Optimal Contracting for Low-Value Care

Angie Acquatella *

September 16, 2025

A significant share of health care spending comes from low-value, or minimal benefit, services. Why do countries pay for these services and what can optimal contracts do to contain this type of spending? I study optimal contracting for 'flat of the curve' medicine in a principal agent framework with altruistic providers. I find that altruism leads to inefficient overprovision of care when governments use cost-based reimbursement, and that treatment caps optimally contain excess spending. Global budgets, as in the U.K., achieve first best outcomes when perfect risk adjustment is possible, or when providers are substantially altruistic.

Keywords: provider payment, contracts, altruism
JEL Codes: I11, I13, D86

*Contact: acqua@bu.edu. I thank Oliver Hart, David Cutler, Stefanie Stancheva, Ed Glaeser, Jerry Green, Dan Brown, Andy Newman, Claudia Goldin, Nathan Hendren, David Sibley, Mike Powell, Grace McCormack, Aileen Devlin, Tamar Oostrom, Adriano Fernandes, Frank Pinter, Samantha Burn, Daniel Ly, Amanda Kreider, Ljubica Ristovska, Chris Walker, Ed Kong, the seminar participants in the Harvard Public Economics seminar, and the National Bureau of Economics Research Aging and Health fellows for helpful suggestions. This material is based upon work supported by the National Institute on Aging under Grant Number T32-AG000186, and by the National Science Foundation under Grant Number DGE 1745303.

1 Introduction

Health care spending constitutes 17% of GDP in the US, and the share of spending coming from low-value, or minimal benefit, services has been estimated to lie between 20% and 30% of total spending (Berwick and Hackbarth, 2012; Institute of Medicine, 2010). Relatively little attention has been paid to low-value care contracting, and high-volumes of these procedures can accrue to significant economic magnitudes.

Why do countries end up paying for low-value services and what can optimal contracts do to contain this spending? 'Low-value' services are those which give little to no marginal benefit to patients, i.e. 'flat-of the curve' medicine. One can hardly argue that there is no inefficient overuse of health care services in the US and Europe, at least to *some* extent, and low-value services are certainly a part of the story. While much of the literature has debated whether 'flat of the curve' medicine explains variations in health care spending (Fuchs, 2004; Chandra and Staiger, 2007), or inefficiencies (Abaluck et. al., 2016; Kyle and Williams, 2017; Silver, 2020), the medical literature seems to agree that there exists a class of services in which the marginal value of care is small.[1]

This article studies optimal contracting for 'flat of the curve' medicine in a principal agent framework with altruistic providers. The model is primarily designed to describe services classified as low-value by the medical literature (weakly positive benefits, but lower than marginal cost), which typically take place in an outpatient setting (e.g. diagnostic services, physical therapy, evaluation and management services) and constitute around $60 billion of Medicare yearly spending.

There are at least two reasons why this contracting problem is important. First, provider payment contracts seem to be all over the map. Table 1 illustrates the variations in contract form across services insured by Medicare: some have treatment caps that set upper limits on administered care, while others issue additional outlier payments for patients that receive unusually high levels of care.[2] Variations in contract form is not just true across types of services, but more broadly, across insurers for the same class of services. Lack of consensus in contracts suggests that this contracting problem is not well understood. My theory rationalizes why we might see features treatment caps on services that have small marginal value at high levels of care, or outlier payments when patient treatment needs are disperse.

Second, standard procurement contract models that do not account for the special features of this market may lead to sub-optimal contracting. Since the beginnings of Medicare in 1965, regulators have been concerned with designing reimbursement policies that compensate providers for their input costs. Cost and cost-plus reimbursement combined with provider moral hazard quickly sky-rocketed health care spending, leading to a series of provider payment contract reforms that targeted incentives so that providers would internalize their marginal costs of care. These reformed high powered incentive contracts ran into adverse selection problems, as they created incentives for providers to keep cheap patients and dismiss complicated patients. Cutler (2010) argues informally

[1] See Mafi et. al. (2017), Schwartz et. al. (2014), Coronini-Cronberg et. al. (2015)

[2] Private insurers use versions of these as well, in addition to contract features like prior authorization, step therapy, and selective contracting.

Table 1: Expenditures Across Provider Administered Treatments

Prospective Payment Services (based on ex-ante expected *treatment costs)*

Health care services	*Medicare Payments 2017 ($ billions)*	*Unit of Payment*	*Treatment Caps*	*Outlier Adjustment*
Acute Inpatient	$ 118	Discharge		X
Physician Services	$ 69	Service		
Outpatient Hospital	$ 51	Service		X
Administered Drugs	$ 32	Dose		
Nursing & Rehabilitation	$ 28	Day of Stay	X	
Hospice	$ 18	Day of Care	X	
Home Health Care	$ 18	60-day episode		X
Dialysis	$ 11	Single treatment		
Outpatient Therapy	$ 8	Service	X	
Inpatient Rehabilitation	$ 8	Discharge		X
Long-term Care	$ 5	Discharge		X
Inpatient Psychiatric	$ 4	Day of Stay		X

Source: MedPAC Payment Basics, 2019.

that insurance and physician cost (plus) reimbursement provide incentives for additional care above what is optimal.

There are two distinguishing features in this market one has to account for: provider altruism and unobserved patient heterogeneity. Providers seem to respond to ethical considerations, in conjunction with financial considerations. Whether motivated by professionalism or the Hippocratic oath, providers seem to value patient health when making treatment decisions, at least to some extent, and the empirical literature rejects the hypothesis that providers are pure profit-maximizers.[3] Accounting for heterogeneity is central for the health care setting, and an old idea tracing back to Arrow (1963): health *care* affects each patient's *health* differently, so health outcomes may look different for observably similar patients that get the same treatment.

By developing a model of provider contracting that accounts for these two features, I arrive at three key results. First, failing to account for altruism and reimbursing providers for their costs creates incentives for inefficient over-treatment: a provider who values patient health and does not bear the costs of treatment will do too much, and at the expense of higher insurance premiums (or taxpayers, for a public insurer). Second, covering a condition that includes some patients with unpredictably higher treatment needs creates incentives for the altruistic provider to treat every patient at the maximum covered level. Unless contained by a treatment cap, covering high-need patients inefficiently raises the level of health care administered to everyone else. Third, systems that pay using global budgets can achieve first best outcomes under two conditions: when health care costs are observable and contractible, or when the distribution of health care costs is known

[3] See Chen and Lakdawalla (2019), Chang and Jacobson (2017), Dranove (2012), Roomkin and Weisbrod (1999); Gregg et al. (2008).

and providers are altruistic. When patients who benefit more from treatment are also more expensive, non-altruistic providers will never have an incentive to give high-need patients sufficient care because these patients are less profitable.

In this paper, I primarily focus on patient heterogeneity in *health benefit* from treatment, which is well suited to describe physician services, nursing, rehabilitation, outpatient therapy, among others. To illustrate heterogeneity in benefit, consider two patients with Chronic Obstructive Pulmonary Disease (COPD) and identical medical histories. COPD impairs breathing and gets worse over time. To mitigate symptoms of COPD, patients can receive treatment of 'pulmonary rehabilitation,' which is an outpatient therapy service that involves time with a respiratory therapist who guides the patient through a series of physical exercises, breathing retraining exercises, as well as nutritional counseling.

In this example, the costs of treating both patients are the same: the provider has to sit with the patient and perform the same exercises. It is also the case that the provider's ability to improve patient health is the same across these two patients. Suppose one of the patients has a health conscious spouse who cooks healthy meals and walks frequently, while the other patient has a lazy spouse. They can both come in for ten visits and be better off. However, the patient with the lazy spouse may marginally benefit more from tenth visit. This is because being periodically reminded to go on walks at therapy may go a long way for the patient who does not have these reminders at home. The different needs of these two patients begs the question of which payment contract can cater to both of them.

In a different example, such as psychiatric care for schizophrenic patients, patient service needs depend on factors beyond diagnosis, such as the severity and duration of the disease, receptivity to treatment, and social support available to promote treatment compliance. These types of patients have the distinctive feature that they are heterogenously costly to treat (with some patients requiring physical restraint and active monitoring), with some of these costs being not contractible. Section 4 generalizes the model to consider such cases, which serves to illustrate the contracting frictions that arise under imperfect risk adjustment and low altruism.

This article is part of a broader research agenda on the optimal design of provider payments. Ellis and McGuire (1986) were the first to evaluate the optimality of provider reimbursement contracts in a theoretical framework with partially altruistic providers.[4] Then followed a number of papers that studied the optimal provider contracting problem under restricted contract forms that are commonly used by regulators, with features that broadly apply to the inpatient hospital care setting and acute care services.[5]

[4]The model in this paper embeds the Ellis and McGuire model, though is more general, differing in two ways. The first is patient heterogeneity, as I study an insurer that has *one* contract for *a heterogeneous patient pool*, while they focus on the optimal contract for a provider-patient pair. The second is the contracting objective function. In their model, there is no 'loss' term for provider payments, which means the insurer faces no financial trade-off from a payment scheme that induces over-treatment.

[5]Ellis (1998) studies provider competition under fixed price (prospective payment) contracts in a model with patient cost and benefit heterogeneity. Choné and Ma (2011) characterize the properties of optimal payment schemes in a general theoretical framework with provider altruism and continuous health benefit heterogeneity, and provide ex-

To my knowledge, this is the first article to focus on contracting for low-value services, and differs from prior work in three aspects. First, I study unobserved heterogeneity across patients and not providers. Second, providers in my model do not 'feed on altruism,' and the altruistic portion of utility does not enter the participation constraint. This less common modeling choice is consistent with the recent literature on corporate social responsibility or market morality that considers agents with altruistic preferences (Broccardo, Hart, and Zingales, 2021; Dewatripont and Tirole, 2021). Third, the generalized asymmetric information in my model describes patients for whom high cost treatment correspond to significant positive benefits, which has not been studied to my knowledge.

This work also relates to the theoretical study of ethical professional norms, and the design of optimal incentives under such settings. Broccardo, Hart, and Zingales (2021) study socially responsible shareholders and their optimal investment strategies. Dewatripont and Tirole (2020) study market behavior and competition when firms and consumers have ethical considerations regarding production (e.g. fair-trade coffee). Besley and Gathak (2005) study firm-worker matching in the context of non-profit organizations with ethical objectives. Bénabou and Tirole (2006) study the interaction of incentives and prosocial behavior. Francois and Vlassopoulos (2008) provide a survey of the various ways the literature has modeled prosocial behavior, and discusses implications for optimal incentives. Finally, Besley (2020) studies tax-compliance of civic-minded citizens and the optimal provision of public goods under general citizen preferences. This article differs from prior work in that it focuses on contracting with a single, ethically motivated, agent in an asymmetric information environment characteristic of the health care setting. The main new theoretical result that arises from this setting is that the optimal contract no longer has 'no distortion at the top': incentive rents with ethically motivated agents make it expensive to implement the efficient action for the unobserved types at either end.

Lastly, this work provides a model that is consistent with the empirical literature, and therefore draws upon the large body of empirical research that estimates how financial incentives affect care delivered and patient health outcomes. Evidence suggests that, on average, providers in the U.S. try to avoid marginally unprofitable patients (Gandhi, 2021; Desai et. al., 2009; Cram et. al., 2008; Ettner, 1993; Greenlees et. al., 1982; Uili, 1995; Ching et al., 2015; Newhouse,1989), despite regulations that make this practice illegal. Nonetheless, provider's care decisions are not entirely a function of patient profitability, as providers are more likely to deliver unprofitable care when patients are from low socioeconomic backgrounds (Chen and Lakdawalla, 2019).

amples of parametrizations that illustrate such properties which could be potentially instrumental for structural work. De Fraja (2000) characterizes the optimal payment contract when providers are heterogeneously efficient and patients have heterogeneous benefits and costs from treatment. Ma (1994) considers the implications of lump sum contracts (prospective) and cost reimbursement on quality of the provider and cost-reduction incentives. Jack (2005) characterizes the optimal contract under heterogeneous provider altruism, with non-contractible quality. Malcomson (2005) studies the problem in a model without provider altruism. Chalkey and Kahlil (2005) study reimbursement contracts that may condition on health outcomes (in addition to treatment) in a model where patient demand also plays a role in health care delivered. Maréchal and Mougeout (2004) study the optimal two part tariff when patient costs are heterogeneous and providers choose can choose unobservable cost-reducing effort. Gaynor, Mehta, Richards-Shubik (2020) estimate the optimal contract parameters using structural methods in a setting where providers face heterogeneous costs of treatment.

The paper will proceed as follows. In Section 2, I present a model for provider reimbursement. In Section 3, the optimal contract is derived and I show that altruism works against the insurer. Cost reimbursement is efficient absent altruism, but creates incentives for inefficient over-provision in the presence of altruism. In Section 4, I consider a generalized version of asymmetric information and study the implications of provider commitment to treating marginally unprofitable patients. I find that altruism can work in favor of the insurer, and show that a global budget contract implements the first best when providers are are sufficiently altruistic, in spite of imperfect risk adjustment. In Section 5, I conclude.

2 A Model for Provider Reimbursement

The model involves three actors: health care providers, patients, and an insurer. The insurer covers medical care for his patients by paying providers directly. Patients are passive and always accept the treatments recommended by their provider.

Patients and health benefit heterogeneity

Patients derive positive health benefits from any level health care, but some patients benefit substantially more than others from a particular level of care. Consider a health production function that is everywhere increasing in care, which means that a lot of health care *never harms* the patient[6] (either because providers follow the "do no harm" clause of the Hippocratic oath, or because the additional health care does not worsen the condition of the patient).

The health production in this model is a good fit for procedures such as diagnostic services, physical therapy, provider office visits, evaluation and management services, diabetes treatment, or dialysis (the procedure itself, not the anemia medications given in parallel), for example, because that patients always weakly benefit from additional care.

The heterogeneous health benefits are known to the provider but not the insurer. The asymmetric information in the model means that the insurer will know there are two patients with COPD, but only the provider knows which patient 'type' will benefit a lot from 10 pulmonary rehabilitation sessions. That said, both patients would be better off with 10 visits, per the assumption on the health function, but it incremental benefit of the patient with the health conscious spouse may just not be worth the additional costs of care. The unobserved patient heterogeneity (to the insurer) is a feature of the model that helps explain why two patients with identical medical record may receive different treatments. That is, even if the insurer collected the best information possible about a patient, it may be hard to know ex-ante the intensity of treatment needed by any particular patient, and ex-post whether the intensity yielded enough health benefits to as to justify the treatment costs.

[6]This health production function excludes treatments of the sort studied in Gaynor, Mehta, Richards-Shubik (2020), where too much treatment *harms* the patient. This modeling choice echoes the Chandra and Skinner (2012) Type II and Type III class of treatments, but is slightly more restrictive.

Formally, let the health production function, $h(x, \theta)$, depend on treatment, x, and patient type, θ, where h describes the dollar value of total health gains from treatment (e.g. value of additional quality adjusted life-years derived from pulmonary rehabilitation therapy). Treatment x may be continuous or discrete, and can encompass intensity of services (e.g. Relative Value Unit), level of treatment (e.g. number of office visits), or probability of major procedure (e.g. catheterization). Suppose θ is private information, known only to the provider, encoding how much benefit a particular patient derives from treatment. Think of the high θ types as patients who get very large benefits from treatment, at all treatment levels.

Assumption 1 *Assume that $h(x, \theta)$ is increasing and concave in treatment, where $h_x(x, \theta) \geq 0$ and $h_{xx}(x, \theta) \leq 0$, $\forall\ x$; and that marginal health gains from treatment are increasing in θ, so that $h_{x\theta}(x, \theta) > 0$, $\forall\ x$.*

Cost of treatment

Suppose that the insurer may observe and contract on[7] the treatment costs incurred by the provider. Let costs of treatment be given by $c(x)$, where costs to the provider depend only on the treatment, and not on the patient type.

Assumption 2 *Assume that costs of treatment $c(x)$ are positive, $c(x) \geq 0$, increasing in treatment and weakly convex, $c_x(x) \geq 0$ and $c_{xx}(x) \geq 0$.*

In the baseline model, not allowing the cost function to depend on θ means that the wedge created by asymmetric information will only emerge through the provider's valuation of patient health, which limits how much we can decompose the separate effect of asymmetric information and imperfect altruism. A provider who does not consider patient health in his health care decision (in addition to financial considerations) will be perfectly indifferent between giving the healthy patient a low level of care versus a high level of care.

This simplification helps illustrate the distortions that arise with altruism when providers are (successfully) fully reimbursed for their costs of care. Since the beginnings of Medicare in 1965, regulators have been chiefly concerned with designing reimbursement policies that compensate providers for their input costs. Absent altruism, such reimbursement policies should lead to efficient outcomes because cost reimbursement should make purely financially motivated agents *indifferent* between intensive care or non-intensive care. The empirical health literature has also been concerned with asymmetric information on costs, and developing methods to improve and perfect risk adjustment. This main specification of the model is pedagogically motivated, and is meant to illustrate the 'costs of altruism' if a world with perfect risk adjustment were feasible. Section 5 extends the model to the more realistic case of asymmetric information about patient costs, which serves to illustrate the contracting frictions that arise under imperfect risk adjustment and low altruism.

[7] While costs of treatment may not be entirely clear to the insurer, in practice, there are some settings where costs are 'more' observable, such as physical therapy, evaluation and management services, or diagnostic services.

Provider treatment decision

Consider a partially altruistic provider values patient health and profits. Denote the reimbursement contract by $r(x)$ and costs of treatment by $c(x)$. Let α be the 'altruism' parameter, which scales the provider's valuation of patient health. Assume $\alpha \geq 0$. [8] Provider utility from treating patient θ is

$$U(x, r) = \alpha h(x, \theta) + r(x) - c(x).$$

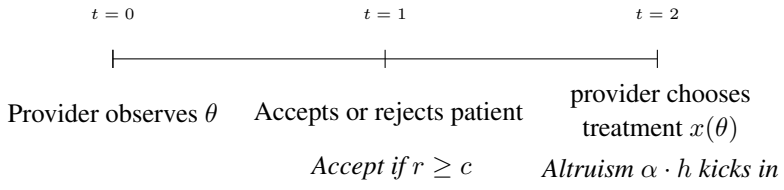
The parameter α is the marginal rate of substitution between profits and patient health. [9] A provider with a $\alpha = 1$ will value patient health gains from treatment at exactly their value for the patient (and for society).

Timing and patient admission decision

The provider knows the patient type at time zero. Suppose that the provider first decides whether or not to treat the patient, and then chooses the treatment quantity to give. The provider is forward looking and knows the costs of treatment for each patient when he observes the patient type.

Assumption 3 *Assume the provider does not treat the patient if reimbursements are less than costs.*

Assumption 3 implies an additional implementation constraint: profits per patient must be non-negative, $r(x) - c(x) \geq 0$. [10] The main specification of the model considers contracts with non-negative profits for two reasons. First, these contracts would be possible under perfect risk adjustment, and I want to highlight the distortions that arise *beyond* imperfect risk adjustment. Second, there is substantial empirical evidence that marginal patient profitability affects health care decisions on the extensive margin, [11] and guaranteeing non-negative profits for every marginal patient ensures a mechanism in which all patient types get seen.



[8] Notice that $\alpha = 0$ shuts down the asymmetric information in this model. A provider with an $\alpha = 0$ is perfectly indifferent between giving high benefit patients too much or too little care. Indeed, the optimal contract when $\alpha = 0$ implements first best outcomes.

[9] The α parameter is the same as in Ellis and McGuire (1986): it is the ratio of the provider's marginal utilities between health and profits. Health and profits are perfect substitutes with a slope α for the provider. Other papers in the literature have applied a more general function on the provider profits component, as opposed to patient health (Chandra and Skinner 2012; Skinner 2012).

[10] The patient admission rule is similar to those in models of 'patient dumping' (i.e. explicit avoidance of high cost patients) considered by Ma (1994), Ellis (1998), and others, which have focused on how fixed price contracts affect the composition of patients receiving care.

[11] Gandhi (2020), Desai et. al. (2009), Cram et. al. (2008), Ettner (1993), Greenlees et. al. (1982), Gruenberg and Willemain (1982), Uili (1995), Ching et al. (2015), Newhouse (1989).

Assumption 3 also implies an effective 'participation constraint' which *differs* from the provider incentive constraint. The assumption mirrors those in other models of agents with altruistic or other-regarding preferences in different settings, such as Broccardo, Hart, and Zingales (2021), Dewatripont and Tirole (2021), or Besley (2020). Similar to how the utility of consumption increases with past consumption in the Becker and Murphy (1988) model of addition, one can argue that the altruistic component of utility is nonexistent prior to the two parties entering a relationship. More broadly, ensuring non-negative financial payoffs could facilitate organizational cohesion within an organization where agents have heterogeneous altruistic preferences. Hansmann's (1998) work on not-for-profit organizations argues that when firm has a collection of owners, divorcing the rights of firm control from the rights to appropriate the firm's earnings can facilitate the formation of such an organization. That is, agreement on financial payoffs may facilitate the formation of the organization, but financial consideration need not guide the activities that the organization undertakes once it is formed if its owners have general preferences.

In this health care setting, providers tend to practice in groups or affiliate to medical institutions, where bottom line profits appear to be a crucial component of organizational cohesion. Providers and hospitals must pay bills and cover costs and remain in operation; when reimbursements are insufficient, they shut down. Even when we consider critical access hospitals or charity care, the government must typically step in and subsidize these institutions for their losses in order to keep them in operation. Legally, there are various laws and regulations that make it difficult for providers to 'fire' a patient. One could view this under a more behavioral angle and argue that, once the provider accepts a patient, he becomes invested in the patient's health and this warm-glow valuation of health 'kicks in' for the consequent treatment decisions.

Finally, this modeling choice is convenient in the characterization of social optima. In particular, it allows one to use the standard contract theory characterization of first best while circumventing 'double-counting' in models of altruism. Modeling altruism is effectively like turning this into an externality problem, where the agent's actions affect not just his own utility, but also that of someone else. There is an extensive theoretical literature in environmental economics that studies similar problems while keeping the standard formulation of the participation constraint. Why the departure from standard? Unlike pollution externalities, where the social cost of an action is the sum of the private cost and the cost on other people, having altruistic providers does not make the value of the treatment any greater than what it already was. Similarly, the value to society of an organization's charitable activities will not be the sum of the intrinsic valuation of all prospective owners plus the value of the activity itself. Nonetheless, one can hardly argue that altruism will not enter the owner's decision making process at the time of choosing the charitable activity. Therefore, by excluding the altruistic component from the participation constraint, one can characterize social optima as the actions that maximize social value, subject to the agent's participation constraint, while simultaneously studying the implementable action space under altruistically motivated agents.

Insurer's problem

Consider a public insurer who chooses a provider reimbursement contract to maximize a social welfare function (SWF). The insurer values both patient health, net of reimbursements, and provider profits, but not the 'warm glow' altruism component, with relative weight $\eta \in [0, 1)$ on provider profits.[12] One can think of η as the social welfare weight on provider profits. Since providers are generally on the right tail of the income distribution, and social welfare weights are inversely proportional to income, a public insurer may place lower weight on provider profits (relative to patients).

Let the SWF be given by

$$SWF = \overbrace{\underbrace{h}_{\substack{\text{value of}\\ \text{treatment}}} - \underbrace{r}_{\substack{\text{provider}\\ \text{payment}}}}^{\text{patient surplus}} + \eta \overbrace{\underbrace{(r-c)}_{\text{profit}}}^{\text{provider surplus}} .$$

There are a couple of advantages of the SWF proposed here.[13] First, it embeds the contracting objectives of previous papers in the literature, which have taken the opted for either $\eta = 0$ or 1. The contracting objective in Ellis and McGuire (1986) corresponds to $\eta = 1$, as their goal is to attain optimal treatment *quantities* when providers are imperfect agents, but the notion of 'saving' reimbursement dollars is beyond the scope of their paper. Private insurers may be thought of as having an $\eta = 0$, trading off achievable health outcomes against the full implementation costs of those outcomes. Since the provider requires incentive rents to implement higher health outcomes, a private insurer may optimally choose lower health outcomes. A public insurer, conversely, may value giving the provider profits (perhaps because it encourages people to become doctors). The second advantage is that it provides a continuous measure by which the public insurer could value provider profits, providing a flexible framework to accommodate the preferences of a particular regulator.

Suppose the government does not observe patient type and can only contract based on the observed treatment cost. The reimbursement contract, $r(x)$, is chosen to maximize the SWF, taking into account that providers choose treatment according to their objective $U(x, r)$. For expositional ease, I will suppose there are only two types of patients–a very responsive patient, θ_H, and a less

[12] Notice that the set of η excludes $\eta = 1$. This is a technical assumption, which I make deliberately: an insurer with $\eta = 1$ will only care about implementing treatments that maximize health gains net of provider treatment costs, *independent* of how costly implementation is. In other words, reimbursement ceases to be part of the objective function and thus is not uniquely determined. If we think making payments is costly, whether it be because they come from tax-payer dollars, or correspond to patient insurance premiums (both outside of the scope of this model), η cannot be equal to one.

[13] Another possible SWF formulation could have included a loss term on payments per the shadow cost of public funds, per Laffont and Tirole (1993). Others including De Fraja (2000), Jack (2005), and Malcomson (2005) have included such term in their welfare specifications. While these models yield interesting comparative statics, I stay away from including the term here because, in my model, it would imply that the social costs of treatment are greater than just the cost of the treatment. That is, the insurer would worry about making the provider internalize the costs of distortionary taxation, in addition to the costs of treatment, adding an additional dimension of welfare distortion to the model which detracts from the main points of the paper.

responsive patient, θ_L–and later show that the results generalize to the N type case. Let there be share γ of patient type H, and $(1-\gamma)$ of patient type L. In the two type case, the provider treatment decision implies two incentive constraints (IC)'s, and the timing assumption implies two participation constraints (PC)'s. The government's problem is to choose (r_H, r_L) according to the following program.

$$\max_{r_L, r_H} \gamma(h(x_H, \theta_H) - r_H + \eta(r_H - c(x_H))) + (1-\gamma)(h(x_L, \theta_L) - r_L + \eta(r_L - c(x_L))) \quad (1)$$

$$\text{s.t. } \alpha h(x_L, \theta_L) + r_L - c(x_L) \geq \alpha h(x_H, \theta_L) + r_H - c(x_H) \quad \text{(IC L)}$$

$$\alpha h(x_H, \theta_H) + r_H - c(x_H) \geq \alpha h(x_L, \theta_H) + r_L - c(x_L) \quad \text{(IC H)}$$

$$r_L - c(x_L) \geq 0 \quad \text{(PC L)}$$

$$r_H - c(x_H) \geq 0. \quad \text{(PC H)}$$

3 Altruistic Providers Give Too Much Low Value Care

The efficient allocation gives each type the level of care at which marginal health benefit equals marginal cost. The first best is not sustainable in the second best because providers have an incentive to over-treat the patients whose efficient level is low: altruism implies that low marginal value care accrues positively to the provider objective function, and cost reimbursement impedes providers from internalizing the costs of their low-value care.

3.1 The efficient level of care

In the first best, the planner solves the problem without the incentive constraints. The two participation constraints bind, and reimbursement is exactly equal to cost. This can be easily seen by looking at the planner's problem; since the objective function will be strictly decreasing when $\eta < 1$, the two participation constraints bind. The first best treatments hence set marginal health benefit of treatment equal to marginal cost, $h_x(x^{FB}, \theta) = c_x(x^{FB})$. Given the assumptions on $h(x, \theta)$ with respect to θ, the first best level of treatment is always increasing in the type. As seen in Figure 1, the first best treatment level for the high type, denoted by x_H^{FB}, is greater than the first best treatment level for the low type, denoted by x_L^{FB}.

However, the first best is not sustainable in the second best when the government must take into account the incentive constraints. At cost based reimbursement, the provider's objective is simply the health production function scaled. Since the treatment levels are increasing in type, and $h(x, \theta)$ is increasing in x,

$$x_H^{FB} > x_L^{FB} \implies \alpha h(x_L, \theta_L) + \underbrace{r_L - c(x_L)}_{=0 \text{ at FB}} \not\geq \alpha h(x_H, \theta_L) + \underbrace{r_H - c(x_H)}_{=0 \text{ at FB}} \quad \text{(IC L)}$$

so (ICL) is violated at the first best. When we try to sustain the first best, we end up with the

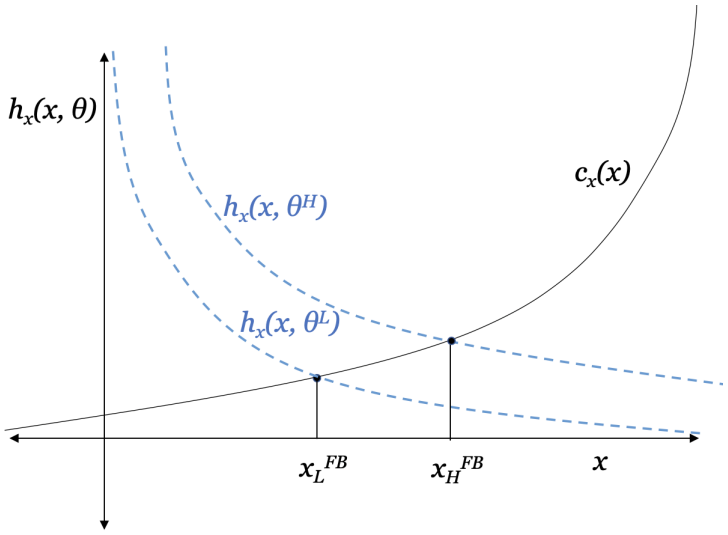


Figure 1: First Best Treatment Levels

provider over-treating the low type. This is suggestive that, in the second best, the contract will have to give rents to the provider for treating the low type at an appropriately lower level.

The altruistic provider in this model will always have an incentive to give more treatment than is socially optimal: more treatment always improves the health of the patient, and the participation constraint of the provider effectively subsidizes costs of treatment to zero. In the COPD example, the models says that the respiratory therapist will want to see the patient for as many hours of respiratory therapy as he can give: he gets positive utility from the small marginal health gains of the patient, no matter how small, without bearing any of the health care costs.

This result may seem surprising– one might expect, ex-ante, that altruism should work in favor of the insurer, and not against. The reason we get over-provision here is cost reimbursement. To my knowledge, few articles have made this point. Mougeout and Naegelen (2013) in an article written in french arrive at a similar conclusion, but other work has been more concerned about imperfect measurement of cost than with the implications of cost reimbursement for low value care.

Altruism, combined with asymmetric information, is the source of the wedge between the insurer and the provider in this model. By giving the provider a profit which exceeds the value of the marginal health gains for patients that need less treatment, however, the insurer can create a financial incentive for the provider to treat the low type at a lower level. By giving a profit on low levels of care, the insurer can contain overall health care supply.

A first concern in implementation is therefore whether the provider's incentives are aligned with the insurer objectives. Lemma 1 shows that this is indeed the case–the altruistic provider has an incentive to give high types more treatment than low types, and the first best treatment levels are monotonic in the type.

Lemma 1 *The two incentive constraints jointly imply monotonicity, and therefore any incentive compatible contract must give a higher treatment level to the high type, relative to the low type.*

Proof. Adding (IC H) and (IC L) and rearranging terms shows that $x_H \geq x_L$ must hold, given that $h_x(x, \theta)$ is increasing in θ by assumption. Otherwise, we would get a contradiction.

$$\alpha h(x_L, \theta_L) + r_L - c(x_L) \geq \alpha h(x_H, \theta_L) + r_H - c(x_H) \quad \text{(IC L)}$$
$$\alpha h(x_H, \theta_H) + r_H - c(x_H) \geq \alpha h(x_L, \theta_H) + r_L - c(x_L) \quad \text{(IC H)}$$
$$\implies h(x_H, \theta_H) - h(x_L, \theta_H) \geq h(x_H, \theta_L) - h(x_L, \theta_L)$$

■

3.2 Distortion at both ends: A new efficiency vs incentive rents trade-off

The optimal contract does not implement the efficient level of treatment *for either type*. This result differs from the standard Mussa and Rosen (1978) price discrimination model with asymmetric information, where one would expect one type to be distorted (due to incentive rents) and the other type to get the efficient (first best) allocation.[14] Altruism distorts the efficiency versus incentive rents trade-off by making efficient allocations more expensive to the principal: he has to dissuade the altruistic provider from doing too much low value (but marginally beneficial) care.

Since the provider over-treats the low type when we try to sustain the first best, the insurer's problem is about designing incentives that keep the provider from over-treating. By pushing down the treatment level of the high type, and pushing up the treatment level of the low type, the temptation to over-treat can be mitigated. This is exactly what the second best contract ends up doing. The following proposition formalizes this result.

Proposition 1 *The optimal contract distorts treatment levels for all types: high types get less treatment and low types get more, relative to their first best levels.*

Proof. Consider a modified problem that has only (ICL), (PCH), and a monotonicity constraint, $x_H \geq x_L$. I will solve this problem instead, and then show its solution coincides with that of the original problem.

The Binding Constraints: In the modified problem, (PCH) must bind; otherwise, one could reduce r_H by $\epsilon > 0$, which would increase the SWF by $\gamma(1-\eta)\epsilon > 0$ while still satisfying all other constraints. So it will be optimal to reduce r_H until (PCH) binds.

Similarly, (ICL) will also bind at the optimum. If it didn't bind, we would have $r_L > c(x_L) + \alpha h(x_H, \theta_L) - \alpha h(x_L, \theta_L)$, and one could reduce r_L by $\epsilon > 0$, still satisfy all the constraints, and increase the SWF by $(1-\gamma)(1-\eta)\epsilon > 0$. Therefore, the payments must be $r_H = c(x_H)$ and $r_L = c(x_L) + \alpha h(x_H, \theta_L) - \alpha h(x_L, \theta_L)$. These two binding constraints imply that $r_H = c(x_H)$ and $r_L = c(x_L) + \alpha h(x_H, \theta_L) - \alpha h(x_L, \theta_L)$.

[14] If $\alpha = 0$ and the private information entered through the cost function as well, the model would be identical to the standard non-linear pricing problem with asymmetric information from Mussa and Rosen (1978). Makris and Siciliani (2013) study a two type model with provider altruism and asymmetric information in the cost function, and they similarly find that $\alpha > 0$ implies an optimal second-best contract with distortion at both ends.

Verifying the solution coincides with the original problem: We have to check that (ICH) and (PCL) are satisfied at the solution. (PCL) is satisfied since r_L has a payment premium above cost, positive by monotonicity. That is, $r_L = c(x_L) + \underbrace{\alpha h(x_H, \theta_L) - \alpha h(x_L, \theta_L)}_{\geq 0} > c(x_L)$. Turning to (ICH), we can evaluate it at the (r_H, r_L) to obtain that,

$$\begin{aligned} & \alpha h(x_H, \theta_H) - \alpha h(x_L, \theta_H) \geq \alpha h(x_H, \theta_L) - \alpha h(x_L, \theta_L) \\ \implies \quad & \alpha h(x_H, \theta_H) + \underbrace{r_H - c(x_H)}_{=0} \geq \alpha h(x_L, \theta_H) + \underbrace{r_L - c(x_L)}_{=\alpha h(x_H, \theta_L) - \alpha h(x_L, \theta_L)} \quad . \end{aligned} \tag{IC H}$$

Characterizing the solution: Suppose that in equilibrium, $x_H > x_L$ with strict inequality; we can characterize the treatment levels implemented by the optimal contract via the first order conditions of the SWF with respect to the treatment levels.

$$\frac{\partial\, SWF}{\partial x_H} = 0 \implies : \qquad h_x(x_H, \theta_H) - c_x(x_H) = (1-\eta)\frac{1-\gamma}{\gamma}\alpha h_x(x_H, \theta_L) \tag{1.1}$$

$$\frac{\partial\, SWF}{\partial x_L} = 0 \implies : \qquad h_x(x_L, \theta_L) - c_x(x_L) = -(1-\eta)\alpha h_x(x_L, \theta_L) \tag{1.2}$$

The left hand side would be zero at the first best treatment levels. Since partial of the SWF with respect to x_H is positive, the right hand side of (1.1) is positive, meaning that the x_H which solves the first order condition is *less than* the first best x_H^{FB}. Via a parallel logic, the x_L which solves the first order condition (1.2) is *greater than* the first best x_L^{FB}.

If the first order conditions (1.1) and (1.2) yield equilibrium x's such that $x_H \leq x_L$, then the solution is NOT characterized by these two conditions. The only way to satisfy the two incentive constraints and the monotonicity condition is by setting $x_H = x_L = x_P$, where x_P denotes the 'pooled' treatment level. The optimal x_P will be such that average health gains are maximized. That is

$$x_P \in \arg\max_x h(x_P, \theta_H) + (1-\gamma)h(x_P, \theta_L) - r_P + \eta(r_P - c(x))$$

which is maximized at $\gamma h_x(x_P, \theta_H) + (1-\gamma)h_x(x_P, \theta_L) = c_x(x_P)$, and $r_P = c(x_P)$. Clearly, both types are distorted from their first best levels, with the high type getting less, and the low type getting more.

■

Since, the provider has an incentive to over-treat the low type, as he derives positive utility for the small marginal health gains (and at no private cost in a world of $r \geq c$), reducing the gap in treatments reduces marginal health gains from over-treating the low type, mitigating the 'temptation' to over-treat. Incentive rents required to keep the provider from over-treating make the first best levels too expensive to implement.

Notice that α is close to zero gets us closer the first best because it shuts down the asymmetric

information distortion. The baseline model is designed to illustrate the *interplay of altruism with asymmetric information* before introducing the additional complications that occur at low levels of altruism. When α is near zero, the provider does not care: he is indifferent between giving low types a lower or higher treatment level. Unless there is some other source of information asymmetry between the provider and the insurer, such as in patient treatment costs, the no altruism case corresponds to a situation with symmetric information. Absent asymmetric information, the insurer can always condition the contract on θ and implement first best treatment levels for all patient types using pure cost reimbursement, which is how we get first best outcomes.

Figure 2 shows graphically the distortion on both types. As η gets close to one, we get closer to first best: the more the planner values provider profits, the less he minds giving incentive rents to the provider. Since the contract *can* implement first best treatment levels, the only reason to distort second best quantities is the large incentive rent required, particularly for providers with high α. In fact, for any $\eta < 1$, the larger the α, the *farther* we get from first best.

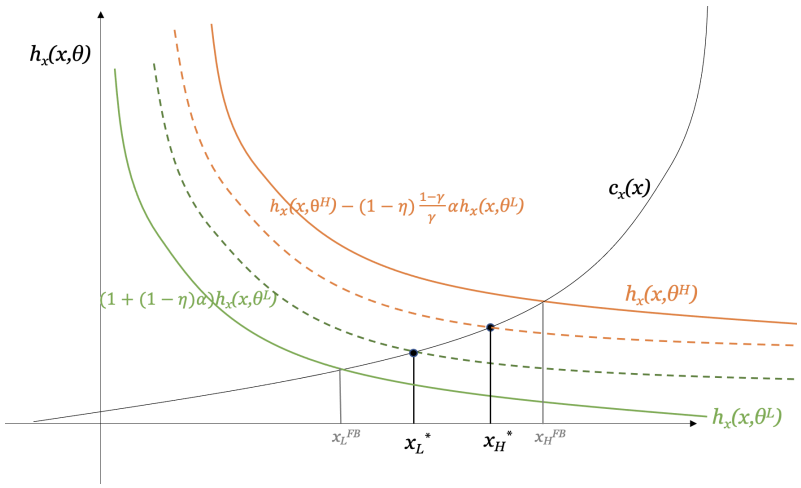


Figure 2: Optimal Contract Implemented Treatments, Unpooled

What about pooling? Since the optimal contract pushes treatments across types closer together, it could be the case that these treatment levels overlap. The proposition below provides a sufficient condition for when the treatment levels do *not* overlap: very large relative health gains for the high type. It is worthwhile for the insurer to pay the incentive rent when the dollar value of health gains from the high types getting higher treatment are sufficiently greater than the cost of paying the incentive rent. Conversely, when the high types do not benefit, health-wise, that much from additional treatment, it becomes optimal to implement a pooling solution.

Proposition 2 *A sufficient condition for the optimal second best contract to NOT pool all types is,*

$$h_x(x,\theta_H) > \left(1+(1-\eta)\frac{\alpha}{\gamma}\right) h_x(x,\theta_L) \tag{2}$$

Proof. Proof: Suppose that (2) holds. Since the cost function across types is the same, we can order the solutions to the first order conditions (1.1) and (1.2), relative to each other. Rearranging

(1.1) and (1.2) to that they both equal $c_x(x)$, it follows that the equilibrium treatment level of x_H will be larger than the equilibrium treatment level for x_L if

$$\underbrace{h_x(x_H,\theta_H)-(1-\eta)\frac{1-\gamma}{\gamma}\alpha h_x(x_H,\theta_L)}_{=c_x(x_H)\text{ from condition }(1.1)} > \underbrace{h_x(x_L,\theta_L)+(1-\eta)\alpha h_x(x_L,\theta_L)}_{=c_x(x_L)\text{ from condition }(1.2)}.$$

By rearranging terms, one immediately obtains condition (2). ■

The pooling solution implements a treatment level at which average marginal health benefits equate to marginal cost. Pooling is *more* likely when the share of high types, γ, is small, as condition (2) becomes harder to satisfy. Intuitively, the result makes sense: if there are not many patients that need high levels of treatment, then the insurer will cater the contract to the patients who need less. Notice that at lower levels of η, it becomes more likely that we are in the pooling solution: since the size of incentive rents depends on the difference in health gains across types, a large difference implies that the insurer needs to give higher profits to the provider on the low type. The less the insurer values provider profits, the less likely he will be willing to pay for a contract that implements a separating equilibrium. Similarly, if α is large, it is more likely we are in the pooling solution. For high α, the incentive rents need to be larger to sustain the separating equilibrium; a contract that implements a separating equilibrium will only be worthwhile to the insurer if the health gains accrued from the high type are sufficiently large.

The properties of the second best contract illustrated in the two type case also generalize to the case with N types: there is distortion for all types, with under-provision of care for the highest type, and over-provision for the lowest, relative to the first best levels. I refer the reader to the Appendix for a detailed derivation of the optimal contract, and for the formal proofs of these results.

3.3 Continuum of types: Optimal treatment caps

When patients have widely disperse treatment needs, treatment caps and outlier payments arise as optimal ways of insuring this patient pool. This subsection involves the study of the continuum of types case, as adding types amplifies the asymmetric information problem between the insurer and the provider. The additional insight of this case is that adding types with disproportionately higher efficient levels of care makes it more expensive to insure all the types below, exacerbating the trade-off between efficiency and incentive rents.

Suppose we are again in a world where reimbursements must be greater than or equal to cost for each patient. Consider a condition like low back pain, where treatment needs for observably similar patients may be very different, with some patients requiring 10, 20, or 30 visits to the physical therapist before they start to see significant improvements, while others may take only 5 visits to see the substantive portion of the health improvement due to physical therapy. If the insurer cannot tell ex-ante which patient is which, and can only choose how much to pay for x total visits in a year, choosing to cover the care of the most complicated patients means that the reimbursement

contract must cover 30 visits in a year. The temptation for the provider then becomes treating all patients with 30 visits: the patients who get better after 5, 10, or 20 visits all benefit (slightly) from coming in for 30 visits. Again, the problem is that the health benefits for the 5, 10, and 20 visit patients from coming in 30 times are not large enough to justify the cost of 30 visits.

The main tension for the insurer dealing with a continuum of types thus comes from the highest level of x covered. The optimal contract will cater to the 30 visit patient only when the value in health gains for that patient type is disproportionately large. Otherwise, the optimal contract will have pooling at the top: there will be threshold patient type, θ_T, above which all patients receive the same treatment amount, x_T.

To provide some intuition, consider first going from two types to three. In the two type case, the reimbursement contract had to pay an incentive rent on the low type in order to keep the provider from over-treating him. Adding a third type at the top means that now the provider will want to treat types 1 and 2 at the level of the highest type. While the incentive rent on type 2 will look a lot like the incentive rent on type L in the previous section, the incentive rent on type 1 will have to be larger. This is because adding types to the right also means that the *lowest* type always benefits, health wise, from receiving the treatment level of the *highest* type.

More formally, suppose now that there is a continuum of patients types on an interval $\theta \in [\underline{\theta}, \overline{\theta}]$. In general, the only set of reimbursement contracts which are consistent with the incentive compatibility constraints must have declining profits in type. This is because a provider who values patient health and does not bear the treatment costs already has an intrinsic motivation to over-treat. By giving large rents on the low types, the insurer dissuades the provider from over-treating at the highest covered level.

In the continuum of types case, the optimal second best contract (derived in the appendix), pools types above the threshold θ_T. All types $\theta \geq \theta_T$ receive a fixed treatment quantity, x_T, and not more, because the contract does not allow the provider to choose treatment levels above x_T. The provider is *not* be on his first order condition for types above θ_T under this contract. The second best contract pays according to the following schedule:

$$r(x^*(\theta)) = \begin{cases} c(x_T) & \text{, for } \theta \geq \theta_T \\ c(x^*(\theta)) + \alpha h(x_T, \theta) - \alpha h(x^*(\theta), \theta) & \text{, for } \theta < \theta_T. \end{cases}$$

Just as in the two type case, the insurer disincentivizes the provider from over-treating low types by giving him incentive rents. If profits were constant across types, or even increasing, the provider would have *both* an intrinsic motive and a financial motive to over-treat. This would result in everyone receiving the maximum treatment level covered. Hence, for a reimbursement contract to implement treatment levels that are lower for the low health benefit patients, and higher for the high benefit patients, it must give declining profits in the unobserved type.

The threshold type, θ_T, crucially determines how expensive it is to insure everyone else. The *dispersion* in the unobserved patient heterogeneity is the key driver of higher insurance costs. In

order to disincentivize the provider from treating the 5 visit patients for 30 visits, the reimbursement for 5, 10, and 20 has to give an incentive rent. It would be cheaper for the insurer to reimburse only for 5 visits, so that all patients receive only 5 visits, but it may not be worthwhile to do so if the value of health gains of the 30 visit patient are large.

By pooling types at the top, the insurer can reduce the costs of insuring *everyone else* because he only has to disincentivize providers from treating lower types at type θ_T's efficient level, which can be done with a smaller incentive rent. The maximum treatment level covered by the insurer will depend on whether the additional unobserved types (who have higher treatment needs) have sufficiently large health gains. Since treatment is monotonic in the type, a treatment cap x_T is equivalent to choosing a threshold patient type, θ_T, above which patients get pooled at the same treatment level.

The size of the treatment cap is determined by equating the average marginal health gains of types at the top against the magnitude of the incentive rents that the insurer must pay on all types at the bottom. More formally, we can observe this in the optimality condition for x_T:

$$\frac{dSWF}{dx_T} = \underbrace{\int_{\theta_T}^{\overline{\theta}} (h_x(x_T,\theta) - c_x(x_T)) f(\theta) d\theta}_{\text{average health gains for types above } \theta_T \text{ (at the capped treatment level)}} \underbrace{- (1-\eta) \int_{\underline{\theta}}^{\theta_T} \alpha h_x(x_T,\theta) f(\theta) d\theta}_{\text{incentive rents on types below } \theta_T} = 0$$

The insurer internalizes that by covering high levels of care (which may be efficient for the high θ types), it runs into a moral hazard problem with its providers: they will want to treat everyone at the highest covered level. Absent incentive rents, this insurer would end up paying for every type at the highest covered level. The incentive rents in the second best contract accrue to a lesser amount than paying for every type to receive x_T.

In COPD, for example, Medicare has a treatment cap of 36 hours of pulmonary rehabilitation therapy per patient, per year. For physical therapy, Medicare has a treatment cap of about $2,000 per patient, per year, currently. From speaking to providers of physical therapy, the impression is that providers are happy to bring in patients for more visits because it can only help. The literature has already studied treatment caps, usually under the term 'supply-side limits', where the insurer constrains the amount of care that a provider can give. Work by Pauly (2000), for instance, argued in favor of treatment caps to solve the moral hazard problem on the patient side. My model provides a slightly more nuanced justification for a treatment cap, rooted in the provider's 'altruistic moral hazard.'

The prevalence of treatment caps in the real world, more generally, can be found in services that contribute to 'flat-of-the-curve' spending when used in excess, such as rehabilitation, nursing, and outpatient therapy. While caps arise as an optimal second-best solution to contain 'flat-of-the-curve medicine,' they do not implement efficient levels of care for anyone. The problem is still rooted in the $r \geq c$ constraint: payments that result in non-negative profits for *all* types means that the incentive rents are all relative to the *most expensive* type, or the type that requires the highest

level of treatment. Outlier contracts relax this $r \geq c$ slightly, offering providers ex-post payment adjustments upon additional documentation that a patient was high need.

When the 'value' of health gains for highest type are disproportionately large, the insurer may want to justify raising the insurance costs of everyone else. If there is an outlier within the heterogeneous patient pool who *really* benefits from higher treatment levels, then the insurer will find it worthwhile to pay a higher profit margin on all the other types. This is because the value of that patient's health from high treatment would be high enough to offset how expensive it is to insure him within the pool.

The main takeaway from this extension is that the patients with the highest treatment needs drive up the costs of insuring everyone. The participation constraint in my model is a key driver of this result: reimbursement must exceed cost *for every patient*. If the insurer can relax $r \geq c$, it becomes less expensive to cater efficient treatment levels to more types.

4 The (first best) promise of global budgets

Suppose it were possible for the provider to commit ex-ante to treat every patient, and did not change care decisions based on ex-post, per patient, profitability. If the only source of asymmetric information is patient health benefit from treatment, the optimal reimbursement contract implements efficient levels of care for all types. This is because the relaxed patient admission constraint allows for contracts in which providers internalize the costs of their care. However, if patients who benefit more from high levels of care are also (unobservably) more expensive to treat, the first best is only achievable when providers are sufficiently altruistic, and the optimal contract has pooling when providers are not altruistic at all. In all cases, knowledge of the provider's marginal rate of substitution between patient health and profits is necessary to implement first best outcomes.

There are two applications where the relaxed participation constraint seems a fitting description of the world: one is a global budget setting, such as in the payment scheme used in the United Kingdom. The second is the setting in which the provider does not know the patient type when deciding to treat or not treat the patient, which could correspond to a condition in which individual treatment needs are hard to predict, ex-ante. In both of these applications, we effectively go back to a world of symmetric information in the baseline specification of the model.

Consider an alternative model with a relaxed patient admission constraint where the provider commits to treat all insured patients as long as average reimbursements are weakly greater than average costs. That is, consider a situation in which the provider can shut down the clinic if he is making losses, but cannot turn away an individual patient if reimbursements are below costs, *for*

that particular patient. The insurer's problem is now given by,

$$\max_{r_L, r_H} \gamma(h(x_H, \theta_H) - c(x_H) - (1-\eta)(r_H - c(x_H))) + (1-\gamma)(h(x_L, \theta_L) - c(x_L) - (1-\eta)(r_L - c(x_L)))$$

$$\text{s.t. } \alpha h(x_L, \theta_L) + r_L - c(x_L) \geq \alpha h(x_H, \theta_L) + r_H - c(x_H) \quad \text{(IC L)}$$

$$\alpha h(x_H, \theta_H) + r_H - c(x_H) \geq \alpha h(x_L, \theta_H) + r_L - c(x_L) \quad \text{(IC H)}$$

$$\gamma(r_H - c(x_H)) + (1-\gamma)(r_L - c(x_L)) \geq 0. \quad \text{(PC)}$$

There will be a continuum of optimal contracts, as the incentive constraints will only pin down the minimum payment wedge between r_H and r_L, but there are many pairs (r_H, r_L) that will satisfy the participation constraint.

4.1 Getting the first best

The relaxed participation constraint helps the insurer because, by effectively removing the asymmetric information wedge, the insurer ends up in the first best . Since the provider now accepts patients based on their *expected* reimbursements, the information set of the principal and the agent coincide. The insurer now has a continuum of fee schedules that he can choose from while satisfying all the constraints. Among such set of fee schedules, the insurer can choose a reimbursement where the provider bears the costs of treating all patients on the margin, removing the incentive to over-treat. Now the insurer has contracts within the implementable set with which he can equate marginal health benefit to marginal cost.

Proposition 3 *If the participation constraint of the provider is such that average reimbursements weakly exceed average costs, then we get the first best.*

Proof. As before, treatments must be monotonic in type for the the incentive constraints to jointly hold. Since the objective is decreasing in payments, we know that (PC) must bind, which pins down a relationship between r_L and r_H. The set of optimal contracts is characterized by pairs of (r_L, r_H) that satisfy $\gamma(r_H - c(x_H)) + (1-\gamma)(r_L - c(x_L)) = 0$. Since there is no unique optimal (r_L, r_H), we can leverage the relationship between (r_L, r_H), which is given by $\gamma(r_H - c(x_H)) + (1-\gamma)(r_L - c(x_L)) = 0$. Internalizing in the relationship of (r_L, r_H) into the objective function, we obtain:

$$\gamma(r_H - c(x_H)) + (1-\gamma)(r_L - c(x_L)) = 0$$
$$\implies (x_L^*, x_H^*) \in \arg\max_{x_L, x_H} \gamma(h(x_H, \theta_H) - c(x_H)) + (1-\gamma)(h(x_L, \theta_L) - c(x_L)).$$

The program coincides with the solution to the first best problem, which means the monotonicity condition is automatically satisfied per the assumptions on h. Therefore, any optimal contract will implement the first best treatments. ■

Among the many contracts that implement first best treatments, an interesting one to focus on is $r(x_i) = t + (1-\alpha)c(x_i)$ for $i = \{L, H\}$. It is easy to see why this contract gets us to the first best by looking at the provider's optimization problem. We know that the equilibrium treatment level for each type θ_i is characterized by,

$$x_i \in \arg\max_x \alpha h(x, \theta_i) + \underbrace{r_i - c(x)}_{t+(1-\alpha)c(x_i)-c(x_i)} \implies \alpha(h_x(x_i, \theta_i) - c_x(x_i)) = 0,$$

which coincides with the solution to the first best treatments. In fact, this is the optimal contract in Ellis and McGuire (1986), as it 'undoes' the wedge of imperfect altruism.

Global budgets effectively shut down the asymmetric information wedge in the extensive margin decision of which patients to treat, and leverage the provider's private information in the intensive margin decision of how intensely to treat them. The importance of knowing α becomes first order in this contracting environment.

Independent of whether α is small or large, the first best is always achievable and optimal if costs of treating patients can be perfectly observed by the insurer, and the insurer knows α. When α gets small, the optimal contract almost fully reimburses providers costs on the margin. This suggests that cost reimbursement payment systems would be optimal in a world where providers are purely profit-maximizing (and perfect risk adjustment were possible!). However, this result breaks down as soon as profit differences across patients arise in ways unknown to the insurer.

4.2 Introducing unobserved cost heterogeneity

The promise of global budgets breaks under a more general version of the model in which treatment costs now also depend on patient type, $C(x, \theta)$, and high θ types are more expensive to treat. Consider a patient that has a sudden schizophrenic flare up episode, requiring a hospitalization that may vary in length of stay (i.e. number of days, x). Patients may differ in the severity of their episode, with some of these patients requiring physical restraint or active monitoring, making them heterogenously costly to the provider in ways that are not contractible.

Assumption 4 *Assume that the patients who are more costly to treat per day also benefit more from a longer hospital stay. Formally, assume that*

$$h_{x\theta}(x, \theta) - C_{x\theta}(x, \theta) > 0.$$

Assumption 4 ensures that the x that maximizes $h(x, \theta) - C(x, \theta)$ is increasing in θ. This assumption makes this model different from the one in De Fraja (2000), who also studies unobserved cost heterogeneity, because he assumes that patients with greater benefit are also cheaper to treat. Assume that costs are non-decreasing and weakly convex in x, with $C_x(x, \theta) \geq 0$ and $C_{xx}(x, \theta) \geq 0$. Suppose also that $C_{x\theta}(x, \theta) > 0$ and $C_\theta(x, \theta) \geq 0$. Figure 3shows graphically

that the assumption implies that increases in marginal costs across types correspond to even larger increases in marginal benefits across types.

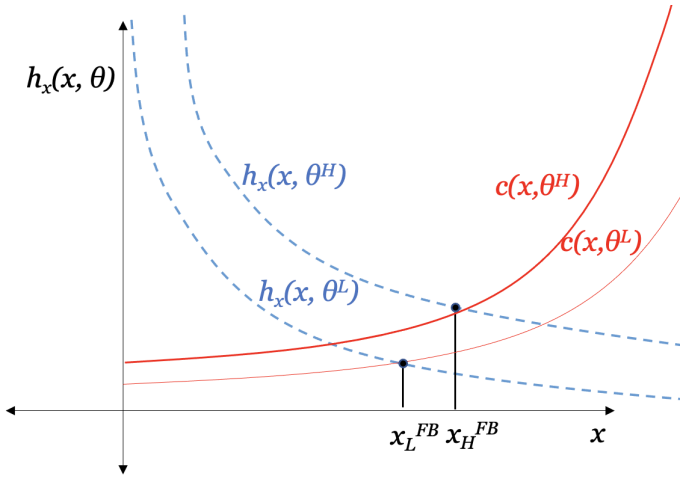


Figure 3: First Best Treatment Levels under Cost and Health Heterogeneity

Suppose that the provider still agrees to the alternative, global budget, participation constraint: he commits ex-ante to admit all patients as long as reimbursements are greater than or equal to costs, on average. For expositional ease, assume that $\eta = 0$ and consider an insurer that does not value provider profits in the social welfare function. The insurer's problem with two patient types is:

$$\max_{r_H, r_L} \gamma(h(x_H, \theta_H) - r(x_H)) + (1-\gamma)(h(x_L, \theta_L) - r(x_L))$$

subject to

$$\alpha h(x_H, \theta_H) + r(x_H) - C(x_H, \theta_H) \geq \alpha h(x_L, \theta_H) + r(x_L) - C(x_L, \theta_H) \quad \text{(IC-H)}$$

$$\alpha h(x_L, \theta_L) + r(x_L) - C(x_L, \theta_L) \geq \alpha h(x_H, \theta_L) + r(x_H) - C(x_H, \theta_L) \quad \text{(IC-L)}$$

$$\gamma(r(x_H) - C(x_H, \theta_H))) + (1-\gamma)(r(x_L) - C(x_L, \theta_L)) \geq 0. \quad \text{(PC-avg)}$$

4.2.1 The case of no altruism

A purely financially motivated provider will only consider the costs of intensive care, without accounting for the benefits of this care. The incentives of this provider are *opposite* to the insurer's objectives: he will want to give less care to patients who benefit from more because the high benefit patients are also more expensive.

Consider a provider with no altruism, with an $\alpha = 0$, so that his treatment decision only takes patient profitability into consideration. Suppose the high type patient's efficient level of care is 10 days, and the low type is 5 days. Suppose the costs of the high type are twice the costs of the low type at any number of days, and suppose that seeing the high type for 10 days costs \$100, while seeing the low type for 5 days costs \$25. If the insurer were to offer a contract that pays \$100 for 10

days or \$25 for 5 days, the non-altruistic provider will have an incentive to over-treat the low type at 10 days: he makes a profit of $\$100 - \$50 > 0$ by doing so. In the two type case, this provider will want to give type L a treatment level x_L that is *greater than* x_H because patient L is cheaper to treat at any level. Lemma 2 formalizes this argument.

Lemma 2 *Suppose that $\alpha = 0$. Then, the provider incentive constraints jointly imply that patient type θ_H receives $x_H^* \leq x_L^*$.*

Proof. Adding the two incentive constraints for a provider with $\alpha = 0$ results in the following inequality:

$$C(x_L, \theta_H) - C(x_H, \theta_H) \geq C(x_L, \theta_L) - C(x_H, \theta_L).$$

Since $C_{x\theta} \geq 0$ by assumption, this inequality can only hold for $x_H \leq x_L$. ■

In order for the provider's incentives to align with the insurer's, α has to be sufficiently large so that the provider's objective values the benefits of giving additional treatment to costlier patients (the θ_H types).[15] The second best contract for a provider with an $\alpha = 0$ pools treatment levels across all types.

Proposition 4 *When $\alpha = 0$, the optimal contract pools both types, and they both receive treatment level such that average marginal health gains are set equal to average marginal costs. That is,*

$$\gamma h_x(x_P, \theta_H) + (1-\gamma) h_x(x_P, \theta_L) = \gamma C_x(x_P, \theta_H) + (1-\gamma) C_x(x_P, \theta_L).$$

Proof. Consider a modified problem with less constraints, (PC-avg) and a monotonicity constraint, $x_H \leq x_L$. We will solve this problem instead, and then show its solution coincides with that of the original problem.

The Binding Constraints: In the modified problem, (PC-avg) must bind because the objective function is decreasing in r_H and r_L. Otherwise, one could reduce r_H by $\epsilon > 0$, strictly increasing the SWF by $\gamma(1-\eta)\epsilon > 0$, while still satisfying the monotonicity constraint, $x_H \leq x_L$. Similarly, one could reduce r_L by $\epsilon > 0$ and increase the SWF by $(1-\gamma)(1-\eta)\epsilon > 0$. So it will be optimal to choose a pair of r_H and r_L so that (PC-avg) binds.

The binding constraint implies that $\gamma r_H + (1-\gamma) r_L = \gamma C(x_H, \theta_H) + (1-\gamma) C(x_L, \theta_L)$, so the problem becomes

$$\begin{aligned} & \max_{x_L, x_H} \gamma(h(x_H, \theta_H) - C(x_H, \theta_H)) + (1-\gamma)(h(x_L, \theta_L) - C(x_L, \theta_L)) \\ s.t. \quad & x_H \leq x_L. \end{aligned}$$

[15] The assumption on $C_{x\theta}$ is a key driver of this result. De Fraja (2000) proposes a model with both health and cost heterogeneity, but assumes that the high benefit patients are also *cheaper* to treat. In that case, the incentives of the provider are never contrary to those of the insurer, regardless of the particular level of altruism. However, the modeling assumption I have made here intends to describe the mental health setting, where the more complicated patients seem to be the ones who benefit the most from intensive care.

By Assumption 4, we know that the x's which solve the objective are $x_H^{FB} > x_L^{FB}$. This means that the monotonicity constraint will bind at the optimum, $x_H = x_L$, and we will be at the pooling solution.

Characterizing the solution: Set $x_H = x_L = x_P$, where x_P denotes the pooled treatment x. The optimal x_P will be such that average marginal health gains are set equal to average marginal costs. That is

$$\gamma h_x(x_P, \theta_H) + (1-\gamma) h_x(x_P, \theta_L) = \gamma C_x(x_P, \theta_H) + (1-\gamma) C_x(x_P, \theta_L).$$

Verifying the solution coincides with the original problem: We have to check that (ICH) and (ICL) are satisfied at the solution. Since $x_H = x_L = x_P$, the two (IC)'s will hold trivially with equality. ■

4.2.2 The case of perfect altruism

On the other hand, a provider who values patient health equally to the insurer will always want to give more care to the high benefit patient, and less care to the low benefit patient. This is because he will internalize the net benefits of the expensive and the cheap patient.

Consider the case of $\alpha = 1$ and the relaxed participation constraint. Continuing with the numerical example, suppose the benefits for the high type, in dollar terms, are \$200 from 10 days and \$100 from 5 days. Suppose the benefits for the low type are \$50 for 5 days and \$60 for 10 days. If the provider's aggregate compensation is equal to aggregate costs of treating one low benefit and one high benefit patient (\$100 + \$25), he will only look at the marginal incentives when deciding how long to see each patient for. The net benefit for the high type patient at 10 days is $\$200 - \$100 = \$100$, which is greater than the net benefit for the high type at 5 days, $\$100 - \$50 = \$50$. For the low type, the net benefit at 5 days is $\$50 - \$25 = \$25$, which is greater than the net benefit at 10 days $\$60 - \$50 = \$10$. In general, the optimal second best contract with an $\alpha = 1$ provider implements the first best.

Proposition 5 *When $\alpha = 1$, the optimal contract implements the first best.*

Proof. Adding the two incentive constraints for an $\alpha = 1$ provider implies monotonicity in the treatment levels, $x_H \geq x_L$. The sum of the two incentive constraints is:

$$h(x_H, \theta_H) - C(x_H, \theta_H) - (h(x_L, \theta_H) - C(x_L, \theta_H)) \geq h(x_H, \theta_L) + r(x_H) - C(x_H, \theta_L) - (h(x_L, \theta_L) - C(x_L, \theta_L)).$$

By Assumption 4, treatment levels (x_H, x_L) satisfying this condition are monotonically increasing in type θ, so the incentive constraints jointly imply that $x_H \geq x_L$.

Consider again a modified problem with the aggregate participation constraint and a monotonicity constraint, $x_H \leq x_L$.

The Binding Constraints: In the modified problem, (PC-avg) must bind because the objective function is decreasing in r_H and r_L. Otherwise, one could reduce r_H by $\epsilon > 0$, strictly increasing

the SWF by $\gamma(1-\eta)\epsilon > 0$, while still satisfying the monotonicity constraint, $x_H \leq x_L$. Similarly, one could reduce r_L by $\epsilon > 0$ and increase the SWF by $(1-\gamma)(1-\eta)\epsilon > 0$. So it will be optimal to choose a pair of r_H and r_L so that (PC-avg) binds.

Characterizing the Solution: The binding constraint implies that $\gamma r_H + (1-\gamma) r_L = \gamma C(x_H, \theta_H) + (1-\gamma) C(x_L, \theta_L)$, so the problem becomes

$$\begin{aligned} & \max_{x_L, x_H} \gamma(h(x_H, \theta_H) - C(x_H, \theta_H)) + (1-\gamma)(h(x_L, \theta_L) - C(x_L, \theta_L)) \\ s.t. \quad & x_L \leq x_H. \end{aligned}$$

The problem coincides with the first best problem, and therefore the solution to the insurer's problem is the first best. ■

At the other corner of perfect altruism, the provider wants to give the first best treatments to each patient type, with high θ types getting more, and at the corner of zero altruism, the provider incentives make him want to reverse the ordering of treatments. These two extremes suggest that there is some minimal level of altruism required so that the provider wants to give more treatment to the high type.

Intuitively, the provider needs to value the health gains of intensive care patients sufficiently so that the benefits, net of costs, are positive from the provider's perspective. Recall that α was the marginal elasticity of substitution between health and profits; a provider who cares too much about profits, relative to patient health, will make decisions based on patient profitability, and the patients who benefit the most from intensive care are unfortunately the most expensive.

4.3 Sufficiently altruistic providers get us to the first best

It is not necessary that providers be purely altruistic to get first best outcomes; we just need providers that value health gains *enough* so that incentives of the provider are aligned with those of the public insurer. Incentives alone are insufficient to get us first best outcomes, but contracts combined with sufficiently altruistic providers can solve this contracting problem.

To illustrate this point, consider the following parametric example.[16] Suppose that treatment affects health according to $h(x, \theta) = \theta(1+\lambda)$ where $\lambda \in \mathbb{R}_+$, and that the costs of treating type θ are given by $C(x, \theta) = \theta x + \frac{1}{2} c x^2$. Socially, the marginal health gains from treatment across types are $h_{x\theta} = 1 + \lambda$, and the marginal cost across types is $C_{x\theta} = 1$, so the first best treatments are increasing in the type. For the provider, the marginal health gains from treatment across types are $\alpha h_{x\theta} = 1 + \lambda$, and the marginal cost across types is $C_{x\theta} = 1$, so treatments may not necessarily be increasing in the type.

Figure 4 depicts first best treatments for each type, as well as the provider's equilibrium choice of treatment levels without incentives for $\alpha \in [\frac{1}{1+\lambda}, 1)$. Handing this provider full autonomy of patient care decisions would result in under-provision for everyone: the provider's equilibrium

[16] Thanks to Oliver Hart for suggesting this specific parametrization, and for extensive discussion about this case.

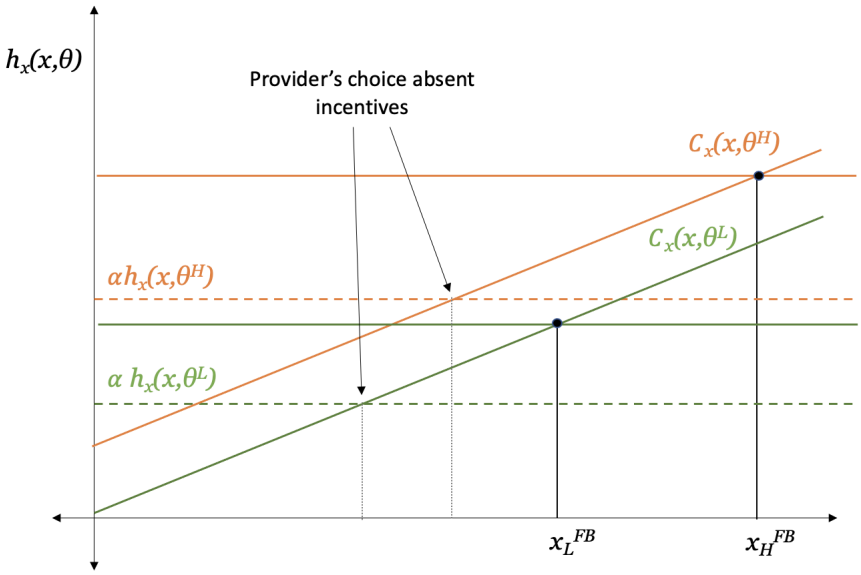


Figure 4: Provider Equilibrium Treatment Choices Relative to First Best

choices for types θ_H and θ_L are below first best when $\alpha < 1$. However, the key component we need in order for contracts to get us to the first best is for the provider to give *more care* to the high types. That is, we need the provider's equilibrium choices to be ordered so they are aligned with the socially optimal treatment choices. Lemma 3 shows that this aligned monotonicity can only hold when providers have some minimal degree of altruism.

Lemma 3 *The provider's incentive constraints imply that $x_H \geq x_L$ only when the provider's valuation of marginal health gains across types ($\alpha h_{x\theta}$) exceeds the marginal cost across types ($C_{x\theta}$), or $\alpha(1+\lambda) \geq 1$.*

Proof. Under this parametrized example, the sum of the two incentive constraints is

$$\underbrace{(\theta_H - \theta_L)}_{\geq 0}(\alpha(1+\lambda) - 1)(x_H - x_L) \geq 0.$$

By construction $(\theta_H - \theta_L) \geq 0$. For the incentive constraints to be jointly satisfied, $(\alpha(1+\lambda) - 1)$ and $(x_H - x_L)$ must have the same sign. Therefore, $x_H \geq x_L \iff \alpha(1+\lambda) \geq 1$ and the reverse, $x_H \leq x_L \iff \alpha(1+\lambda) \leq 1$. ■

Notice that when incremental health benefits of care across types, $(1+\lambda)$, are very large this monotonicity condition is likely still satisfied at low levels of altruism, α. In other words, if the differences across types is such that there obvious benefits for the costlier patients, it becomes less critical to have a provider that places a lot of weight on patient health, because even a provider to cares very little about the patient will see the benefit to give more care to the more expensive patient.

When the incentives of the provider are aligned with society's, a non-linear forcing contract can implement first best outcomes. If the first best treatment levels for the high type and the low

type are 10 and 5, respectively, a provider with an $\alpha < 1$ may optimally choose 8 and 3 if given an unrestricted treatment choice set. This is because the trade-off that characterizes the provider's choice is different from the social trade-off. However, if the insurer only allows the provider to administer 10 or 5, a sufficiently altruistic provider will appropriately give 10 to the high type and 5 to the low type. The problem with this contract arises when the provider does not value patient health by enough, and is tempted to give the high (and more expensive) type 5 instead of 10.

Proposition 6 *When $\alpha \geq \frac{1}{1+\lambda}$, the optimal contract implements first best, and when $\alpha < \frac{1}{1+\lambda}$, the optimal contract pools both types.*

Proof. Suppose $\alpha \geq \frac{1}{1+\lambda}$. Consider a modified problem with only the aggregate participation constraint and the monotonicity constraint, $x_H \geq x_L$. In the modified problem, x_L and x_H maximize

$$\max_{r_L, r_H} \gamma(\theta_H(1+\lambda)x_H - r_H + \eta(r_H - \theta_H x_H - \frac{cx_H^2}{2})) + (1-\gamma)(\theta_H(1+\lambda)x_L - r_L + \eta(r_L - \theta_L x_L - \frac{cx_L^2}{2})) \quad (6)$$

$$\text{s.t.} \qquad \gamma(r_H - \theta_H x_H - \frac{1}{2}cx_H^2) + (1-\gamma)(r_L - \theta_L x_L - \frac{1}{2}cx_L^2) \geq 0 \qquad \text{(PC-agg)}$$

$$x_H \geq x_L. \qquad \text{(monotonicity)}$$

Since the objective is decreasing in the payments to providers, (PC-agg) binds in the optimum, and the problem reduces to:

$$\max_{x_L, x_H} \gamma(\theta_H(1+\lambda)x_H - (\theta_H x_H + \frac{1}{2}cx_H^2)) + (1-\gamma)(\theta_H(1+\lambda)x_L(\theta_L x_L + \frac{1}{2}cx_L^2))$$

s.t. $x_H \geq x_L$. The solution coincides with the first best, $x_H = \frac{\theta_H \lambda}{c}$ and $x_L = \frac{\theta_L \lambda}{c}$, and satisfies the monotonicity condition. There is a continuum of pairs (r_H, r_L), that satisfy (PC-agg) at the first best treatments, described by the following relation:

$$r_H - r_L = \left(\frac{\theta_H^2 \lambda}{c} + \frac{\theta_H^2 \lambda^2}{2c}\right) + \frac{(1-\gamma)}{\gamma}\left(\frac{\theta_L^2 \lambda}{c} + \frac{\theta_L^2 \lambda^2}{2c}\right) - \frac{r_L}{\gamma}$$

One such pair is such that (IC-H) is binding, corresponding to the contract:

$$r_L = \left(\gamma\frac{\theta_H^2 \lambda}{c} + (1-\gamma)\frac{\theta_L^2 \lambda}{c}\right) + \frac{1}{2}\frac{\theta_L^2 \lambda^2}{c} + \gamma\left(\alpha(1+\lambda) - 1\right)\theta_H \frac{\lambda}{c}(\theta_H - \theta_L)$$

$$r_H = \left(\gamma\frac{\theta_H^2 \lambda}{c} + (1-\gamma)\frac{\theta_L^2 \lambda}{c}\right) + \frac{1}{2}\frac{\theta_H^2 \lambda^2}{c} - (1-\gamma)\left(\alpha(1+\lambda) - 1\right)\theta_H \frac{\lambda}{c}(\theta_H - \theta_L)$$

Now we have to check that the solution to the modified problem satisfies the two incentives constraints of the original problem. Evaluating (IC-H) and (IC-L) at the conjectured (r_H, r_L) and the first best treatments, we obtain after some algebra that:

$$0 \geq 0 \qquad \text{(IC-H)}$$

$$\left(\alpha(1+\lambda) - 1\right)\frac{\lambda}{c}(\theta_H - \theta_L)^2 \geq 0 \qquad \text{(IC-L)}$$

which is always true given that we assumed at the beginning $\alpha \geq \frac{1}{1+\lambda}$. Therefore, this particular solution to the modified problem is also a solution to the original problem.

Now suppose $\alpha \geq \frac{1}{1+\lambda}$. Consider a modified problem with only the aggregate participation con-

straint and the monotonicity constraint, $x_H \leq x_L$. The planner again maximizes the objective given by equation (2), now subject to $x_H \leq x_L$. Since the solution to (2) is increasing in type θ, the monotonicity constraint in this case binds, and $x_L = x_H = x_P$. The optimal contract has a pooling solution, $x_P = \gamma\lambda\theta_H + (1-\gamma)\lambda\theta_L$. At the pooling solution, the incentive constraints are trivially satisfied and the provider receives payment r_P equal to the average cost of treating patients at x_P.

■

Proposition 6 gives us a hopeful conclusion. When ethical professional norms push providers towards valuing patient health gains, we can get the first best. Interestingly enough, recent developments in the business model used by Kaiser Permanente seems to be moving in this direction. Kaiser is an integrated system of payors, hospitals, and physicians, who announced in 2015 they would be launching their own teaching hospital, and it was inaugurated in June, 2020. They try to foster an organizational culture that values patient health, while also internalizing the costs of care decisions.

In practice, it may be hard for an insurer to know the α of providers it contracts with. It may well be the case that providers have heterogeneous α's, as some of the popular work by Atul Gawande may suggest. Indeed, if providers put more or less weight on financial incentives, the design of the reimbursement contract can have heterogeneous impacts on the patients insured. But α need not be an exogenous primitive from the insurer's perspective. Payors can look beyond financial incentives and aim to foster an organizational culture that rewards providers for patient-centered decisions.

If α were the only unobservable for the insurer, a forcing contract would implement first best treatments for everyone (since there would only be a single optimal treatment level for every patient in the pool), even if heterogeneous across doctors. In other words, altruism heterogeneity alone would not be a problem if patients do not have heterogenous treatment needs and costs

The more relevant policy implication of this portion of the analysis is the idea to 'endogenize' the altruistic component of preferences, via ethical professional norms that are shaped by an organizational narrative or mission. This idea is by no means new, and has already been put forth in Besley and Gathak (2005) and Bidner and Francois (2010) in other contexts. Perhaps a more holistic approach to business in health care, such as one that fosters a patient-centered organizational culture in addition to the right set of incentives, is better suited to circumvent the asymmetric information problems that are so characteristic of this market.

5 Conclusion

To conclude, I review what I consider the three main insights from this article. First, cost-reimbursement is particularly inefficient when providers are altruistic because it creates a moral hazard problem in which providers give too much low-value care. If providers were not altruistic at all, cost-reimbursement would implement efficient outcomes under perfect risk adjustment. Second, , as patient treatment needs become more widely dispersed, covering care for the high utilization patients raises insurance costs on everyone else in the group: directly via raising average costs, and

indirectly by raising utilization of care by everyone else in the group.

The third is that financial incentives *coupled* with the right set of professional ethical norms can get us the first best. Contracts that move away from cost-based reimbursement and towards cost-reduction incentives (possibly through global budgets) are taking a step in the right direction, but are insufficient absent a better understanding on how professional ethical norms affect health care delivery. Interestingly enough, recent developments in the business model used by Kaiser Permanente seems to be moving in this direction. Kaiser is an integrated system of payors, hospitals, and physicians, who announced in 2015 they would be launching their own teaching hospital, and it was inaugurated in June, 2020. They try to foster an organizational culture that values patient health, while also internalizing the costs of care decisions.

Health is seen (by many) as a human right, and the provider contracting problem merits special attention. It is an important for future work to continue asking how to design better contracts in health care from new and innovative angles. Flat of the curve medicine, while of significant economic magnitude, is only a small part of the inefficiency story in health care. A unifying framework that accommodates the complex health care setting is still necessary, but precisely because health care is so nuanced, there is unlikely to be a 'one size fits all' solution.

References

Abaluck, Jason, Leila Agha, Chris Kabrhel, Ali Raja, and Arjun Venkatesh, "The Determinants of Productivity in Medical Testing: Intensity and Allocation of Care," *American Economic Review*, December 2016, *106* (12), 3730–3764.

Allen, Robin and Paul J. Gertler, "Regulation and the Provision of Quality to Heterogenous Consumers: The Case of Prospective Pricing of Medical Services," *Journal of Regulatory Economics*, 1991, *3* (4), 361–75. Publisher: Springer.

Arrow, Kenneth J., "Uncertainty and the Welfare Economics of Medical Care," *The American Economic Review*, 1963, *53* (5), 941–973. Publisher: American Economic Association.
A Theory of Incentives in Procurement and Regulation | The MIT Press

***A Theory of Incentives in Procurement and Regulation | The MIT Press*.**

Baicker, Katherine, Sendhil Mullainathan, and Joshua Schwartzstein, *"Behavioral Hazard in Health Insurance *," The Quarterly Journal of Economics, November 2015, 130 (4), 1623–1667.*

Becker, Gary S. and Kevin M. Murphy, *"A Theory of Rational Addiction," Journal of Political Economy, 1988, 96 (4), 675–700. Publisher: University of Chicago Press.*

Benabou, Roland and Jean Tirole, *"Incentives and Prosocial Behavior," THE AMERICAN ECONOMIC REVIEW, 2006, 96 (5), 27.*

Berwick, Donald M. and Andrew D. Hackbarth, *"Eliminating Waste in US Health Care," JAMA, April 2012, 307 (14), 1513–1516.*

Besley, Timothy, *"State Capacity, Reciprocity, and the Social Contract," Econometrica, 2020, 88 (4), 1307–1335. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.3982/ECTA16863.*

_ and Maitreesh Ghatak, *"Competition and Incentives with Motivated Agents," American Economic Review, May 2005, 95 (3), 616–636.*

Bidner, Chris and Patrick Francois, *"Cultivating Trust: Norms, Institutions and the Implications of Scale*," The Economic Journal, 2011, 121 (555), 1097–1129. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1468-0297.2010.02398.x.*

Broccardo, Eleonora, Oliver D. Hart, and Luigi Zingales, *"Exit vs. Voice," Working Paper 27710, National Bureau of Economic Research August 2020. Series: Working Paper Series.*

Chalkley, Martin and Fahad Khalil, *"Third party purchasing of health services: Patient choice and agency," Journal of Health Economics, November 2005, 24 (6), 1132–1153.*

_ and James M. Malcomson, *"Chapter 15 - Government Purchasing of Health Services," in Anthony J. Culyer and Joseph P. Newhouse, eds., Handbook of Health Economics, Vol. 1 of* Handbook of Health Economics*, Elsevier, January 2000, pp. 847–890.*

Chang, Tom and Mireille Jacobson, *"What do Nonprofit Hospitals Maximize? Evidence from California's Seismic Retrofit Mandate," p. 61.*

Chen, A. and D. Lakdawalla, *"Saving Lives or Saving Money? Understanding the Dual Nature of Physician Preferences," Innovation in Aging, June 2017, 1 (Suppl 1), 1343.*

Chen, Alice and Darius N. Lakdawalla, *"Healing the poor: The influence of patient socioeconomic status on physician supply responses," Journal of Health Economics, March 2019, 64, 43–54.*

Choné, Philippe and Ching-To Albert Ma, *"Optimal Health Care Contract under Physician Agency," Annals of Economics and Statistics, 2011, (101-102), 229–256. Publisher: GENES.*

Clemens, Jeffrey and Joshua D. Gottlieb, *"Do Physicians' Financial Incentives Affect Medical Treatment and Patient Health?," American Economic Review, April 2014, 104 (4), 1320–1349.*

Coronini-Cronberg, Sophie, Honor Bixby, Anthony A. Laverty, Robert M. Wachter, and Christopher Millett, *"English National Health Service's Savings Plan May Have Helped Reduce The Use Of Three 'Low-Value' Procedures," Health Affairs, March 2015, 34 (3), 381–389. Publisher: Health Affairs.*

Coulam, R. F. and G. L. Gaumer, *"Medicare's prospective payment system: a critical appraisal," Health Care Financing Review. Annual Supplement, 1991, pp. 45–77.*

Cram, Peter, Hoangmai H. Pham, Levent Bayman, and Mary S. Vaughan-Sarrazin, *"Insurance status of patients admitted to specialty cardiac and competing general hospitals: are accusations of cherry picking justified?," Medical Care, May 2008, 46 (5), 467–475.*

Cutler, David M., *"Where Are The Health Care Entrepreneurs? The Failure of Organizational Innovation in Health Care," Working Paper 16030, National Bureau of Economic Research May 2010. Series: Working Paper Series.*

_ **and Richard J. Zeckhauser**, *"Chapter 11 - The Anatomy of Health Insurance," in Anthony J. Culyer and Joseph P. Newhouse, eds., Handbook of Health Economics, Vol. 1 of* Handbook of Health Economics*, Elsevier, January 2000, pp. 563–643.*

Desai, Amar, Roger Bolus, Allen Nissenson, Glenn Chertow, Sally Bolus, Matthew D. Solomon, Osman S. Khawar, Jennifer Talley, and Brennan M. R. Spiegel, *"Is there "cherry picking" in the ESRD Program? Perceptions from a Dialysis Provider Survey," Clinical journal of the American Society of Nephrology: CJASN, April 2009, 4 (4), 772–777.*

Dewatripont, Mathias and Jean Tirole, *"Incentives and Ethics: How Markets and Organizations Shape our Moral Behavior," p. 39.*

Dranove, David and Paul Wehner, *"Physician-induced demand for childbirths," Journal of Health Economics, 1994, 13 (1), 61–73. Publisher: Elsevier.*

Ellis, R. P., *"Creaming, skimping and dumping: provider competition on the intensive and extensive margins," Journal of Health Economics, October 1998, 17 (5), 537–555.*

Ellis, Randall P. and Thomas G. McGuire, *"Provider behavior under prospective reimbursement," Journal of Health Economics, June 1986, 5 (2), 129–151.*

_ **and** _ , *"Supply-Side and Demand-Side Cost Sharing in Health Care," Journal of Economic Perspectives, December 1993, 7 (4), 135–151.*

Ettner, Susan L., *"Do elderly Medicaid patients experience reduced access to nursing home care?," Journal of Health Economics, 1993, 12 (3), 259–280. Publisher: Elsevier.*

Fraja, Gianni De, *"Contracts for health care and asymmetric information," Journal of Health Economics, 2000, 19 (5), 663–677. Publisher: Elsevier.*

Francois, Patrick and Michael Vlassopoulos, *"Pro-Social Motivation and the Delivery of Social Services," CESifo Economic Studies, March 2008, 54, 22–54.*

Fuchs, Victor R., *"More Variation In Use Of Care, More Flat-Of-The-Curve Medicine," Health Affairs, December 2018. Publisher: Project HOPE - The People-to-People Health Foundation, Inc.*

Gandhi, Ashvin, *"Picking Your Patients: Selective Admissions in the Nursing Home Industry," SSRN Scholarly Paper ID 3613950, Social Science Research Network, Rochester, NY May 2020.*

Garber, Alan M. and Jonathan Skinner, *"Is American Health Care Uniquely Inefficient?," The journal of economic perspectives : a journal of the American Economic Association, September 2008, 22 (4), 27–50.*

Gaumer, Gary L., Eugene L. Poggio, Craig G. Coelen, Cary S. Sennett, and Robert J. Schmitz, *"Effects of State Prospective Reimbursement Programs on Hospital Mortality," Medical Care, 1989, 27 (7), 724–736. Publisher: Lippincott Williams & Wilkins.*

Gaynor, Martin, Nirav Mehta, and Seth Richards-Shubik, *"Optimal Contracting with Altruistic Agents: A Structural Model of Medicare Payments for Dialysis Drugs," Working Paper 27172, National Bureau of Economic Research May 2020. Series: Working Paper Series.*

Greenlees, John S., John M. Marshall, and Donald E. Yett, *"Nursing Home Admissions Policies under Reimbursement," Bell Journal of Economics, 1982, 13 (1), 93–106. Publisher: The RAND Corporation.*

Gregg, Paul, Paul Grout, Anita Ratcliffe, Sarah Smith, and Frank Windmeijer, *"How important is pro-social behaviour in the delivery of public services?," The Centre for Market and Public Organisation, Department of Economics, University of Bristol, UK May 2008.*

Gruber, Jonathan and Maria Owings, *"Physician Financial Incentives and Cesarean Section Delivery," The RAND Journal of Economics, 1996, 27 (1), 99–123. Publisher: [RAND Corporation, Wiley].*

Guterman, S. and A. Dobson, *"Impact of the Medicare prospective payment system for hospitals," Health Care Financing Review, 1986, 7 (3), 97–114.*

Hart, Oliver and Luigi Zingales, *"Companies Should Maximize Shareholder Welfare Not Market Value," Journal of Law, Finance, and Accounting, 2017, 2 (2), 247–274.*

Hodgkin, Dominic and Thomas G. McGuire, *"Payment levels and hospital response to prospective payment," Journal of Health Economics, March 1994, 13 (1), 1–29.*

Institute of Medicine (US) Roundtable on Evidence-Based Medicine, *The Healthcare Imperative: Lowering Costs and Improving Outcomes: Workshop Series Summary, Washington (DC): National Academies Press (US), 2010.*

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

Jacobson, Mireille, Craig C. Earle, Mary Price, and Joseph P. Newhouse, *"How Medicare's payment cuts for cancer chemotherapy drugs changed patterns of treatment," Health Affairs (Project Hope), July 2010, 29 (7), 1391–1399.*

Kahn, Katherine L., Lisa V. Rubenstein, David Draper, Jacqueline Kosecoff, William H. Rogers, Emmett B. Keeler, and Robert H. Brook, *"The Effects of the DRG-Based Prospective Payment System on Quality of Care for Hospitalized Medicare Patients: An Introduction to the Series," JAMA, October 1990, 264 (15), 1953–1955. Publisher: American Medical Association.*

Kesternich, Iris, Heiner Schumacher, and Joachim Winter, *"Professional norms and physician behavior: Homo oeconomicus or homo hippocraticus?," Journal of Public Economics, November 2015, 131, 1–11.*

Lurie, N., J. Christianson, M. Finch, and I. Moscovice, *"The effects of capitation on health and functional status of the Medicaid elderly. A randomized trial," Annals of Internal Medicine, March 1994, 120 (6), 506–511.*

Ma, Albert, *"Cost and Quality Incentives in Health Care: Altruistic Providers," Technical Report 0084, Boston University - Industry Studies Programme December 1997. Publication Title: Papers.*

Mafi, John N., Kyle Russell, Beth A. Bortz, Marcos Dachary, William A. Hazel, and A. Mark Fendrick, *"Low-Cost, High-Volume Health Services Contribute The Most To Unnecessary Health Spending," Health Affairs, October 2017, 36 (10), 1701–1704. Publisher: Health Affairs.*

Malcomson, James, *"Supplier Discretion over Provision: Theory and an Application to Medical Care," Technical Report 1407, CESifo Group Munich 2005. Publication Title: CESifo Working Paper Series.*

Maréchal, François and Michel Mougeot, *"Risk sharing and moral hazard under prospective payment to hospitals: how to reimburse services for outlier patients," April 2004.*

Miller, George, Corwin Rhyan, Beth Beaudin-Seiler, and Paul Hughes-Cromwick, *"A Framework for Measuring Low-Value Care," Value in Health: The Journal of the International Society for Pharmacoeconomics and Outcomes Research, April 2018, 21 (4), 375–379.*

Miller, R. H. and H. S. Luft, *"Managed care plan performance since 1980. A literature analysis," JAMA, May 1994, 271 (19), 1512–1519.*

Mougeot, Michel and Florence Naegelen, *"Hospital price regulation: A normative analysis," Revue d'Economie Politique, March 2013, 123, 179–210.*

Mussa, Michael and Sherwin Rosen, *"Monopoly and product quality," Journal of Economic Theory, August 1978, 18 (2), 301–317.*

Newhouse, Joseph P., *"Do unprofitable patients face access problems?," Health Care Financing Review, 1989, 11 (2), 33–42.*

_ , *"Reimbursing Health Plans and Health Providers: Efficiency in Production Versus Selection," Journal of Economic Literature, 1996, 34 (3), 1236–1263. Publisher: American Economic Association.*

Nguyen, N X and F W Derrick, *"Physician behavioral response to a Medicare price reduction.," Health Services Research, August 1997, 32 (3), 283–298.*

Pauly, Mark V., *"Optimal Health Insurance," The Geneva Papers on Risk and Insurance. Issues and Practice, 2000, 25 (1), 116–127. Publisher: Palgrave Macmillan Journals.*

Perlroth, Daniella J., Dana P. Goldman, and Alan M. Garber, *"The potential impact of comparative effectiveness research on U.S. health care expenditures," Demography, 2010, 47 Suppl, S173–190.*

Rice, T. H., *"The impact of changing medicare reimbursement rates on physician-induced demand," Medical Care, August 1983, 21 (8), 803–815.*

Roomkin, Myron J. and Burton A. Weisbrod, *"Managerial Compensation and Incentives in For-Profit and Nonprofit Hospitals," Journal of Law, Economics, & Organization, 1999, 15 (3), 750–781. Publisher: Oxford University Press.*

Rossiter, Louis F. and Gail R. Wilensky, *"A Reexamination of the Use of Physician Services: The Role of Physician-Initiated Demand," Inquiry, 1983, 20 (2), 162–172. Publisher: Sage Publications, Inc.*

Schwartz, Aaron L., Bruce E. Landon, Adam G. Elshaug, Michael E. Chernew, and J. Michael McWilliams, *"Measuring Low-Value Care in Medicare," JAMA Internal Medicine, July 2014, 174 (7), 1067–1076.*

Silver, David, *"Haste or Waste? Peer Pressure and Productivity in the Emergency Department," SSRN Scholarly Paper 3588769, Social Science Research Network, Rochester, NY May 2020.*

Ulli, Robin Moremen, *"Resident Selection in a Connecticut Nursing Home: A Hew from within," Journal of Aging and Health, February 1995, 7 (1), 139–159. Publisher: SAGE Publications Inc.*

Wynn, Barbara O, *"Medicare Payment for Hospital Outpatient Services: A Historical Review of Policy Options," WR-267-MedPAC, May 2005, p. 103.*

Yip, Winnie C., *"Physician response to Medicare fee reductions: changes in the volume of coronary artery bypass graft (CABG) surgeries in the Medicare and private sectors," Journal of Health Economics, 1998, 17 (6), 675–699. Publisher: Elsevier.*

6 Appendix

6.1 Insurer's Problem: Two-type Case

The insurer's problem is to choose (r_H, r_L) according to the following program.

$$\max_{r_L, r_H} \gamma(h(x_H, \theta_H) - r_H + \eta(r_H - c(x_H))) + (1-\gamma)(h(x_L, \theta_L) - r_L + \eta(r_L - c(x_L))) \tag{2}$$

$$\text{s.t. } \alpha h(x_L, \theta_L) + r_L - c(x_L) \geq \alpha h(x_H, \theta_L) + r_H - c(x_H) \tag{IC L}$$

$$\alpha h(x_H, \theta_H) + r_H - c(x_H) \geq \alpha h(x_L, \theta_H) + r_L - c(x_L) \tag{IC H}$$

$$r_L - c(x_L) \geq 0 \tag{PC L}$$

$$r_H - c(x_H) \geq 0. \tag{PC H}$$

6.2 Non-linear Contract: Continuum of Types Case

For a continuum of types, the insurer's objective is as follows.

$$SWF = \int_{\underline{\theta}}^{\overline{\theta}} (h(x, \theta) - r(x) + \eta(r(x) - c(x)) f(\theta) d\theta$$

The government does not observe patient type, but does observe treatment x. There's a continuum of (IC) constraints, such that the physician prefers to give x^*_θ to patient type $x^*(\theta)$, instead of giving them $x^*(\theta')$, for all θ, θ'.

That is,

$$\alpha h(x(\theta), \theta) + r(x(\theta)) - c(x(\theta)) \geq \alpha h(x(\theta'), \theta) + r(x(\theta')) - c(x(\theta')) \; \forall \; \theta, \theta'. \tag{IC}$$

Lemma 4 *The incentive constraints jointly imply that equilibrium treatment levels are increasing in type. That is, equilibrium treatment levels are monotonic in θ, with $\frac{\partial x^*(\theta)}{\partial \theta} \geq 0$.*

Proof. Consider two adjacent incentive constraints, for $\theta > \theta'$.

$$\alpha h(x(\theta), \theta) + r(x(\theta)) - c(x(\theta)) \geq \alpha h(x(\theta'), \theta) + r(x(\theta')) - c(x(\theta'))$$
$$\alpha h(x(\theta'), \theta') + r(x(\theta')) - c(x(\theta')) \geq \alpha h(x(\theta), \theta') + r(x(\theta)) - c(x(\theta))$$

If we add them, we obtain

$$h(x(\theta), \theta) - h(x(\theta'), \theta) \geq h(x(\theta), \theta') - h(x(\theta'), \theta').$$

Since $h(x, \theta)$ is increasing in x and in θ, this can only hold if $x(\theta) \geq x(\theta')$. Therefore, the equilibrium treatment level $x(\theta)$ will be increasing in the type, θ. QED. ■

6.2.1 Solving for the optimal contract

Consider a modified problem with only the participation constraints, the local incentive constraints, and the monotonicity constraint. We will characterize the solution to this problem, and then show it satisfies the solution to the original problem.

Let $\theta \sim F(\theta)$ where $\theta \in [\underline{\theta}, \overline{\theta}]$. Replacing the continuum of incentive constraints with the monotonicity constraint, the government's problem is:

$$\max_{r(x)} \int_{\underline{\theta}}^{\overline{\theta}} [h(x(\theta), \theta) - r(x(\theta)) + \eta(r(x(\theta)) - c(x(\theta)))] f(\theta) d\theta$$

subject to

$$r(x(\overline{\theta})) - c(x(\overline{\theta})) \geq 0 \qquad \text{(PC-}\overline{\theta}\text{)}$$

$$\frac{\partial x(\theta)}{\partial \theta} \geq 0 \qquad \text{(monotonicity)}$$

$$\alpha h_x(x(\theta), \theta) + r_x(x(\theta)) - c_x(x(\theta)) = 0. \qquad \text{(provider FOC)}$$

The standard procedure for solving this type of problem is to ignore the monotonicity constraint and solve the relaxed problem with only the (PC) of the highest type, $\overline{\theta}$, and the equilibrium treatment condition. We verify that the monotonicity condition is satisfied at the end.

Step 1: Show that participation constraint of the highest covered type is binding.

Proposition 7 *(PC-$\overline{\theta}$) is binding.*

Proof. Suppose, for the sake of contradiction, that PC of $\bar{\theta}$ is not binding, and $r(x(\overline{\theta})) > c(x(\overline{\theta}))$. Since the objective is strictly decreasing in r, there exists an $\epsilon > 0$ such that $r(x(\overline{\theta})) - \epsilon > c(x(\overline{\theta}))$ still holds, and the objective will increase by $(1-\eta)\epsilon > 0$. Therefore, it is optimal to keep reducing $r(x(\overline{\theta}))$ until (PC) of $\overline{\theta}$ binds. ■

Step 2: Integrate physician FOC over x to solve for incentive compatible contract, $r(x^*(\theta))$.

Note that the type θ that enters the argument of the health benefit function remains fixed at θ, as we are just integrating over the x's.

Proposition 8 *For any type $\theta < \overline{\theta}$, the incentive compatible payment scheme will pay a margin above cost given by the following expression:*

$$r(x^*(\theta)) = c(x^*(\theta)) + \underbrace{\alpha h(x^*(\overline{\theta}), \theta) - \alpha h(x^*(\theta), \theta)}_{\textit{incentive rent}}$$

where $x^*(\theta)$ is the equilibrium treatment, and an implicit function of only θ defined by the physician optimization problem, for a fixed $r(x)$.

Proof. Fixing θ, integrate the physician FOC from $x^*(\theta)$ to $x^*(\overline{\theta})$.

$$\int_{x_\theta}^{x_{\overline{\theta}}} \alpha h_x(x,\theta) + r_x(x) - c_x(x)dx = 0$$

It follows that,

$$\alpha h(x^*(\overline{\theta}),\theta) - \alpha h(x^*(\theta),\theta) - (r(x^*(\theta)) - c(x^*(\theta))) + \underbrace{r(x^*(\overline{\theta})) - c(x^*(\overline{\theta}))}_{=0 \text{ by PC}\overline{\theta}} = 0.$$

This gives us an explicit expression for $r(x^*(\theta))$.

$$r(x^*(\theta) = c(x^*(\theta)) + \alpha h(x^*(\overline{\theta}),\theta) - \alpha h(x^*(\theta),\theta)$$

■

The local incentive constraint and the participation constraint of the high type jointly determine the full schedule of $r(x^*(\theta))$.

$$r(x^*(\theta)) = \begin{cases} c(x^*(\overline{\theta})) & \text{, for } \overline{\theta} \\ c(x^*(\theta)) + \alpha h(x^*(\overline{\theta}),\theta) - \alpha h(x^*(\theta),\theta) & \text{, for } \theta < \overline{\theta}. \end{cases}$$

Step 3: Show that fully separating equilibrium contract violates the monotonicity condition.

Since there is a unique contract $r(x^*(\theta))$ which satisfies the constraints of the modified problem, we can plug it into our objective and characterize the solution of the equilibrium treatments.

Lemma 5 *Ignoring the monotonicity condition, the derivative of the SWF with respect to $x^*(\theta)$ is*

$$\frac{dSWF}{dx^*(\theta)} = \begin{cases} f(\theta)[h_x(x^*(\theta),\theta) - c_x(x^*(\theta) + (1-\eta)\alpha h_x(x^*(\theta),\theta)] & ,\theta < \bar{\theta} \\ f(\overline{\theta})[h_x(x^*(x^*(\overline{\theta}),\overline{\theta}) - c_x(x^*(\overline{\theta}))] - (1-\eta)\int_{\underline{\theta}}^{\overline{\theta}} \alpha h_x(x^*(\overline{\theta}),y)f(y)dy & ,\theta = \bar{\theta}. \end{cases}$$

Proof. Evaluating the social welfare function at the contract implied by the constraints results in the following equation:

$$SWF = \int_{\underline{\theta}}^{\overline{\theta}} [h(x^*(\theta),\theta) - c(x^*(\theta)) - (1-\eta)\alpha(h(x^*(\overline{\theta}),\theta) - h(x^*(\theta),\theta)]f(\theta)d\theta. \tag{5}$$

Differentiating with respect to $x^*(\theta)$ for types $\theta < \overline{\theta}$, we obtain:

$$\frac{dSWF}{dx^*(\theta)} = f(\theta)[h_x(x^*(\theta),\theta) - c_x(x^*(\theta) + (1-\eta)\alpha h_x(x^*(\theta),\theta)].$$

For the high type, however, the first order condition is different, as $x^*(\overline{\theta})$ shows up in for every payment premium of types $\theta < \overline{\theta}$, and the incentive rent is zero for the highest type. Differentiating the social welfare function with respect to $x^*(\overline{\theta})$ gives us:

$$\frac{dSWF}{dx^*(\overline{\theta})} = f(\overline{\theta})[h_x(x^*(x^*(\overline{\theta}),\overline{\theta}) - c_x(x^*(\overline{\theta}))] - (1-\eta)\int_{\underline{\theta}}^{\overline{\theta}} \alpha h_x(x^*(\overline{\theta}),y)f(y)dy.$$

■

Proposition 9 *The fully separating equilibrium contract characterized by the optimality conditions derived in Lemma 5 violates the monotonicity condition,* $\frac{dx^*(\theta)}{d\theta} \geq 0$.

Proof. Suppose, for the sake of contradiction, that $x^*(\theta)$ is monotonically increasing in θ, $\frac{dx^*(\theta)}{d\theta} > 0$. For the menu of $x^*(\theta)$'s characterized in Lemma 5 to be monotonic and continuous at the top, it must be the case that the limit of the solution to $dSWF/dx^*(\theta) = 0$ as $\theta \to \overline{\theta}$ equals the solution to $dSWF/dx^*(\bar{\theta}) = 0$. Taking the limit of the difference, $\lim_{\theta\to\bar{\theta}}[dSWF/dx^*(\theta) - dSWF/dx^*(\bar{\theta})]$, reduces to the following expression:

$$\lim_{\theta\to\bar{\theta}}\left[\alpha h_x(x^*(\bar{\theta}),\bar{\theta}) + \int_{\underline{\theta}}^{\overline{\theta}} \alpha h_x(x^*(\overline{\theta}),y)f(y)dy\right] > 0$$

which is strictly positive. The positive sign implies that the solution to $dSWF/dx^*(\theta) = 0$ as $\theta \to \overline{\theta}$ is greater than the solution to $dSWF/dx^*(\bar{\theta}) = 0$, which implies a contraction with the monotonicity condition supposed at the beginning. Alternatively, we can see that these two first order conditions violate the monotonicity condition by evaluating them both at the first best $x^{FB}(\bar{\theta})$.

$$\left.\frac{dSWF}{dx^*(\theta)}\right|_{x^{FB}(\bar{\theta})} = \underbrace{h_x(x^{FB}(\bar{\theta}),\bar{\theta}) - c_x(x^{FB}(\bar{\theta})}_{=0} = \underbrace{-(1-\eta)\alpha h_x(x^{FB}(\bar{\theta}),\bar{\theta})}_{\leq 0} \qquad \text{(FOC } \theta < \bar{\theta})$$

$$\left.\frac{dSWF}{dx^*(\bar{\theta})}\right|_{x^{FB}(\bar{\theta})} = \underbrace{h_x(x^{FB}(\bar{\theta}),\bar{\theta}) - c_x(x^{FB}(\bar{\theta})}_{=0} = \underbrace{(1-\eta)\int_{\underline{\theta}}^{\overline{\theta}} \alpha h_x(x^*(\overline{\theta}),y)f(y)dy}_{\geq 0} \qquad \text{(FOC } \bar{\theta})$$

The solution to $x(\theta) < x(\overline{\theta})$ in the neighborhood of $x(\overline{\theta})$ is to the right of the optimum of the characterized by $dSWF/dx^*(\bar{\theta}) = 0$. ■

Step 4: Conjecture a pooling solution and verify the constraints (of the modified problem).

The discontinuity in the contract for the highest type, $\overline{\theta}$ is indicative of a solution with pooling at the top. Let θ_T be the threshold type, above which everyone gets pooled at treatment level x_T. Consider the following pooling contract, where all types $\theta \geq \theta_T$ receive a fixed treatment quantity, and not more, because the contract does not allow the provider to choose treatment levels above x_T. The provider will *not* be on his first order condition for types above θ_T under this contract.

$$r(x^*(\theta)) = \begin{cases} c(x_T) & \text{, for } \theta \geq \theta_T \\ c(x^*(\theta)) + \alpha h(x_T, \theta) - \alpha h(x^*(\theta), \theta) & \text{, for } \theta < \theta_T. \end{cases}$$

Consider now a similar modified problem in which the insurer maximizes the SWF subject to the monotonicity constraint $\frac{dx^*(\theta)}{d\theta} > 0$, the participation constraint of the threshold type θ_T, and the local incentive constraints. The constraints of the modified problem are now:

$$r(x(\theta_T)) - c(x(\theta_T)) \geq 0 \qquad \text{(PC-}\theta_T\text{)}$$

$$\frac{\partial x(\theta)}{\partial \theta} \geq 0 \ \ \forall\, \theta \qquad \text{(monotonicity)}$$

$$\alpha h_x(x(\theta), \theta) + r_x(x(\theta)) - c_x(x(\theta)) = 0 \ \forall\, \theta. \qquad \text{(provider FOC)}$$

Lemma 6 *The pooling contract satisfies the participation constraint of the threshold type θ_T, and the local incentive constraints.*

Proof. The participation constraint of the threshold type is trivially satisfied since $r(x^*(\theta_T)) = c(x^*(\theta_T))$ implies $r(x^*(\theta_T)) - c(x^*(\theta_T)) = 0 \geq 0$.

To show the contract satisfies the local incentive constraints, first consider $\theta \leq \theta_T$. Differentiating the pooling contract with respect to x, we obtain that $r_x(x^*(\theta)) = c_x(x^*(\theta)) - \alpha h_x(x^*(\theta), \theta)$. Plugging $r_x(x^*(\theta))$ into the provider FOC, we obtain that

$$\alpha h_x(x(\theta), \theta) + \underbrace{c_x(x^*(\theta)) - \alpha h_x(x^*(\theta), \theta)}_{r_x(x(\theta))} - c_x(x(\theta)) = 0.$$

Now consider $\theta > \theta_T$. The provider's problem is constrained for $\theta > \theta_T$ to choose

$$x \in \arg\max_x \alpha h(x, \theta) + \underbrace{r(x(\theta_T))}_{\text{fixed}} - c(x) \ \ \forall \theta > \theta_T$$

s.t. $x \leq x(\theta_T)$. The complementary slackness condition determines the solution, and $x^*(\theta) = x(\theta_T)$ for $\theta > \theta_T$, with $\alpha h_x(x^*(\theta), \theta) - c_x(x^*(\theta)) < 0$. ■

Step 5: Show that the pooling contract satisfied the monotonicity condition in the original modified problem.

Proposition 10 *The pooling contract satisfies the monotonicity condition,* $\frac{dx(\theta)}{d\theta} \geq 0$.

Proof. The social welfare function under the pooling contract is given by:

$$SWF = \int_{\underline{\theta}}^{\theta_T} [h(x^*(\theta),\theta) - c(x^*(\theta)) - (1-\eta)(\alpha h(x_T,\theta) - \alpha h(x^*(\theta),\theta)]f(\theta)d\theta + \int_{\theta_T}^{\overline{\theta}} [h(x_T,\theta) - c(x_T)f(\theta)d\theta$$

For all types above the threshold type, $\theta \geq \theta_T$, the treatment level will be characterized by the following expression:

$$\frac{dSWF}{dx_T} = \int_{\theta_T}^{\overline{\theta}} (h_x(x_T,\theta) - c_x(x_T))f(\theta)d\theta - (1-\eta)\int_{\underline{\theta}}^{\theta_T} \alpha h_x(x_T,\theta)f(\theta)d\theta \tag{10.1}$$

For types $\theta < \theta_T$, treatment level will be characterized by our same condition from Lemma 5 for low types,

$$\frac{dSWF}{dx^*(\theta)} = (h_x(x^*(\theta),\theta) - c_x(x^*(\theta)) + (1-\eta)\alpha h_x(x^*(\theta),\theta))f(\theta). \tag{10.2}$$

For $x(\theta)$ to be monotonic at x_T, solution to the limit of (10.2) as $\theta \to \theta_T$ must less than or equal to the solution to (10.1). Notice that (10.2) at the limit coincides with the limits of integration in (10.1). Taking the limit of the difference, we obtain that:

$$\lim_{\theta \to \theta_T} \left[\frac{dSWF}{dx_T} - \frac{dSWF}{dx^*(\theta)}\right] = \int_{\theta_T}^{\overline{\theta}} (h_x(x_T,\theta) - c_x(x_T))f(\theta)d\theta - (1-\eta)\int_{\underline{\theta}}^{\theta_T} \alpha h_x(x_T,\theta)f(\theta)d\theta.$$

This limit is equal to zero at the x_T which solves $\frac{dSWF}{dx_T} = 0$. This tells us that the solution characterized by equations (10.1) and (10.2) is not just monotonic, but also continuous. ■

Step 6: Show that the solution to the modified problem is also a solution to the original problem

Lemma 7 *The participation constraints of all types* $\theta < \theta_T$ *are jointly implied by (PC-*θ_T*) and the local downwards incentive constraints. The participation constraints of* $\theta \geq \theta_T$ *are trivially satisfied at the pooling solution.*

Proof. By monotonicity in treatments, type θ_T receives the highest level of treatment, x.

$$x(\theta_T) \geq x(\theta), \quad \forall\, \theta < \theta_T.$$

Consider the participation constraint of θ_T, and the incentive constraint of θ, for $\theta < \theta_T$.

$$r(x_T) - c(x_T) \geq 0 \qquad \text{(PC } \theta_T\text{)}$$
$$\alpha h(x(\theta), \theta) + r(x(\theta)) - c(x(\theta)) \geq \alpha h(x_T, \theta) + r(x_T) - c(x_T) \qquad \text{(IC } \theta \to \theta_T\text{)}$$

If we add them, we obtain

$$r(x(\theta)) - c(x(\theta)) \geq \underbrace{\alpha h(x_T, \theta) - \alpha h(x(\theta), \theta)}_{\geq 0 \text{ by monotonicity}} + \underbrace{r(x_T) - c(x_T)}_{\geq 0 \text{ by PC } \theta_T} \geq 0$$

Therefore, the participation constraints of all $\theta < \theta_T$ are jointly implied by the PC of θ_T and the local downwards incentive constraints, and participation constraints of types below are non-binding. For types $\theta \geq \theta_T$, equilibrium treatment is x_T, and the reimbursement contract pays $r(x_T) = c(x_T)$, and all the participation constraints are satisfied. ■

Lemma 8 *The incentive constraints are all satisfied at the pooling solution.*

Proof. By Proposition 10, the pooling solution satisfies the monotonicity condition. Consider the two adjacent incentive constraints of an arbitrary type $\theta < \theta_T$, and consider $\theta' < \theta$. Evaluating these at the pooling contract, we obtain:

$$\alpha h(x_T, \theta) - \alpha h(x(\theta'), \theta) \geq \alpha h(x_T, \theta') - \alpha h(x^*(\theta'), \theta')) \qquad \text{(IC } \theta \to \theta'\text{)}$$

which is always true by monotonicity and the assumption that $h(x, \theta)$ is increasing in θ, and that

$$\alpha h(x(\theta'), \theta') + \alpha(h(x_T, \theta') - h(x(\theta'), \theta')) \geq \alpha h(x(\theta), \theta') + \alpha(\underbrace{h(x_T, \theta)}_{\geq h(x_T, \theta')} - h(x(\theta), \theta))$$
$$\alpha h(x(\theta), \theta) - \alpha h(x(\theta), \theta') \geq 0 \qquad \text{(IC } \theta' \to \theta\text{)}$$

which is always true by the assumption that $h(x, \theta)$ is increasing in θ. ■

6.2.2 Characterizing the Solution

Lemma 9 *Under the pooling contract, the lowest type gets over-treated relative to first best.*

Proof. For type $\underline{\theta}$, the solution characterized by $\frac{dSWF}{dx^*(\underline{\theta})} = 0$ can be written as:

$$\underbrace{h_x(x^*(\underline{\theta}), \underline{\theta}) - c_x(x^*(\underline{\theta}))}_{=0 \text{ at first best}} = \underbrace{-(1-\eta)\alpha h_x(x^*(\underline{\theta}), \underline{\theta})}_{\leq 0}$$

Therefore, the solution to $\frac{dSWF}{dx^*(\underline{\theta})} = 0$ is to the right of the first best optimum. ■

Lemma 10 *Under the pooling contract, the highest type gets under-treated relative to first best.*

Proof. For type $\overline{\theta}$, treatment level received is characterized by $\frac{dSWF}{dx_T} = 0$. Since $h_x(x, \theta)$ is increasing in θ by construction, it follows that

$$h_x(x_T, \overline{\theta}) - c_x(x_T) > \int_{\theta_T}^{\overline{\theta}} (h_x(x_T, \theta) - c_x(x_T)) f(\theta) d\theta$$

for a non-empty interval $\theta \in [\theta_T, \overline{\theta}]$. From (10.1), we know that this implies $h_x(x_T, \overline{\theta}) - c_x(x_T)$ is strictly greater than zero.

$$h_x(x_T, \overline{\theta}) - c_x(x_T) > \int_{\theta_T}^{\overline{\theta}} (h_x(x_T, \theta) - c_x(x_T)) f(\theta) d\theta - (1 - \eta) \int_{\underline{\theta}}^{\theta_T} \alpha h_x(x_T, \theta) f(\theta) d\theta = 0$$

At the first best, $h_x(x^{FB}(\overline{\theta}), \overline{\theta}) - c_x(x^{FB}(\overline{\theta})) = 0$. Therefore, the solution to $\frac{dSWF}{dx^*(\theta)T)} = 0$ is to the right of the first best optimum for type $\overline{\theta}$. ■