined (McNabb 1997). Currie et al. (2012) find that 39% of patients are non-adherent if we consider clinical non-adherence (appointment no-shows), but only 4.4% if we look

at medication non-adherence (not taking prescribed medicine). Clinical non-adherence is the one which corresponds better to our model, because these are the patients who miss out on benefiting from a GP's treatment. More so, specific sub-populations can have vastly different adherence rates. Hospitalized patients have virtually full adherence as appointment no-shows are not an issue, and adherence to the medication regime is ensured by the nurses. On the other hand, adolescents have an exceedingly low adherence rate, which can go as low as 0.1 (Taddeo et al. 2008, Borus and Laffel 2010). To ensure that we consider the full range of possible adherence values, while still having more parameter sets with moderate adherence, we draw the adherence rate ϕ from a triangular distribution with support $[0.1, 1]$ and mode 0.61 (the point estimate for clinical adherence rate).

Complication rates. Mathur et al. (2017) study the 338,390 type-2 diabetes patients in the UK's Clinical Practice Research Datalink (representative of the whole population) and find the yearly probability of developing retinopathy to be 3.22%. However, they do not take adherence into account. Currie et al. (2012) find that clinically non-adherent patients have 10% higher mortality while medication-non-adherent ones have 30% higher. García-Pérez et al. (2013) find that medication non-adherent patients have $38\% - 58\%$ more hospitalizations. We use the extremes of these estimates (10%-58% higher) combined with a baseline complication rate of 3.22% (Mathur et al. 2017) to get the bounds for $\overline{\lambda}$. The calculation $[110\% * 3.22\%, 158\% * 3.22\%]$ yields $\overline{\lambda} \in [3.54\%, 5.08\%]$.

Service level. The lower bound for the service level ($\underline{\mu}$) can be found from the steady state condition $\underline{\mu} > n\Lambda(0)$. We ensure the steady state condition holds by setting the lower bound at 1% over the strict bound when $\bar{\lambda}$ is at the highest level in our parameter set ($\bar{\lambda} = 5.08\%$), i.e., $\underline{\mu} = 1.01n\Lambda(0) = 1.01n\bar{\lambda} = n * 5.1308\%$.

The health impact of complications. The effect of blindness on quality of life is debated (and depends on the person's ability to adjust) with estimates ranging from 34% to 64% loss of QALY (Javitt and Aiello 1996, Rein et al. 2007), we use $W \in [0.34, 0.64]$ to denote this value and will consider the entire range. Untreated diabetic retinopathy will progress to legal blindness in an estimated 3.2 years (Ferris 1993),[11] from which we have $b = W/3.2$. While early-stage retinopathy is asymptomatic, we use $a = b/2$ to reflect the expected time to initial diagnosis of 6 months (NHS conducts yearly retinopathy screening for diabetes patients).

Cost functions. There is good data availability on the realized costs of treating diabetes in the UK. Managing a single diabetes patient *without* complications costs $\sim$ £550 per year (Diabetes UK 2014). For retinopathy, there are two main treatment options: 1) laser therapy (pan-retinal

[11] In the absence of treatment, vision continues deteriorating even after the onset of legal blindness, progressing to total blindness eventually, with a QALY loss of 74% (Rein et al. 2007).

laser photocoagulation), which costs £823 for the treatment itself and £1112 for the ophthalmologist services, and 2) Ranibizumab (monoclonal antibody fragment, injected directly into the eye), costing £7422 for the medicine itself (10 injections), and £2149 for the ophthalmologist services (cost data from Mitchell et al. 2012). These treatments are also sometimes combined with total material costs of £7503 and ophthalmologist costs of £1853.[12]

This still leaves us with two challenges for calibrating costs. The first challenge is separating the variable from the fixed costs for the specialist. While the cost of drugs administered is certainly variable, physician time likely reflects both fixed and variable costs. We address this by modeling their sum and checking if the sum is reasonable. The other challenge is determining the shape of the cost function from the realized costs. We consider the range of variables costs in $[£823, £9356]$; the bottom of this range is the material cost of the cheapest treatment option (laser therapy), while the upper end of the range is the cost of the most expensive treatment option (laser plus ranibizumab) including ophthalmologist costs. We assume that the cost functions are scaled power functions ($k_G(d_G) = \gamma_1(d_G)^{\gamma_2}$, $k_{SF}(d_S) = \delta_1(d_G)^{\delta_2}$).

Parameter plausibility check. The question here is how to find a reasonable range for the hyper-parameters $\gamma_1, \gamma_2, \delta_1, \delta_2$. The parameter ranges are set to the wide range of $\gamma_1, \delta_1 \in [1, 10^6]$ and $\gamma_2, \delta_2, \in [1, 4]$, after which the following three plausibility checks are conducted: i) verifying whether the first-best complication rates under such parameters are really lower than the complication rates observed in the literature (Khalid et al. 2014), ii) verifying that waiting times for treatment under first-best are at least somewhat consistent with the waiting times in the literature (literature estimates range from a few days to several months, so parameter sets resulting in waiting times of more than a year or less than a day are deemed to be unrealistic), and iii) lastly, we verify that the costs under the first-best are no more than a factor of 4 away from the observed realized costs (Mitchell et al. 2012). Parameter draws that did not fit these criteria were dropped.

Algorithm 1 presents the simulation algorithm that we used to generate random samples of the values of all parameters for the model, to drop samples that do not meet plausibility checks such as those described in the preceding paragraph, and to assess the performance of each contract. For numerical comparisons reported in this paper, 20,000 parameter sets were generated, out of which 13,843 were dropped due to failing the plausibility checks, leaving 6,157 plausible parameter sets for analysis. Performance statistics for each contract were computed for results in the main paper and online companion (unless specified otherwise). If we look at the commonalities of dropped

[12] There is additional complexity with treatment options as two more options exist. One option is the use of intraocular corticosteroid drugs, including implants that slowly release them. The other option is ocular surgery (vitrectomy). Furthermore, two other drugs can be used in place of ranibizumab: aflibercept and bevacizumab (Maniadakis and Konstantakopoulou 2019).

Algorithm 1 Simulate Scenarios for Assessing Contract Performance

procedure SAMPLE PARAMETERS; PERFORM PLAUSIBILITY CHECK; ASSESS CONTRACTS

Set *size*; (number of parameter sets/scenarios to sample)

Set QALY; (the monetary value of one QALY)

Set GPdistrib; (the distribution of patients per FTE GP, imported from NHS data UK NHS (2019))

for $i = 1$ to *size* **do**

Step 1: draw pseudo-random parameter values from specified distributions

$m \leftarrow$ Random from GPdistrib; $n \leftarrow$ Random from Bin$(m, 0.064)$; $\overline{\lambda} \leftarrow$ Random$[3.54\%, 5.08\%]$;

$\phi \leftarrow$ Random from Triangular$(0.1, 0.61, 1)$; $\gamma_1 \leftarrow$ Random$[1, 10^6]$; $\gamma_2 \leftarrow$ Random$[1, 4]$;

$\delta_1 \leftarrow$ Random$[1, 10^6]$; $\delta_2 \leftarrow$ Random$[1, 4]$; $k_{SV} \leftarrow$ Random$[823, 9356]$; $b \leftarrow$ Random$[0.106, 2]$;$a \leftarrow b/2$;

Step 2: define costs and minimum treatment intensity needed to guarantee steady state

$\underline{\mu} \leftarrow 1.01n * 5.08\%$; $k_G(d_G) \leftarrow \gamma_1(d_G)^{\gamma_2}$; $k_{SF}(d_S) \leftarrow \delta_1(d_S)^{\delta_2}$;

Step 3: conduct a plausibility check for the generated parameter set

Find (d_G^*, d_S^*) by numerically maximizing the function $u(d_G, d_S)$ given by (5);

if $\lambda^* > 0.0322$ **or** $1/(\mu(d_S^*) - n\Lambda(d_G^*)) < 1/365$ **or** $1/(\mu(d_S^*) - n\Lambda(d_G^*)) > 1$ **or** $k_G(d_G^*) < n550/4$ **or** $k_G(d_G^*) > n550 * 4$ **or** $k_{SV} n\Lambda(d_G^*) + k_{SF}(d_S^*) < n\Lambda(d_G^*)1935/4$ **or** $k_{SV} n\Lambda(d_G^*) + k_{SF}(d_S^*) > n\Lambda(d_G^*)9356 * 4$ **then**

drop the parameter set from consideration and goto Step 1;

Step 4: find equilibrium decisions for each possible scenario of naïvite consequences

$(\tilde{d}_{G,A}^{FR}, \tilde{d}_{S,A}^{FR})$ are found as the intersection of best response functions (18) and (19);

for each $j \in \{C, P\}, l \in \{C, P\}$ **do**

find $(\tilde{d}_{G,j,l}^{C}, \tilde{d}_{S,j,l}^{C})$ by numerically maximizing the best responses of colluding agents (23)-(24);

Step 5: Get performance metrics for each possible scenario of naïvite consequences

Total population health $\mathbb{E}Q_P(d_G, d_S)$, given by (4), evaluated at decisions generated in Step 4;

System efficiency $u(d_G, d_S)$, given by (5), evaluated at Step 4 decisions;

Profit of providers, from sum of obj. functions in (9)–(10), evaluated at Step 4 decisions;

Government expenditure, from the objective function of (8) for individual contracts ((15) for group contracts), minus $\mathbb{E}Q_P(d_G, d_S)$, both evaluated at Step 4 decisions;

Compute statistics over all *size* runs for each reimbursement scenario

parameter sets, the main one appears to be an overly steep function for the GP's costs, which would then fail two checks: it would cause GP costs that far exceed £550 per patient, and it would cause a higher complication rate in the first-best than is observed in practice.

Appendix C: Other Extensions

We now present several variations and extensions of our model in order to assess the sensitivity of conclusions to assumptions of the model and to explore the scope of applicability of the results. In Appendix C.1, we consider provider decisions directly impacting the health of their patients, as

well as patients endogenously deciding on whether to be adherent. In Appendix C.2, we relax our assumption that measures of health are unbiased. In Appendix C.3, we extend our model to include multiple complications and multiple specialists who treat them. In Appendix C.4, we consider a different model of collusion: colluding agents making decisions jointly as a single decision maker. In Appendix C.5, we consider which results can be replicated without having an explicit expression for population health available. Some interesting special cases in Appendix C.5 are different specifications for the health impact of treatment delay (the $w(t)$ function) and a change of queueing discipline to GI/G/1 under heavy traffic. Appendix C.6 numerically tests the sensitivity of optimal contracts to misestimation of the cost functions. Finally, Appendix C.7 proves mathematical claims regarding those variations and extensions.

One result of these extensions is the apparently remarkable robustness of some of our conclusions: the beneficial properties of outcomes-based capitation contracts appear to hold throughout these extensions. However, this analysis also points out two important limitations. First, the performance of conventional contract types is likely to be underestimated in our model, as we do not model non-monetary drivers of physician behavior such as reputation or altruism. Second, we encounter a lack of tractability if we make the providers risk-averse (thus sensitive to the *variance* of noise), or if we make the noise term correlated with provider decisions. Thus, the effect of noise is likely deeper and more nuanced in reality than in our model, posing a direction for further research.

C.1. Other Effects of Provider Decisions

In our model, we focus on the complication rate, the service rate and, the resulting queueing dynamics as the main effect of the provider decisions. However, it is reasonable to expect that quality of care has effects along other dimensions as well. Thus, here we consider other potential consequences of provider decisions, and how this complexity affects the model.

Firstly, the quality of GP care could affect the health of patients directly, not only through the complication rate. Secondly, the quality of SP care can influence the amount of health generated by specialist care, not just the service rate μ. Thirdly, d_G could impact the adherence decisions of the patients. We model these three additional effects by introducing alternative assumptions:

Assumption C.1.1 *GP care directly improves health: at the end of the period, the health capital of each adherent patient is improved by a factor $\omega(d_G)$, an increasing, twice differentiable function.*

Assumption C.1.2 *The amount of health capital restored by the SP depends on the quality of SP care: the factor ζ is replaced by $\zeta(d_S)$, an increasing, twice differentiable function.*

Assumption C.1.3 *The adherence rate is a function of the quality of GP care: the constant adherence rate ϕ is replaced by $\phi(d_G)$, a twice differentiable function.*

Note that it is not clear if $\phi(d_G)$ is monotonic, and if so, with which slope, as there are arguments to be made for both directions. On the one hand, patients should be more likely to adhere the more health benefit they derive from treatment (suggesting $\phi(d_G) \uparrow d_G$). On the other hand, there is evidence that more rigorous treatment is associated with lower adherence (suggesting $\phi(d_G) \downarrow d_G$), due to the increased effort needed from the patient (Schectman et al. 2002, Borus and Laffel 2010).

There is both theoretical and empirical evidence of a trade-off between speed and quality of care in healthcare (Anand et al. 2011, Kc and Terwiesch 2011, Alizamir et al. 2013). We do not directly model this trade-off, but rather assume that the specialist will be able to optimally resolve the trade-off, with higher d_G implying both higher service rate $\mu(d_S)$ and health improvement $\zeta(d_S)$.

The main effect of this more nuanced model is on the expected health of patients, which is now:

$$\begin{aligned}\mathbb{E}Q_P(d_G,d_S) =& q_P^0\phi(d_G)\omega(d_G)\left(1-\lambda(d_G)a(1-\zeta(d_S))-\frac{b\lambda(d_G)(1-\zeta(d_S))}{\mu(d_S)-n(\phi(d_G)(\lambda(d_G)-\bar{\lambda})+\bar{\lambda})}\right)\\ &+q_P^0(1-\phi(d_G))\left(1-\bar{\lambda}a(1-\zeta(d_S))-\frac{b\bar{\lambda}(1-\zeta(d_S))}{\mu(d_S)-n(\phi(d_G)(\lambda(d_G)-\bar{\lambda})+\bar{\lambda})}\right). \qquad \text{(EC.46)}\end{aligned}$$

We can similarly rewrite the expected health of the GP's patients and the SP's patients as

$$\mathbb{E}Q_G(d_G,d_S) = q_P^0\phi(d_G)\omega(d_G)\left(1-a\lambda(d_G)(1-\zeta(d_S))-\frac{b\lambda(d_G)(1-\zeta(d_S))}{\mu(d_S)-n(\phi(d_G)(\lambda(d_G)-\bar{\lambda})+\bar{\lambda})}\right), \qquad \text{(EC.47)}$$

$$\begin{aligned}\mathbb{E}Q_S(d_G,d_S) =& q_P^0(\phi(d_G)\lambda(d_G)\omega(d_G)+(1-\phi(d_G))\bar{\lambda})\\ &\times\left(1-a(1-\zeta(d_S))-\frac{b(1-\zeta(d_S))}{\mu(d_S)-n(\phi(d_G)(\lambda(d_G)-\bar{\lambda})+\bar{\lambda})}\right). \qquad \text{(EC.48)}\end{aligned}$$

We also introduce one technical assumption needed to keep this model tractable:

Assumption C.1.4 *$\mathbb{E}Q_P(d_G,d_S)$ is increasing and (jointly) concave.*

The assumption enables us to continue to use first-order conditions to optimize, and it is a relatively weak assumption. It assumes that the health of patients increases in the quality of care, and that the efforts to improve health have diminishing marginal returns.

Notice that this alternative set of assumptions only changes our model of the care pathway and how it affects health; it does not alter the formulation of any of the contracting problems in the paper. With this in mind, most of the main results of the paper will replicate in this setting. (Proofs for all formal statements in Section C are given in Section C.7.)

Theorem C.1.1 *Let Assumptions C.1.1, C.1.2, and C.1.4 hold. Then, Propositions 1 and 2, as well as Theorems 3 and 4, still hold as stated. The same is true if Assumption C.1.3 also holds, but with one limitation: in Theorems 3 and 4.1, contracts that use q_A as a signal cannot be replicated.*

Thus, the key insights from the naïve and main models are the same as in the main paper. What does not replicate in this setting are the directions of deviations of the sub-optimal contracts, as those now suffer from numerous countervailing effects.

C.2. Biased Measures of Health

Our model assumes the existence of an unbiased measure of health. Here, we will examine the consequences of such an assumption being violated. Recall, the measure of population health is $\mathcal{Q}_P = Q_P + \varepsilon_P$, where Q_P is the deterministic actual population health, and ε_P is the noise term, which is a zero mean random variable. So, consider the following alternative assumption:

Assumption C.2.1 *For every $i \in \{P, G, S, A\}$, the measurement noise ε_i has a nonzero expectation ($\mathbb{E}\varepsilon_i = \mathcal{b}_i \neq 0$). None of the decision makers in the model are aware of this bias.*

Here, we can derive the following.

Proposition C.2.1 *Let Assumption C.2.1 hold. (i) The contract offered by the principal, the decisions made by providers, and the health of patients all remain the same for every contracting problem in the paper (NICP and NGCP in Section 3, MICP in Section 5.2, MGCP in Section 5.1). (ii) There is a wealth transfer between the principal and the agents. If the contract uses the contractible variable $\mathcal{Q}_i$, where $i \in \{P, G, S, A\}$, the wealth transfer favors the agent if $\mathcal{b}_i > 0$ or the principal if $\mathcal{b}_i < 0$.*

Lack of awareness about the bias is essential to its existence as it would be easy to de-bias the measure of health if $\mathcal{b}$ was known. Yet, being unaware of the bias, both the principal and the agents make their decisions as if the bias did not exist. The sole impact of the bias takes effect when the measures of health are realized, and results in under-payment of the agents (in case of a negative bias) or over-payment (in case of a positive one).

Also worth considering is the situation where the agents are aware of the bias, but the principal is not. This can arise if the agents have more direct information about data measurements.

Remark C.2.1 *If the agents are aware of the bias, Prop. C.2.1 holds,* but *the agent will reject the contract if $\mathcal{b}_i < 0$. This follows from the same proof as Prop. C.2.1, by noticing that the agent's incentive compatibility constraint does not change as the* $\arg\max$ *of $\mathbb{E}\mathcal{Q}_P(d_G, d_S)$ and $\mathbb{E}Q_P(d_G, d_S)$ is the same (they differ only by a constant); however, the participation constraint is violated any time $\mathcal{b}_i < 0$.*

Thus asymmetric knowledge of the bias can cause either contract failure (if it leads to rejection) or a wealth transfer from the principal to the agents.

C.3. Multiple Complications and Specialists

In our desire for parsimony, we model a single specialist who treats complications. Reality is more complex than the model. For example, diabetes has several possible complications (e.g., diabetic retinopathy, nephropathy, neuropathy) that may require different specialists. Most of our results will extend in a relatively straightforward way – with the same proof structures still working – if there are multiple complications and multiple specialists who treat them. The exception to this is Theorem 4, where complexity arises in a multi-specialist situation about who could potentially collude with whom. The theorem still replicates (see Theorem C.3.2), but requires assumptions about the formation of collusive coalitions.

Consider a model that uses the following alternative assumption.

Assumption C.3.1 *There are k specialists, each of which treats a set of possible complications (they are indexed by $i \in \{1,..,k\}$). These sets are mutually exclusive. Each set of complications has a different rate at which it occurs (thus $\lambda_i(d_G)$,$\overline{\lambda}_i$,$\Lambda_i(d_G)$), different health impact (a_i, b_i), different cost of treatment (k_{SVi}, $k_{SFi}(\cdot)$), and different time needed to treat (μ_i). To ensure well defined probabilities, we assume $\sum_{i=1}^{k} \overline{\lambda}_i \leq 1$.*

This has a direct impact on the health of patients, where the health of patients needs to be recalculated in accordance with Assumption C.3.1. Denote by $\mathbf{d}_S$ the vector of specialist decisions ($\mathbf{d}_S \doteq (d_{S1},...,d_{Sk})$). Then, we have

$$\mathbb{E}Q_P(d_G,\mathbf{d}_S) = q_P^0\left(1-\sum_{i=1}^{k}\left(a_i\Lambda_i(d_G)(1-\zeta)+\frac{b_i\Lambda_i(d_G)(1-\zeta)}{\mu_i(d_{Si})-n\Lambda_i(d_G)}\right)\right) \tag{EC.49}$$

$$\mathbb{E}Q_G(d_G,\mathbf{d}_S) = q_P^0\phi\left(1-\sum_{i=1}^{k}\left(a_i\lambda_i(d_G)(1-\zeta)+\frac{b_i\lambda_i(d_G)(1-\zeta)}{\mu_i(d_{Si})-n\Lambda_i(d_G)}\right)\right) \tag{EC.50}$$

$$\mathbb{E}Q_{Si}(d_G,\mathbf{d}_S) = q_P^0\Lambda_i(d_G)\left(1-a_i(1-\zeta)-\frac{b_i(1-\zeta)}{\mu_i(d_{Si})-n\Lambda_i(d_G)}\right). \tag{EC.51}$$

Assumption C.3.1 affects not only the care pathway, but the whole contracting problem (unlike the relatively elegant analysis of Section C.1). Denote by $\mathbf{v}_S(v_{S1}(\mathcal{Q}_1),v_{S2}(\mathcal{Q}_2),...,v_{Sk}(\mathcal{Q}_k))$ the vector of contracts offered to the different specialists, where $Q_i \in \{\mathcal{Q}_P,\mathcal{Q}_{Si}\}$. Similarly denote by $\tilde{\mathbf{d}}_S \doteq (\tilde{d_{S1}},\tilde{d_{S2}},...,\tilde{d_{Sk}})$ the vector of specialist decisions in the equilibrium, and by $\tilde{\mathbf{d}}_S(d_{Si}) \doteq (\tilde{d_{S1}},\tilde{d_{S2}},...,d_{Si},...,\tilde{d_{Sk}})$ the vector of specialist decisions where only specialist i deviates from the equilibrium, making decision d_{Si}. Then, the NICP in this setting is:

$$\max_{v_G(\mathcal{Q}_0),\mathcal{Q}_0\in\{\mathcal{Q}_P,\mathcal{Q}_G\},\mathbf{v}_S} \quad \mathbb{E}\left[Q_P(\tilde{d_G},\tilde{\mathbf{d}}_S)-v_G(\mathcal{Q}_0(\tilde{d_G},\tilde{\mathbf{d}}_S))-\sum_{i=1}^{k}v_{Si}(\mathcal{Q}_i(\tilde{d_G},\tilde{\mathbf{d}}_S))\right] \tag{EC.52}$$

$$\text{subject to} \quad \tilde{d_G} \in \underset{d_G \in [0,\infty)}{\arg\max}\, \mathbb{E}\left[v_G(\mathcal{Q}_0(d_G, \tilde{\mathbf{d}_S})) - k_G(d_G)\right] \tag{EC.53}$$

$$\tilde{d_{Si}} \in \underset{d_{Si} \in [0,\infty)}{\arg\max}\, \mathbb{E}\left[v_{Si}(\mathcal{Q}_i(\tilde{d_G}, \tilde{\mathbf{d}_S}(d_{Si})) - k_{SFi}(d_{Si}) - \mathcal{S}_i(\tilde{d_G})k_{SVi}\right], \forall i \in \{1,..,k\} \tag{EC.54}$$

$$V_G \leq \mathbb{E}\left[v_G(\mathcal{Q}_0(\tilde{d_G}, \tilde{\mathbf{d}_S})) - k_G(\tilde{d_G})\right] \tag{EC.55}$$

$$V_{Si} \leq \mathbb{E}\left[v_{Si}(\mathcal{Q}_i(\tilde{d_G}, \tilde{\mathbf{d}_S})) - k_{SFi}(\tilde{d_{Si}}) - \mathcal{S}_i(\tilde{d_G})k_{SVi}\right], \forall i \in \{1,...,k\}. \tag{EC.56}$$

An analogue of Prop. 1 holds here:

Proposition C.3.1 (Optimal naïve individual contracts) *There exists an* outcomes-adjusted capitation contract *that is optimal and achieves first-best. Under this contract, the provider is paid a capitation fee c_j, which is paid for every patient in the population, irrespective of whether that person receives treatment (here $j \in \{G, S1, S2, ..., Sk\}$). This fee is outcomes-adjusted according to the measure of health at the population level:*

$$c_j(\boldsymbol{q}_P) = f_{jC} + r_{jC}(\boldsymbol{q}_P - t_{jC}). \tag{EC.57}$$

The first-best can also be achieved by outcomes-adjusted per-patient contracts, *which award providers a fee p_j per every patient they treat. Here, the per-patient fee is outcomes-adjusted based on the health of patients treated by that provider: $p_j(\boldsymbol{q}_j) = f_{iP} + r_{iP}(\boldsymbol{q}_j - t_{jP})$.*

The two contract types can also be combined, giving some providers a capitation contract, while the others receive a per-patient one; this combination also achieves the first-best. (Closed-form expressions for optimal values of all the parameters are given in Appendix C.7.)

We can also reconsider and solve the NGCP in this setting:

$$\max_{\mathcal{Q} \in \{\mathcal{Q}_P, \mathcal{Q}_A\}, v(\mathcal{Q})} \mathbb{E}\left[Q_P(\tilde{d_G}, \tilde{\mathbf{d}_S}) - v(\mathcal{Q}(\tilde{d_G}, \tilde{\mathbf{d}_S}))\right] \tag{EC.58}$$

$$\text{subject to} \quad (\tilde{d_G}, \tilde{\mathbf{d}_S}) \in \underset{(d_G, \mathbf{d}_S) \in [0,\infty)^{k+1}}{\arg\max}\, \mathbb{E}\left[v(\mathcal{Q}(d_G, \mathbf{d}_S)) - k_G(d_G) - \sum_{i=1}^{k}(k_{SFi}(d_{Si}) + \mathcal{S}_i(d_G)k_{SVi})\right] \tag{EC.59}$$

$$\mathbb{E}\left[v(\mathcal{Q}(\tilde{d_G}, \tilde{\mathbf{d}_S})) - k_G(\tilde{d_G}) - \sum_{i=1}^{k}\left(k_{SFi}(\tilde{d_{Si}}) + \mathcal{S}_i(\tilde{d_G})k_{SVi}\right)\right] \geq V_G + V_S. \tag{EC.60}$$

Solutions to this problem are characterized by Prop. C.3.2.

Proposition C.3.2 (Optimal naïve group contracts) *Optimal group contracts are an outcomes-adjusted capitation contract (v_{AC}) with per-capita fee $c_A(\boldsymbol{q}_P) = f_{AC} + r_{AC}(\boldsymbol{q}_P - t_{AC})$, and an outcomes-adjusted per-patient contract (v_{AP}) with per-patient fee $p_A(\boldsymbol{q}_A) = f_{AP} + r_{AP}(\boldsymbol{q}_A - t_{AP})$. Both contracts achieve the first-best. (Closed-form expressions for optimal values of all the parameters are given in Appendix C.7.)*

We can also consider group contracting situations with endogenous free-riding (MGCP), by using the Multilateral Nash Bargaining solution (Bennett 1987). Applying the Multilateral Nash Bargaining, the revenue from the contract is split in the following way: each agent takes away the value of their outside option, and the remainder is split equally amongst the providers. This gives us the MGCP for this setting:

$$\max_{\mathcal{Q}\in\{\mathcal{Q}_P,\mathcal{Q}_A\},v(\mathcal{Q})} \quad \mathbb{E}\left[Q_P(\tilde{d}_G^{FR},\tilde{\mathbf{d}}_S^{FR})-v(\mathcal{Q}(\tilde{d}_G^{FR},\tilde{\mathbf{d}}_S^{FR}))\right], \tag{EC.61}$$

$$\text{subject to} \quad \tilde{d}_G^{FR}\in\arg\max_{d_G\in[0,\infty)}\mathbb{E}\left[\frac{1}{k+1}\left(v(\mathcal{Q}(d_G,\tilde{\mathbf{d}}_S^{FR}))-V_G-\sum_{i=1}^{k}V_{Si}\right)+V_G-k_G(d_G)\right], \tag{EC.62}$$

$$\forall i\in 1,...,k: \quad \tilde{d}_{Si}^{FR}\in\arg\max_{d_{Si}\in[0,\infty)}\mathbb{E}\left[\frac{1}{k+1}\left(v(\mathcal{Q}(\tilde{d}_G^{FR},\mathbf{d}_S(d_{Si}))-V_G-\sum_{i=1}^{k}V_{Si}\right)+V_{Si}-k_{SFi}(d_{Si})-\mathcal{S}_i(\tilde{d}_G^{FR})k_{SVi}\right], \tag{EC.63}$$

$$\mathbb{E}\left[\frac{1}{k+1}\left(v(\mathcal{Q}(\tilde{d}_G^{FR},\tilde{\mathbf{d}}_S^{FR}))-V_G-\sum_{i=1}^{k}V_{Si}\right)+V_G-k_G(\tilde{d}_G^{FR})\right]\geq V_G, \tag{EC.64}$$

$$\forall i\in 1,...,k: \quad \mathbb{E}\left[\frac{1}{k+1}\left(v(\mathcal{Q}(\tilde{d}_G^{FR},\tilde{\mathbf{d}}_S^{FR}))-V_G-\sum_{i=1}^{k}V_{Si}\right)+V_{Si}-k_{SFi}(\tilde{d}_{Si}^{FR})-\mathcal{S}_i(\tilde{d}_G^{FR})k_{SVi}\right]\geq V_{Si}. \tag{EC.65}$$

Under this setup, we can derive an analogue of Theorem 3:

Theorem C.3.1 (Group contracts) *Let $v(\boldsymbol{q})$ be a differentiable contract and let $\tilde{d_G},\tilde{\mathbf{d}_S}$ be the interior decisions that it induces. Then there exists a linear outcomes-adjusted capitation contract $v^{\dagger}(\boldsymbol{q}_P)$ such that it induces the same decisions $(\tilde{d_G},\tilde{\mathbf{d}_S})$, at an equal or lower cost to the principal.*

The problem of contracting with individual providers who can possibly collude (MICP) does not extend in such a straightforward way, and will require some additional modeling choices. There are two questions that arise.

The first question is, how will a coalition of colluding providers split the gains from the collusion? For group contracting, this question was easily resolved as it was just a straightforward application of multilateral Nash Bargaining. This is because the providers faced a binary choice: either they accept the contract or they will receive their outside option. If we want to consider collusive agreements, the outside option is no longer a fixed value (e.g., V_{Si}) but instead depends on what other collusive agreements can be formed. In the spirit of trying to maintain generality while preserving tractability, we model it in the following way.

Denote the set of colluding providers by $C\subseteq\{S1,S2,...,Sk\}\cup\{G\}$. The colluding providers will split revenue from collusion in the following way: each provider $l\in C$ will take from the revenue a constant $a_l(C)$, and the remaining revenue will be split amongst the provider so that each provider

receives fraction $1/|C|$ of the remainder. (Here, $|C|$ is the cardinality of the set C, i.e., the number of agents in the collusive coalition.)

Note that this definition is a generalization of Multilateral Nash Bargaining (Bennett 1987), which is, in turn, a generalization of the classical model of Nash.

Thus, the decisions of all providers when there is a coalition C are given by an analogue of our main collusion problem (CP). Define $S0 = G$ to ease notation in the summation for cases of collusive coalitions that include the GP. The CP for this setting is:

$$\tilde{d}_G \in \begin{cases} \arg\max_{d_G \in [0,\infty)} \mathbb{E}\left[\frac{1}{|C|}\sum_{Si\in C}\left(v_{Si}(\mathcal{Q}_i(d_G,\tilde{\mathbf{d}_S})) - a_{Si}(C)\right) + a_G(C) - k_G(d_G)\right] & \text{if } G \in C \\ \arg\max_{d_G \in [0,\infty)} \mathbb{E}\left[v_G(\mathcal{Q}_0(d_G,\tilde{\mathbf{d}_S})) - k_G(d_G)\right] & \text{otherwise} \end{cases}$$

$$\tilde{d}_{Si} \in \underset{d_{Si}\in[0,\infty)}{\arg\max}\, \mathbb{E}\left[\frac{1}{|C|}\sum_{Si\in C}\left(v_{Si}(\mathcal{Q}_i(\tilde{d_G},\tilde{\mathbf{d}_S})) - a_{Si}(C)\right) + a_{Si}(C) - k_{SFi}(d_{Si}) - \mathcal{S}_i(\tilde{d_G})k_{SVi}\right], \forall i \geq 1 | S_i \in C$$

$$\tilde{d}_{Si} \in \underset{d_{Si}\in[0,\infty)}{\arg\max}\, \mathbb{E}\left[v_{Si}(\mathcal{Q}_{Si}(\tilde{d_G},\tilde{\mathbf{d}_S}(d_{Si})) - k_{SFi}(d_{Si}) - \mathcal{S}_i(\tilde{d_G})k_{SVi}\right], \forall i \geq 1 | S_i \notin C$$

The second question that arises is, which collusive coalition is going to form? This is a core question in cooperative game theory, but also an open one, with multiple approaches and solution concepts (Kahan and Rapoport 2014). However, the results we are interested in proving are not sensitive to how the process of coalition formation works, so we do not need to constrain ourselves to a particular solution concept. Thus, let us define the abstract coalition formation function

$$\mathcal{cf}(v_G, \mathbf{v}_S) \mapsto C,$$

which maps the contracts given to the providers to the coalition that will form under those contracts (possibly an empty set). Denote by $\pi_l(v_G, \mathbf{v}_S, C)$ the profit of provider l under contracts $v_G, \mathbf{v}_S$ and coalition C – this vector is given by the solution of the CP above. The only assumption we make about $\mathcal{cf}$ is that the members of the resulting coalition are better off under the coalition than on their own, or formally

$$\forall l \in \mathcal{cf}(v_G, \mathbf{v}_S) : \pi_l(v_G, \mathbf{v}_S, \mathcal{cf}(v_G, \mathbf{v}_S)) \geq V_l, \tag{EC.66}$$

where for at least one member of C, this inequality is strict. With this in mind, we can formulate the MICP for this setting:

$$\max_{v_G(\mathcal{Q}_0),\mathcal{Q}_0\in\{\mathcal{Q}_P,\mathcal{Q}_G\},\mathbf{v}_S} \mathbb{E}\left[Q_P(\tilde{d_G},\tilde{\mathbf{d}_S}) - v_G(\mathcal{Q}_0(\tilde{d_G},\tilde{\mathbf{d}_S})) - \sum_{i=1}^{k} v_{Si}(\mathcal{Q}_i(\tilde{d_G},\tilde{\mathbf{d}_S}))\right] \tag{EC.67}$$

$$\text{subject to} \quad C = \mathcal{cf}(v_G, \mathbf{v}_S) \tag{EC.68}$$

$$\text{if } G \in C: \quad \tilde{d}_G \in \underset{d_G\in[0,\infty)}{\arg\max}\, \mathbb{E}\left[\frac{1}{|C|}\sum_{Si\in C}\left(v_{Si}(\mathcal{Q}_i(d_G,\tilde{\mathbf{d}_S})) - a_{Si}(C)\right) + a_G(C) - k_G(d_G)\right] \tag{EC.69}$$

$$\text{if } G \notin C: \qquad \tilde{d}_G \in \underset{d_G \in [0,\infty)}{\arg\max}\, \mathbb{E}\left[v_G(\mathcal{Q}_0(d_G, \tilde{\mathbf{d}}_S)) - k_G(d_G)\right] \tag{EC.70}$$

$$\forall i \geq 1 | Si \in C: \qquad \tilde{d}_{Si} \in \underset{d_{Si} \in [0,\infty)}{\arg\max}\, \mathbb{E}\left[\frac{1}{|C|}\sum_{Si \in C}\left(v_{Si}(\mathcal{Q}_i(\tilde{d}_G, \tilde{\mathbf{d}}_S)) - a_{Si}(C)\right) + a_{Si}(C) - k_{SFi}(d_{Si}) - \mathcal{S}_i(\tilde{d}_G)k_{SVi}\right] \tag{EC.71}$$

$$\forall i \geq 1 | Si \notin C: \qquad \tilde{d}_{Si} \in \underset{d_{Si} \in [0,\infty)}{\arg\max}\, \mathbb{E}\left[v_{Si}(\mathcal{Q}_{Si}(\tilde{d}_G, \tilde{\mathbf{d}}_S(d_{Si})) - k_{SFi}(d_{Si}) - \mathcal{S}_i(\tilde{d}_G)k_{SVi}\right] \tag{EC.72}$$

$$\forall i \in \{0, 1, .., k\}: \qquad \pi_{Si}(\tilde{v}_G, \tilde{\mathbf{v}}_S, C) \geq V_{Si} \tag{EC.73}$$

Here, we will refer to contracts such that $\mathcal{cf}(v_G, \mathbf{v}_S) = \emptyset$ as collusion-proof contracts. Finally, with the setup complete, we are in a position to derive the analogue of Theorem 4 for this setting:

Theorem C.3.2 (Individual contracts)

1. *The health outcome of optimal group contracts – the ones solving (EC.61)-(EC.65) – can be replicated at a lower cost, by using $k+1$ individual linear outcomes-adjusted capitation contracts instead.*
2. *If the naïve individual contracts (as given in Proposition C.3.1) are collusion proof, they are also optimal and achieve first-best in the MICP given by (EC.67)-(EC.73).*

The above is not a complete replication of Theorem 4; what is missing is part 2 of Theorem 4. Thus, the theoretical optimally of individual contracts replicates in this setting, but linear capitation contracts might not be the optimal ones. However, this setting also gives rise to some potentially interesting properties about coalitions that can form.

Proposition C.3.3 *For all the individual contracts $(v_G, \mathbf{v}_S)$ given in Prop. C.3.1, it holds that either $G \in \mathcal{cf}(v_G, \mathbf{v}_S)$, or $\mathcal{cf}(v_G, \mathbf{v}_S) = \emptyset$.*

Essentially, any collusive coalition needs the GP to be part of it. The reason for this is different for different contract types. Consider a coalition consisting of specialists only.

If they hold per-patient contracts, the specialists are not able to affect each other's compensation directly due to them treating non-overlapping patient pools.

Under optimal capitation contracts, every specialist receives the full marginal benefits created for the system (from the optimal reimbursement rates in the proof of Prop. C.3.1). If the coalition was able to perfectly coordinate, this would create an incentive to over-treat patients as the coalition would receive more than the full marginal returns (see Section C.4 for details); however, the free-riding effect exactly counteracts this incentive to over-treat, creating a situation where the marginal benefits for the specialist are the same irrespective of him or her being part of the coalition.

C.4. Coordinated Collusion in Individual Contracts

A classical economic approach to modeling collusion is to consider colluding agents that make decisions as a single agent that maximizes the sum of the two profits. In our main model, we did not follow this approach, as we wanted a consistent set of assumptions that would give rise to both collusion and free-riding. The result of our assumptions is that collusive agreements are equally inefficient as formal alliances, with both suffering from free-riding issues.

While our belief is that the classical way of modeling collusion is less realistic for the situation at hand, there is no technical obstacle to modeling it in that way – actually, the classical modeling approach improves tractability. So, it might be a useful exercise to also explore how things change under that approach, which we do in this section. At the very least, it paints a picture of the most extreme results that could happen under collusion. We will refer to this type of collusion as *coordinated* collusion.

After being given individual contracts $v_G(q_1), v_S(q_2)$, agents engaged in coordinated collusion will not make the non-cooperative equilibrium decisions ($\tilde{d}_G$, $\tilde{d}_S$) given by (9)–(10), but instead choose $\tilde{d}_G^{CC}$, $\tilde{d}_S^{CC}$ to maximize their joint benefit by solving the *coordinated collusion problem* (CCP):

$$(\tilde{d}_G^{CC}, \tilde{d}_S^{CC}) \in \underset{(d_G,d_S)\in[0,\infty)^2}{\arg\max}\ \mathbb{E}\left[v_G(\mathcal{Q}_1(d_G,d_S)) + v_S(\mathcal{Q}_2(d_G,d_S)) - k_G(d_G) - k_{SF}(d_S) - \mathcal{S}(d_G)k_{SV}\right]. \tag{EC.74}$$

The way colluding agents deviate from the first-best will depend on which contracts they have. Theorem C.4.1 formalizes the effects of these deviations, and the results of the theorem are summarized in Table EC.2.

Theorem C.4.1 (Coordinated collusion) *Suppose the GP and SP are given individual contracts in Prop. 1, with the GP holding either contract v_{Gk} where $k \in \{C, P\}$, and the SP holding v_{SC}. If such agents engage in coordinated collusion by solving the CCP in (EC.74), then the GP and SP will fail to achieve first-best and* (i) *System efficiency u is lower,* (ii) *complication rate is lower,* (iii) *expected population health is higher,* (iv) *expected joint profit of agents is higher,* (v) *expected government expenditure is higher. If the SP holds the per-patient contract v_{SP} instead of the capitation contract v_{SC}, then coordinated collusion will fail to achieve first-best and will result in* (vi) *lower system efficiency, and* (vii) *higher expected joint profit of agents. Coordinated collusion when the SP holds the per-patient contract v_{SP} compared to coordinated collusion when he or she holds the capitation contract v_{SC} results in* (viii) *higher complication rate,* (ix) *higher service rate, and* (x) *lower expected population health when the GP holds v_{GC}.*

In the absence of collusion, the proof of Prop. 1 showed that, with optimal contracts, both agents receive the full value of marginal benefits created by the system. If agents act in their own interest,

Table EC.2 Summary of findings about the effects of Coordinated Collusion

	FB vs (v_{Gk},v_{SC})	FB vs (v_{Gk},v_{SP})	(v_{Gk},v_{SC}) vs (v_{Gk},v_{SP})
System efficiency (u)	$>$	$>$	
Expected population health ($\mathbb{E}Q_P$)	$<$		$>$[1]
Complication rate (Λ)	$>$		$<$
Service rate (μ)			$<$
Joint profit of agents ($\pi_G+\pi_S$)	$<$		
Government expenditure (v_G+v_S)	$<$	$<$	

FB denotes first-best. $<$ ($>$) denotes that the corresponding metric is higher (lower), e.g., $>$ in the first column first row denotes that the system efficiency u is higher in FB than if the agents hold contracts v_{Gk},v_{SC}. Empty fields are inconclusive. [1] This holds only if $k=C$.

such compensation is beneficial, serving the purpose of aligning the agents' interests. The presence of collusion, however, creates an incentive to over-treat patients because colluding parties can be compensated at a higher marginal rate than the value they create. If the SP holds a capitation contract, Theorem C.4.1 shows that this effect exists and is independent of the type of contract held by the GP. While increased treatment intensity and health might look appealing, note that it comes paired with increased health care costs and decreased system efficiency.

If the SP holds a per-patient contract, the same incentive to over-treat patients noted above is still present; however, there is an additional perverse incentive in play. An SP with such a contract earns more the more patients he or she has, creating an incentive for a colluding GP to decrease treatment intensity so as to increase complication rates and generate demand for the SP.

Theorem C.4.1 also gives conclusive results for coordinated collusion when the SP holds a per-patient contract. However, several of the performance comparisons between the first-best solution and the solution under coordinated collusion are inconclusive in this case, due to the countervailing effects that cause intractability when making comparisons with first-best performance. We *can* identify which contract type results in more severe effects on patients, and give these results in parts (viii)-(x) of the theorem.

Part (x) of Theorem C.4.1 holds in general only if the GP holds a capitation contract. The reason for that lies in the fact that the GP holding a per-patient contract adds another layer of complexity. If the GP holds that contract, the colluding parties' payout depends on the health of all of their patients, but they earn a higher marginal return from improving the health of the GP's patients than for other ones.

This distortion of incentives will be stronger the lower the adherence rate ϕ is, as r_{GP} is decreasing in ϕ. Intuitively, the mechanism behind this behavior of r_{GP} in ϕ is that the GP's treatment helps even the patients he or she is not treating due to reducing congestion for the SP. The optimal GP reimbursement needs to take these second-order effects into account, and it is higher if there are more people that benefit from these effects (low ϕ).

C.5. Robustness of conclusions on Outcomes-adjusted Capitation Contracts

An important technical characteristic of several of our proofs about the properties of outcomes-adjusted capitation contracts is that the proofs are not reliant on having the closed-form expression for population health in (4) available, only that the function $\mathbb{E}Q_P$ satisfies some basic properties. Thus, we consider the setting where:

Assumption C.5.1 *Instead of the model of queueing dynamics and impact of care on health as given by Section 2, assume only that $\mathbb{E}Q_P(d_G, d_S)$ is increasing, jointly concave, and twice differentiable.*

Theorem C.5.1 *Let Assumption C.5.1 hold. Then,*

1. *The part of Proposition 1, which speaks of individual outcomes-adjusted capitation contracts, still holds. Specifically, there exists an* outcomes-adjusted capitation contract *that is optimal (in the NICP) and achieves first-best. For $i \in \{G, S\}$, under this contract, provider i is paid a capitation fee c_i, which is paid for every patient in the population, irrespective of whether that person receives treatment. This fee is outcomes-adjusted according to the measure of health on the population level:*
$$c_i(\boldsymbol{q}_P) = f_{iC} + r_{iC}(\boldsymbol{q}_P - t_{iC}).$$
2. *Proposition 2 and Theorems 3 and 4 still hold as stated.*

Thus under Assumption C.5.1, basically *all* of our insight about the performance of outcomes-based capitation contracts still holds. This has far-reaching implications because many possible generalizations of our model are just specific cases of Assumption C.5.1.

The only limitation of this general setting is that the consequences of ignoring free-riding and collusion in contract design (Section 4) and the numerical analysis (Section 6) are reliant on having an explicit expression for $\mathbb{E}Q_P$ available.

If we look at (3), we can easily consider more complicated ways in which $\mathbb{E}Q_P$ could be determined, such as a) non-linear impact of treatment delays on health ($w(t)$), b) queueing disciplines other than M/M/1, and c) referral process having an impact on treatment delays. All of these could be included without impacting the key insight, as long as they do not destroy the structural properties contained in Assumption C.5.1.

Remark 2 *For an example of a non-linear effect of treatment delay that satisfies this assumption, consider a model where $w(t) \doteq a + \mathbb{1}(t \geq T)b$, where T is a critical response time. Then, inserting $w(t)$ into (2) and solving the integral yields*

$$\mathbb{E}Q_P(d_G, d_S) = q_P^0 \left(1 - a\Lambda(d_G)(1-\zeta) - b\Lambda(d_G)(1-\zeta)\exp(-T(\mu(d_S) - n\Lambda(d_G)))\right).$$

Here, the joint concavity of $\mathbb{E}Q_P(d_G, d_S)$ follows directly from the convexity of $\Lambda(d_G)$ and linearity of $\mu(d_S)$ by applying the rules for the preservation of convexity under compositions (Boyd and Vandenberghe 2004, p. 86). Thus, Assumption D7 holds in this setting.

For a less straightforward example, we can use the heavy traffic approximations (Kingman 1961, Whitt 1993) to show that our results extend to GI/G/1 queues under heavy traffic, as follows.

Proposition C.5.1 (GI/G/1 heavy traffic queue) *Let the arrival process of patients with complications be partially specified by the mean number of arrivals $\Lambda(d_G)$ and a constant coefficient of variation of inter-arrival times cv_A. Similarly, let the SP's service time be distributed according to a distribution with mean $\mu(d_S)$ and a constant coefficient of variation cv_S. Then, under heavy traffic ($\mu(d_G) \to^+ n\Lambda(d_g)$) the expected population health is*

$$\mathbb{E}Q_P(d_G, d_S) \approx q_P^0 \left(1 - \Lambda(d_G)a - \left(\frac{\mathrm{cv}_A^2 + \mathrm{cv}_S^2}{2} \frac{n\Lambda^2(d_G)}{n\Lambda(d_G)a - \mu(d_S)} + 1\right) \frac{b}{\mu(d_S)}\right). \tag{EC.75}$$

Furthermore, such $\mathbb{E}Q_P(d_G, d_S)$ satisfies Assumption C.5.1, and Theorem C.5.1 holds in this setting.

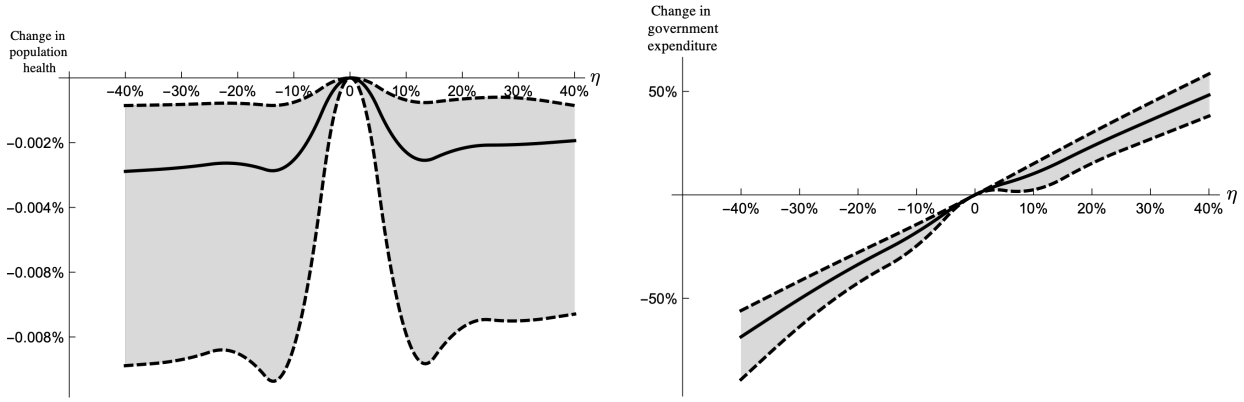
C.6. Estimation of Cost Functions

As discussed in Section 8, if the cost function of the providers is unknown, one option for the principal is to resort to estimation of this function. This approach introduces the possibility of estimation errors.

Here, we conduct a sensitivity analysis of how the performance of our contracts will be impacted by systematic underestimation or overestimation of cost functions, using the setting of Section 6 for illustration. Suppose that all cost parameters are misestimated by a factor η (so that the principal incorrectly believes the costs functions are $(1+\eta)k_G(\cdot), (1+\eta)k_{SF}(\cdot), (1+\eta)k_{SV}$ and optimizes accordingly). Figure EC.2 reports the effects of such misestimation.

In order to generate Figure EC.2, we considered the potential effect of misestimating the cost functions systematically high or low. For the purpose of this illustration, 200 parameter sets and related metrics were generated using the same algorithm as in Section 6. For each generated parameter set, 8 additional scenarios were considered (for cost misestimations $\eta \in \{-40\%, -30\%, -20\%, -10\%, 10\%, 20\%, 30\%, 40\%\}$). For every one of those scenarios, the parameters of the outcomes-adjusted capitation contracts in Prop. 1 were calculated according to the misestimated cost values. For calculating provider decisions that the misestimated contract would induce, we do not have a closed-form solution available, nor can the problem be reduced to numerical optimization. Our approach was to compute these decisions by exploiting the property of Nash equilibrium that if a sequence of mutual exploitation converges, the point of convergence

will be an equilibrium. This was a computationally demanding process and the reason why only 200 parameter sets were used for this illustration. Population health, system efficiency, and government expenditure were calculated for each of these misestimated contracts using the decisions they induce. Finally, for other values of $\eta \in [-40,\%,40\%]$, values were numerically approximated using Mathematica's native InterpolatingFunction method.



(a) Change in population health (b) Change in government expenditure

Figure EC.2 Change in population health and government expenditure as a result of cost misestimation by η. The full line is the mean, while dotted lines are the 5th and 95th percentile.

For the most part, the cost functions and the first-best decisions only show up in the computation of the fixed part of the capitation rate, i.e., f_{GC} and f_{SC} in Prop. 1. Yet, it is the variable part of the compensation scheme (r_{GC} and r_{SC}) that completely drives the providers' decisions, whereas the fixed part merely serves to ensure the providers' participation in the system. The variable compensation of the SP is not dependent on these inputs, and the variable compensation of the GP only has a single additive component, which depends on the first-best decisions d_G^* and d_S^*, while having no direct dependency on k_G and k_{SF}. Consequently, such misestimation of k_G and k_{SF} will hardly impact the decisions being made by providers, and thus, also the health of patients.

The consequences of misestimation will primarily manifest as a wealth transfer between the principal and the providers, as can be seen in panel (b) of Figure EC.2, which has no bearing on the ystem efficiency or the health of patients. Overestimation of costs can result in a significant cost increase to the principal. Underestimation can result in contracts that break the providers' participation constraint, which will cause them to be rejected.

C.7. Proofs of Mathematical Claims in Appendix C.

Proof of Theorem C.1.1. The proof follows the steps of Propositions 1 and 2 as well as Theorems 3 and 4, with the following key differences. In all situations where Lemma 1 was guaranteeing the existence of interior optimum and sufficiency of first-order conditions, the same holds in this setting

as well, but by virtue of Assumption C.1.4. In Proposition 1, the closed-form expressions for the reimbursement rates of the optimal individual contracts for the GP, given in (EC.13) and (EC.16), are no longer correct. The correct values follow from inserting (EC.46),(EC.47), and (EC.48) into (EC.12) and (EC.15). In Proposition 2, if Assumption C.1.1 holds, then ϕ can no longer be used in computing the optimal reimbursement rate as $\phi(d_G)$ is endogenous and not contractible. In that case, the same result can be obtained by replacing the ϕ in (EC.23) and all of the following equations with $\phi(d_G^*)$. In Theorem 3, the identity $\frac{\partial}{\partial d_G}\mathbb{E}Q_A(d_G,d_S)=\left(\frac{\partial}{\partial d_G}\mathbb{E}Q_P(d_G,d_S)\right)$, does not hold under Assumption C.1.3; this is what prevents the results of Theorems 3 and 4.1 from completely replicating in this setting. □

Proof of Prop. C.2.1. Notice that due to lack of knowledge about the bias, the problems NICP, NGCP, MICP, MGCP remain exactly as they are. However, the outcome of those problems does not. The true outcome of the contract is given by replacing $\mathcal{Q}_i$ with $\mathcal{Q}_i+\varepsilon_i$ in both the principal's and the agent's payout functions. The property that positive bias favors the agents (and negative favors the principal) then follows from the observation that it is necessary for the optimal contract to be increasing in realized health. □

The following Lemma is an analogue of Lemma 1 (from Appendix A) for application to the setting with multiple complications and specialists in Section C.3.

Lemma C.7.1 *Let Assumption C.3.1 hold and let* $\alpha_j\geq 0,\forall j\in\{0,...,2k+2\}$. *The function* $f:[0,\infty)^{k+1}\to\mathbb{R}$, *with mapping rule*

$$f(d_G,\mathbf{d}_S)=\alpha_0\mathbb{E}Q_P(d_G,\mathbf{d}_S)+\alpha_1\mathbb{E}Q_G(d_G,\mathbf{d}_S)-\alpha_2k_G(d_G)-\sum_{i=1}^{k}\left(\alpha_{i+2}k_{SFi}(d_{Si})+\alpha_{k+i+2}k_{SVi}n\Lambda_i(d_G)\right),$$

is jointly concave. If $\max\{\alpha_0,\alpha_1\}>0$ *and* $\max\{\alpha_j|j\in\{2,..,k+2\}\}>0$, *then* $f(d_G,\mathbf{d}_S)$ *has an interior maximum.*

Proof of Lemma C.7.1. Denote by $\mathbb{E}Q_P^i(d_G,d_{Si})$ (resp. $\mathbb{E}Q_G^i(d_G,d_{Si})$) the expected health of all patients (resp. the GP's patients) in a model where only the i-th set of complications and the i-th SP exist. We can then rewrite (EC.49)-(EC.50) as

$$\mathbb{E}Q_P(d_G,\mathbf{d}_S)=-(k-1)q_P^0+\sum_{i=1}^{k}\mathbb{E}Q_P^i(d_G,d_{Si}),\quad \mathbb{E}Q_G(d_G,\mathbf{d}_S)=-(k-1)q_P^0\phi+\sum_{i=1}^{k}\mathbb{E}Q_G^i(d_G,d_{Si}).$$

Applying Lemma 1, all of the summands above are concave, thus so are $\mathbb{E}Q_P(d_G,\mathbf{d}_S)$ and $\mathbb{E}Q_G(d_G,\mathbf{d}_S)$. The existence of interior maximum then follows from $k_G'(0),k_{Si}'(0)=0,\forall i\in\{1,...,k\}$.

Proof of Prop. C.3.1. The structure of the proof follows the one of Prop. 1. The first-best decisions $(d_G^*,\mathbf{d}_S^*)$ are solutions of the optimization problem

$$\underset{(d_G,\mathbf{d}_S)\in[0,\infty)^{k+1}}{\arg\max}\ \mathbb{E}Q_P(d_G,\mathbf{d}_S)-k_G(d_G)-\sum_{i=1}^{k}\left(k_{SFi}(d_{Si})+k_{SVi}n\Lambda_i(d_G)\right). \tag{EC.76}$$

As this objective function is jointly concave and has an interior maximum (from Lemma C.7.1), the first-best decisions are unique solutions to FOCs

$$k'_G(d_G) = \frac{\partial \mathbb{E}Q_P(d_G, \mathbf{d}_S)}{\partial d_G} - \sum_{i=1}^{k} k_{SVi} n \Lambda'_i(d_G), \tag{EC.77}$$

$$k'_{SF}(d_{Si}) = \frac{\partial \mathbb{E}Q_P(d_G, \mathbf{d}_S)}{\partial d_{Si}}, \forall i \in \{1, ..., k\}. \tag{EC.78}$$

Part 1: deriving the optimal capitation contract for the GP ($\boldsymbol{v_{GC}}$). Assume there exists a linear contract $v_{GC}(q_P) = a_{GC} + b_{GC} q_P$ for the GP, under which the first-best can be achieved. Then, when holding the aforementioned contract, the GP's best response to SP decisions $\mathbf{d}_S^*$ needs to be d_G^*, i.e., $d_G^* \in \arg\max_{d_G \in [0,\infty)} \mathbb{E}\left[a_{GC} + b_{GC} \mathcal{Q}_P(d_G, \mathbf{d}_S^*) - k_G(d_G)\right]$, which can be replaced by its FOC (using Lemma C.7.1):

$$k'_G(d_G) = b_{GC} \frac{\partial \mathbb{E}Q_P(d_G, \mathbf{d}_S^*)}{\partial d_G}. \tag{EC.79}$$

Thus, $(d_G, \mathbf{d}_S) = (d_G^*, \mathbf{d}_S^*)$ needs to simultaneously solve (EC.77) and (EC.79). Equating the RHS of those two equations and solving for b_{GC} yields

$$b_{GC} = 1 - \frac{\sum_{i=1}^{k} k_{SVi} n \Lambda'_i(d_G^*)}{\frac{\partial}{\partial d_G} \mathbb{E}Q_P(d_G^*, \mathbf{d}_S^*)} \tag{EC.80}$$

as the unique solution. For the contract to achieve FB, it is also necessary that the individual rationality constraint (EC.55) is binding, thus solving a binding (EC.55) for a_{GC} yields $a_{GC} = V_G + k_G(d_G^*) - b_{GC} \mathbb{E}Q(d_G^*, \mathbf{d}_S^*)$ as the unique solution. Setting

$$r_{GC} = \frac{1}{n} - \frac{\sum_{i=1}^{k} k_{SVi} \Lambda'_i(d_G^*)}{\frac{\partial}{\partial d_G} \mathbb{E}Q_P(d_G^*, \mathbf{d}_S^*)}, \quad f_{GC} = k_G(d_G^*)/n, \quad t_{GC} = \mathbb{E}Q_P(d_G^*, \mathbf{d}_S^*) - \frac{V_G}{n r_{GC}}, \tag{EC.81}$$

ensures that $v_{GC}(q_P) = a_{GC} + b_{GC} q_P = n\left(f_{GC} + r_{GC}(q_P - t_{GC})\right)$.

Part 2: deriving the optimal per-patient contract for the GP ($\boldsymbol{v_{GP}}$). Assume there exists a linear contract $v_{GP}(q_G) = a_{GP} + b_{GP} q_G$ for the GP, under which the first-best can be achieved. Then, when holding the aforementioned contract, $d_G^* \in \arg\max_{d_G \in [0,\infty)} \mathbb{E}\left[a_{GP} + b_{GP} \mathcal{Q}_G(d_G, \mathbf{d}_S^*) - k_G(d_G)\right]$, a maximization problem that can be reduced (using Lemma C.7.1) to its FOC:

$$k'_G(d_G) = b_{GP} \frac{\partial \mathbb{E}Q_G(d_G, \mathbf{d}_S^*)}{\partial d_G}. \tag{EC.82}$$

Thus, $(d_G, \mathbf{d}_S) = (d_G^*, \mathbf{d}_S^*)$ needs to simultaneously solve (EC.77) and (EC.82). Equating the RHS of those two equations and solving for b_{GP} yields

$$b_{GP} = \frac{\frac{\partial}{\partial d_G} \mathbb{E}Q_P(d_G^*, \mathbf{d}_S^*) - \sum_{i=1}^{k} k_{SVi} n \Lambda'_i(d_G^*)}{\frac{\partial}{\partial d_G} \mathbb{E}Q_G(d_G^*, \mathbf{d}_S^*)} \tag{EC.83}$$

as the unique solution. For the contract to achieve FB, it is also necessary that the individual rationality constraint (EC.55) is binding and, therefore, $a_{GP} = V_G - b_{GP}\mathbb{E}Q_G(d_G^*, \mathbf{d}_S^*) + k_G(d_G^*)$. Set

$$r_{GP} = \frac{\frac{\partial}{\partial d_G}\mathbb{E}Q_P(d_G^*, \mathbf{d}_S^*) - \sum_{i=1}^{k} k_{SVi} n \Lambda_i'(d_G^*)}{\phi n \frac{\partial}{\partial d_G}\mathbb{E}Q_G(d_G^*, \mathbf{d}_S^*)}, \tag{EC.84}$$

$$f_{GP} = k_G(d_G^*)/(\phi n), \quad t_{GP} = \mathbb{E}Q_G(d_G^*, d_S^*) - \frac{V_G}{n\phi r_{GP}}, \tag{EC.85}$$

to ensure that $v_{GP}(q_G) = a_{GP} + b_{GP} q_G = n\phi\,(f_{GP} + r_{GP}(q_G - t_{GP}))$.

Part 3: deriving the optimal capitation contracts for the i-th SP ($\boldsymbol{v_{SCi}}$). Assume there exists a linear contract $v_{SCi}(q_P) = a_{SCi} + b_{SCi} q_P$ for the i-th SP , under which the first-best can be achieved. Then, when holding the aforementioned contract, $d_{Si}^* \in \arg\max_{d_{Si}\in[0,\infty)} \mathbb{E}\left[a_{SCi} + b_{SCi}\mathcal{Q}_P(d_G^*, \mathbf{d}_S) - k_{SFi}(d_{Si}) - \mathcal{S}_i(d_G^*)k_{SVi}\right]$, a maximization problem that can be reduced (using Lemma C.7.1) to its FOC:

$$k_{SFi}'(d_{Si}) = b_{SCi}\frac{\partial \mathbb{E}Q_P(d_G^*, \mathbf{d}_S)}{\partial d_{Si}}. \tag{EC.86}$$

Thus, $(d_G, \mathbf{d}_S) = (d_G^*, \mathbf{d}_S^*)$ needs to simultaneously solve (EC.78) and (EC.86). Equating the RHS of those two equations and solving for b_{SCi} yields $b_{SCi} = 1$ as the unique solution. For the contract to achieve FB, it is also necessary that the individual rationality constraint (EC.56) is binding, from which we get $a_{SCi} = V_{Si} + k_{SVi} n \Lambda_i(d_G^*) + k_{SFi}(d_{Si}^*) - b_{SCi}\mathbb{E}Q_P(d_G^*, \mathbf{d}_S^*)$. Setting

$$r_{SCi} = \frac{1}{n}, \quad f_{SCi} = k_{SVi}\Lambda_i(d_G^*) + k_{SFi}(d_{Si}^*)/n, \quad t_{SCi} = \mathbb{E}Q_P(d_G^*, \mathbf{d}_S^*) - V_{Si} \tag{EC.87}$$

ensures that $v_{SCi}(q_P) = a_{SCi} + b_{SCi} q_P = n\,(f_{SCi} + r_{SCi}(q_P - t_{SCi}))$.

Part 4: deriving the optimal per-patient contract for the i-th SP ($\boldsymbol{v_{SPi}}$). Note $\mathbb{E}Q_{Si}(d_G, \mathbf{d}_S)$ is concave in d_{Si} but not in d_G. Assume there exists a linear contract $v_{SPi}(q_S) = a_{SPi} + b_{SPi} q_{Si}$ for the i-th SP, under which the first-best can be achieved. Then, when holding the aforementioned contract, $d_{Si}^* \in \arg\max_{d_{Si}\in[0,\infty)} \mathbb{E}\left[a_{SPi} + b_{SPi}\mathcal{Q}_{Si}(d_G^*, \mathbf{d}_S) - k_{SFi}(d_{Si}) - \mathcal{S}_i(d_G^*)k_{SVi}\right]$, a maximization problem that can be reduced (using Lemma C.7.1) to its FOC:

$$k_{SFi}'(d_{Si}) = b_{SPi}\frac{\partial \mathbb{E}Q_{Si}(d_G^*, \mathbf{d}_S)}{\partial d_{Si}}. \tag{EC.88}$$

Thus, $(d_G, \mathbf{d}_S) = (d_G^*, \mathbf{d}_S^*)$ needs to simultaneously solve (EC.78) and (EC.88). Equating the RHS of those two equations and solving for b_{SPi} yields $b_{SPi} = 1$ as the unique solution. For the contract to achieve FB, it is also necessary that the individual rationality constraint (EC.56) is binding, using which yields $a_{SPi} = V_{Si} + k_{SVi} n \Lambda_i(d_G^*) + k_{SFi}(d_{Si}^*) - b_{SPi}\mathbb{E}Q_{Si}(d_G^*, \mathbf{d}_S^*)$ as the unique solution for a_{SPi}. Setting

$$r_{SPi} = \frac{1}{n\Lambda_i(d_G^*)}, \quad f_{SPi} = k_{SVi} + \frac{k_{SFi}(d_{Si}^*)}{n\Lambda_i(d_G^*)}, \quad t_{SPi} = \mathbb{E}Q_{Si}(d_G^*, \mathbf{d}_S^*) - V_{Si}, \tag{EC.89}$$

ensures that $v_{SPi}(q_{Si}) = a_{SPi} + b_{SPi} q_{Si} = n\Lambda_i(d_G^*)\,(f_{SPi} + r_{SPo}(q_{Si} - t_{SPi}))$.

Part 5: sufficiency. Finally, giving the GP either of the contracts v_{GC} or v_{GP} as given in parts 1 and 2 of this proof, while giving every SP either of the contracts v_{SCi} or v_{SPi} (where i is the SP's index), as defined in parts 3 and 4 of this proof, will result in the following. GP's best response to the SP choosing $\mathbf{d}_S^*$ will be to choose d_G^*, as shown in parts 1 and 2 of this proof, while the i-th SP's best response to the GP choosing d_G^* will be to choose d_{Si}^* (and the best response will not depend on what other specialists are deciding), as shown in parts 3 and 4 of this proof. Thus, $(d_G^*, \mathbf{d}_S^*)$ is a Nash Equilibrium and solves the individual compatibility constraints (EC.53)–(EC.54). The system efficiency function is maximized for $(d_G^*, \mathbf{d}_S^*)$, from the definition of the first-best. The individual rationality constraints (EC.55)–(EC.56) are binding, as shown in parts 1-4; thus, agents will accept the contracts, and all the value generated will be appropriated by the principal. Hence all such contract pairs achieve the first-best and thus solve the NICP in (EC.52)–(EC.56). □

Proof of Prop. C.3.2. Proof follows the structure of Prop. 2. We first derive an expression for $\mathbb{E}Q_A$. Denote by M the (random) number of patients who are treated neither by neither the GP nor any of the SPs. Thus, M is distributed according to $\text{Bin}(n(1-\phi), 1 - \sum_{i=1}^k \overline{\lambda}_i)$ as M is equal to the number of non-adherent patients ($n(1-\phi)$) who do not develop complications (each has a $1 - \sum_{i=1}^k \overline{\lambda}_i$ probability of not developing a complication). Using Wald's equation gives $\mathbb{E}M = n(1 - \sum_{i=1}^k \overline{\lambda}_i)(1-\phi)$, applying which to (EC.49), we obtain

$$\mathbb{E}Q_A(d_G, \mathbf{d}_S) = \mathbb{E}Q_P(d_G, \mathbf{d}_S) - q_P^0(1 - \sum_{i=1}^k \overline{\lambda}_i)(1-\phi). \tag{EC.90}$$

The first-best decisions (d_G^*, d_S^*) are solutions to the optimization problem (EC.76), or equivalently (from Lemma C.7.1) to FOCs (EC.77) and (EC.78). Equilibrium decisions of a unified provider under group contract v are given by the incentive compatibility constraint (EC.59). Specifically, for the linear group contracts $v_{AC}(q_P) = a_{AC} + q_P$ and $v_{AP}(q_A) = a_{AP} + q_A$, the objective function in the incentive compatibility constraint (EC.59) differs from the system efficiency function–the objective of (EC.76)–only by a constant, thus the set of maximizers is the same. Consequently, a group under such a contract will make first-best decisions. The optimal a_{AC} and a_{AP} are then derived from (EC.60) to ensure individual rationality is binding, which yields $a_{AC} = V_G + k_G(d_G^*) + \sum_{i=1}^k (V_{Si} + k_{SFi}(d_{Si}^*) + k_{SVi} n\Lambda_i(d_G^*)) - \mathbb{E}Q_P(d_G^*, \mathbf{d}_S^*)$ and $a_{AP} = V_G + k_G(d_G^*) + \sum_{i=1}^k (V_{Si} + k_{SFi}(d_{Si}^*) + k_{SVi} n\Lambda_i(d_G^*)) - \mathbb{E}Q_A(d_G^*, \mathbf{d}_S^*)$. Then, set

$$r_{AC} = \frac{1}{n}, \quad f_{AC} = \frac{k_G(d_G^*) + \sum_{i=1}^k (k_{SFi}(d_{Si}^*) + k_{SVi} n\Lambda_i(d_G^*))}{n}, \quad t_{AC} = \mathbb{E}Q_P(d_G^*, \mathbf{d}_S^*) - V_G - \sum_{i=1}^k V_{Si},$$

$$r_{AP} = \frac{1}{n(1 - (1 - \sum_{i=1}^k \overline{\lambda}_i)(1-\phi))}, \quad f_{AP} = \frac{k_G(d_G^*) + \sum_{i=1}^k (k_{SFi}(d_{Si}^*) + k_{SVi} n\Lambda_i(d_G^*))}{n(1 - (1 - \sum_{i=1}^k \overline{\lambda}_i)(1-\phi))},$$

$$t_{AP} = \mathbb{E}Q_A(d_G^*, \mathbf{d}_S^*) - V_G - \sum_{i=1}^{k} V_{Si},$$

to ensure that $v_{AC}(\boldsymbol{q}_P) = a_{AC} + n\boldsymbol{q}_P = n(t_{AC} + r_{AC}(\boldsymbol{q}_P - t_{AC}))$ and $v_{AP}(\boldsymbol{q}_A) = a_{AP} + n(1-(1-\sum_{i=1}^{k}\overline{\lambda}_i)(1-\phi))\boldsymbol{q}_A = n(1-(1-\sum_{i=1}^{k}\overline{\lambda}_i)(1-\phi))((t_{AP} + r_{AP}(\boldsymbol{q}_P - t_{AP})))$. □

Proof of Theorem C.3.1. Denote the index of the contract's metric of choice as $j \in \{P, A\}$. The induced GP decision $\tilde{d}_G$ satisfies the FOC of the first incentive compatibility constraint (EC.62):

$$\frac{\partial}{\partial d_G}\mathbb{E}v(\mathcal{Q}_j(\tilde{d_G}, \tilde{\mathbf{d}}_S)) = (k+1)k_G'(\tilde{d_G}).$$

As the error term is independent of provider decisions, we can express the signal as $\mathcal{Q}_j(\tilde{d_G}, \tilde{\mathbf{d}}_S) = \mathbb{E}Q_j(\tilde{d_G}, \tilde{\mathbf{d}}_S) + \varepsilon_j$, where ε_j is the stochastic error term. Thus, we can rewrite the FOC as

$$\mathbb{E}\left[v'(\mathcal{Q}_j(\tilde{d_G}, \tilde{\mathbf{d}}_S))\frac{\partial}{\partial d_G}\left(\mathbb{E}Q_j(\tilde{d_G}, \tilde{\mathbf{d}}_S) + \varepsilon_j\right)\right] = (k+1)k_G'(\tilde{d_G}),$$

$$\mathbb{E}\left[v'(\mathcal{Q}_j(\tilde{d_G}, \tilde{\mathbf{d}}_S))\right] = \frac{(k+1)k_G'(\tilde{d_G})}{\frac{\partial}{\partial d_G}\mathbb{E}Q_j(\tilde{d_G}, \tilde{\mathbf{d}}_S)}. \tag{EC.91}$$

Doing the same for the other k incentive compatibility constraints, as given by (EC.63), yields

$$\mathbb{E}\left[v'(\mathcal{Q}_j(\tilde{d_G}, \tilde{\mathbf{d}}_S))\right] = \frac{(k+1)k_{SFi}'(\tilde{\mathbf{d}}_S)}{\frac{\partial}{\partial d_{Si}}\mathbb{E}Q_j(\tilde{d_G}, \tilde{\mathbf{d}}_S(d_{Si}))}, \forall i \in \{1, .., k\}. \tag{EC.92}$$

If $j = P$, the performance of this contract can be replicated by a linear contract $v^\dagger(\boldsymbol{q}_P) \doteq a + b\boldsymbol{q}_P$ by setting $b \doteq (1+k)k_G'(\tilde{d_G})/\frac{\partial}{\partial d_G}\mathbb{E}Q_P(\tilde{d_G}, \tilde{\mathbf{d}}_S)$. Using Lemma C.7.1, such $v^\dagger$ is going to induce the unique solution of the very same FOCs, and is thus guaranteed to also induce decisions $\tilde{d_G}, \tilde{\mathbf{d}}_S$. From (EC.90), we have that $\frac{\partial}{\partial d_G}\mathbb{E}Q_A(d_G, \mathbf{d}_S) = \frac{\partial}{\partial d_G}\mathbb{E}Q_P(d_G, \mathbf{d}_S)$, thus identical contract works even if $i = A$. We can also ensure that this is done at the lowest possible cost to the principal by lowering the fixed pay component a until one of the participation constraints (EC.64)-(EC.65) is binding. Analogously to proof of Prop. C.3.1, $v^\dagger$ can also be expressed as an outcomes-adjusted capitation contract so that $v^\dagger(\boldsymbol{q}_P) = n\,(f + r(\boldsymbol{q}_P - t))$. □

Proof of Theorem C.3.2.

Part 1. Let $v(\boldsymbol{q}_P)$ be the optimal group contract. Note that we can assume without loss of generality that its argument is $\boldsymbol{q}_P$, as any group contract that is a function of q_A can be expressed using (EC.90) as a function of q_P instead. Denote by $v^\dagger(\boldsymbol{q}_P) = a + b\boldsymbol{q}_P$ the linear capitation contract that replicates the performance of $v(\boldsymbol{q}_P)$ (as introduced in the proof of Theorem C.3.1). Then, consider the situation when the following individual contracts are given to the agents: $v_G(\boldsymbol{q}_P) \doteq g_G + b\frac{1}{|C|}\boldsymbol{q}_P$ for the GP and $v_{Si}(\boldsymbol{q}_P) \doteq g_{Si} + b\frac{1}{|C|}\boldsymbol{q}_P$ for the SPs (for the moment, g_G and g_{Si}-s are undefined constants). Observe that under these contracts, the incentive compatibility constraint for each

provider – as given by (EC.69)-(EC.72) – is the same irrespective of that provider's membership in the collusive coalition; thus, this set of contracts will induce the same decisions, no matter which coalition forms. The incentive compatibility constraints are also equivalent to incentive compatibility constraints in the group problem (EC.62)-(EC.63), when the group holds $v^\dagger(\boldsymbol{q}_P)$, so this set of individual contracts will induce the same decisions as $v^\dagger(\boldsymbol{q}_P)$ (thus also $v(\boldsymbol{q}_P)$). Because the induced decisions are the same, irrespective of which coalition forms, then so will be the sum of all agent's incomes, and consequently, these contracts will be collusion-proof as there is no way that the same joint income can be split in such a way to make both of the agents better off when colluding. The constants g_G and g_{Si} can be set to make the participation constraint (EC.73) binding. Finally, with these g_G, g_{Si} in mind, the contracts v_G and $\mathbf{v}_S$ induce the same decisions and outcomes as v, but unlike v, do so in a cost efficient way, completing part 1 of the theorem.

Part 2. The naïve contracts solve the non-collusive incentive compatibility constraints (EC.70), (EC.72), because those are equivalent to constraints (EC.53)-(EC.54) in the naïve problem. Naïve participation constraints (EC.55)-(EC.56) imply the MICP participation constraint (EC.73) irrespective of whether collusion occurs. Additionally, if the contracts are collusion-proof, then if (EC.55)-(EC.56) are binding, then so is (EC.73). Finally, constraints (EC.69) and (EC.71) are redundant for collusion-proof contracts. Thus, collusion-proof solutions of the NICP also solve the MICP, and the same incentive compatibility constraints induce the same decisions (the first-best ones). Because the participation constraints are binding, first-best is achieved in the MICP as well. □

Proof of Prop. C.3.3. For the per-patient contracts, the statement follows immediately from noticing that the compensation of each specialist does not depend on the actions of another. For capitation contracts, the statement follows from inserting the optimal capitation contracts given by (EC.89) into the MICP and noticing the resulting equivalence between the incentive compatibility constraints (EC.71) and (EC.71). □

Proof of Theorem C.4.1. The proof uses the τ, μ substitution as given in (EC.1)-(EC.2), and is done in two parts, first for the effects of collusion when the SP holds a capitation contract v_{SC}, then for per-patient contract v_{SP}.

Part 1: SP holds v_{SC}. Without collusion, the agents will make first-best decisions (τ^*, μ^*), which follows from Prop. 2. By definition of first-best decisions, $(\tau^*, \mu^*) \in \arg\max_{\tau,\mu} \mathbb{E}Q_P(\tau,\mu) - k_G(\tau) - k_{SF}(\mu) - k_{SV} n \Lambda(\tau)$. Equivalently $(\tau^*, \mu^*) \in \arg\max_{\tau,\mu} g(\tau, \mu; 1, 0)$, with g given by Lemma 3. Applying Lemma 3 and using its notation for $\tilde{\tau}(\alpha, \beta)$ and $\tilde{\mu}(\alpha, \beta)$ gives us $\tau^* = \tilde{\tau}(1,0)$, $\mu^* = \tilde{\mu}(1,0)$.

Under coordinated collusion, the agents will choose decisions $(\tilde{\tau}^{CC}, \tilde{\mu}^{CC})$ that maximize joint profits, as given by (EC.74). If the GP holds the capitation contract v_{GC} as given by

Prop. 1, the optimization problem given by (EC.74) is equivalent to $(\tilde{\tau}^{CC}_{C,C}, \tilde{\mu}^{CC}_{C,C}) \in \arg\max_{\tau,\mu}(1+nr_{GC})\mathbb{E}Q_P(\tau,\mu) - k_G(\tau) - k_{SF}(\mu) - k_{SV}n\Lambda(\tau) = \arg\max_{\tau,\mu} g(\tau,\mu;1+nr_{GC},0)$, thus by Lemma 3: $\tilde{\tau}^{CC}_{C,C} = \tilde{\tau}(1+nr_{GC},0) > \tilde{\tau}(1,0) = \tau^*$ and $\mathbb{E}Q_P(\tilde{\tau}^{CC}_{C,C}, \tilde{\mu}^{CC}_{C,C}) = \mathbb{E}Q_P(\tilde{\tau}(1+nr_{GC},0), \tilde{\mu}(1+nr_{GC},0)) > \mathbb{E}Q_P(\tilde{\tau}(1,0), \tilde{\mu}(1,0)) = \mathbb{E}Q_P(\tau^*,\mu^*)$.

Analogously, if the GP holds the per-patient contract v_{GP} as given by Prop. 1, the optimization problem given by (EC.74) is equivalent to $(\tilde{\tau}^{CC}_{P,C}, \tilde{\mu}^{CC}_{P,C}) \in \arg\max_{\tau,\mu} \mathbb{E}Q_P(\tau,\mu) + r_{GP}n\phi\mathbb{E}Q_G(\tau,\mu) - k_G(\tau) - k_{SF}(\mu) - k_{SV}n\Lambda(\tau) = \arg\max_{\tau,\mu} g(\tau,\mu;1,r_{GP}\phi n)$, thus by Lemma 3: $\tilde{\tau}^{CC}_{P,C} = \tilde{\tau}(1, r_{GP}\phi n) > \tilde{\tau}(1,0) = \tau^*$ and $\mathbb{E}Q_P(\tilde{\tau}^{CC}_{P,C}, \tilde{\mu}^{CC}_{P,C}) = \mathbb{E}Q_P(\tilde{\tau}(1,r_{GP}\phi n), \tilde{\mu}(1,r_{GP}\phi n)) > \mathbb{E}Q_P(\tilde{\tau}(1,0), \tilde{\mu}(1,0)) = \mathbb{E}Q_P(\tau^*,\mu^*)$, which completes the proof for parts (ii) and (iii) of the theorem.

Part (i) of the theorem follows from $\lambda(\tilde{\tau}^{CC}_{P,C}) < \lambda(\tau^*)$ and $\lambda(\tilde{\tau}^{CC}_{C,C}) < \lambda(\tau^*)$ as (τ^*,μ^*) is the sole maximizer of system efficiency function u (by Lemma 1). Likewise, agents' joint profit is higher under collusion because the decisions under collusion are the sole maximizers of the joint profit function as shown above, showing part (iv) of the theorem. Let $(\tilde{\tau}^{CC}, \tilde{\mu}^{CC}) \in \{(\tilde{\tau}^{CC}_{C,C}, \tilde{\mu}^{CC}_{C,C}), (\tilde{\tau}^{CC}_{P,C}, \tilde{\mu}^{CC}_{P,C})\}$. Because $u(\tilde{\tau}^{CC}, \tilde{\mu}^{CC}) < u(\tau^*,\mu^*)$ and $\mathbb{E}Q_P(\tilde{\tau}^{CC}, \tilde{\mu}^{CC}) > \mathbb{E}Q_P(\tau^*,\mu^*)$, from (5), we have

$$\mathbb{E}\left[k_G(\tilde{\tau}^{CC}) + k_{SF}(\tilde{\mu}^{CC}) + k_{SV}\mathcal{S}(\tilde{\tau}^{CC})\right] > \mathbb{E}\left[k_G(\tau^*) + k_{SF}(\mu^*) + k_{SV}\mathcal{S}(\tau^*)\right],$$

that is: the agents' total costs go up under collusion. Because the agents' joint profit increases under collusion despite the increased cost, it can only be due to contract payouts (and thus government expenditure) being higher under collusion, showing part (v) of the theorem.

Part 2: SP holds v_{SP}. We show parts of the theorem (vi)-(x) for the case when the GP holds a capitation contract ($k = C$); the case when the GP holds a per-patient contract ($k = P$) is analogous.

From (EC.74): if the GP holds contract v_{GC}, then agents engaged in coordinated collusion will make decisions $\tilde{\tau}^{CC}_{C,P}$,$\tilde{\mu}^{CC}_{C,P}$, which solve

$$\underset{(\tau,\mu)\in[1/\overline{\lambda},\infty)\times[\underline{\mu},\infty)}{\arg\max} \quad \mathbb{E}\left[v_{GC}(\mathcal{Q}_P(\tau,\mu)) + v_{SP}(\mathcal{Q}_S(\tau,\mu)) - k_G(\tau) - k_{SF}(\mu) - \mathcal{S}(\tau)k_{SV}\right]. \tag{EC.93}$$

Using (4), (7), the expressions for v_{GC}, v_{SP} in Prop. 1, and the notation of Prop. 1, the objective function above can be decomposed:

$$\Big((1+b_{GC})\mathbb{E}Q_P(\tau,\mu) - k_G(\tau) - k_{SF}(\mu) - k_{SV}n\Lambda(\tau)\Big) + a_{GC} + a_{SP} + q^0_P\frac{\phi}{\tau} + q^0_P\frac{1-\phi-\underline{\tau}}{\underline{\tau}}. \tag{EC.94}$$

The term in brackets of (EC.94) is equal to $g(\tau,\mu,1+b_{GC},0)$ as defined in Lemma 3, a jointly concave function with an interior maximum (applying Lemma 1), which is maximized at $(\tilde{\tau}^{CC}_{C,C}, \tilde{\mu}^{CC}_{C,C})$, as shown in Part 1 of this proof. The term $\frac{\phi q^0_P}{\tau}$ is a univariate, decreasing and convex function of τ, while the remaining terms in (EC.94) are constants. Consequently, as the objective function is a sum of jointly concave and univariate decreasing convex functions, it either has an interior maximum with $\tilde{\tau}^{CC}_{C,P} \in (\underline{\tau}, \tilde{\tau}^{CC}_{C,C})$ or a corner maximum, in which case $\tilde{\tau}^{CC}_{C,P} = \underline{\tau}$. In either case

$\tilde{\tau}^{CC}_{C,P} < \tilde{\tau}^{CC}_{C,C}$, showing part (viii) of the theorem. By Topkis (1978), we have $\tilde{\mu}^{CC}_{C,P} > \tilde{\mu}^{CC}_{C,C}$, showing part (ix). To show part (vi), it is sufficient to demonstrate that $(\tilde{\tau}^{CC}_{C,P}, \tilde{\mu}^{CC}_{C,P}) \neq (\tau^*, \mu^*)$ as (τ^*, μ^*) is the sole maximizer of system efficiency function u (by Lemma 1). Assume $\tau^{CC}_{C,P} = \tau^*$, then from (EC.94), we have $\tilde{\mu}^{CC}_{C,P} \in \arg\max_\mu (1 + nr_{GC})\mathbb{E}Q_P(\tau^*, \mu) - k_{SF}(\mu)$, whereas from (5), we have $\mu^* \in \arg\max_\mu \mathbb{E}Q_P(\tau^*, \mu) - k_{SF}(\mu)$. Applying Lemma 3 and using its notation gives $\tilde{\mu}^{CC}_{C,P} = \tilde{\mu}(1 + nr_{GC}, 0) > \tilde{\mu}(1, 0) = \mu^*$, thus $(\tilde{\tau}^{CC}_{C,P}, \tilde{\mu}^{CC}_{C,P}) \neq (\tau^*, \mu^*)$. Note from (20) and (21) that (EC.93) is equivalent to $\arg\max_{(\tau,\mu)\in[1/\bar{\lambda},\infty)\times[\underline{\mu},\infty)} \pi_G(v_{GC}, \tau, \mu) + \pi_S(v_{SP}, \tau, \mu)$. Then, part (vii) of the theorem follows from the fact that $(\tilde{\tau}^{CC}_{C,P}, \tilde{\mu}^{CC}_{C,P})$ maximizes (EC.93) but (τ^*, μ^*) does not. Applying the function $\hat{\mu}$ as defined in Lemma 3 gives us: $\tilde{\mu}^{CC}_{C,P} = \hat{\mu}(\tilde{\tau}^{CC}_{C,P}, 1 + nr_{GC}, 0)$ and $\tilde{\mu}^{CC}_{C,C} = \hat{\mu}(\tilde{\tau}^{CC}_{C,C}, 1 + nr_{GC}, 0)$. Part (x) of the theorem then follows from $\tilde{\tau}^{CC}_{C,C} > \tilde{\tau}^{CC}_{C,P}$ and $\mathbb{E}Q_P(\tau, \hat{\mu}(\tau, 1 + nr_{GC}, 0)) \uparrow \tau$ (by Lemma 3). □

Proof of Proposition C.5.1. Here, we find it useful to express waiting times and health as functions of inter-arrival time $\tau \doteq 1/(n\Lambda(d_G)) = \phi\bar{\lambda}/(1 + d_G) + (1 - \phi)\bar{\lambda}$ and service rate $\mu \doteq \mu(d_G) = \underline{\mu} + \theta d_S$ instead of d_G and d_S. Then, using the heavy traffic approximation of Kingman (1961), it follows that the expected waiting time in the queue is

$$\mathbb{E}W(\tau, \mu) \approx \frac{\mathrm{cv}_A^2 + \mathrm{cv}_S^2}{2} \frac{1}{\mu\tau - 1} \frac{1}{\mu},$$

from where the Hessian of $\mathbb{E}W(\tau, \mu)$ is

$$\mathbf{H}(\mathbb{E}W(\tau, \mu)) = \begin{bmatrix} \dfrac{\mu(\mathrm{cv}_A^2 + \mathrm{cv}_S^2)}{(\mu\tau - 1)^3} & \dfrac{\tau(\mathrm{cv}_A^2 + \mathrm{cv}_S^2)}{(\mu\tau - 1)^3} \\ \dfrac{\tau(\mathrm{cv}_A^2 + \mathrm{cv}_S^2)}{(\mu\tau - 1)^3} & \dfrac{(3\mu\tau(\mu\tau - 1) + 1)(\mathrm{cv}_A^2 + \mathrm{cv}_S^2)}{\mu^3(\mu\tau - 1)^3} \end{bmatrix}.$$

Here we can use the steady state condition $(\mu\tau - 1 > 0)$ to see that the leading principal minor is positive $(\mathbf{H}_{1,1}(\mathbb{E}W(\tau, \mu)) > 0)$ and that $|\mathbf{H}(\mathbb{E}W(\tau, \mu))| = (\mathrm{cv}_A^2 + \mathrm{cv}_S^2)^2(2\mu\tau - 1)/(\mu^2(\mu\tau - 1)^5) > 0$, thus the Hessian is positive definite, and hence, $\mathbb{E}W(\tau, \mu)$ is jointly convex. Thus, noting that we need to add the expected service time to the waiting time to arrive at the sojourn time, we can express the expected average population health as

$$\mathbb{E}Q_P(\tau, \mu) = q_P^0 \left(1 - \frac{a}{n\tau} - \frac{b}{n\tau}\mathbb{E}W(\tau, \mu) - \frac{b}{\mu n\tau}\right). \tag{EC.95}$$

Finally, the expression for $\mathbb{E}Q_P(d_G, d_S)$ in (EC.75) follows from (EC.95) by revering the τ, μ substitution and the concavity of $\mathbb{E}Q_P(d_G, d_S)$ follows from the rules for the preservation of concavity under compositions (Boyd and Vandenberghe 2004, p. 86). □

References

Alizamir S, De Véricourt F, Sun P (2013) Diagnostic accuracy under congestion. *Management Science* 59(1):157–171.

Anand KS, Paç MF, Veeraraghavan S (2011) Quality–speed conundrum: Trade-offs in customer-intensive services. *Management Science* 57(1):40–56.

Bennett E (1987) Nash bargaining solutions of multiparty bargaining problems. *The Logic of Multiparty Systems*, 67–78 (Springer).

Borus JS, Laffel L (2010) Adherence challenges in the management of type 1 diabetes in adolescents: prevention and intervention. *Current Opinion in Pediatrics* 22(4):405.

Boyd S, Vandenberghe L (2004) *Convex optimization* (Cambridge Univ. Press).

Clarke P, Gray A, Holman R (2002) Estimating utility values for health states of type 2 diabetic patients using the EQ-5D (UKPDS 62). *Medical Decision Making* 22(4):340–349.

Currie CJ, Peyrot M, Morgan CL, et al. (2012) The impact of treatment noncompliance on mortality in people with type 2 diabetes. *Diabetes care* 35(6):1279–1284.

Cutler DM, Richardson E (1998) The value of health: 1970-1990. *American Economic Review* 88(2):97–100.

Diabetes UK (2014) The cost of diabetes report. Diabetes UK, https://www.diabetes.org.uk/resources-s3/2017-11/diabetes%20uk%20cost%20of%20diabetes%20report.pdf, Accessed 12/05/2021.

Diabetes UK (2019) Diabetes: Facts and stats. https://www.diabetes.org.uk/professionals/position-statements-reports/statistics, Accessed 1 May 2021.

Ferris FL (1993) How effective are treatments for diabetic retinopathy? *JAMA* 269(10):1290–1291.

García-Pérez LE, Álvarez M, Dilla T, Gil-Guillén V, Orozco-Beltrán D (2013) Adherence to therapies in patients with type 2 diabetes. *Diabetes Therapy* 4(2):175–194.

Grossman M (1972) On the concept of health capital and the demand for health. *J. Political Economy* 80(2):223–255.

Javitt JC, Aiello LP (1996) Cost-effectiveness of detecting and treating diabetic retinopathy. *Annals of Internal Medicine* 125(11):939.

Kahan JP, Rapoport A (2014) *Theories of coalition formation* (Psychology Press).

Kc DS, Terwiesch C (2011) The effects of focus on performance: Evidence from California hospitals. *Management Science* 57(11):1897–1912.

Khalid JM, et al. (2014) Rates and risk of hospitalisation among patients with type 2 diabetes: retrospective cohort study using the UK general practice research database linked to English hospital episode statistics. *International Journal of Clinical Practice* 68(1):40–48.

Kingman JFC (1961) The single server queue in heavy traffic. *Proc. Cambridge Philosophical Society* 57:902–904.

Leal J, Gray AM, Clarke PM (2009) Development of life-expectancy tables for people with type 2 diabetes. *European Heart Journal* 30(7):834–839.

Lee CP, Chertow GM, Zenios SA (2009) An empiric estimate of the value of life: Updating the renal dialysis cost-effectiveness standard. *Value in Health* 12(1):80–87.

Maniadakis N, Konstantakopoulou E (2019) Cost effectiveness of treatments for diabetic retinopathy: a systematic literature review. *PharmacoEconomics* 37(8):995–1010.

Mathur R, et al. (2017) Population trends in the 10-year incidence and prevalence of diabetic retinopathy in the uk: a cohort study in the clinical practice research datalink 2004–2014. *BMJ Open* 7(2).

McNabb WL (1997) Adherence in diabetes: can we define it and can we measure it? *Diabetes Care* 20(2):215.

Mitchell P, et al. (2012) Cost-effectiveness of ranibizumab in treatment of diabetic macular oedema (dme) causing visual impairment: evidence from the RESTORE trial. *Brit. Jour. Ophthalmol.* 96(5):688–693.

Nash JF (1950) The bargaining problem. *Econometrica* 18(2):155–162.

Neumann PJ, Cohen JT, Weinstein MC (2014) Updating cost-effectiveness: the curious resilience of the $50,000-per-qaly threshold. *New England Journal of Medicine* 371(9):796–797.

Rein DB, Wirth KE, Johnson CA, Lee PP (2007) Estimating quality-adjusted life year losses associated with visual field deficits using methodological approaches. *Ophthalmic epidemiology* 14(4):258–264.

Schectman JM, Nadkarni MM, Voss JD (2002) The association between diabetes metabolic control and drug adherence in an indigent population. *Diabetes care* 25(6):1015–1021.

Taddeo D, Egedy M, Frappier J (2008) Adherence to treatment in adolescents. *Paediatrics & Child Health* 13(1):19.

Topkis DM (1978) Minimizing a submodular function on a lattice. *Operations Research* 26(2):305–321.

UK NHS (2019) General Practice Workforce, Official Statistics, 30 September 2019. NIHR, https://digital.nhs.uk/data-and-information/publications/statistical/general-and-personal-medical-services/final-30-september-2019, Accessed 10/05/2021.

Whitt W (1993) Approximations for the GI/G/M queue. *Production and Operations Management* 2(2):114–161.