Optimal Reimbursement Contracts for Provider Administered Treatments*

Angie Acquatella

September 14, 2020

How should an insurer contract with health care providers? I propose a theoretical framework to study optimal incentive schemes taking into account two distinctive features: unobserved patient heterogeneity and provider altruism. These two combined shed new insights on the practical considerations for the insurer's reimbursement decision, and yields new theoretical results in asymmetric information models. I find that altruism is a source of inefficient over-treatment. I also find that unobserved heterogeneity matters through the patient with the highest health care needs. Theoretically, adding altruism results in a different solution to the standard non-linear pricing problem. To evaluate the optimality of existing reimbursement schemes, I derive a sufficient statistics formula that depends on two empirical objects which can be computed in observational data. In an empirical application to Chronic Obstructive Pulmonary Disease (COPD) treatment, I find that Medicare's current fee-for-service payment scheme for this condition is close to optimal.

*I thank Oliver Hart, David Cutler, Stefanie Stancheva, Ed Glaeser, Jerry Green, Dan Brown, Andy Newman, Claudia Golding, Nathan Hendren, David Sibley, Scott Kominers, Mike Powell, Leemore Dafny, Mark Shepard, Rafaella Sadun, Matt Weinzierl, Ariel Stern, Laura Keohane, Mark Duggan, Grace McCormack, Aileen Devlin, Tamar Oostrom, Adriano Fernandes, Frank Pinter, Samantha Burn, Daniel Ly, Rebecca Sachs, Ashvin Gandhi, Amanda Kreider, Ambra Seck, Ljubica Ristovska, Pierfrancesco Mei, Leonardo D'Amico, 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.

1 Introduction

The way we pay health care providers matters: it affects the intensity of health care services and set of patients treated. I focus on provider-administered treatments, where the insurer's reimbursement contract can have potentially large consequences on the care delivered and the patient's health. I propose a theoretical framework to think about how an insurer may want to design a reimbursement scheme optimally. There are two distinguishing features in my model: unobservable patient heterogeneity and provider altruism.

These two features combined shed new insights on the practical considerations for the insurer's reimbursement contract design, and produces new theoretical results in asymmetric information models. Practically, I find that altruism is a source of inefficient over-treatment: if the provider values patient health and also get fully reimbursed for costs, he will treat as if treatment costs were zero (i.e. too much). Moreover, he will do so at the expense of higher insurance premiums (or taxpayers for a government insurer). I also find that if there is a patient with disproportionately larger health care utilization within the unobserved heterogeneity, this patient critically raises the costs of insuring everyone else. Theoretically, adding altruism means that, unlike the standard non-linear pricing problem, the optimal contract distorts treatment on *all* types.

Why are these two features important? The first–patient heterogeneity–has always been a distinguishing element of the health care setting. As Arrow (1963) pointed out, health *care* affects each patient's *health* distinctly, so health outcomes may look different for observably similar patients that get the same treatment. Here, I focus on patient heterogeneity in *health benefit* from treatment. An illustration may be helpful. Consider two patients with Chronic Obstructive Pulmonary Disease (COPD) and identical medical histories. COPD is a progressive condition that impairs breathing and has no cure. To mitigate symptoms of COPD, patients can receive treatment of 'pulmonary rehabilitation,' which involves time with a respiratory therapist or pulmonologist who guides the patient through a series of physical exercises, breathing retraining exercises, as well as nutritional counseling.

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 benefit a lot more from tenth visit. Why? 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. I want to find a payment contract that caters to the different needs of these two patients, who are observably equivalent to the insurer but not to the provider. 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.

In regards to altruism, one can hardly argue that the provider's ethical compulsion does not dictate part of the treatment choice, in conjunction with financial considerations. Whether motivated by the Hippocratic oath or social obligation, providers seem to value patient health when choosing treatment course, at least to *some* extent. There is some empirical evidence of altruism in the strand of the health literature about not-for-profit hospitals (Dranove, 2012; Roomkin and Weisbrod, 1999; Gregg et al., 2008).

How do insurers design provider reimbursement contracts, currently? Similar to a government procurement contract, both private and public (Medicare) insurers 'cover' a condition by reimbursing the costs incurred by the provider in providing treatment. These sometimes give positive profits, as the intent of a reimbursement contract is to offset the provider's input cost. While cost-based reimbursement contracts are very common in government procurement problems, such as a state commissioned bridge or a military defense project, it is not obvious why an insurer would opt for

such contract. Altruism is common in provider agency models, but less so in the government contracting models, as we generally do not think that a trash pick-up driver intrinsically enjoys picking up the trash above and beyond his fair wage compensation.

To characterize the optimal reimbursement contract, I incorporate the two discussed features into the standard government procurement contracting framework. While my model was designed specifically for the health care setting, my framework is general and can apply to settings such as an employer's decision to pay on hourly wage or fixed monthly salary, or a procurement problem in which the contractor intrinsically values the product for which he was hired. My framework is also adaptable to study many forms of unobserved heterogeneity, such as heterogeneity in provider ability, costs of treatment, altruism, or risk-aversion, which I leave for future work.

In the spirit of the optimal tax literature, I use my model to arrive at sufficient statistics formulas that may help an insurer decide how to pay for a particular service while maintaining fairly general modeling assumptions. I focus on linear contracts that reimburse a share of costs plus a lump sum payment. I find that the optimal cost reimbursement share depends critically on two empirical objects that can be measured in observational data: the range of treatment heterogeneity, and a measure that captures the extent to which there are outlier patients. I then show how to calculate these empirical objects to evaluate pulmonary rehabilitation therapy for COPD, using observational data from the Medicare outpatient claims.

1.1 Related Literature

Related work has studied optimal provider payment contracts in more restrictive environments. The canonical paper in the provider contracting literature is Ellis and McGuire (1986), who were the first to evaluate the optimality of existing Medicare reimbursement contracts in a theoretical framework with partially altruistic providers. Their model has been widely used in papers on provider agency, and they were the first to provide insight that marginal lus reimbursement is never optimal, particularly when providers are altruistic. We have in common that we both study optimal provider cost-sharing in a world where costs are observable and verifiable.

My model embeds the Ellis and McGuire model, but differs in two ways. The first is the patient heterogeneity piece. I study the provider contracting problem for an insurer that has *one* contract for *a heterogeneous patient pool*. They study the provider contracting problem when there is one contract available for each provider-patient pair. The second is the contracting objective function. In the Ellis and McGuire contracting objective, there is no 'loss' term for provider payments. These two features combined drive the different result on the optimality of prospective payment. In their model, PPS reimbursement is optimal when providers are perfectly altruistic because, by making the provider incur the full cost of treatment at the margin, the provider chooses treatments to that marginal health benefit equates to marginal cost. When patients are heterogeneously costly in ways the insurer cannot contract on, opting for this type of scheme becomes too expensive for the insurer.

There are a few other papers which have studied similar versions of the provider contracting problem with heterogeneity that also merit mention. De Fraja (2000) characterize the optimal payment contract when providers have heterogeneous costs. Jack (2005) solves for the optimal contract under heterogeneous provider altruism, with non-contractible quality. Malcomson (2005) studies the problem in a model without provider altruism. Gaynor, Mehta, Richards-Shubik (2020) estimate the optimal contract parameters using structural methods in a setting where providers face heterogeneous costs of treatment.

Choné and Ma (2011) study the effect of altruism on optimal payment schemes in a very general and elegant theoretical framework, characterizing general properties of optimal payment contracts for an insurer that does not observe provider altruism and also does not know the patient's value

of treatment on health. Chalkey and Kahlil (2005) study the provider contracting problem under a restricted, yet somewhat broader, contracting space, where reimbursement may condition on health outcomes in addition to treatment.

An overwhelming body of empirical research has found that the structure of the contract affects treatment volume, and there is some evidence that finds real effects on patient health outcomes. The reimbursement contracts seen in practice calculate costs at different levels of aggregation, with the most common being treatment inputs (fee-for-fervice) or diagnosis (prospective payment). The general consensus from this empirical literature is that prospective payment (PPS) led providers to cut back on care (Coulam and Gaumer, 1991; Hodgkin and McGuire, 1994; Ellis and McGuire, 1993), but there is mixed evidence on whether the cut back constituted under provision of care with adverse effects on mortality (Cutler and Zeckhauser, 2000; Gaumer et. al., 1989; Kahn et al., 1990), or whether it constituted a reduction of inefficient care without adverse effects on patient health (Chandra, Cutler, and Song, 2012; Miller and Luft, 1994; Lurie et al., 1994; Cutler, 2004; Berwick, 1996). On fee-for-service (FFS), which reimburses treatment inputs at the margin, the evidence is also mixed. Some have found that reductions in marginal reimbursement lead to increases in service volume (Rossiter and Wilensky, 1983; Dranove and Wehner, 1994; Gruber and Owings, 1996; Nguyen and Derrick, 1997; Yip, 1998; Jacobson et al., 2010; Rice, 1983), while others have found that increases in payments by 1% lead to an increase in service volume of 1.5% (Clemens and Gottlieb, 2014).

The paper will proceed as follows. In Section 2, I present the model and discuss its implications about the socially optimal allocation of treatment. In Section 3 I derive the optimal non-linear reimbursement contract in order to benchmark the 'best' attainable outcomes absent the practical restrictions of reimbursement contracts seen in the real world, discuss the optimality of treatment caps, and the potential of global budgets in achieving socially optimal outcomes. In Section 4, I restrict attention to linear contracts and discuss the optimality of fee-for-service and a prospective payment, which are the two types of reimbursement contracts seen in practice. I also derive a sufficient statistics formula that evaluates the optimality of payment schemes for existing conditions. In Section 5, I present an empirical exercise that computes the objects in the sufficient statistics formula in order to evaluate the payment scheme for pulmonary rehabilitation therapy, which is given to patients with COPD. In Section 6, I conclude.

2 Model

The model involves three actors: health care providers, patients, and the 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

I restrict attention to treatments that *do not harm* the patient. This assumption merits some discussion, as it confines the set of treatments and health conditions to which the model applies. While one may think that *all* treatments satisfy the "do no harm" clause of the Hippocratic oath, it is the case that some treatments, when given in excess, are toxic to the patient. An example of such treatments may include some provider-administered drugs, where an excessive dosage may potentially kill the patient. By assuming a health production function that is non-decreasing in treatment, I am automatically excluding treatments of the sort studied in Gaynor, Mehta, Richards-Shubik (2020), where too much treatment *harms* the patient.

Formally, I assume that the health production function is everywhere increasing in treatment, which means that patients will always benefit, at least slightly, from additional care. This health production is a good fit for procedures such as diagnostic services, physical therapy, provider office visits, evaluation and management services, diabetes treatment, dialysis (the procedure itself, not the anemia medications given in parallel), etc. It could even describe chemotherapy if we believe providers will not administer dosages above the toxicity threshold. This modeling choice echoes the Chandra and Skinner (2012) Type II and Type III class of treatments, but is slightly more restrictive.

Patients derive heterogeneous health benefits from treatment, and the benefits are known to the provider but not the insurer. This is the asymmetric information: the insurer will only know there are two patients with COPD, but only the provider knows which patient 'type' will benefit a lot from ten pulmonary rehabilitation sessions. That said, both patients would be better off with ten 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.

Formally, let the health production function, $h(x, \theta)$, depend on treatment, x, and patient type, θ, where h describes the dollar value of health (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.

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 both health gains and marginal health gains from treatment are increasing in θ, meaning that $h(x, \theta) \geq h(x, \theta')$ and that $h_x(x, \theta) \geq h_x(x, \theta')$, $\forall\ \theta > \theta'$.

Costs of treatment

Assume that costs of treatment, $c(x) \geq 0$ are observable, verifiable, and that $c_x(x) \geq 0$ and $c_{xx}(x) \geq 0$. Assume that the insurer may observe and contract on $c(x)$.[1] Notice that the cost depends only on the treatment, and not on the patient type. This assumption is convenient because it allows us to focus on one specific type of unobserved heterogeneity–in the health benefit– but it does restrict us to the types of treatments whose costs are the same across patients, conditional on treatment. In practice, we may think that patient type enters both the health production function and patient costs. Going back to the COPD example, a patient who really benefits from pulmonary rehabilitation may also require more attention during the session. However, among diagnostic services, evaluation and management services, diabetes screenings, and catheterizations, heterogeneity in costs is likely much smaller than heterogeneity in health benefit in the provider treatment decision.

Shutting down private information in the cost function means that asymmetric information can only create a wedge between the provider and the insurer via their respective valuations of the

[1] In practice, costs of treatment may not be entirely clear to the insurer. Typically, insurers have to impute cost information from charges, which are at best a noisy signal of costs. Nonetheless, there are some settings where costs are 'more' observable– such as physical therapy, evaluation and management services, or diagnostic services– and I focus on such settings for this paper.

patient health. On the one hand, the modeling choice means that the unobserved patient type will map one-to-one into the observed equilibrium treatment by entering only via the health production function, and not the cost function. On the other hand, it limits how much we can decompose the separate effect of asymmetric information and provider altruism. I will return to this second point when I describe the provider's treatment decision.

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 μ be the 'altruism' parameter, which scales the provider's valuation of patient health. Assume $\mu > 0$ with strict inequality. Provider utility from treating patient θ is

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

The parameter μ is the marginal rate of substitution between profits and patient health. A provider with a $\mu = 1$ will value patient health gains from treatment at exactly their value for the patient (and for society). I exclude $\mu = 0$ from the set for a couple of reasons. The first is that one could hardly argue that a health care provider enters the profession purely for financial reasons. While some providers may care more about financial incentives than others, it seems unrealistic to model a provider who does not value patient health at all. If one is to take the Hippocratic oath as an inherent feature of the health care profession, then it only seems appropriate to say that the provider cares about patient health.

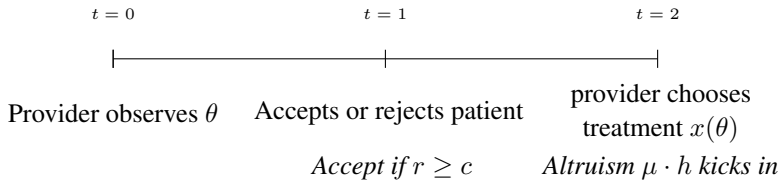
The second is more technical: $\mu = 0$ would shut down the asymmetric information in the model, which is one of the main interesting features I wish to study. Since I modeled patient heterogeneity through the health benefit function, and the provider's private information is over the patient's health benefit from treatment, including $\mu = 0$ in the set effectively shuts down the private information piece from the agent's decision. This would not be the case if the private information entered through the cost function, for instance. In a spirit of transparency, I note that working out the math with $\mu = 0$ results in first best outcomes, which may seem somewhat ludicrous. This happens because, in my model, a provider who does not put any weight on patient health will be perfectly indifferent between giving the healthy patient less treatment versus more.

Notice also that, if $\mu = 0$, the model would be identical to the standard non-linear pricing problem with asymmetric information from Maskin and Riley (1984). It is only when $\mu > 0$ that this model yields different predictions from the standard non-linear pricing problem. In fact, this term is precisely what drives the new theoretical results about second-best distortion for all types. Unlike the standard asymmetric information model, this agent values both his compensation *and* the principal's objective (health). The general flavor of the trade-offs in this contracting problem involve designing incentives to keep this agent from providing *too much care*, particularly if he gets reimbursed 100% of the costs of inputs on the margin.

The μ is isomorphic to the 'α' parameter in the Ellis and McGuire (1986) model: it is the ratio of the provider's marginal utilities between health and profits. My modeling choice is slightly more restrictive. By writing the objective function as I have done, I have implicitly imposed that health and profits are perfect substitutes with a slope μ 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). Nonetheless, the assumption that the marginal rate of substitution is constant may be thought of as a first order approximation, and has the benefit of simplifying the mechanics.

Timing

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.



Assume that the provider does not treat the patient if reimbursements are less than costs. Formally, assume that the participation constraint (PC) is such that profits per patient must be non-negative, $r(x) - c(x) \geq 0$. The provider is forward looking and the costs of treating each patient when he observed the patient type.

The assumption I have made implies a participation constraint which *differs* from the incentive constraint, which is non-standard and therefore merits some discussion.[2] There are many reasons why I think this is the appropriate modeling choice, both in this specific health care setting and more broadly when considering partially altruistic agents.

In this health care setting, providers typically practice in groups or affiliate to a larger medical care institution. At the group or institutional level, the 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. Empirically, there is strand of papers on selective admissions suggesting that provider practices avoid admitting patients who are unprofitable (Gruenberg and Willemain, 1982; Ettner, 1993; Uili, 1995; Ching et al., 2015; Ghandi, 2018). 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.

More broadly, this modeling choice may apply to other settings where altruistic agents that are making decisions within an organization. Excluding the altruistic component from the ex-ante participation constraint resonates more with what we see in the world than the standard alternative. Echoing the ideas from Hansmann's (1998) work on not-for-profit organizations, it is sometimes the case that the firm has a collection of owners, divorcing the rights of firm control from the rights to appropriate the firm's earnings. In such settings, one could argue that agreement on financial payoffs may facilitate the formation of the organization, even when the prospective owners intrinsically value the activities that the organization will undertake once it is formed. Agreement on financial returns may be of particular importance in the case where the intrinsic valuation of the organization's future activities differs among the potential owners.

Finally, the main theoretical results about treatment distortions for all types do not depend on this assumption. Nonetheless, this modeling choice turns out to be very convenient in the characterization of social optima. In particular, this choice 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

[2]The formulation has some overlap with Hart and Zingales (2019). Thank you to Oliver Hart, Edward Glaeser, Jerry Green, and David Sibley for providing extensive examples and insightful discussion on this modeling choice.

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, then, 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.

The Insurer's Problem

In the spirit of the government procurement literature, I formulate the provider contracting problem as that of a public insurer, and hence define a social welfare function (SWF) which will be the contracting objective for the insurer. In the rest of the paper, the problem of a private insurer will always be embedded in that of a public insurer, and I will discuss how the optimal contract for these different types of insurers may differ. I will refer to the public insurer as 'the government' interchangeably.

Consider a public insurer that 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. One can think of η as the social welfare weight on provider profits. Why might this be less than the social welfare weight on the patient? In the optimal tax literature, we typically think of social welfare weights as being inversely proportional to income. Since providers are generally on the right tail of the income distribution, this may be a reason to have small.

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. 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 into which a regulator can plug in his preferences. 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.

Tangentially, another possible SWF formulation could have included a loss term on payments per the shadow cost of public funds (Laffont and Tirole 1993). However, this formulation then implies that the social costs of treatment are greater than just the cost of the treatment. Why? Because in this case, 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. It would be interesting to study this alternative formulation in future work, but I do not pursue it here.

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 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.

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

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

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

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

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

2.1 First Best

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,

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

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

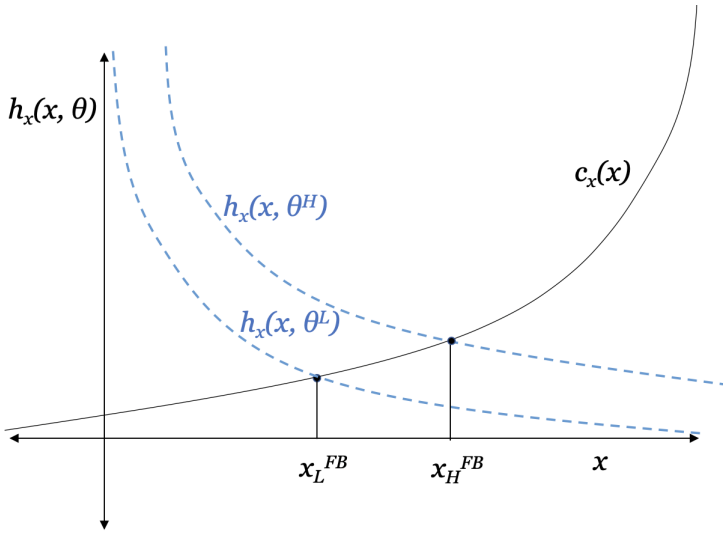


Figure 1: First Best Treatment Levels

3 Second Best: Optimal Non-Linear Contract

The provider in my model will always have an incentive to give more treatment than is socially optimal. Why? Because more treatment always improves the health of the patient, and the participation constraint of the provider effectively subsidizes costs of treatment. 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.

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 such that the high type gets more treatment than the low type, i.e. whether treatments are monotonic in the unobserved type. The first best treatment levels are already monotonic. If this provider's incentives are such that equilibrium treatments are *not* monotonic, then the second best contract will pool the two types.

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.

$$\begin{aligned} \mu h(x_L, \theta_L) + r_L - c(x_L) &\geq \mu h(x_H, \theta_L) + r_H - c(x_H) && \text{(IC L)} \\ \mu h(x_H, \theta_H) + r_H - c(x_H) &\geq \mu h(x_L, \theta_H) + r_L - c(x_L) && \text{(IC H)} \\ \implies h(x_H, \theta_H) - h(x_L, \theta_H) &\geq h(x_H, \theta_L) - h(x_L, \theta_L) \end{aligned}$$

QED. ■

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 $r_L > c(x_L) + \mu h(x_H, \theta_L) - \mu h(x_L, \theta_L)$, and one could reduce r_L by $\epsilon > 0$, still satisfy all the the constraints, and in turn increase the SWF by $(1-\gamma)(1-\eta)\epsilon > 0$. Therefore, the payments must be $r_H = c(x_H)$ and $r_L = c(x_L) + \mu h(x_H, \theta_L) - \mu h(x_L, \theta_L)$. These two binding constraints imply that $r_H = c(x_H)$ and $r_L = c(x_L) + \mu h(x_H, \theta_L) - \mu h(x_L, \theta_L)$.

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. Turning to (ICH), we can evaluate it at the (r_H, r_L) to obtain that,

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

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.

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

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

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. QED. ■

The optimal contract does not implement the first best treatment *for either type*. 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.

The reason this model distorts both types is altruism–the more this provider values health, the larger the incentive rent needs to be on low health benefit patients. In the standard asymmetric information model, the agent does not care about the principal's surplus, whereas in this model, he does. As I mentioned in the model section, this result differs from the standard 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. Notice that μ is close to zero gets us closer the first best because it shuts down the asymmetric information distortion. The model in this paper is not designed to study low levels of altruism, but rather the interplay of altruism with asymmetric information between the insurer and the provider. When μ is near zero, the provider does not care: he is indifferent between giving low types a lower or higher treatment level.[3]

Figure 2 shows graphically the distortion on both types. As η gets close to one, we get closer to first best. Why? 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 μ. In fact, for any $\eta < 1$, the larger the μ, the *farther* we get from first best.

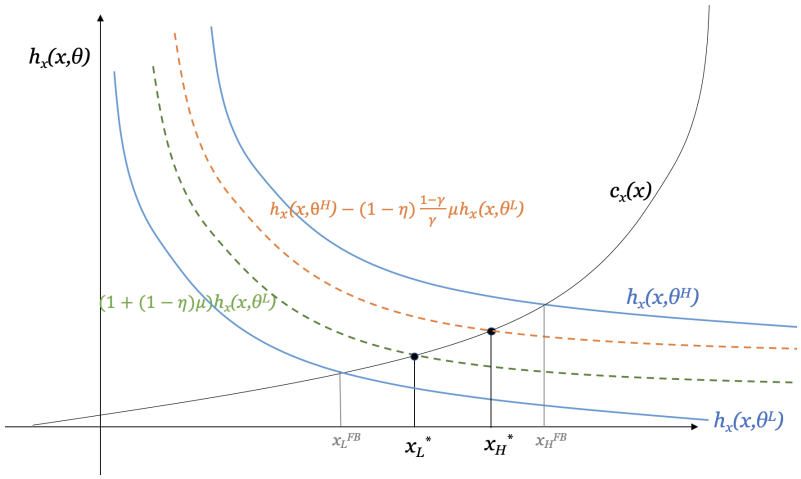


Figure 2: Optimal Contract Implemented Treatments, Unpooled

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 providers 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.

[3] A formal study of low altruism and asymmetric information would model the asymmetric information through both the health function and the cost function, which I leave for future work. 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.

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

$$h_x(x, \theta_H) > \left(1 + (1-\eta)\frac{\mu}{\gamma}\right) h_x(x, \theta_L) \qquad (*)$$

Proof. Proof: Suppose that $(*)$ 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

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

By rearranging terms, one immediately obtains condition $(*)$. QED. ■

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 $(*)$ 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 μ is large, it is more likely we are in the pooling solution. For high μ, 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 N-type case: Treatment Cap and Outlier Patients

The results in the two type case carry forth into the N type case. The additional insight from the N type case is that adding types with higher treatment needs makes it more expensive to insure all the types below, so the optimal contract end up capping treatment at some level. Adding types is like magnifying the wedge from asymmetric information, because it means that treatment needs for patients within a diagnosis group are more volatile.

Consider a condition like schizophrenia, for example, where treatment needs for observably similar patients may be very different, with some inpatients requiring severe physical restraining and intensive medical care, while the majority of the other inpatients just require a pharmaceutical prescription and light monitoring. If the insurer cannot tell ex-ante which patient is which, and can only choose how much to pay per day for an inpatient stay, choosing to cover the care of the most complicated patients means that the reimbursement contract must cover, say, a 30 day inpatient stay. The temptation for the provider then becomes keeping all the simple patients for 30 days, as they all will benefit slightly from the longer stay, though not as much as the most complicated patients.

The main tension for the insurer dealing with N types thus comes from the highest level of x covered, or equivalently, choosing a threshold patient type above which there is pooling in the treatment level. 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. Why? Since health is increasing in treatment, adding types to the right also means that the *lowest* type always benefits, health wise, from receiving the treatment level of the *highest* type.

In order to disincentivize the provider from keeping the simpler schizophrenic patients around, the reimbursement for one, two, or even five days has to yield a high profit margin. It would be cheaper for the insurer to cap treatment at five days, but it would not be worthwhile to do so if the marginal health gains of the most complicated patients are much larger if they stay the thirty days. 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: when additional treatment cannot hurt any patient, and the provider gets fully reimbursed for costs, the altruistic provider will always want to give more treatment to his patients.

The prevalence of treatment caps in the real world is also consistent with the predictions of my model. 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 bottom line, however, is that when the provider delivers care as if costs were zero, covering the costs of that care eventually circles back to the patient or to taxpayers.

More formally, suppose now that there are N types of patients, $\theta_i \in \{\theta_1, ..., \theta_N\}$, where $h(x, \theta_i)$ and $h_x(x, \theta_i)$ are both increasing in θ. 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.

Definition 1 *Let θ_T denote the threshold type, above which there is pooling, and let x_T denote the maximum level of treatment covered. That is, let treatment level for all $\theta_i \leq \theta_T$ be given by the equilibrium treatment $x_i^*(r, \theta_i)$, and the treatment level for all $\theta_i > \theta_T$ be fixed at x_T.*

In the N type case, the optimal contract involves both choosing a maximum treatment level x_T (which has a one-to-one correspondence to a threshold type θ_T), and a non-linear fee schedule for all coverage levels below. Why? Because the $r \geq c$ constraint on *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. When we begin to add patient types to the right, the treatment needs of the highest type begin to look significantly different (and larger) than the treatment needs of the average type. This comes from the type of heterogeneity in my model: patients with higher θ derive higher benefits from treatment, and thus benefit from receiving more treatment than the rest. From the provider's perspective, there is no downside to treating the lower patient types with as much care as the high types, as the low types benefit slightly (though not enough to make it worth the incremental treatment cost, socially).

Proposition 3 *For all types $\theta_j < \theta_T$, the optimal contract gives an incentive rent on type j such that*

$$\pi(x_j) = \mu \sum_{i=j}^{T-1} h(x_{i+1}, \theta_i) - h(x_i, \theta_i) \qquad , j \in \{1, ..., T-1\}.$$

For all types above θ_T, the optimal contract pays $\pi(x_T) = 0$.

Proof. In the N type case using the profits notation, the participation constraint of the threshold type T is $\pi(x_T) \geq 0$. First, $\pi(x_T)$ has to be zero because $(PC\ T)$ will be binding at the optimum. For the sake of contradiction, suppose that $(PC\ T)$ is not binding. Then, there exists an $\epsilon > 0$ such that $\pi(x_T) - \epsilon \geq 0$. It is the case that this $\pi(x_T) - \epsilon$ will also satisfy all the local upward incentive constraints.

Let $(IC\ j \rightarrow j+1)$ denote the local upward incentive constraint for type j, which requires that $\mu h(x_j, \theta_j) + \pi(x_j) \geq \mu h(x_{j+1}, \theta_j) + \pi(x_{j+1})$. Writing the local upwards incentive constraints recursively for every $j \in \{1, ...T-1\}$ yields that $\pi(x_j) \geq \pi(x_T) + \mu \sum_{i=j}^{T-1} h(x_{i+1}, \theta_i) - h(x_i, \theta_i)$. Reducing $\pi(x_T)$ by ϵ, one can see that

$$\pi(x_j) \geq \pi(x_T) + \mu \sum_{i=j}^{T-1} h(x_{i+1}, \theta_i) - h(x_i, \theta_i) \implies \pi(x_j) \geq \pi(x_T) - \epsilon + \mu \sum_{i=j}^{T-1} h(x_{i+1}, \theta_i) - h(x_i, \theta_i).$$

Further, reducing $\pi(x_T)$ by ϵ will strictly increase the SWF by ϵ. Therefore, it is optimal to continue reducing $\pi(x_T)$ until $(PC\ T)$ is binding.

Second, we have to show that for an arbitrary $j < T$, the local upward incentive constraint for type j is binding. For the sake of contradiction, suppose it is not binding. Then, there exists an $\epsilon > 0$ such that $\pi(x_j) - \epsilon \geq \mu \sum_{j}^{N-1} h(x_{j+1}, \theta_j) - h(x_j, \theta_j)$, meaning that local upward incentive constraint is still satisfied, and reducing $\pi(x_j)$ by ϵ strictly increases the SWF. Therefore, it is optimal to continue reducing $\pi(x_N)$ until $(IC\ j \rightarrow j+1)$ is binding. QED. ■

Proposition 3 highlights that the threshold patient, θ_T, crucially determines how expensive it is to insure everyone. Reducing the number of types pooled at the top raises the magnitude of the incentive rents required on all the types below.

The only case where the optimal non-linear contract *does not* pool the types at the top is when the incremental health benefit of the highest type is significantly higher than the incremental health benefit across all the other types. In other words, 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.

Proposition 4 *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. It is sufficient to require that the average health gains of all types above the threshold type T satisfy the following condition.*

$$\frac{\sum_{j=T}^{N} h_x(y, \theta_j) - h_x(y, \theta_{T-1})}{h_x(y, \theta_{T-1}) - h_x(y, \theta_{T-2})} > \mu(1-\eta)\left((T-1) + T \cdot \frac{h_x(y, \theta_{T-1})}{h_x(y, \theta_{T-1}) - h_x(y, \theta_{T-2})}\right) \qquad (**)$$

Proof. One can plug in the contract derived in Proposition 3 into the SWF and characterize

the solution for each x_i.

$$\frac{\partial SWF}{\partial x_T} = \sum_{j=T}^{N}(h_x(x_T,\theta_j) - c_x(x_T) - (1-\eta)\cdot\mu(T-1)h_x(x_T,\theta_{T-1}) = 0 \quad (2.\text{T})$$

....

$$\frac{\partial SWF}{\partial x_j} = h_x(x_j,\theta_j) - c_x(x_j) + \mu(1-\eta)[j\cdot h_x(x_j,\theta_j) - (j-1)\cdot h_x(x_j,\theta_{j-1})] = 0 \quad (2.\text{j})$$

....

$$\frac{\partial SWF}{\partial x_1} = h_x(x_1,\theta_1) - c_x(x_1) + \mu(1-\eta)h_x(x_1,\theta_1) = 0 \quad (2.1)$$

Since the solution characterized must be monotonic in the treatment x's, we can leverage the fact that the cost function is the same. We can order the x_i's that solve conditions $(2.T)$, $(2.T-1)$, ..., (2.1) by solving for the equilibrium $c_x(x_i)$ in each first order condition and ranking them in order. The monotonicity condition for the threshold type comes from establishing an inequality between the expression that come from $(2.T)$ and $(2.T-1)$.

$$\underbrace{h_x(x_{T-1},\theta_{T-1}) + \mu(1-\eta)[j\cdot h_x(x_{T-1},\theta_{T-1}) - (T-2)\cdot h_x(x_{T-1},\theta_{T-2})]}_{=c_x(x_{T-1}) \text{ from condition } (2.T-1)} < \underbrace{\sum_{j=T}^{N}(h_x(x_T,\theta_j) - (1-\eta)\cdot\mu(T-1)h_x(x_T,\theta_{T-1})}_{=c_x(x_T) \text{ from condition } (2.T)}$$

By rearranging the expression, one immediately arrives at condition $(**)$. QED. ■

The condition derived in Proposition 4 parallels that of the two type case, derived in Proposition 2: it is a sufficient condition for there to be a separating equilibrium between type T and all types below. Notice that as T gets large and starts to approach N, the condition becomes harder to satisfy. In fact, the only way for the condition to hold as $T \to N$ is if the incremental marginal health gains between type T and type $T-1$ is very large, and much larger than the rest. In other words, the value accrued to the insurer in 'health benefit' dollar terms has to be very large.

Conversely, if the highest patient type has small incremental health gains relative to everyone else, the financial effect dominates, and it becomes too expensive to insure them to receive high levels of care. This is because catering to the high types raises the insurance costs for everyone (via the larger incentive rents), so the incremental health benefits of the few patient types at the top would have to contribute a lot to the average health level of the patient pool for it to be worthwhile. When these are incremental health gains of types at the top are small, it becomes optimal for the insurer to cap their treatment level at some fixed amount in order to contain insurance costs for everyone else. In the optimal contract, the threshold type gets over-treated relative to his first best levels.

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.

In general, the only set of reimbursement contracts which are consistent with the incentive compatibility constraints and monotonicity must have declining profits in type. Why? Because a provider who values patient health and does not bear the treatment costs already has an intrinsic motivation to over-treat. Just as in the two type case, the insurer keeps the provider from over-treating low types via 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 following Lemma formalizes this logic, and proves to be useful later in section 4.

Lemma 2 *Let $\pi(x_i) = r(x_i) - c(x_i)$. The incentive compatible contract must have profits decreasing in type.*

Proof. Rearranging the local upwards incentive constraint for any arbitrary type j yields that,

$$\mu h(x_j, \theta_j) + \pi(x_j) \geq \mu h(x_{j+1}, \theta_j) + \pi(x_{j+1}) \implies \pi(x_j) - \pi(x_{j+1}) \geq \underbrace{\mu h(x_{j+1}, \theta_j) - \mu h(x_j, \theta_j)}_{\geq 0 \text{ by monotonicity}}.$$

Hence, $(IC\ j \to j+1)$ and monotonicity in x jointly imply that profits must be decreasing in the type. QED. ■

3.1 Relaxing the Participation Constraint: Promise of Global Budgets

One of the main takeaways from the previous section 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 did not have to worry about designing an incentive scheme with $r \geq c$ on the most expensive patient, then the contracting problem can be solved with many different reimbursement fee schedules, all which implement first best outcomes.

Consider an alternative model with a relaxed participation constraint, such that the provider treats patients as long as average reimbursements are weakly greater than average costs. While I already discussed the reasons why I think the $r \geq c$ implementation requirement is realistic, there are two applications in which I think the relaxed participation constraint would fit well: one is a global budget setting, where the insurer can agree to compensate the provider ex-ante under the premise that the provider will treat every patient, independent of the ex-post, per patient, profitability. The second is the setting in which the provider does not know the patient type when deciding to treat or not treat the patient. In both of these applications, we effectively go back to a world of symmetric information.[4] The first could correspond to a global budget system as in the United Kingdom. The second could correspond to any condition in which individual treatment needs are hard to predict, ex-ante.

More formally, consider an alternative model in which the participation constraint of the provider is such that he treats every patient, as long as reimbursements are greater than or equal to costs on average. That is, consider a world 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,

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

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

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

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

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] Thank you to David Cutler and Dan Barron for insightful discussions that helped develop and interpret this specification.

Lemma 3 *There is a continuum of contracts that satisfy all the constraints and maximize the objective.*

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 $\gamma(r_H - c(x_H)) + (1-\gamma)(r_L - c(x_L)) = 0$. QED. ■

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

Proof. Since there is no unique optimal (r_L, r_H), we can leverage the relationship between (r_L, r_H), which is given by $\gamma(r_H - c(x_H)) + (1-\gamma)(r_L - c(x_L)) = 0$ per Lemma 3. Plugging in the relationship of (r_L, r_H) into the objective function, we obtain:

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

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. QED. ■

Among the many contracts that implement first best treatments, an interesting one to focus on is $r(x_i) = t + (1-\mu)c(x_i)$ 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,

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

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. By shutting down the asymmetric information wedge, the importance of knowing μ rises to the surface. In practice, it may be hard for the insurer to know μ. It may well be the case that providers have heterogeneous μ'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.

While I leave a more formal study of altruism heterogeneity for future work, I would like to highlight that if μ was 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). In other words, altruism heterogeneity alone would not be a problem is patients do not have heterogenous treatment needs. Further, since the participation constraint in

my model does not depend on the provider's altruism, altruism heterogeneity can only affect the levels of treatment given to patients, and not the set of patients seen. The interplay of altruism heterogeneity and other forms of patient heterogeneity, however, is much more nuanced and merits independent study.

4 Linear Contracts

Linear contracts merit special attention given they are both practical and very prevalent in the real world. In this section, I go back to contracts that require reimbursement to be at least cost for every patient, and conduct a detailed study of linear contracts when the choice set for the provider is *unrestricted*. What does this mean? I think of an insurer that can not dictate the exact treatment level x for which he pays $r(x)$, but rather sets a fee schedule where providers get reimbursed for claims with any possible x administered. I focus on unrestricted choice spaces because the administrative burden of such contracts for the insurer is somewhat smaller, and contracts observed in the world tend to have this feature.

The goal of this section two-fold. The first is to delineate a set of conditions in which prospective payment, fee-for-service, or a mixed reimbursement system may be optimal under minimal assumptions. Beyond the health care setting, one could apply these results to an employer's decision on whether to pay with a fixed salary or an hourly wage when he has altruistically motivated employees. Notice that fee-for-service corresponds to a wage contract that pays only the margin, conditioning on costly inputs, while prospective payment corresponds to a fixed salary that is independent of input costs.

The second goal is more specific to the health care setting, and consists of deriving sufficient statistics formulas that may be useful to an insurer (public or private) in setting reimbursement rates for provider services. The latter will involve more restrictive assumptions that are sensible in this particular health care setting. In the spirit of the optimal tax theory literature, I will derive formulas from the model that depend on measurable and observable descriptives about health care costs.

Suppose the insurer reimbursement contract, $r(x)$, is restricted to be linear in $c(x)$, $r(x) = t + \phi c(x)$. Suppose also that there are N types. The insurer's objective thus is

$$SWF = \sum_{i=1}^{N} \gamma_i [h(x_i^*, \theta) - c(x_i^*) - (1-\eta)\eta(t + \phi c(x_i^*) - c(x_i^*))],$$

where x_i^* is optimally chosen by the provider. Since the insurer dislikes making payments but values health outcomes for all patients, the trade off will be to choose a t that is sufficiently high to cover costs of the high type, and balancing it against a ϕ such that our insurer does not have to reimburse every patient at the costs of the highest type. In effect, the ϕ allows the insurer to save on the t, which a lump amount paid *for every patient*, but at the cost of pushing everyone towards *too much* treatment.

The table below illustrates the correspondence between the model's linear contract and the reimbursement schemes seen in the world: prospective payment pays nothing at the margin, while fee-for-service reimburses a share of costs.

Payment Scheme	*Share of costs reimbursed,* ϕ	*Lump sum transfer per patient,* t
Prospective Payment (PPS)	$\phi = 0$	$t > 0$
Fee-for-Service (FFS)	$\phi \geq 1$	$t = 0$
Mixed Reimbursement	$\phi < 1$	$t > 0$

4.1 Characterizing the optimal contract

The contracting tensions for the insurer in the linear case involve trading off the incentives for over-treatment against the magnitude of the incentive rent. The ϕ affects the provider's marginal incentives by subsidizing a share of treatment costs. This means that higher values of ϕ push equilibrium treatment levels up for everyone, driving up overall health care utilization. The t determines the highest treatment coverage level. Since the insurer has to issue the flat payment amount for every patient, it can get very expensive to set a t that is sufficiently large to cover the treatment costs of high patient types. Just as in the N type non-linear contract, the insurer will have to choose the maximum level of treatment he is willing to cover.

Definition 2 *Let $x_i^*(\phi)$ be the equilibrium treatment quantity for patient type θ_i, at a fixed contract (t, ϕ). That is, denote $x_i^*(\phi)$ the solution to*

$$\mu h_x(x_i^*(\phi), \theta) - c_x(x_i^*(\phi))(1 - \phi) = 0. \qquad \text{(provider FOC)}$$

It will never be optimal for the insurer in my model to reimburse more than one hundred percent of costs. Why? Doing so implies that the provider makes a profit on each unit of treatment. Since the provider values both profits and patient health, and we have assumed patient health is increasing in treatment quantity, reimbursing above 100% of costs makes the provider want to give infinite treatment in the model.[5]The Lemma below formalizes this logic.

Lemma 4 *The optimal ϕ has to be less than or equal to one.*

Proof. Per Lemma 2, we know that any incentive compatible contract must give declining profits in type. Since $x_i^*(\phi)$ is increasing in θ_i, and $c(x)$ is increasing and convex, the only way for $\pi(x) = t + \phi c(x) - c(x)$ to be decreasing in θ_i is if $\phi \leq 1$. QED. ■ Lemma 4 implies that the only possible fee-for-service system that could be optimal is $\phi = 1$. Notice that mixed reimbursement schemes are constrained to have $t > 0$ and $\phi < 1$ via the incentive and participation constraints; these restrictions are not imposed, but rather fall out of the implementation conditions.

Corollary 1 *If $\phi < 1$, any optimal mixed reimbursement scheme must have $t > 0$.*

Proof. If $\phi < 1$, the set of t's that satisfy the participation constraint, $t + (\phi - 1)c(x) \geq 0$ must be such that $t > 0$. QED. ■

The lump transfer, t, will be pinned down by the binding participation constraint of the highest type. Intuitively, the flat rate component of the contract is there to compensate the provider for total treatment costs, in the case that the share of costs reimbursed is less than one-hundred percent. The benefit of the flat payment is that it does not distort incentives on the margin. The downside is that it is paid on every patient. This contracting problem echoes the one of a monopolist choosing a two-part tariff who is fixated on catering to entire market, as we *do* want to make sure *all patients* receive treatment. Hence, the flat rate must be such that the most expensive patient receives treatment.

Lemma 5 *For any $\phi \leq 1$, the fixed payment of the contract, t, is always pinned down by the treatment cost of the highest type covered, θ_N.*

$$t = (1 - \phi)c(x_N)$$

[5]In practice, there surely are capacity constraints that prevent provider or provider from giving infinite care, and these can be easily put into the model, if we want. Independent of whether the model's equilibrium treatments are infinite or finite at some level of the capacity constraint, one can hardly argue that reimbursing above 100% of costs does not create additional incentives for over-treatment.

Proof. The participation constraint requires that $t + \phi c(x_i) - c(x_i) \geq 0$, $\forall\, i \in \{1, ..., N-1\}$. By monotonicity in x_i, it follows that $c(x_N) \geq c(x_i)$ for all patient types θ_i in $i \in \{1, ..., N-1\}$.

$$t \geq (1-\phi)c(x_N) \qquad \implies t \geq (1-\phi)c(x_i),\ \forall\, i \in \{1, ..., N-1\}.$$

Since the SWF is decreasing in t, $t \geq (1-\phi)c(x_N)$ will bind at the optimum. QED. ■

The result in lemma 5 means that the insurer's optimization problem is really just over the cost sharing parameter, ϕ. We can think of the ϕ as a high powered incentive, as it affects treatment decisions on the margin, for *all* patients. To gather some intuition, note that ϕ is doing two things at the same time: it subsidizes the treatment costs borne by the provider–pushing every patient toward higher treatment levels–but it also reduces the size of the flat payment. The first effect affects total insurer expenditures through a higher overall health care utilization channel, while the second effect affects expenditures purely through a financial channel.

Definition 3 *Define the semi-elasticity of treatment with respect to reimbursement rate for a patient type θ_i as*

$$\varepsilon_\phi^i \equiv (1-\phi)\frac{dx_i^*}{d\phi}.$$

The semi-elasticity, ε_ϕ^i, describes how much a provider changes the equilibrium administered treatment with a proportional reduction in treatment cost-sharing. In other words, ε_ϕ^i describes the *level* changes in treatment with *proportional* changes in provider cost sharing. It is important to note that the semi-elasticity depends on ϕ via $\frac{dx_i^*}{d\phi}$, which may be some complicated function of h and c.

Proposition 6 *The derivative of the social welfare function with respect to the optimal slope of the linear contract, ϕ, is*

$$\frac{dSWF}{d\phi} = \left(\frac{1-\phi}{\mu} - 1\right)\bar{\varepsilon}_\phi + (1-\eta)(1-\phi)(\bar{\varepsilon}_\phi + c(x_N) - \bar{c}) - (1-\eta)c_x(x_N)\varepsilon_\phi^N, \qquad (2)$$

where $\bar{\varepsilon}_\phi \equiv \sum_{i=1}^{N} \gamma_i c_x(x_i)\varepsilon_\phi^i$ is the average cost *semi-elasticity, and $\bar{c} \equiv \sum_{i=1}^{N} \gamma_i c(x_\theta)$ the average cost.*

Proof. See Appendix. ■

4.2 On the Optimality of Prospective Payment and Fee-for-Service

Consider a world in which providers are perfectly altruistic, meaning $\mu = 1$. At a first glance, one might be inclined to say that a fixed payment contract with $\phi = 0$ would be 'optimal' for this provider: he will choose treatment such that marginal health benefit equals marginal cost, and reimbursements on the margin would only make him deviate from this. This would certainly be true if there was only one patient type, this is what Ellis and McGuire (1986) call as 'the promise of prospective payment'. But as soon as we consider an insurer who dislikes making payments, this 'promise of prospective payment' breaks. I find that prospective payment becomes too expensive when we consider both that making payments to providers is costly and that there is heterogeneity in treatment costs across observably equivalent patients. Why? Because the insurer has to pay a fixed rate on every patient, and this amount must be sufficiently large to cover the costs of the most expensive patient.

An important insight from my model is highlighting the central role of cost spreads, in addition to the provider agency problem, when designing an optimal payment system. If we think there

patient heterogeneity in treatment in the world, and the insurer has interest in a contract that gets all of his patients at least seen, then the fixed payment contract may be too expensive. Note that an implicit assumption is that the insurer in my model wants all of his patients to be seen by the provider.[6]

Proposition 7 *If an change in ϕ shifts treatment costs uniformly across types, then prospective payment will not be optimal when $\mu = 1$.*

Proof. If increasing ϕ affects treatment costs uniformly across types, then $c_x(x_N)(\varepsilon_\phi^N \approx \bar{\varepsilon}_\phi)$. Evaluating the derivative of the SWF from Proposition 6 at $\phi = 0$ and $\mu = 1$ results in the following expression.

$$\left.\frac{dSWF}{d\phi}\right|_{\phi=0} = (1-\eta)(\underbrace{c(x_N) - \bar{c}}_{\geq 0 \text{ by monotonicity}} - \underbrace{(c_x(x_N)\varepsilon_\phi^N - \bar{\varepsilon}_\phi)}_{\approx 0 \text{ by assumption}}) \geq 0$$

If the marginal impact of raising ϕ by a small amount is positive, the means that there was a welfare increasing deviation at $\phi = 0$, so $\phi = 0$ could not have been optimal. QED. ■

Proposition 6 provides a new motive for mixed reimbursement scheme (i.e. flat payment plus share of cost reimbursement), *beyond* provider altruism: cost spreads. If we were not worried about providers being imperfect agents for their patients, accounting for implementation costs provides a good reasons to shy away from prospective payment. Now, if the government were to value provider rents at a high value of η, the welfare gains from moving away from a purely prospective payment system become very small.

What about fee-for-service? Recall that the only possible optimal fee-for-service has $\phi = 1$ because $\phi > 1$ is never optimal, and $\phi < 1$ violates the participation constraint.

Proposition 8 *A fee-for-service system with reimbursement at cost is never optimal, for any level of altruism μ.*

Proof. Evaluating the derivative of the SWF at $\phi = 1$, we obtain that

$$\left.\frac{dSWF}{d\phi}\right|_{\phi=1} = -\bar{\varepsilon}_\phi - (1-\eta)c_x(x_N)\varepsilon_\phi^N.$$

This derivative is always negative because $\varepsilon_\phi^i = (1-\phi)\frac{dx_i^*}{d\phi}$. From the physician optimization problem, we know that $\frac{dx_i^*}{d\phi} = \frac{c_x(x)}{-\mu h_{xx}(x,\theta_i)+(1-\phi)c_{xx}(x)} \geq 0$. Therefore, $\frac{dSWF}{d\phi} \leq 0$, which means that $\phi = 1$ could not have been optimal. QED. ■

The insurer can always increase social welfare by exposing the provider to some cost sharing and compensating him with flat fee. That is, the insurer can always do better by switching to a mixed reimbursement contract, and this result does not depend on the level of provider altruism or the social welfare weight placed on physician profits.

[6] One could embark on a separate exercise and use this model to study whether it is socially efficient to treat all types, which would be an interesting direction for future work. This could inform the direction of reimbursement reform for Medicaid schizophrenic patients, where costs are volatile across patients, and the prevailing concern is that DRG payments are too low for providers to accept them for treatment.

4.3 Sufficient statistics for optimal reimbursement

Moving forward, I will assume that costs are linear in treatment level, so $c(x) = cx$. One could argue that linear costs is well suited for treatments like physical therapy, diagnostics, or other evaluation and management services, as costs scale linearly with the intensity of care. Ultimately, it depends on how we specify that x is measured–if x corresponds to number of visits, frequency of arthritis injections, or frequency of dialysis treatment per week, we can think of the additional unit as having similar treatment costs (from the provider perspective) as the rest. However, if x corresponds to the probability of catheterization or some other major procedure (as in Clemens and Gottlieb, 2014), then higher values of x may correspond to higher treatment costs, and a convex cost function may be more appropriate.

The trade offs from restrictive assumptions here are similar to those in the optimal tax theory literature; there, the typical assumption is that the labor supply elasticty is constant across unobserved types. Here, I do not need to go as far as assuming a constant elasticity, but my assumptions about cost do confine the types of care where my formulas apply. Before the presenting the formula of the optimal linear contract, I discuss the objects on which the optimal ϕ depends on in order to motivate the functional restrictions that follow. Under linear costs, however, this elasticity is simpler, and in fact corresponds one-to-one and monotonically with the treatment heterogeneity across types, $\frac{dx_i^*}{d\theta_i}$.

Lemma 6 *The derivative of equilibrium treatment x_i^* with respect to ϕ is a monotonic transformation of the derivative of x_i^* with respect to the unobserved type, θ_i. In particular,*

$$\frac{dx_i^*}{d\phi} = \frac{dx_i^*}{d\theta_i} \cdot \frac{c}{\mu h_{x\theta}(x_i^*, \theta_i)}.$$

Proof. Differentiating the provider first order condition totally with respect to ϕ and θ_i results in the following two expressions.

$$\begin{aligned} \frac{d\ FOC}{d\phi}: &\qquad \mu h_{xx}(x_i^*, \theta_i)\frac{dx_i^*}{d\phi} + c = 0 \\ \frac{d\ FOC}{d\theta_i}: &\qquad \mu h_{xx}(x_i^*, \theta_i)\frac{dx_i^*}{d\theta_i} + \mu h_{x\theta}(x_i^*, \theta_i) = 0 \end{aligned}$$

By rearranging, one can immediately see that $\frac{dx_i^*}{d\phi} = \frac{dx_i^*}{d\theta_i} \cdot \frac{c}{\mu h_{x\theta}(x_i^*, \theta_i)}$. Since x_i^* is monotonically increasing in type, and we know by assumption that $h_{x\theta} \geq 0$, we conclude this relationship is monotonically increasing. QED. ■

This lemma turns out to be very useful in mapping the model to an empirical counterpart, as it tells us that the observed variation in equilibrium action corresponds monotonically to the unobserved health benefit heterogeneity in the model. This means we can learn about the provider's optimization decision by looking at the variation in the data of treatment levels within observably similar types (e.g. same diagnosis code and risk profile).

If one further assumes that the heterogeneity in health benefits across patients types is multiplicatively separable in θ, meaning $h(x, \theta) = \theta h(x)$, we can finally arrive at a working formula for the optimal contract. What does this assumption imply about unobserved heterogeneity? That patients health benefit varies *proportionately* across types. Going back to the COPD example, this means that if the patient with the health conscious spouse gains one quality adjusted life year by going to 10 sessions of pulmonary rehabilitation, the single patient will gain *twice as much*. This class of benefit functions resonates with the ones in the standard non-linear pricing problem.

Lemma 7 *Assume that the health production function is of the form $h(x, \theta_i) = \theta_i h(x)$. Then, the semi elasticity, ε^i_ϕ will not depend directly on ϕ, and will be given by*

$$\varepsilon^i_\phi = \theta_i \frac{dx^*_i}{d\theta_i}.$$

Proof. If the health production function is given by $h(x, \theta_i) = \theta_i h(x)$, it follows that $h_{x\theta}(x, \theta) = h_x(x)$. Under this class of $h(\cdot)$ functions, the provider's equilibrium treatment decision implies that $h_x(x, \theta_i) = \frac{(1-\phi)c}{\mu\theta_i}$. Per the lemma from above, it follows that,

$$\frac{dx^*_i}{d\phi} = \frac{dx^*_i}{d\theta_i} \cdot \frac{c}{\mu h_x(x^*_i)} = \frac{dx^*_i}{d\theta_i} \cdot \frac{c}{\frac{(1-\phi)c}{\theta_i}}. = \frac{dx^*_i}{d\theta_i} \cdot \frac{\theta_i}{1-\phi}.$$

Therefore, the semi-elasticity under this functional form assumption for $h(x, \theta)$ is equal to to the expression above. QED. ■

Optimal contract

The formula for the optimal contract depends primarily on two empirical objects: the range of administered treatment levels observed in equilibrium, and the discrepancy of outliers with the rest of the patients within the payment group. To gain some intuition, let's fix the provider altruism at $\mu = 1$ and benchmark the 'efficient' level of care with the prospective payment contract (e.g. flat payment). The range matters because the flat payment must adjust to cover the treatment costs of the patient with highest medical care. If the range is large, it becomes worthwhile to 'distort' care from the efficient level by subsidizing provider costs on the margin because, otherwise, the flat payment would have to be very large. The outlier component, on the other hand, tells us how different the patient with the highest care needs looks from the rest; if he looks very different from the rest, it is not worthwhile to distort everyone else from the efficient level of care.

Definition 4 *Let the normalized range by denoted by σ, where*

$$\sigma \equiv \sum_{i=1}^{N} \gamma_i \frac{(x^*_N - x^*_i)}{\sum_{i=1}^{N} \gamma_i [\theta_i \frac{dx^*_i}{d\theta_i}]}.$$

The range, σ, describes the difference in equilibrium treatment levels between the highest patient type and the average patient, normalized by the average incremental treatment across types (which is equal to the average semi-elasticity if the health function is multiplicative in the unobserved type). The denominator of σ will be large if the equilibrium treatment levels observed in equilibrium are very spaced out across patient types. What does this correspond to in the data? Say we have three patient types. If the first receives one visit per month, the second receives two, and the third receives ten, then the average incremental treatment will be five. If the third patient type received, instead, three visits per month, the average incremental treatment will be one. Therefore, the magnitude range is 'penalized' if the types happen to be very spaced out. Conversely, if the types are close together, the magnitude of the range will be 'amplified'.

Definition 5 *Let the normalized outlier ratio by denoted by ω, where*

$$\omega \equiv \frac{\theta_N \frac{dx^*_N}{d\theta_N}}{\sum_{i=1}^{N} \gamma_i [\theta_i \frac{dx^*_i}{d\theta_i}]}.$$

The outlier ratio captures how different the highest patient type is from the rest. The outlier ratio will be large if the incremental treatment for the highest patient type, relative to the second highest, is much larger than the incremental treatment across the other types. In other words, if the distribution of treatment quantities observed in equilibrium has a tail on the right, the ω will be big. Conversely, if this distribution is concentrated around the mean, the ω will be small.

Proposition 9 *Assume costs are linear and that the health production function is multiplicative in the unobserved patient type. The optimality condition for the linear contract is given by*

$$1 - \phi = \frac{1}{\frac{1}{\mu} + (1-\eta)(\sigma - \omega + 1)}. \tag{9.1}$$

The optimal linear contract will tend towards:

1. *fee-for-service if the range, σ is large.*
2. *prospective payment if the outlier ratio, ω is large.*

Proof. When costs are linear, the derivative of the SWF with respect to ϕ is

$$\frac{dSWF(\phi)}{d\phi} = \sum_{i=1}^{N} \gamma_i [(\frac{1-\phi}{\mu} - 1) c \frac{dx_i^*}{d\phi} - (1-\eta) \left[(1-\phi) \left(c\frac{dx_N^*}{d\phi} - c\frac{dx_i^*}{d\phi} \right) - (cx_N^* - cx_i^*) \right]].$$

Setting $\frac{dSWF}{d\phi} = 0$ and substituting in for $(1-\phi)\frac{dx_i^*}{d\phi} = \theta_i \frac{dx_i^*}{d\theta_i}$, we can establish a relationship between $(1-\phi)$ and our empirical objects.

$$1 - \phi = \frac{1}{\frac{1}{\mu} + (1-\eta) \sum_{i=1}^{N} \gamma_i [\frac{(x_N^* - x_i^*)}{\sum_{i=1}^{N} \gamma_i [\theta_i \frac{dx_i^*}{d\theta_i}]}] - (1-\eta)(\frac{\theta_N \frac{dx_N^*}{d\theta_N}}{\sum_{i=1}^{N} \gamma_i [\theta_i \frac{dx_i^*}{d\theta_i}]} - 1)}$$

Since $\sum_{i=1}^{N} \gamma_i [\theta_i \frac{dx_i^*}{d\theta_i}] > 0$, the formula is well defined, though not an explicit solution for ϕ given that the semi-elasticity still depends indirectly on ϕ via the equilibrium x_i^*, even if it does not depend *directly* on ϕ. Applying the definitions from σ and ω, we arrive at the desired formula. The comparative statics on μ, σ, and ω immediately follow. QED. ■

Looking at proposition 9 qualitatively, if the range is large, provider cost sharing yields savings the insurer on the flat payment amount. However, if the outlier ratio is very large, meaning that the patient with the highest utilization of medical services looks very different from the rest, it is not worthwhile to use provider cost sharing to save on the prospective payment because subsidizing provider costs raises utilization of the highest type by too much. Instead, it is better for the insurer to cap maximum treatment with a prospective payment.

What are the advantages of the assumptions on $c(\cdot)$ and $h(\cdot)$? The linear cost assumption has the advantage that the level of marginal costs factors out of the social welfare function when evaluated at the equilibrium treatment levels, reducing to a formula that depends only on equilibrium treatment level, $x_i^*(\phi, t)$, and contract's effect on treatment $\frac{dx_i^*}{d\phi}$. The separability assumption on h has the advantage that it simplifies our semi-elasticity. At the optimal contract, the derivative of the social welfare function with respect to ϕ will be equal to zero. These two assumptions, combined, do *not* give us an explicit formula for the optimal contract, though they have the advantage of simplifying the linear contract's optimality condition.

Exploiting these two assumptions allow us to study the tendencies of the optimal contract, based on the value of just a few empirical objects: the range σ and the outlier ratio ω. Further, the assumptions make σ and ω empirical objects that are easily available in *observational data*, as they depend on the dispersion of the treatment quantities observed in equilibrium. Hence, I would argue that imposing a little bit more structure on the problem can take us a long way. Nonetheless, we must be mindful about model mispecification, as there are settings where differences in health benefit from treatment are really *not* proportional across types.

If the researcher has a way to estimate the semi-elasticity of treatment with respect to reimbursement, such as an exogenous reimbursement change in the data, then one could drop the assumption on the health function. The comparative statics with respect to μ, ω, and σ will be the same, but the data that goes into computing the last two will differ. One could considering validating the multiplicative health assumption by looking at the correlation between the semi-elasticity and the average incremental treatment across types.

5 Empirical Application

I compute the normalized range, σ, and the outlier ratio, ω, for pulmonary rehabilitation therapy. Pulmonary rehabilitation is a preventative treatment for patients with Chronic Obstructive Pulmonary Disease (COPD). COPD is a condition of the lungs that impairs a person's breathing, making every day life activities more difficult. COPD is a progressive, chronic condition that affects about 12 percent of Medicare beneficiaries. There are two main treatments available to manage the symptoms of COPD. If the patient sees a pulmonologist, he may be prescribed a broncho-dialator medication or referred to pulmonary rehabilitation therapy. The former is a prescription inhaler that dilates the blood vessels to facilitate breathing, and it is typically given to patients with more severe COPD. The latter is more 'preventative', and it involves going to a respiratory therapist, who is a certified technician licensed only for this type of care.

In pulmonary rehabilitation, the patient performs a number of exercises to retrain their breathing, and also gets educated on lifestyle changes that mitigate the symptoms of COPD. Patients typically receive pulmonary rehab in an outpatient setting, though some patients may receive this care in their own home, as some home health agencies have respiratory therapists. I look at pulmonary rehab that takes place in the outpatient setting.

COPD patients who do not manage their condition well may present complications or have flare up that make them end up in the hospital for cardiac failures. Medicare only began covering pulmonary rehabilitation therapy in 2010 with the Medicare Improvements for Patients and Providers Act (MIPPA). Respiratory therapists argue that pulmonary rehab reduces the likelihood of a hospital re-admission down the line, an argument which become primary motive for expanding pulmonary care coverage in MIPPA. The coverage expansion resulted in the introduction of a new HCPCS code for pulmonary rehab, G0424, in addition to the three prior existing codes: G0237-9.

5.1 Data Description

The data being used is the 100 percent Medicare Outpatient File from 2013 to 2016. I construct the sample by first extracting all the outpatient claims for patients who have ever been diagnosed with COPD, and received pulmonary rehab. I define a patient to have received pulmonary rehab is he has at least a claim line with any of HCPCS codes G0424, G0237, G0238, or G0239. The table below shows summary statistics of my analysis sample.

Summary Statistics				
	Mean	*S.d.*	*Min*	*Max*
Length of Treatment (days)	110	176	1	1,008
Visits per Month	5.2	3.1	1	13
Hours per Month	5	4.3	0.25	18
Age	72.6	9	–	–
Number of Claims	625,836			
Number of Providers	38,949			
Number of Patients	153,896			
Female Share	53%			
	G0237	G0238	G0239	G0424
Number of Claims	97,675	42,538	93,516	397,107
Number of Patients	15,373	13,914	21,263	103,346
Number of Providers	9,589	4,192	6,313	18,855

What is the difference between the HCPCS codes? The G0424 is for patients with severe COPD, and they must qualify for this type of care by showing medical evidence of a limited forced expiratory volume (FEV). The G0237-9 is for patients that do not meet the diagnosis criteria for COPD, but nonetheless need to see a respiratory therapist. This may include patients with milder forms of COPD. The G0237-9 codes have existed in the Medicare billing space long before MIPPA, and were typically billed for patients with cardiopulmonary complications, though Part B coverage was limited. These codes describe 15 minute units of face-to-face time with a provider in which the provider helps the patient perform therapeutic procedures that increase strength or endurance of the respiratory muscles, improve respiratory function. The G0239 is used for group therapy sessions, only, whereas the G0237-8 are used for one-on-one sessions.

5.2 Range and Outlier Ratio

Assuming that provider altruism is equal to one, evaluating the right hand side of $(*)$ requires estimating σ and ω in the data. If the right hand side is close to zero, this is indicative that a fee-for-service system is optimal; if it is close to one, then it would be indicative that a prospective payment system is optimal.

To compute my two empirical objects, I define a unit of x_i to be an hour of pulmonary rehabilitation. Within a diagnosis (ICD-9) code, I count the frequency of patients receiving 1, 2, 3, etc.. hours of care in a month. Then, I calculate the average hours and the average incremental treatment within a diagnosis code, weighted by number of patients, as well as the maximum hours and maximum incremental treatment within that diagnosis code. Finally, I aggregate across years and diagnosis codes, weighting by the number of patients diagnosed with that code.

The challenge in computing ω and σ as defined above is that both of there depend on the incremental marginal health gain, θ_i. I make the identifying assumption that $\frac{dx_i^*}{d\theta_i}$ is an good proxy for $\theta_i \frac{dx_i^*}{d\theta_i}$. What does this mean? That all the information about the incremental health gains from treatment for an unobserved type are captured in the incremental treatment for that type.

Definition 6 *Define the approximated normalized range,* $\hat{\sigma} \equiv \sum_{i=1}^{N} \gamma_i \frac{(x_N^* - x_i^*)}{\sum_{i=1}^{N} \gamma_i [\frac{dx_i^*}{d\theta_i}]}$, *and the approximated outlier ratio,* $\hat{\omega} \equiv \frac{\frac{dx_N^*}{d\theta_N}}{\sum_{i=1}^{N} \gamma_i [\frac{dx_i^*}{d\theta_i}]}$.

The table below shows the estimated values for $\hat{\omega}$ and $\hat{\sigma}$ across the four types of pulmonary rehabilitation therapy. The results are suggestive evidence that the fee-for-service payment scheme for pulmonary rehabilitation is not to far from optimal. Since the right hand side of the optimal contract formula $(*)$ is not statistically significant from zero, this suggests that it would *not* be a welfare increasing deviation to move towards a prospective payment for this type of care.

Parameter	*Calibration*			
Altruism μ	1			
Provider SWF weight η	0			
	G0424	G0237	G0238	G0239
Approx Normalized Range $\hat{\sigma}$	126.19	170.23	188.88	78.81
	(65.24)	(133.8)	(136.78)	(42.25)
Approx Outlier Ratio $\hat{\omega}$	24.62	23.12	19.04	21.99
	(17.59)	(19.78)	15.41)	(18.54)
Optimal Payment Scheme	$1-\phi^* = \frac{1}{\frac{1}{\mu}+(1-\eta)(\sigma-\omega+1)}$			
$\hat{RHS}$	.022	.083	.068	.044
	(.072)	(.181)	(.165)	(.105)

Qualitatively, the implications are consistent with the wide treatment heterogeneity within diagnosis codes for COPD patients. For a prospective payment system to get the same set of patients seen for treatment in equilibrium, it would have to set the rate large enough to cover costs of the most expensive patient in the group.

6 Conclusion

To conclude, I review what I consider the four main insights from this paper, followed by discussion of limitations and directions for future work. First, altruistic providers want to give too much treatment when the insurer reimburses his costs. Second, as the unobserved patient heterogeneity increases, altruistic providers are tempted to treat everyone at the highest covered level of care, and contracts without treatment caps become way too expensive. The third is that moving away from systems that require reimbursement to exceed costs for every patient within a group (possibly through global budgets) solves the contracting frictions and implements first best treatments. The fourth is that having outlier patients raises costs of insurance for everyone, but it may be worthwhile to cater the insurance contract to higher treatment quantities when the health gains accrued from the outlier patients are sufficiently large.

While the type of heterogeneity in the model I presented here fits some settings very well, it is an important for future work to study other forms of unobserved heterogeneity, as these may have very different implications for the optimal contract. A main one to start with is heterogeneity in provider ability, which has been one of the top explanations for regional variations in health care utilization. The work by Atul Gawande is also indicative that there may be substantial heterogeneity in provider altruism. Putting unobserved heterogeneity in the cost function also gives rise to selective admissions distortions, which have very different welfare consequences than those discussed here.

One can hardly argue that the shape of financial incentives does not play a major role in determining provider treatment choice. Since Gaynor and Pauly (1990)'s early evidence, the empirical

literature has only reaffirmed this point. Health is seen (by many) as a human right, and the provider contracting problem merits special attention and customization to the different health care types. The model I proposed here attempts to offer a unifying framework that accommodates the complex health care setting while embedding the more salient ideas from the theoretical health economics literature. But precisely because health care is so nuanced, the framework must be adapted and extended if it is to be applied more generally. Ultimately, the goal of this research agenda is to take a step back and evaluate not just the design of insurance payment contracts, but also the overall design of a government insurance payment system, both in the United States and abroad.

References

Arrow, Kenneth J., "Uncertainty and the Welfare Economics of Medical Care," *The American Economic Review*, 1963, *53* (5), 941–973. Publisher: American Economic Association.

Berwick, D. M., "A primer on leading the improvement of systems.," *BMJ : British Medical Journal*, March 1996, *312* (7031), 619–622.

_ , "Quality of health care. Part 5: Payment by capitation and the quality of care," *The New England Journal of Medicine*, October 1996, *335* (16), 1227–1231.

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.

Chandra, Amitabh, David Cutler, and Zirui Song, "Chapter Six - Who Ordered That? The Economics of Treatment Choices in Medical Care," in Mark V. Pauly, Thomas G. Mcguire, and Pedro P. Barros, eds., *Handbook of Health Economics*, Vol. 2 of *Handbook of Health Economics*, Elsevier, January 2011, pp. 397–432.

Chon, 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.

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.

Cutler, David M., *Your Money or Your Life: Strong Medicine for America's Health Care System.*, New York: Oxford University Press, 2004.

_ **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.

Dranove, David, "Chapter Ten - Health Care Markets, Regulators, and Certifiers," in Mark V. Pauly, Thomas G. Mcguire, and Pedro P. Barros, eds., *Handbook of Health Economics*, Vol. 2 of *Handbook of Health Economics*, Elsevier, January 2011, pp. 639–690.

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

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.

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

Gaumer, G. L., E. L. Poggio, C. G. Coelen, C. S. Sennett, and R. J. Schmitz, "Effects of state prospective reimbursement programs on hospital mortality," *Medical Care*, July 1989, *27* (7), 724–736.

Gaynor, Martin and Mark V. Pauly, "Compensation and Productive Efficiency in Partnerships: Evidence from Medical Groups Practice," *Journal of Political Economy*, 1990, *98* (3), 544–573. Publisher: University of Chicago Press.

_ , 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.

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].

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

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.

Laffont, Jean-Jacques and Jean Tirole, *A theory of incentives in procurement and regulation*, Cambridge, Mass: MIT Press, 1993.

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.

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.

Maskin, Eric and John Riley, "Monopoly with Incomplete Information," *The RAND Journal of Economics*, 1984, *15* (2), 171.

McGuire, Thomas G., "Chapter Five - Demand for Health Insurance," in Mark V. Pauly, Thomas G. Mcguire, and Pedro P. Barros, eds., *Handbook of Health Economics*, Vol. 2 of *Handbook of Health Economics*, Elsevier, January 2011, pp. 317–396.

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

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.

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.

Sloan, Frank A., "Chapter 21 Not-for-profit ownership and hospital behavior," in "Handbook of Health Economics," Vol. 1, Elsevier, January 2000, pp. 1141–1174.

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.

Appendix

Non-linear Contract: N type Case

Suppose that there are N types, of equal shares; $\theta \in \{\theta_1, ..., \theta_N\}$. As in the two type case, any incentive compatible contract must pay a premium for treating the low type at the lower level. Redefine variables so that everything is in terms of profits and costs.

Let $\pi(x) = r(x) - c(x)$. Denote the incentive constraint of type j pretending to be $j+1$ by $(IC\ j \to j+1)$.

The planner's problem is

$$\begin{aligned}
&\max_{x_i} \sum_{i=1}^{N} h(x_i, \theta_i) - (1-\eta)\pi(x_i) - c(x_i) \\
\text{s.t. } &\mu h(x_1, \theta_1) + \pi(x_1) \geq \mu h(x_2, \theta_1) + \pi(x_2) && \text{(IC 1}\to\text{ 2)} \\
&\mu h(x_2, \theta_2) + \pi(x_2) \geq \mu h(x_1, \theta_2) + \pi(x_1) && \text{(IC 2} \to \text{ 1)} \\
&\mu h(x_2, \theta_2) + \pi(x_2) \geq \mu h(x_3, \theta_2) + \pi(x_3) && \text{(IC 2} \to \text{ 3)} \\
&.... \\
&\mu h(x_j, \theta_j) + \pi(x_j) \geq \mu h(x_{j+1}, \theta_j) + \pi(x_{j+1}) && \text{(IC j} \to \text{ j+1)} \\
&\mu h(x_{j+1}, \theta_{j+1}) + \pi(x_{j+1}) \geq \mu h(x_j, \theta_{j+1}) + \pi(x_j) && \text{(IC j+1} \to \text{ j)} \\
&.... \\
&\mu h(x_N, \theta_N) + \pi(x_N) \geq \mu h(x_{N-1}, \theta_N) + \pi(x_{N-1}) && \text{(IC N-1} \to \text{ N)} \\
&\pi(x_1) \geq 0 && \text{(PC 1)} \\
&.... \\
&\pi(x_N) \geq 0 && \text{(PC N)}
\end{aligned}$$

Step 1: Show that the incentive constraints jointly imply monotonicity.

If we add the two adjacent incentive constraints, we can easily see that the x_j's must be monotonically increasing in the type, meaning that $x_{j+1} \geq x_j\ \forall\ j$.

$$\begin{aligned}
&\mu h(x_j, \theta_j) + \pi(x_j) \geq \mu h(x_{j+1}, \theta_j) + \pi(x_{j+1}) && \text{(IC j} \to \text{ j+1)} \\
+&\mu h(x_{j+1}, \theta_{j+1}) + \pi(x_{j+1}) \geq \mu h(x_j, \theta_{j+1}) + \pi(x_j) && \text{(IC j+1} \to \text{ j)} \\
\implies &h(x_{j+1}, \theta_{j+1}) - h(x_j, \theta_{j+1}) \geq h(x_{j+1}, \theta_j) - h(x_j, \theta_j)
\end{aligned}$$

Otherwise, we would get a contradiction.

Step 2: Consider a modified problem.

Consider instead a modified problem with only the local upwards incentive constraints, $(IC\ j \to j+1)$, a monotonicity constraint, $x_{j+1} \geq x_j\ \forall\ j$, and (PCN). We will characterize the solution to this problem, and later show that it is also a solution to the original problem.

The modified problem is

$$\begin{aligned}
&\max_{x_i} \sum_{i=1}^{N} h(x_i,\theta_i) - c(x_i) - (1-\eta)\pi(x_i) \\
\text{s.t. } &\mu h(x_j,\theta_j) + \pi(x_j) \geq \mu h(x_{j+1},\theta_j) + \pi(x_{j+1}) \quad , \ j \in \{1,...,N-1\} \quad \text{(IC j} \to \text{j+1)} \\
&x_{j+1} \geq x_j \quad , \ j \in \{1,...,N-1\} \quad \text{(monotonicity)} \\
&\pi(x_N) \geq 0. \quad \text{(PC N)}
\end{aligned}$$

Step 3: Show that the local upward incentive constraints imply that profits must be decreasing in type.

The local upward incentive constraint for type θ_j and monotonicity in x jointly imply that profits must be decreasing in the type.

$$\begin{aligned}
&\mu h(x_j,\theta_j) + \pi(x_j) \geq \mu h(x_{j+1},\theta_j) + \pi(x_{j+1}) \quad \text{(IC j} \to \text{j+1)} \\
\implies &\pi(x_j) - \pi(x_{j+1}) \geq \underbrace{\mu h(x_{j+1},\theta_j) - \mu h(x_j,\theta_j)}_{\geq 0 \text{ by monotonicity}}
\end{aligned}$$

We can once again see that, as in the two type case, any incentive compatible contract must pay a premium for treating the low type at the lower level. Otherwise, the provider has an incentive to misreport the low types and high types.

Step 4: Show that the local upwards constraints imply a recursive relationship for profits across types.

Writing the local upwards incentive constraints for types $N-1$ and $N-2$ shows that they are both bounded below by the profits of the highest type, $\pi(x_N)$.

$$\begin{aligned}
&\pi(x_{N-1}) \geq \pi(x_N) + \mu h(x_N,\theta_{N-1}) - \mu h(x_{N-1},\theta_{N-1}) \\
&\pi(x_{N-2}) \geq \pi(x_{N-1}) + \mu h(x_{N-1},\theta_{N-2}) - \mu h(x_{N-2},\theta_{N-2}) \\
\implies &\pi(x_{N-2}) \geq \pi(x_N) + \mu h(x_N,\theta_{N-1}) - \mu h(x_N,\theta_{N-1}) + \mu h(x_{N-1},\theta_{N-2}) - \mu h(x_{N-2},\theta_{N-2}) \\
&\qquad = \pi(x_N) + \mu \sum_{i=N-2}^{N-1} h(x_{i+1},\theta_i) - h(x_i,\theta_i)
\end{aligned}$$

We can keep substituting local upward constraints to derive that, for every $j \in \{1,...N-1\}$, and show that the local upwards constraints can all be written as functions of $\pi(x_N)$ and $h(\cdot,\theta_j)$.

$$\begin{aligned}
&\pi(x_j) \geq \pi(x_N) + \mu \sum_{i=j}^{N-1} h(x_{i+1},\theta_i) - h(x_i,\theta_i) \\
&\qquad \ldots\ldots \\
&\pi(x_1) \geq \pi(x_N) + \mu \sum_{i=1}^{N-1} h(x_{i+1},\theta_i) - h(x_i,\theta_i).
\end{aligned}$$

Step 5: Show that $(PC\ N)$ is binding.

Suppose it is not binding, so that $\pi(x_N) > 0$. Then, there exists an $\epsilon > 0$ such that $\pi(x_N) - \epsilon \geq 0$. We can show that $\pi(x_N) - \epsilon$ will also satisfy all the local upward incentive constraints, since

$$\pi(x_j) \geq \pi(x_N) + \mu \sum_{i=j}^{N-1} h(x_{i+1}, \theta_i) - h(x_i, \theta_i)$$
$$\implies \pi(x_j) \geq \pi(x_N) - \epsilon + \mu \sum_{i=j}^{N-1} h(x_{i+1}, \theta_i) - h(x_i, \theta_i).$$

Further, reducing $\pi(x_N)$ by ϵ will strictly increase the objective function.

$$SWF = h(x_N, \theta_N) - c(x_N) - (1-\eta)\pi(x_N) + \sum_{i=1}^{N-1} h(x_i, \theta_i) - c(x_i) - (1-\eta)\pi(x_i)$$
$$< h(x_N, \theta_N) - (1-\eta)\pi(x_N) + \underbrace{(1-\eta)\epsilon}_{>0} - c(x_N) + \sum_{i=1}^{N-1} h(x_i, \theta_i) - \pi(x_i) - c(x_i).$$

Therefore, it is optimal to continue reducing $\pi(x_N)$ until $(PC\ N)$ is binding. QED.

Step 6: Show that the local upwards constraints are binding.

For an arbitrary j, suppose $(IC\ j \to j+1)$ is not binding. This means that,

$$\pi(x_j) > \mu \sum_{j}^{N-1} h(x_{j+1}, \theta_j) - h(x_j, \theta_j),$$

since $(PC\ N)$ is binding and $\pi(x_N) = 0$. There exists an $\epsilon > 0$ such that

$$\pi(x_j) - \epsilon \geq \mu \sum_{j}^{N-1} h(x_{j+1}, \theta_j) - h(x_j, \theta_j),$$

meaning that $(IC\ j \to j+1)$ is still satisfied, and reducing $\pi(x_j)$ by ϵ strictly increases the objective function.

$$SWF = h(x_j, \theta_j) - c(x_j) - (1-\eta)\pi(x_j) + \sum_{i \neq j}^{N} h(x_i, \theta_i) - c(x_i) - (1-\eta)\pi(x_i)$$
$$< h(x_j, \theta_j) - (1-\eta)\pi(x_j) + \underbrace{(1-\eta)\epsilon}_{>0} - c(x_j) + \sum_{i \neq j}^{N} h(x_i, \theta_i) - \pi(x_i) - c(x_i).$$

Therefore, it is optimal to continue reducing $\pi(x_N)$ until $(IC\ j \to j+1)$ is binding. QED.

Step 7: Characterize the solution.

The binding constraints fully determine the schedule of profits for types $j \in \{1, ..., N-1\}$.

$$\pi(x_j) = \mu \sum_{i=j}^{N-1} h(x_{i+1}, \theta_i) - h(x_i, \theta_i) \qquad , j \in \{1, ..., N-1\}$$
$$\pi(x_N) = 0.$$

We can plug the schedule of profits into the social welfare function to characterize equilibrium treatment levels.

$$\begin{aligned} SWF &= \sum_{j=1}^{N} h(x_j, \theta_j) - c(x_j) - (1-\eta)\pi(x_j) \\ &= \sum_{j=1}^{N} h(x_j, \theta_j) - c(x_j) - (1-\eta)\left(\mu \sum_{i=j}^{N-1} h(x_{i+1}, \theta_i) - h(x_i, \theta_i)\right) \\ &= \sum_{j=1}^{N} h(x_j, \theta_j) - c(x_j) - (1-\eta) \cdot \mu \sum_{j=1}^{N-1} j \cdot (h(x_{j+1}, \theta_j) - h(x_j, \theta_j)) \end{aligned}$$

It follows that

$$\max_{x_j} h(x_N, \theta_N) - c(x_N) + \sum_{j=1}^{N-1} h(x_j, \theta_j) - c(x_j) - (1-\eta)\mu \cdot j \cdot (h(x_{j+1}, \theta_j) - h(x_j, \theta_j))$$

$$\Longrightarrow$$

$$\frac{\partial SWF}{\partial x_N} = h_x(x_N, \theta_N) - c_x(x_N) - (1-\eta)\mu(N-1)h_x(x_N, \theta_{N-1}) = 0$$

....

$$\frac{\partial SWF}{\partial x_{j+1}} = h_x(x_{j+1}, \theta_{j+1}) - c_x(x_{j+1}) + \mu(1-\eta)[(j+1) \cdot h_x(x_{j+1}, \theta_{j+1}) - j \cdot h_x(x_{j+1}, \theta_j)] = 0$$

....

$$\frac{\partial SWF}{\partial x_1} = h_x(x_1, \theta_1) - c_x(x_1) + \mu(1-\eta)h_x(x_1, \theta_1) = 0$$

Step 8: Verifying the monotonicity condition

As before, we need to constraint the relative health gains so that the set of first order conditions above yield a monotonic sequence of x_j.

It is sufficient to require that:

$$\frac{h_x(y, \theta_{j+1}) - h_x(y, \theta_j)}{h_x(y, \theta_j) - h_x(y, \theta_{j-1})} \geq \frac{\mu(1-\eta)(j-1)}{\mu(1-\eta)(j+1)+1}.$$

Proof. Monotonicity for the high type–

$$\frac{h_x(y, \theta_N) - h_x(y, \theta_{N-1})}{h_x(y, \theta_{N-1}) - h_x(y, \theta_{N-2}))} > \mu(1-\eta)\left((N-2) + \frac{N h_x(y, \theta_{N-1})}{h_x(y, \theta_{N-1}) - h_x(y, \theta_{N-2})}\right)$$

Monotonicity for all types below—

$$\frac{h_x(y,\theta_{j+1})-h_x(y,\theta_j)}{h_x(y,\theta_j)-h_x(y,\theta_{j-1})} > \underbrace{\frac{\mu(1-\eta)(j-1)}{1+\mu(1-\eta)(j+1)}}_{\in(0,1)}$$

One can immediately see that the monotonicity condition is likely to fail for the high type and the second highest type, because the incremental health gains required for type N relative to type $N-1$ would need to be significantly larger than the incremental health gains across the other types.

Step 9: Characterizing the pooling solution when monotonicity fails

If the monotonicity condition fails, there is pooling between the type $N-1$ and N. Let T denote the top coverage level, above which everyone gets pooled.

The contract now has to be

$$\pi(x_j) = \mu \sum_{i=j}^{N-1} h(x_{i+1},\theta_i) - h(x_i,\theta_i) \qquad ,j \in \{1,...,T-1\}$$
$$\pi(x_T) = 0.$$

$$SWF = \frac{1}{N}\sum_{j=1}^{T} h(x_j,\theta_j) - c(x_j) + \frac{1}{N}\sum_{j=T}^{N}(h(x_T,\theta_j)-c(x_T)) - (1-\eta)\cdot\mu\cdot\frac{1}{N}\sum_{j=1}^{T-1} j\cdot(h(x_{j+1},\theta_j)-h(x_j,\theta_j))$$

$$\Longrightarrow$$

$$\frac{\partial SWF}{\partial x_T} = \sum_{j=T}^{N}(h_x(x_T,\theta_j) - c_x(x_T) - (1-\eta)\cdot\mu(T-1)h_x(x_T,\theta_{T-1}) = 0$$

....

$$\frac{\partial SWF}{\partial x_j} = h_x(x_j,\theta_j) - c_x(x_j) + \mu(1-\eta)[j\cdot h_x(x_j,\theta_j) - (j-1)\cdot h_x(x_j,\theta_{j-1})] = 0$$

....

$$\frac{\partial SWF}{\partial x_1} = h_x(x_1,\theta_1) - c_x(x_1) + \mu(1-\eta)h_x(x_1,\theta_1) = 0$$

The monotonicity condition for the top becomes–

$$\frac{\sum_{j=T}^{N} h_x(y,\theta_j) - h_x(y,\theta_{T-1})}{h_x(y,\theta_{T-1}) - h_x(y,\theta_{T-2})} > \mu(1-\eta)\left((T-1) + T\cdot\frac{h_x(y,\theta_{T-1})}{h_x(y,\theta_{T-1}) - h_x(y,\theta_{T-2})}\right)$$

Treatment level for the threshold type. θ_T is above his first best level:

$$h_x(x_T,\theta_T) - c_x(x_T) + \underbrace{\sum_{j=T+1}^{N} h_x(x_T,\theta_j)}_{\geq(N-1-T)h_x(x_T,\theta_{T+1})} - \underbrace{(1-\eta)\mu}_{\leq 1}(T-1)h_x(x_T,\theta_{T-1}) = 0 \qquad (\tfrac{\partial SWF}{\partial x_T})$$

$$\Longrightarrow\ h_x(x_T,\theta_T) - c_x(x_T) \leq (1-\eta)\mu\left(\underbrace{(T-1)h_x(x_T,\theta_{T-1}) - (N-1-T)h_x(x_T,\theta_{T+1})}_{\leq 0}\right)$$

$$\leq 0$$

Treatment level for the highest type. θ_N, is below first best level:

$$\sum_{j=T}^{N}(h_x(x_T,\theta_j)-c_x(x_T)-(1-\eta)\cdot\mu(T-1)h_x(x_T,\theta_{T-1})=0 \qquad (\tfrac{\partial SWF}{\partial x_T})$$

$$\implies h_x(x_T,\theta_N)-c_x(x_T)=(1-\eta)\cdot\mu(T-1)\underbrace{h_x(x_T,\theta_{T-1})}_{\geq 0}+\underbrace{\left(h_x(x_T,\theta_N)-\sum_{j=T}^{N}(h_x(x_T,\theta_j)\right)}_{\geq 0 \text{ by } h_{x\theta}>0}$$

$$\geq 0$$

■

Step 10: Checking that the solution to this problem satisfies the other constraints.

Revisting $(IC\ j+1\to j)$, and plugging in the solution of the modified problem, it follows that,

$$\mu h(x_{j+1},\theta_{j+1})+\pi(x_{j+1})\geq \mu h(x_j,\theta_{j+1})+\pi(x_j) \qquad \text{(IC j+1} \to \text{j)}$$

$$\mu h(x_{j+1},\theta_{j+1})-\mu h(x_j,\theta_{j+1})\geq \pi(x_j)-\pi(x_{j+1})$$

$$=\mu\left(\sum_{i=j}^{N-1}h(x_{i+1},\theta_i)-h(x_i,\theta_i)-\sum_{i=j+1}^{N-1}h(x_{i+1},\theta_i)-h(x_i,\theta_i)\right)$$

$$h(x_{j+1},\theta_{j+1})-h(x_j,\theta_{j+1})\geq h(x_{j+1},\theta_j)-h(x_j,\theta_j)$$

which will always hold for monotonic x_j's.

The ignored participation constraints are also trivially satisfied, since

$$\pi(x_j)=\mu\sum_{i=j}^{N-1}h(x_{i+1},\theta_i)-h(x_i,\theta_i)$$

$$\geq 0$$

for monotonic x_j's. QED.

6.1 Linear Contract: Derivative of the Social Welfare Function

Proposition 10 *The derivative of the social welfare function with respect to the optimal slope of the linear contract, ϕ, is*

$$\frac{dSWF}{d\phi} = \left(\frac{1-\phi}{\mu} - 1\right)\bar{\varepsilon}_\phi + (1-\eta)(1-\phi)(\bar{\varepsilon}_\phi + c(x_N) - \bar{c}) - (1-\eta)\varepsilon_\phi^N, \tag{3}$$

where $\bar{\varepsilon}_\phi \equiv \sum_{i=1}^{N} \gamma_i \varepsilon_\phi^i$ is the average cost elasticity, and $\bar{c} \equiv \sum_{i=1}^{N} \gamma_i c(x_\theta)$ the average cost.

Proof. We characterize the optimal ϕ via the first order condition of the SWF with respect to ϕ, over the support $\phi \leq 1$, and we've ruled out $\phi > 1$ per the lemma above. The insurer's problem can be reduced to the following,

$$SWF = \sum_{i=1}^{N} [h(x,\theta) - \underbrace{t + \phi c(x)}_{r(x)} + \eta(\underbrace{t + \phi c(x)}_{r(x)} - c(x))] f(\theta)$$

$$\text{s.t.} \qquad \mu h_x(x_\theta, \theta) - c_x(x_\theta)(1-\phi) = 0 \quad , \phi \leq 1 \qquad \text{(LIC)}$$

$$t + \phi c(x_\theta) - c(x_\theta) \geq 0 \; \forall \; x_\theta, \qquad \text{(PC)}$$

where (LIC) stands for the local incentive constraint of the physician. Since the set of incentive compatible x_θ's are monotonically increasing in θ, and costs are non-decreasing in treatment level x, it follows that the set of (PC) constraints for all types below θ_N will not be binding.

Regarding the participation constraint, note that, if costs are increasing in treatment, the (PC) for type N will imply that the (PC) holds for all types below. This is because equilibrium treatment levels are increasing in type, per the earlier proposition. Thus, the t is always determined by the binding (PC) of type θ_N.

Differentiating the SWF with respect to the slope of contract, we get

$$\frac{dSWF}{d\phi} = \sum_{i=1}^{N} \gamma_i [h_x(x,\theta)\frac{dx}{d\phi} - c_x(x)\frac{dx}{d\phi} - (1-\eta)\left[(\phi-1)c_x(x)\frac{dx}{d\phi} - c(x) - \frac{dt}{d\phi}\right]].$$

Since x_θ is increasing in θ, the only binding (PC) will be that of type N.

$$t = (1-\phi)c(x_N) \implies \qquad \frac{dt}{d\phi} = (1-\phi)c_x(x_N)\frac{dx_N}{d\phi} - c(x_N)$$

Substituting in for the physician first order condition yields the following expression.

$$\frac{dSWF}{d\phi} = \sum_{i=1}^{N} \gamma_i [(\frac{1-\phi}{\mu} - 1 + (1-\eta)(1-\phi))c_x(x)\frac{dx}{d\phi} - (1-\eta)\left[(1-\phi)c_x(x_N)\frac{dx_N}{d\phi} + c(x) - c(x_N)\right]]$$

which can be factored and rearranged to get the expression as stated above. QED. ■