This paper proposes a theoretical framework for designing insurer reimbursement contracts. There is little consensus among insurers on how to reimburse providers, suggesting that relevant contracting metrics are not well understood. My theory suggests that dispersion in treatment needs is a critical determinant for the efficiency of the payment system, explaining why prospective payment contracts may issue outlier payment adjustments. My model combines two elements that yield new theoretical results: provider altruism and unobserved patient heterogeneity. Provider altruism is a source of inefficient over-treatment, and treatment caps can efficiently contain overall health care delivery when patient benefits are unobservably disperse. Turning to linear schemes, I provide new reasons for why insurers may want partially retrospective reimbursements when there is patient heterogeneity. Finally, I propose a way to evaluate the optimality of payments using a sufficient statistics approach. In an empirical application to Medicare’s Outpatient Prospective Payment System, I find that Medicare payments for a small set of services is too prospective.

*I thank Oliver Hart, David Cutler, Stefanie Stancheva, Ed Glaeser, Jerry Green, Dan Brown, Andy Newman, Claudia Goldin, Nathan Hendren, David Sibley, Scott Kominers, Mike Powell, Leemore Dafny, Mark Shepard, Rafaela Sadun, Matt Weinzierl, Ariel Stern, Laura Keohane, Mark Duggan, Grace McCormack, Aileen Devlin, Tamar Oostrom, Adrian Fernandez, 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, and by the National Science Foundation under Grant Number DGE 1745303.
1 Introduction

The way we pay health care providers matters. There is little consensus among insurers on how to reimburse providers, and vast empirical evidence that reimbursement contracts have consequences on care delivered. Provider payment contracts seem to be all over the map: some have treatment caps that set upper limits on administered care, while others issue additional outlier payments for patients that receive unusually high levels of care. This paper proposes a novel theoretical framework that can explain when and why these contract features may arise optimally, as well as guide an insurer’s decision on how to design a provider payment contract optimally.

There are two distinguishing features in this market that one has to account for: provider altruism and unobserved patient heterogeneity. First, providers seem to respond to ethical considerations, 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. Second, accounting for heterogeneity is central for the health care setting, and an old idea tracing back to Arrow (1963): health care affects each patient's health differently, so health outcomes may look different for observably similar patients that get the same treatment.

By developing a model for optimal provider payments that accounts for these two features, I arrive a three key results. First, failing to account for altruism and reimbursing providers for their costs creates incentives for inefficient over-treatment: a provider who values patient health and does not bear the costs of treatment will do too much, and at the expense of higher insurance premiums (or taxpayers, for a public insurer). Second, covering a condition that includes some patients with unpredictably higher treatment needs creates incentives for the altruistic provider to treat every patient at the maximum covered level. Unless contained by a treatment cap, covering high need patients inefficiently raises the level of health care administered to everyone else. Third, a partially retrospective payment system, such as one that includes outlier adjustment payments, may be a cheaper way of insuring conditions in which patient treatment needs are widely dispersed.

The theory can also help shed some light for insurers and policy makers on the evaluate the efficiency of existing payment schemes. Insurer reimbursements tend to make a distinction between prospective payment and retrospective payment. The first determines the payment amount from ex-ante expected (or average) costs, while the second determines it from ex-post reported costs. While insurers have largely shifted towards prospective payment schemes in the past fifty years, I provide reasons for why a partially retrospective system may be better for welfare.

Table I shows an array of provider administered treatments covered by Medicare, all of which are paid, in principle, on a prospective payment system. Some services have outlier adjustments, which mean that Medicare issues additional payment when a provider shows that the treatment costs were unusually high for a particular patients. These outlier adjustments make the payment scheme more retrospective, in practice. Other services have treatment caps, which set an upper

\[1\] 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).
limit on the number of services that the provider can administer to a patient in a given year. The theory rationalizes why we see treatment caps on services that have small marginal value of care at high levels of care.

Table 1: Expenditures Across Provider Administered Treatments

<table>
<thead>
<tr>
<th>Health care services</th>
<th>Medicare Prospective Payment Services (based on ex-ante expected treatment costs)</th>
<th>Treatment Caps</th>
<th>Outlier Adjustment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acute Inpatient</td>
<td>$118</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Physician Services</td>
<td>$69</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Outpatient Hospital</td>
<td>$51</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Administered Drugs</td>
<td>$32</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nursing &amp; Rehabilitation</td>
<td>$28</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Hospice</td>
<td>$18</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Home Health Care</td>
<td>$18</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>Dialysis</td>
<td>$11</td>
<td>Single treatment</td>
<td></td>
</tr>
<tr>
<td>Outpatient Therapy</td>
<td>$8</td>
<td>Service</td>
<td></td>
</tr>
<tr>
<td>Inpatient Rehabilitation</td>
<td>$8</td>
<td>Discharge</td>
<td></td>
</tr>
<tr>
<td>Long-term Care</td>
<td>$5</td>
<td>Discharge</td>
<td></td>
</tr>
<tr>
<td>Inpatient Psychiatric</td>
<td>$4</td>
<td>Day of Stay</td>
<td></td>
</tr>
</tbody>
</table>


This work fits into several strands of the literature, starting with provider payment models. Ellis and McGuire (1986) were the first to evaluate the optimality of provider reimbursement contracts in a theoretical framework with partially altruistic providers, and their model still is widely used in papers on provider agency. My model embeds the Ellis and McGuire model, though is more general, differing in two ways. The first is patient heterogeneity, as I study an insurer that has one contract for a heterogeneous patient pool, while they focus on the optimal contract for a provider-patient pair. The second is the contracting objective function. In their model, there is no ‘loss’ term for provider payments, which means the insurer faces no financial trade-off from a payment scheme that induces over-treatment.

In this paper, I focus on patient heterogeneity in health benefit from treatment, which is well suited to describe physician services, nursing, rehabilitation, outpatient therapy, among others. There are a handful of other papers which have studied similar versions of the provider contracting problem with heterogeneity, but have focused on other types of heterogeneity. De Fraja (2000) characterizes 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.

To illustrate heterogeneity in benefit, consider two patients with Chronic Obstructive Pulmonary
Disease (COPD) and identical medical histories. COPD impairs breathing and gets worse over time. To mitigate symptoms of COPD, patients can receive treatment of ‘pulmonary rehabilitation,’ which is an outpatient therapy service that involves time with a respiratory therapist who guides the patient through a series of physical exercises, breathing retraining exercises, as well as nutritional counseling. Suppose one of the patients has a health conscious spouse who cooks healthy meals and walks frequently, while the other patient has a lazy spouse. They can both come in for ten visits and be better off. However, the patient with the lazy spouse may marginally benefit more from tenth visit. This is because being periodically reminded to go on walks at therapy may go a long way for the patient who does not have these reminders at home. 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.

Choné and Ma (2011) study a similar kind of heterogeneity as my paper; they characterize the mathematical properties of optimal payment schemes in a general theoretical framework with provider altruism and continuous health benefit heterogeneity. They provide examples of parametrizations that illustrate such properties which could be potentially instrumental for structural work. My paper abstracts away from such a mathematically technical approach, and focuses instead on the discrete case and two-type case to illustrate some interesting properties of optimal payment contracts. Chalkey and Kahlil (2005) study a slightly different version of the provider contracting problem, focusing on reimbursement contracts that may condition on health outcomes (in addition to treatment) in a model where patient demand also plays a role in health care delivered.

This model also relates to the government procurement problem in Laffont and Tirole (1993). Altruism is not a common feature in 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. 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. Nonetheless, reimbursement payments seem to have been designed in a similar spirit to government procurement contracts: Medicare issues provider payments with the intent of offsetting input costs. I draw upon the tools from this literature, accounting for the feature that agents in health care are altruistic, to characterize optimal provider payments for a government insurer (in addition to a private insurer), which makes my theory applicable beyond just the US health care system. Theoretically, adding altruism means that, unlike the standard non-linear pricing problem, the optimal contract distorts treatment on all types. This relates my paper to the canonical Maskin and Riley (1984) asymmetric information model and provides a new theoretical result, as I have a case in which no unobserved type receives the efficient action.

Lastly, this work draws upon the large body of empirical research showing that financial incentives have real consequences on care delivered, and even patient health outcomes. The general consensus from this empirical literature is that prospective payment 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, 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 for provider payments and discuss why cost reimbursement is not socially optimal. In Section 3, I derive the optimal non-linear reimbursement contract and discuss why treatment caps may optimally arise in a payment scheme. In Section 4, I focus on prospective payment contracts and when they may or may not be optimal for an insurer. In Section 5, I estimate the extent to which Medicare’s Outpatient Prospective System is prospective, and derive a sufficient statistics formula that is informative on whether the observed payment scheme is too prospective or too retrospective. In Section 6, I conclude.

2 A Model for Provider Reimbursement

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. While one may think that all treatments satisfy the “do no harm” clause of the Hippocratic oath, some treatments given in excess can be toxic to the patient. Assuming a health production function that is non-decreasing in treatment excludes treatments of the sort studied in Gaynor, Mehta, Richards-Shubik (2020), where too much treatment harms the patient.

A health production function that is everywhere increasing in treatment means that patients always weakly benefit from additional care. My health production is a good fit for procedures such as diagnostic services, physical therapy, provider office visits, evaluation and management services, diabetes treatment, or dialysis (the procedure itself, not the anemia medications given in parallel), for example. 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, \( \theta \), where \( h \) describes the dollar value of total health gains from treatment (e.g. value of additional quality adjusted life-years derived from pulmonary rehabilitation therapy). Treatment \( x \) may be continuous or discrete, and can encompass intensity of services (e.g. Relative Value Unit), level of treatment (e.g. number of office visits), or probability of major procedure (e.g. catheterization). Suppose \( \theta \) is private information, known only to the provider, encoding how much benefit a particular patient derives from treatment. Think of the high \( \theta \) 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 \( \theta \), meaning that \( h(x, \theta) \geq h(x, \theta') \) and that \( h_x(x, \theta) \geq h_x(x, \theta'), \forall \theta > \theta' \).

**Cost 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 \( c(x) \).

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. Shutting down private information in the cost function means that the wedge created by asymmetric information will only emerge through the provider’s valuation of patient health. On the one hand, the modeling choice results in an convenient, monotonic mapping from the unobserved patient type and the observed equilibrium treatment. 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.

---

\(^2\)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.
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 \( \mu \) be the ‘altruism’ parameter, which scales the provider’s valuation of patient health. Assume \( \mu > 0 \) with strict inequality. Provider utility from treating patient \( \theta \) is

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

The parameter \( \mu \) 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.

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.

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 \( \mu \) is isomorphic to the ‘\( \alpha \)’ 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 \( \mu \) 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.

\[3\] If one were to work out the math with \( \mu = 0 \) results, one would get 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.
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.

\[ t = 0 \quad \text{Provider observes} \quad \theta \quad \text{Accepts or rejects patient} \quad t = 2 \quad \text{provider chooses treatment} \quad x(\theta) \]

\[ \text{Accept if } r \geq c \quad \text{Altruism } \mu \cdot h \text{ kicks in} \]

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.\(^4\) 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

\(^4\)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.
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 actions affect not just his own utility, but also that of someone else. There is an extensive theoretical literature in environmental economics that studies similar problems while keeping the standard formulation of the participation constraint. Why the departure from standard? Unlike pollution externalities, where the social cost of an action is the sum of the private cost and the cost on other people, having altruistic providers does not make the value of the treatment any greater than what it already was. Similarly, the value to society of an organization’s charitable activities will not be the sum of the intrinsic valuation of all prospective owners plus the value of the activity itself. Nonetheless, one can hardly argue that altruism will not enter the owner’s decision making process at the time of choosing the charitable activity. Therefore, by excluding the altruistic component from the participation constraint, one can characterize social optima as the actions that maximize social value, subject to the agent’s participation constraint, while simultaneously studying the implementable action space under altruistically motivated agents.

**Insurer’s problem**

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 \( \eta \) as the social welfare weight on provider profits. Since providers are generally on the right tail of the income distribution, and social welfare weights are inversely proportional to income, a public insurer may place lower weight on provider profits (relative to patients).

Let the SWF be given by

\[
SWF = \frac{h}{h} - \frac{r}{r} + \eta \left( \frac{r - c}{c} \right).
\]

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 $\eta$ 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), $\eta$ 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. That is, the insurer would worry about making the provider internalize the costs of distortionary taxation, in addition to the costs of treatment, adding an additional dimension of welfare distortion to the model. 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, $\theta_H$, and a less responsive patient, $\theta_L$—and later show that the results generalize to the $N$ type case. Let there be share $\gamma$ 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.

\[
\begin{align*}
\max_{r_L, r_H} & \gamma(h(x_H, \theta_H) - r_H + \eta(r_H - c(x_H))) + (1 - \gamma)(h(x_L, \theta_L) - r_L + \eta(r_L - c(x_L))) \\
\text{s.t.} & \quad \mu h(x_L, \theta_L) + r_L - c(x_L) \geq \mu h(x_H, \theta_L) + r_H - c(x_H) \quad \text{(IC L)} \\
& \quad \mu h(x_H, \theta_H) + r_H - c(x_H) \geq \mu h(x_L, \theta_H) + r_L - c(x_L) \quad \text{(IC H)} \\
& \quad r_L - c(x_L) \geq 0 \quad \text{(PC L)} \\
& \quad r_H - c(x_H) \geq 0. \quad \text{(PC H)}
\end{align*}
\]

(1)

2.1 First Best: Efficient Levels of Care

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

![Figure 1: First Best Treatment Levels](image)

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 \( x \),

\[
x^{FB}_H > x^{FB}_L \implies \mu h(x_L, \theta_L) + r_L - c(x_L) \geq \mu h(x_H, \theta_L) + r_H - c(x_H)
\]

(1)

This is not sustainable at the first best because the government would incentivize levels of care that provide a lower marginal health benefit than the first best levels, resulting in less optimal outcomes.
so \((IC\ L)\) 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.

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

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 \(h(x, \theta)\) is increasing in \(\theta\) by assumption. Otherwise, we would get a contradiction.

\[
\begin{align*}
\mu h(x_L, \theta_L) + r_L - c(x_L) &\geq \mu h(x_H, \theta_L) + r_H - c(x_H) \\
\mu h(x_H, \theta_H) + r_H - c(x_H) &\geq \mu h(x_L, \theta_H) + r_L - c(x_L)
\end{align*}
\]

\[\implies h(x_H, \theta_H) - h(x_L, \theta_H) \geq h(x_H, \theta_L) - h(x_L, \theta_L)\]

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,

\[
\mu h(x_H, \theta_H) - \mu h(x_L, \theta_H) \geq \mu h(x_H, \theta_L) - \mu h(x_L, \theta_L) \\
\implies \mu h(x_H, \theta_H) + r_H - c(x_H) \geq \mu h(x_L, \theta_H) + r_L - c(x_L) = 0
\]

\[= \mu h(x_H, \theta_L) - \mu h(x_L, \theta_L) \]

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.

\[
\frac{\partial SWF}{\partial x_H} = 0 \implies h_x(x_H, \theta_H) - c_x(x_H) = (1-\eta) \frac{1-\gamma}{\gamma} \mu h_x(x_H, \theta_L) \quad (1.1)
\]
\[
\frac{\partial SWF}{\partial x_L} = 0 \implies h_x(x_L, \theta_L) - c_x(x_L) = -(1-\eta)\mu h_x(x_L, \theta_L) \quad (1.2)
\]

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.

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 \( \mu \) is close to zero gets us closer to 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 \( \mu \) is near zero, the provider does not care: he is indifferent between giving low types a lower or higher treatment level.

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

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

---

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 \( \theta \) and implement first best treatment levels for all patient types using pure cost reimbursement, which is how we get first best outcomes.
additional treatment, it becomes optimal to implement a pooling solution.

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

\[ h_x(x, \theta_H) > (1 + (1 - \eta) \frac{\mu}{\gamma}) h_x(x, \theta_L) \]  

(2)

**Proof.** Suppose that (2) holds. Since the cost function across types is the same, we can order the solutions to the first order conditions (1.1) and (1.2), relative to each other. Rearranging (1.1) and (1.2) to that they both equal \( c_x(x) \), it follows that the equilibrium treatment level of \( x_H \) will be larger than the equilibrium treatment level for \( x_L \) if

\[ h_x(x_H, \theta_H) - (1 - \eta) \frac{\mu}{\gamma} h_x(x_H, \theta_L) > h_x(x_L, \theta_L) + (1 - \eta) \mu h_x(x_L, \theta_L) \]

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

The pooling solution implements a treatment level at which average marginal health benefits equate to marginal cost. Pooling is more likely when the share of high types, \( \gamma \), is small, as condition (2) becomes harder to satisfy. Intuitively, the result makes sense: if there are not many patients that need high levels of treatment, then the insurer will cater the contract to the patients who need less. Notice that at lower levels of \( \eta \), 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 \( \mu \) is large, it is more likely we are in the pooling solution. For high \( \mu \), 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. This is because adding types to the right also means that the lowest type always benefits, health wise, from receiving the treatment level of the highest type.

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 $h_+(x, \theta_i)$ are both increasing in $\theta$. Since treatment is monotonic in the type, a treatment cap $x_T$ is equivalent to choosing a threshold patient type, $\theta_T$, above which patients get pooled at the same treatment level.

**Definition 1** Let $\theta_T$ 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 $x_T^*(r, \theta_i)$, 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 $\theta_T$), and a non-linear fee schedule for all coverage levels below. This is 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 $\theta$ 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** Define profits $\pi(x_j) \equiv r(x_j) - c(x_j)$. For all types $\theta_j < \theta_T$, the optimal contract gives an incentive rent on type $j$ such that

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

For all types above $\theta_T$, 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 $\pi(x_j) \geq \pi(x_T) + \mu \sum_{i=j}^{T-1} h(x_{i+1}, \theta_i) - h(x_i, \theta_i)$.  

17
Reducing $\pi(x_T)$ by $\epsilon$, 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 $\epsilon$ will strictly increase the $SWF$ by $\epsilon$. 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 $\epsilon$ strictly increases the $SWF$. Therefore, it is optimal to continue reducing $\pi(x_N)$ until $(IC_j \rightarrow j+1)$ is binding.

Proposition 3 highlights that the threshold patient, $\theta_T$, 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.

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

**Proof.** One can plug in the contract derived in Proposition 3 into the $SWF$ and characterize
the solution for each \( x_i \).

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

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

\[
\frac{\partial SWF}{\partial x_1} = h_x(x_1, \theta_1) - c_x(x_1) + \mu(1 - \eta) h_x(x_1, \theta_1) = 0 \quad (d.1)
\]

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 \( (d.T) \), \( (d.T - 1) \), \( ..., (d.1) \) 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 \( (d.T) \) and \( (d.T - 1) \).

\[
\sum_{j=T}^{N} (h_x(x_T, \theta_j) - (1 - \eta) \cdot \mu(T - 1) h_x(x_T, \theta_{T-1}) > h_x(x_{T-1}, \theta_{T-1}) + \mu(1 - \eta)[j \cdot h_x(x_{T-1}, \theta_{T-1}) - (T - 2) \cdot h_x(x_{T-1}, \theta_{T-2})]
\]

By rearranging the expression, one immediately arrives at condition \( (4) \). □

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 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. This is 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,

\[
\mu_h(x_j, \theta_j) + \pi(x_j) \geq \mu_h(x_{j+1}, \theta_j) + \pi(x_{j+1}) \implies \pi(x_j) - \pi(x_{j+1}) \geq \mu_h(x_{j+1}, \theta_j) - \mu_h(x_j, \theta_j).
\]

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

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\[^6\] The first could correspond to a global budget system as in the United Kingdom. 

[^6]: Thank you to David Cutler and Dan Barron for insightful discussions that helped develop and interpret this specification.
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,

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

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.

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

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:

\[
\gamma (r_H - c(x_H)) + (1 - \gamma) (r_L - c(x_L)) = 0
\]

\[
\implies (x_L^*, x_H^*) \in \arg \max_{x_L, x_H} \gamma (h(x_H, \theta_H) - c(x_H)) + (1 - \gamma) (h(x_L, \theta_L) - c(x_L)).
\]

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

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

\[
x_i \in \arg \max_x \mu h(x, \theta_i) + \frac{r_i - c(x)}{t + (1 - \mu) c(x_i) - c(x_i)} \implies \mu (h(x_i, \theta_i) - c(x_i)) = 0,
\]

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 \(\mu\) rises to the surface. In practice, it may be hard for the insurer to know \(\mu\). It may well be the case that providers have heterogeneous \(\mu\)’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 \(\mu\) 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 if 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 ‘Prospective Payment’ with Retrospective Adjustments

The remaining sections focus on the empirical implications of the theory, with the purpose of evaluating existing reimbursement schemes. I study the financial incentives of these schemes via studying the optimal linear contract, and characterize the settings where making payments partially retrospective may be socially desirable.
As illustrated earlier on Table 1, provider payment contracts take different shapes across different types of services. As a basis of payment, insurers use either reported (retrospective) treatment costs or expected (prospective) treatment costs. While Medicare and most insurers have shifted away from retrospective payment contracts, features such as outlier adjustments or ad hoc claim appeals make insurer payments, in practice, more retrospective. I mimic the financial effects of these adjustments by studying a linear contract with two components: a flat (prospective) payment plus a (retrospective) share of reported treatment costs.

The linear contract has the advantage of reducing the complexity of the various payment forms into two parameters. I consider reimbursement contracts of the form,

\[ r(x) = t + \phi c(x). \]

Since the linear contract is intended to replicate potentially non-linear contracts (e.g. with outlier adjustments), I interpret \( \phi \) as the coefficient estimate from a regression of reimbursement on treatment costs. A purely retrospective payment system would pay the provider an amount exactly equal to the costs (corresponding to a \( \phi = 1 \) and a \( t = 0 \)), while a purely prospective payment will disregard reported costs (\( \phi = 0 \) and a \( t > 0 \)).

4.1 Institutional Background: the shift to prospective payment

When Medicare was first established in 1965, reimbursement contracts were retrospective. That is, Medicare would pay the provider the lesser amount of hospital reasonable cost or customary charges for the service furnished. Over the first decade of the program, health care utilization and costs rose dramatically, as cost-based reimbursement rewarded the health care providers who incurred the highest health care costs. This lead policy makers to reform the payment system into a prospective payment system, with the goal of creating payment incentives for efficient care delivery. Payment reform occurred first for hospital inpatient services, followed shortly after by hospital outpatient services.

In the Omnibus Budget Reconciliation Acts of 1986 (OBRA), the Secretary of Health and Human Services began to develop the model prospective payment system for outpatient hospital services, which began a series of reviews, committees, and proposed rules which ultimately concluded April, 2000. The final rule was implemented in August, 2000. This new payment system is the Outpatient Prospective Payment System (OPPS), which is still in use today. There are various different kinds of payment methodologies used within OPPS, which vary across type of service.

---

\(^7\)In the medical terminology, reported treatment costs are referred to as charges.

\(^8\)The baseline methodology for setting medical and surgical payment rates assigns services to an Ambulatory Procedure Code (APC) group and multiplies the APC rate by a relative weight and a conversion factor. Then there are a number of potential payment adjustments, based on whether the service qualifies for additional pass-through payments (new technologies) and outlier payments, or whether the service was rendered in a rural, cancer, or children’s hospital. There is a different payment methodology for ‘separately payable’ services, which include drugs and bio-
Sometimes a group of services are bundled together for payment. This means that if two bundled services get billed on the same day, OPPS only pays for one of those services.

Though OPPS is ‘prospective’, in principle, there are a number of possible adjustments allowed for in the legislation (e.g. outlier payments and standard facility adjustments) that make payments issued in practice different from the prospective ‘statutory rate’. The unit at which the prospective payment rate is set also affects provider incentives, and changes the extent to which a payment is actually retrospective. For instance, if one were to think of a prospective payment rate for a lung cancer chemotherapy session, a separate payment for the chemotherapy drug injection may induce the provider to choose the most expensive drug available (or at the very least make him indifferent between a high cost drug and a low cost drug).

**Unit of Payment**

There are three units of payment typically used by insurers: individual procedure or service (e.g. fee schedule), diagnosis (DRG), or individual patient (capitated rate). In the discussion of the model so far, the implicit unit of payment has been an individual service, where the \( x \) corresponded to the quantity of that service, and the treatment decisions were made by the provider of that service. However, one could interpret the provider choice variable \( x \) to be the intensity of health care services used by a patient. In this case, the decision maker could be the primary care physician, or the specialist in charge of determining course of treatment (which may be rendered by other providers in the team). One could also interpret \( x \) to be the overall volume of health care services used by a patient in any given year. In this case, the decision maker may be a group of health care providers, such as an Accountable Care Organization (ACO).

<table>
<thead>
<tr>
<th>Unit of Payment</th>
<th>Interpretation of ( x )</th>
<th>Decision maker</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatment/Service</td>
<td>Quantity</td>
<td>Service Provider</td>
</tr>
<tr>
<td>Diagnosis</td>
<td>Intensity of Care</td>
<td>Head Physician</td>
</tr>
<tr>
<td>Patient</td>
<td>Overall Care Utilization</td>
<td>ACO</td>
</tr>
</tbody>
</table>

While the model was designed to study settings of care in which the payment unit is a particular treatment, one could easily map the model to different units of payment by giving the appropriate interpretation to the entity who is making health care decisions, and what the choice variable \( x \) captures. Since the subsequent empirical section applies the theory to the outpatient hospital setting, I will continue to think of the unit of payment as an individual service, as this is the payment unit used for this class of services.²

---

²Logicals, brachytherapy sources, or therapeutic radiopharmaceuticals. Drugs and biologicals payment rates are based on their national average sales price, which is collected from pharmaceutical companies directly. Brachytherapy and radiopharmaceutical payment is based on cost-adjusted charges.

³In order to apply the theory in other health care settings, it is important to think about the correct measure for the provider choice variable, and how this may vary across unobservably heterogeneous patient types. The model I have laid out takes the payment unit as given, and does not provide formal guidance about the optimal level of aggregation for payments. A formal study of the optimal unit of payment would require multiple dimensions of heterogeneity, which I leave for future work.
4.2 Characterizing the optimal linear contract

As in the original set up of the model, I focus on contracts that require reimbursement to be greater than or equal to costs for every patient. This means that a purely prospective payment contract will need to issue a payment that covers the costs of the patient with the highest level or intensity of care. I also assume that the treatment choice set for the provider is unrestricted. That is, the provider can file a claim for any possible \( x \) administered.

The contracting tensions for the insurer in the linear case involve trading off the retrospective payment incentives for over-treatment against the magnitude of the prospective incentive rent. The retrospective component, \( \phi \), affects the provider’s marginal incentives by subsidizing a share of treatment costs. The prospective amount, \( t \) determines the highest treatment coverage level, e.g. the treatment cap. Since the insurer has to pay the prospective amount on every patient, it can get very expensive to set a \( t \) that is sufficiently large to cover the treatment costs of high patient types. The retrospective component can help offset the magnitude of the prospective payment, but it also drives up equilibrium treatment levels up for everyone.

Optimality Condition for the Linear Contract

Suppose there are \( N \) types, and the insurer reimbursement contract is linear in \( c(x) \). The insurer’s objective is

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

where \( x^*_i \) is optimally chosen by the provider.

The prospective payment amount, \( t \), will be pinned down by the binding participation constraint of the highest type covered. Just as in the non-linear case, there may be an optimal treatment cap, meaning that the highest type covered may be \( \theta_T < \theta_N \). For the purposes of the subsequent discussion, however, I will think of types at the top as having a large incremental value of health gains so that the insurer always ‘wants’ to cover type \( N \).

Lemma 4 For any \( \phi \leq 1 \), the prospective payment component of the contract, \( t \), is always pinned down by the treatment cost of the highest type covered, \( \theta_T \).

\[
t = (1 - \phi)c(x_T)
\]

Proof. Let \( x_T \) be the maximum treatment level the insurer decides to cover. The participation constraint requires that \( t + \phi c(x_i) - c(x_i) \geq 0 \), \( \forall \ i \in \{1,...,N\} \). By monotonicity in \( x_i \), it follows that \( c(x_T) \geq c(x_i) \) for all patient types \( \theta_i \) in \( i \in \{1,...,T\} \).

\[
t \geq (1 - \phi)c(x_T) \implies t \geq (1 - \phi)c(x_i), \forall \ i \in \{1,...,T\}.
\]

For patient types \( i \in \{T,...,N\} \), the physician FOC will be slack, so \( x^*_i = x_T \). Since the SWF is
decreasing in \( t \), \( t \geq (1 - \phi)c(x_T) \) will bind at the optimum.

The derivative of the social welfare function captures the main trade-offs that arise from making a payment more retrospective, while highlighting the additional insights from embedding the provider contracting problem into a social welfare function. The main insight from it is that having payments be slightly retrospective need not always be bad. The welfare effects of a partially retrospective payment will depend on dispersion of treatment costs across types and on the provider’s behavioral response to the payment contract.

**Definition 2** Define the semi-elasticity of treatment with respect to reimbursement rate for a patient type \( \theta_i \) as

\[
\varepsilon^i_\phi \equiv (1 - \phi) \frac{dx^*_i}{d\phi}.
\]

The semi-elasticity, \( \varepsilon^i_\phi \), describes how much a provider changes the equilibrium administered treatment with a proportional reduction in treatment cost-sharing. In other words, \( \varepsilon^i_\phi \) describes the level changes in treatment with proportional changes in provider cost sharing.

**Proposition 6** The derivative of the social welfare function with respect to the optimal retrospective component, \( \phi \), is

\[
\frac{dSWF}{d\phi} = \left( \frac{1}{\mu} - \frac{1}{1 - \phi} \right) \bar{\varepsilon} + (1 - \eta)(c(x_N) - \bar{c}) - (1 - \eta)(c(x_N)\varepsilon^N_\phi - \bar{\varepsilon}),
\]

where \( \bar{\varepsilon} \equiv \sum_{i=1}^{N} \gamma_i c(x_i) \varepsilon^i_\phi \) 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.

The first term resonates with the ideas in the literature about supply-side moral hazard: making a payment more retrospective distorts treatment decisions to inefficient levels. The second and third terms are new, as they capture the financial trade-offs of retrospective payment.

The second term captures the benefits of making payments more retrospective. Since a prospective payment system pays for every patient as if they were the most expensive patient, allowing contracts to be slightly more retrospective generates savings on the prospective payment amount, while still ensuring that costs of every patient in the group are covered. This deviation, however, can only improve welfare if supply-side moral hazard is small. That is, making a payment slightly more retrospective can only be desirable by an insurer if the provider supply response to it is small. The argument is formalized in the following section.

Nonetheless, it will never be optimal for the insurer in my model to opt for a purely retrospective system. One can rule out that \( \phi > 1 \) almost immediately. The intuition is that a provider who

\[\text{Note that the semi-elasticity depends on } \phi \text{ via } \frac{dx^*_i}{d\phi}, \text{ which may be some complicated function of } h \text{ and } c.\]
already wants to over-treat does not need an additional profit motive to provide more treatment. Since the provider values both profits and patient health, reimbursing above 100% of costs makes the provider want to give infinite treatment in the model.\footnote{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.}

**Lemma 5** The optimal $\phi$ cannot be greater than 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 $\theta_i$, and $c(x)$ is increasing and convex, the only way for $\pi(x) = t + \phi c(x) - c(x)$ to be decreasing in $\theta_i$ is if $\phi \leq 1$.

Lemma 5 does not rule out at-cost reimbursement ($\phi = 1$), which requires a slightly more involved argument. A cost-based retrospective payment system is never optimal because the costs of supply-side moral hazard are unambiguously too large under this contract. The insurer can always increase social welfare by exposing the provider to some costs on the margin to contain healthcare utilization. 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.

**Proposition 7** A retrospective payment system with reimbursement at cost is never optimal, for any level of altruism $\mu$.

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

$$\left. \frac{dSWF}{d\phi} \right|_{\phi=1} = -\bar{\epsilon}_\phi - (1 - \eta)c_x(x_N)\bar{\epsilon}_N.$$

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

### 4.3 On the Optimality of Prospective Payment

Consider a world in which providers are perfectly altruistic, meaning $\mu = 1$. At a first glance, one might be inclined to say that a prospective payment contract (with $\phi = 0$) would be ‘optimal’ for this provider: he will choose treatment such that marginal health benefit equals marginal cost. This would certainly be true if there was only one patient type, which is what Ellis and McGuire (1986) find and label as ‘the promise of prospective payment’. But in my more general model, this intuition is no longer optimal.
I find that prospective payment becomes too expensive when there is value to covering higher treatment quantities for high benefit patient types. This is 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 highest patient type covered. If covered treatment needs are disperse, a prospective payment contract gives large profit margins on low type patients.

The Optimal Prospective Payment Amount is Not Average Cost

The prospective payment contracts seen in practice reimburse the provider at average treatment cost. Typically, Medicare calculates this amount by looking at claims data on charges from previous years, applying ‘cost-to-charge’ ratios, and aggregating across providers nationally. However, if providers indeed accept or reject patients based on whether reimbursement is greater than or equal to treatment cost, as I have modeled here, it is not obvious why the prospective payment should be set at average cost. Moreover, the dynamic effects of the reimbursement determination formula that uses lagged charges data may not be desirable to an insurer, particularly when there are patients that derive large health benefits from additional treatment.

There is an optimal treatment cap which depends on the marginal value of health gains of patient types at the top, and the magnitude of the prospective payment determines the treatment cap in the linear contract. Suppose the insurer is constrained to a prospective payment contract, but optimally chooses the highest covered type, $\theta_T$. The insurer’s problem becomes,

$$x_T^* \in \arg \max_{x_T} \left( \sum_{i=1}^{T-1} \gamma_i (h(x_i^*, \theta_i) + \eta (c(x_T) - c(x_i^*))) + \sum_{i=T}^{N} \gamma_i h(x_T, \theta_i) - c(x_T) \right).$$

The optimal prospective payment amount trades off the health gains of ‘capped’ types (i.e. patient types that receive the maximum level of covered treatment because their marginal health gains exceed marginal costs of treatment) against a discounted marginal cost of treatment. The marginal costs of treatment for capped types is discounted when $\eta > 0$, meaning that the insurer values giving profits to the provider. This is because the insurer positively internalizes part of the provider margins given on the non-capped types. Formally, the optimality condition for the optimal treatment cap is

$$\frac{dSWF}{dx_T} = \sum_{i=T}^{N} \gamma_i h_x(x_T, \theta_i) - \left( 1 - \eta \sum_{i=1}^{T-1} \gamma_i \right) c_x(x_T).$$

The solution to this is unlikely to coincide with the equilibrium average cost. In fact, if the prospective payment amount is determined from the average cost resulting from equilibrium treatment decisions at the optimal treatment cap, the new prospective payment amount will be too low, and no longer optimal. This is because the average of $c(x_T)$ and treatment costs for types below will result
in an average cost $\bar{c} \leq c(x_T)$. Consequentially, the treatment cap will have been ‘reduced’ (since costs are monotonically increasing and weakly convex).

Cost Spreads within Payment Groups

An important insight from my model is highlighting the role of treatment dispersion within a payment group, in addition to the provider agency problem, when designing an optimal reimbursement scheme. If there is heterogeneity in health benefits that results in disperse first-best treatment needs, and the insurer has interest in a contract that gets all of his patients at least seen, then the prospective payment contract may be too expensive.

In the same way that the most expensive patient determined the magnitude of incentive rents in the non-linear case, the most expensive patient determines the size of the prospective payment. A prospective payment contract, in effect, gives a strong profit motive for the provider to treat low-types at lower levels: providers keep a margin equal to the difference between the highest cost type and whatever type they are treating.

**Proposition 8** \(\mu = 1\). Suppose that

1. the semi-elasticity of treatment with respect to reimbursement is small, so that \(\varepsilon^i_\phi \approx 0\); and that
2. \(\varepsilon^i_\phi\) is constant across types, so that \(\varepsilon^i_\phi = \varepsilon_\phi\) for all \(i \in \{1, \ldots, N\}\).

Then, prospective payment will not be optimal.

**Proof.** Suppose, for the sake of contradiction, that \(\phi = 0\) is optimal. Denote average marginal cost by \(\bar{c}_x \equiv \sum_{i=1}^{N} \gamma_i c(x^*_i)\). Having \(\varepsilon^i_\phi\) constant across types means that \(c_x(x_N)\varepsilon^N_\phi - \varepsilon_\phi = (c_x(x_N) - \bar{c}_x)\varepsilon_\phi\). Since costs are weakly convex, the first factor is always positive. If \(\varepsilon_\phi\) is close to zero, then the derivative of the SWF from Proposition 6 at \(\phi = 0\) and \(\mu = 1\) is positive.

\[
\frac{dSWF}{d\phi} \bigg|_{\phi=0} = (1 - \eta)(c(x_N) - \bar{c}) - (c_x(x_N) - \bar{c}_x)\varepsilon_\phi \geq 0
\]

That means that raising \(\phi\) by a small amount is a welfare increasing deviation, so \(\phi = 0\) could not have been optimal.

Proposition 8 provides a new motive for a partially retrospective payment scheme: cost spreads. When supply-side moral hazard is small, opting for a prospective payment contract may be more expensive that adopting a partially retrospective system. The reason is patient heterogeneity. The empirical implication of this proposition is that we may expect a positive relationship between the cost spreads and how retrospective a payment is, particularly in services where we expect little supply-side moral hazard.
5  Empirical Implications for Medicare OPPS

In this section, I use the model to evaluate the optimality of the Medicare Outpatient Prospective Payment System. I propose a way to measure the extent to which a payment is retrospective, and develop a way to evaluate the optimality of payments. The latter involves a sufficient statistics approach, in which I make additional assumptions that lead to a formula which depends on measurable empirical objects available in claims data.

5.1  Background and Data Description

Medicare Part B covers a wide range of hospital outpatient services, paid for under OPPS. These services include evaluation visits, laboratory tests, ambulatory surgical procedures, mental health care, preventative and screening services, provider-administered drugs and biologicals, imaging and radiology, outpatient therapy, and emergency services.

The data contains 100% Medicare outpatient claim lines from 2016. For each claim line, the unit of payment is an individual service or procedure, identified by a Healthcare Common Procedure Coding System (HCPCS) code\textsuperscript{12} This means that I have transaction-level data for every service paid under OPPS. In the construction of my sample, I pull all claim lines for services covered under OPPS and have non-missing the HCPCS code\textsuperscript{13} The table below shows some descriptive statistics of the data set used in the analysis.

5.2  Estimating ‘Retrospectiveness’ in OPPS

Conceptually, the retrospectiveness of a reimbursement contract is the slope of the relationship between insurer payments and provider costs of treatment. The main challenge in estimating retrospectiveness is that actual costs are not observed, but only charges (which are akin to reported costs). Provider charges are unlikely to describe actual treatment costs, and are often largely inflated. One can think of charges as the ‘list prices’ of this health care setting which are never transacted on, at face value.

Insurers do take charges into account, however. Charges are often deflated using a provider specific cost-to-charge ratio (CCR) when calculating reimbursement for some services. For instance, Medicare issues outlier payments when cost-adjusted charges exceed some threshold amount\textsuperscript{14} A cost-to-charge ratio describes the ratio of operating costs to service charges, which can be thought of as the proportional markup. Due to limited information about provider input costs, Medicare and private insurers calculate CCRs at the provider level, and use them to deflate charges as a way to infer cost by implicitly assuming a constant markup across services, within providers.

\textsuperscript{12} 42 CFR §419.2 (a)
\textsuperscript{13} The set of services paid for under OPPS are updated yearly. According to the 2016 OPPS Proposed Rule, Addendum D1, the set of services paid under OPPS in 2016 included all claim lines with revenue status indicators G, J1, J2, K, N, P, Q1, Q2, Q3, R, S, T, and V.
\textsuperscript{14} 42 CFR §419.43 (d)
<table>
<thead>
<tr>
<th>Service Category</th>
<th>Number of Patients Using Service (M)</th>
<th>Share of Medicare OPPS Spending</th>
<th>Medicare Spending ($ M)</th>
<th>Number of HCPCS codes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Surgical Procedures</td>
<td>11.42</td>
<td>0.332</td>
<td>$ 14,990</td>
<td>3,601</td>
</tr>
<tr>
<td>Administered Drugs</td>
<td>9.65</td>
<td>0.214</td>
<td>$ 9,678</td>
<td>565</td>
</tr>
<tr>
<td>Evaluation &amp; Management</td>
<td>14.05</td>
<td>0.176</td>
<td>$ 7,931</td>
<td>72</td>
</tr>
<tr>
<td>Imaging &amp; Radiology</td>
<td>14.28</td>
<td>0.128</td>
<td>$ 5,759</td>
<td>621</td>
</tr>
<tr>
<td>Cardiovascular Procedures</td>
<td>8.79</td>
<td>0.065</td>
<td>$ 2,944</td>
<td>181</td>
</tr>
<tr>
<td>Medical and Surgical Supplies</td>
<td>1</td>
<td>0.029</td>
<td>$ 1,288</td>
<td>342</td>
</tr>
<tr>
<td>Psychiatry and Psychology</td>
<td>0.35</td>
<td>0.009</td>
<td>$ 399</td>
<td>35</td>
</tr>
<tr>
<td>Neurological</td>
<td>0.61</td>
<td>0.007</td>
<td>$ 298</td>
<td>104</td>
</tr>
<tr>
<td>Vaccine</td>
<td>3.13</td>
<td>0.006</td>
<td>$ 260</td>
<td>72</td>
</tr>
<tr>
<td>Pulmonary Services</td>
<td>2.04</td>
<td>0.005</td>
<td>$ 248</td>
<td>50</td>
</tr>
<tr>
<td>Dialysis Services</td>
<td>0.09</td>
<td>0.005</td>
<td>$ 218</td>
<td>5</td>
</tr>
<tr>
<td>Blood Services</td>
<td>0.3</td>
<td>0.005</td>
<td>$ 209</td>
<td>45</td>
</tr>
<tr>
<td>Preventative &amp; Screening</td>
<td>0.84</td>
<td>0.004</td>
<td>$ 187</td>
<td>48</td>
</tr>
<tr>
<td>Medical Devices</td>
<td>1.91</td>
<td>0.003</td>
<td>$144</td>
<td>120</td>
</tr>
<tr>
<td>Outpatient Therapy</td>
<td>0.66</td>
<td>0.003</td>
<td>$ 136</td>
<td>58</td>
</tr>
<tr>
<td>Admin, Misc, &amp; Investigational</td>
<td>2.32</td>
<td>0.002</td>
<td>107</td>
<td>73</td>
</tr>
<tr>
<td>Pathology and Laboratory</td>
<td>12.38</td>
<td>0.002</td>
<td>$ 106</td>
<td>1,043</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>84.63</strong></td>
<td><strong>0.994</strong></td>
<td><strong>$ 45,079</strong></td>
<td><strong>7,201</strong></td>
</tr>
</tbody>
</table>

*Note:* The service categorization was done according to the classification of HCPCS code ranges by the American Academy of Professional Coders, plus manual classification of temporary codes used in the data period. Rows do not add to total because total includes Emergency, Ophthalmology, Gastroenterology, and level III codes, in addition to the service categories shown.

Medicare reimbursement for a service HCPCS may vary with the kind of services billed in the same day, the patient medical documentation submitted with the claim, the provider involved in the rendering of the service, or case-by-case payment review. This means that the same HCPCS code can get paid differently within OPPS, even when the performed service is the same. Medicare publishes a fee-schedule with statutory reimbursement amounts for each HCPCS code from which providers learn the expected payment for a service.

**Definition 3** For each treatment, \( t \) administered in hospital \( h \), let \( R_{t,h} \) denote the per unit reimbursement amount\(^{15} \), \( Ch_{t,h} \) the reported per unit charges, and \( C_{t,h} \) the actual per unit costs of treatment. Let \( CCR_h \) be the cost-to-charge ratio used by hospital \( h \), so that \( Ch_{t,h} \cdot CCR_h = C_{t,h} \).

Define the estimated retrospectiveness parameter, \( \hat{\phi} \), according to the coefficient in the following OLS regression:

\[
R_{t,h} = \tau_t + \phi(Ch_{t,h} \cdot CCR_h) + \epsilon_{t,h}. \tag{3}
\]

I estimate the retrospectiveness across all OPPS services by estimating equation\(^{3} \) constructing a CCR from the claims data. Notice that this equation implicitly assumes a constant markup across

\(^{15}\)By reimbursement, I refer to the total amount the provider receives for the service, which includes the patient coinsurance, deductible, Medicare payment, and any payments from a secondary insurer.
provider services, which is the same assumption insurers make when issuing reimbursement, but it need not be the correct model of the world. One could consider an alternative approach involving cost apportionment, which would involve regressing provider aggregate cost on a long vector of all service quantities provided. Computationally, this alternative approach would be very taxing, as there are over 7,600 services in OPPS alone, and about 41,000 providers. I leave this alternative approach for future work, and focus here on the cost-adjustment method that is most commonly used by practitioners.

The CCR is obtained by dividing aggregate costs by aggregate charges, and aggregating across all services administered by a provider. The data does not have accounting operating costs at the provider level\textsuperscript{16} However, I can impute provider costs by using information about the service mix offered by each provider.

Since Medicare reimbursement is set based on an estimate of cost, assign a ‘cost’ to each service unit by calculating the average Medicare reimbursement for that service. Then, aggregate across all services offered by that provider, weighting by the corresponding administered service quantities. This yields a single number which I interpret as aggregate provider costs. Then, aggregate all charges from that provider, and obtain the CCR by taking the ratio of costs to charges. Cost-adjusted charges are then the product of this CCR and provider charges for each service.

Table 3 shows the coefficient from estimating $\phi$ in equation 3 under two specifications. The baseline specification in the first row estimates the coefficient of cost-adjusted charges on reimbursement, with HCPCS service fixed effects. Standard errors are clustered at the HCPCS level. Given the possibility of coding and reporting errors in the claims data, I also estimate the coefficient $\phi$ using a subsample in which I remove reimbursement outliers. To identify potential outliers, I first demean unit reimbursement and charges for each HCPCS code, and then remove values outside of the 1st and 99th percentile. The second specification uses raw charges, instead of deflated charges. Since charges are likely a noisier measure of cost, it is not surprising that the coefficient is attenuated in the second specification.

The main takeaway is that payments look pretty prospective in OPPS. This is slightly different from what McClellan (2004) found for the Inpatient Prospective Payment System, which sets prospective payment rates using patient diagnosis code as unit of payment. Perhaps the more granular unit of aggregation allows for payment to be prospective within that unit because there may be less treatment heterogeneity within more granular units. In the section that follows, I evaluate whether the prospectiveness of OPPS is optimal.

5.3 Sufficient statistics for optimal reimbursement

In this section, I propose a way to evaluate the derivative of the social welfare function for each service in OPPS. The main challenge is to find the set of assumptions that make the elasticity of

\textsuperscript{16}The accounting costs are available in the Hospital Cost Reports published by Medicare. However, most hospital outpatient departments also have an inpatient department, and report a single set of costs for both sets of services. Therefore, I do not observe accounting costs for outpatient hospital departments.
costs with respect to reimbursement measurable in the data. While finding an exogenous shock to measure this elasticity may be a difficult exercise, the idea is to measure how much ‘room’ providers have for supply-side moral hazard. My additional assumptions ultimately map the elasticity to the average incremental treatment, which I will explain in detail below.

Moving forward, I will assume that costs are linear in treatment level, so \( c(x) = cx \). Under linear costs, the elasticity corresponds one-to-one and monotonically with the treatment heterogeneity across types, \( \frac{dx^*_i}{d\theta_i} \). 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. These are the types of services that fall within OPPS. 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.

**Lemma 6**  
The derivative of equilibrium treatment \( x^*_i \) with respect to \( \phi \) is a monotonic transformation of the derivative of \( x^*_i \) with respect to the unobserved type, \( \theta_i \). 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 \( \phi \) and \( \theta_i \) results in the following two expressions.

\[
\frac{d}{d\phi} FOC = \mu h_{xx}(x^*_i, \theta_i) \frac{dx^*_i}{d\phi} + c = 0
\]

\[
\frac{d}{d\theta_i} FOC = \mu h_{xx}(x^*_i, \theta_i) \frac{dx^*_i}{d\theta_i} + \mu h_{x\theta}(x^*_i, \theta_i) = 0
\]
By rearranging, one can immediately see that \( \frac{dx^*_i}{d\phi} = \frac{dx^*_i}{d\theta_i} \cdot \frac{c}{\rho h_{x\theta}(x^*_i, \theta^*_i)} \). 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. ■

This lemma is 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 \( \theta \), 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 patient with the less health-conscious spouse 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, \( \varepsilon_{\phi}^i \) will not depend directly on \( \phi \), and will be given by

\[
\varepsilon_{\phi}^i = \theta_i \frac{dx^*_i}{d\theta_i}.
\]

**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 \( h_{x}(x, \theta_i) = (1 - \phi) c \mu_{\theta_i} \). Per the lemma from above, it follows that,

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

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

Notice that the assumptions substantially simplify the estimation objects in the optimality condition for the optimal linear contract. While finding the empirical counterpart of an elasticity may be difficult, and require exogenous reimbursement rate experiments unavailable to researchers, finding an empirical counterpart to treatment increments is a lot more achievable. That said, the implications of these assumptions are strong, and may not be broadly applicable. In effect, the assumptions say that we can infer something about the behavioral response of the provider to a change in reimbursement from the dispersion in the provider’s equilibrium treatment decisions.

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. One could considering validating the multiplicative health assumption by looking at the correlation between the semi-elasticity and the average incremental treatment across types.

**Optimality condition**

The optimality condition for the linear contract depends primarily on two empirical objects: the range of administered treatment levels observed in equilibrium, and the incremental treatments across types within the payment group. Let’s fix the provider altruism at $\mu = 1$ and benchmark the ‘efficient’ level of care with the prospective payment contract. The range matters because a prospective payment amount large enough to cover the patient who benefits from the highest level of treatment also gives substantial rents on all the patients with small benefits. The incremental treatments across types, under the assumptions I made, informs us about the scope for supply-side moral hazard. If the range is large, and there is not much scope for supply-side moral hazard, the model suggests it is worthwhile to ‘distort’ care from the efficient level by making payment partially retrospective.

**Definition 4** Let the average incremental treatment be denoted by $\bar{\Delta}$, where

$$\bar{\Delta} = \sum_{i=1}^{N} \gamma_i \left[ \theta_i \frac{dx_{i}^*}{d\theta_i} \right]$$

The average incremental treatment across types is equal to the average semi-elasticity if the health function is multiplicative in the unobserved type. This $\bar{\Delta}$ 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. The idea is that types of treatments for which observed equilibrium quantities are bunched together may also have less scope for supply-side moral hazard.

**Definition 5** Let the outlier ratio be denoted by $\omega$, where

$$\omega \equiv \theta_N \frac{dx_{N}^*}{d\theta_N}$$

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 large. In other words, if the distribution of treatment quantities observed in equilibrium has a tail on the right, the $\omega$ will be big. Conversely, if this distribution is concentrated around the mean, the $\omega$ will be similar to $\bar{\Delta}$.
Proposition 9 Assume costs are linear, that the health production function is multiplicative in the unobserved patient type, that $\mu = 1$, and that $\eta = 0$. The optimality condition for the linear contract is given by

$$\frac{dSWF}{d\phi} \propto \left(1 - \frac{1}{1 - \phi}\right) \bar{\Delta} + (x_N - \bar{x}) - (\omega - \bar{\omega}).$$

(9)

**Proof.** When costs are linear, the derivative of the SWF with respect to $\phi$ is

$$\frac{dSWF(\phi)}{d\phi} = \sum_{i=1}^{N} \gamma_i \left[\left(\frac{1 - \phi}{\mu} - 1\right) c \frac{dx^*_i}{d\phi} - (1 - \eta) \left(c \frac{dx^*_N}{d\phi} - c \frac{dx^*_i}{d\phi}\right) - (cx^*_N - cx^*_i)\right].$$

Substituting in for $(1 - \phi) \frac{dx^*_i}{d\phi} = \theta_i \frac{dx^*_i}{d\theta_i}$, and applying the definitions from $\bar{\Delta}$ and $\bar{\omega}$, we arrive at the desired formula. ■

These two assumptions combined simplify the linear contract’s optimality condition. Evaluating the right hand side of equation 9 requires estimating $\phi$, $\bar{\Delta}$ and $\omega$ in the data. The challenge in computing $\bar{\Delta}$ and $\omega$ as defined above is that both of there depend on the incremental marginal health gain, $\theta_i$. I make the identifying assumption that $\frac{dx^*_i}{d\theta_i}$ is an good proxy for $\theta_i \frac{dx^*_i}{d\theta_i}$. This means 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 incremental treatment, $\hat{\Delta} \equiv \sum_{i=1}^{N} \gamma_i \left[\frac{dx^*_i}{d\theta_i}\right]$, and the approximated outlier ratio, $\hat{\omega} \equiv \frac{dx^*_N}{d\theta_N}$

To compute $\hat{\Delta}$ and $\hat{\omega}$ in the data, I use the following procedure. First, define a unit of $x$ to be one unit of the HCPCS service code[$^{17}$]. Sort all the claims lines for a given HCPCS and compute the first difference across claim lines; this describes the incremental treatment amount, $\frac{dx^*_i}{d\theta_i}$, in the model. Then compute the average incremental treatment across all claim lines for that HCPCS code and the 99th percentile incremental treatment. There two numbers are $\hat{\Delta}$ and $\hat{\omega}$, respectively. Compute the average and 99th percentile treatment levels for $x_N$ and $\bar{x}$.

The reason to use the 99th percentile and not the maximum is to pick up real outliers and avoid coding errors. There are sometimes billing errors that report units which would be unfeasible for a patient to receive in one visit or one day. These errors are particularly frequent in drug codes because providers confuse revenue units with dosage. For instance, Medicare standardized the dosage of an immune globulin injection to 500 milligrams per unit. If a patient receives 1,000 milligrams, the provider should bill two units, not 1,000. However, there are frequent instances of large numbers like 1,000 within the drug service category. For this reason, I exclude drug codes from the analysis.

$^{17}$Units may vary across HCPCS codes, but are consistent within each HCPCS (e.g. one unit may be 15 min of face-to-face interaction for a physical therapy code, but one full visit of undefined length for a new patient evaluation).
The empirical estimate of the derivative of the social welfare function with respect to the retrospective component, substitutes into equation $\hat{\Delta}$ for $\bar{\Delta}$, and $\hat{\omega}$ for $\omega$. I calibrate the retrospective component $\phi$ for all services using the overall estimate of retrospectiveness from Table 3 from the sample without outliers. This gives us a value of the derivative of the social welfare function for each individual HCPCS code. The figure below plots the empirical derivative of the social welfare function against the aggregate Medicare provider payments by HCPCS code, for all non-drug services.

![Outpatient Hospital Services (Non-Drug) Optimality Condition by HCPCS code](image)

Calibrated with $\mu=1$ and $\eta=0$

The positive derivative of the social welfare function suggests that payment for a group of services is too prospective, and there is a welfare increasing deviation in making the payment for these HCPCS codes slightly more retrospective. The services with large potential welfare gains are primarily from the Medical and Surgical Supplies category. That said, these services constitute a small share of Medicare’s bill. For the HCPCS codes that constitute a larger share of the Medicare provider payments, payment appears to be close to optimal.

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 the corner solution of prospective payment is not always optimal: there are welfare reasons for why an insurer may opt for a partially retrospective system in light of patient heterogeneity.

While the type of heterogeneity in the model I presented here fits some settings very well, it is important for future work to study other forms of unobserved heterogeneity, as these may have very different implications for the optimal contract. My framework is adaptable to study heterogeneity in provider ability, costs of treatment, altruism, or risk-aversion. Heterogeneity in provider ability has been one of the top explanations for regional variations in health care utilization, and the work by Atul Gawande suggests 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.

Lastly, though the model here was designed specifically for the health care setting, it is a general framework that can apply to other settings such as an employer’s decision to pay on hourly wage or fixed monthly salary. It would be interesting to explore how an employer’s wage contracting decisions may create inefficiencies when failing to account for altruistic motivation in his employees, particularly within non-profit institutions.
References


Appendix

6.1 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 \((IC \ j \rightarrow j + 1)\).

The planner’s problem is

\[
\max_{x_i} \sum_{i=1}^{N} h(x_i, \theta_i) - (1 - \eta)\pi(x_i) - c(x_i)
\]

s.t.  \( \mu h(x_1, \theta_1) + \pi(x_1) \geq \mu h(x_2, \theta_1) + \pi(x_2) \) \quad (IC 1 \rightarrow 2)

\[
\mu h(x_2, \theta_2) + \pi(x_2) \geq \mu h(x_1, \theta_2) + \pi(x_1) \quad \text{(IC 2 \rightarrow 1)}
\]

\[
\mu h(x_2, \theta_2) + \pi(x_2) \geq \mu h(x_3, \theta_2) + \pi(x_3) \quad \text{(IC 2 \rightarrow 3)}
\]

\[
\mu h(x_j, \theta_j) + \pi(x_j) \geq \mu h(x_{j+1}, \theta_j) + \pi(x_{j+1}) \quad \text{(IC j \rightarrow j+1)}
\]

\[
\mu h(x_{j+1}, \theta_{j+1}) + \pi(x_{j+1}) \geq \mu h(x_j, \theta_{j+1}) + \pi(x_j) \quad \text{(IC j+1 \rightarrow j)}
\]

\[
\mu h(x_N, \theta_N) + \pi(x_N) \geq \mu h(x_{N-1}, \theta_N) + \pi(x_{N-1}) \quad \text{(IC N-1 \rightarrow N)}
\]

\[
\pi(x_1) \geq 0 \quad \text{(PC 1)}
\]

\[
\pi(x_N) \geq 0 \quad \text{(PC N)}
\]

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 \( x_{j+1} \geq x_j \ \forall \ j \).

\[
\mu h(x_j, \theta_j) + \pi(x_j) \geq \mu h(x_{j+1}, \theta_j) + \pi(x_{j+1}) \quad \text{(IC j \rightarrow j+1)}
\]

\[
+ \mu h(x_{j+1}, \theta_{j+1}) + \pi(x_{j+1}) \geq \mu h(x_j, \theta_{j+1}) + \pi(x_j) \quad \text{(IC j+1 \rightarrow j)}
\]

\[
\implies h(x_{j+1}, \theta_{j+1}) - h(x_j, \theta_{j+1}) \geq h(x_{j+1}, \theta_{j}) - h(x_j, \theta_{j})
\]

Otherwise, we would get a contradiction.
Step 2: Consider a modified problem.

Consider instead a modified problem with only the local upwards incentive constraints, \((IC \ j \rightarrow \ j+1)\), 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

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

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

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

\[
\begin{align*}
\mu h(x_j, \theta_j) + \pi(x_j) &\geq \mu h(x_{j+1}, \theta_j) + \pi(x_{j+1}) \quad (IC \ j \rightarrow j+1) \\
\implies \pi(x_j) - \pi(x_{j+1}) &\geq \underbrace{\mu h(x_{j+1}, \theta_j) - \mu h(x_j, \theta_j)}_{\geq 0 \text{ by monotonicity}}
\end{align*}
\]

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, \(\pi(x_N)\).

\[
\begin{align*}
\pi(x_{N-1}) &\geq \pi(x_N) + \mu h(x_N, \theta_{N-1}) - \mu h(x_{N-1}, \theta_{N-1}) \\
\pi(x_{N-2}) &\geq \pi(x_{N-1}) + \mu h(x_{N-1}, \theta_{N-2}) - \mu h(x_{N-2}, \theta_{N-2}) \\
\implies \pi(x_{N-2}) &\geq \pi(x_N) + \mu h(x_N, \theta_{N-1}) - \mu h(x_{N-1}, \theta_{N-1}) + \mu h(x_{N-1}, \theta_{N-2}) - \mu h(x_{N-2}, \theta_{N-2}) \\
&= \pi(x_N) + \mu \sum_{i=N-2}^{N-1} h(x_{i+1}, \theta_i) - h(x_i, \theta_i)
\end{align*}
\]
We can keep substituting local upward constraints to derive that, for every \( j \in \{1, \ldots, N - 1\} \), and show that the local upwards constraints can all be written as functions of \( \pi(x_N) \) and \( h(\cdot, \theta_j) \).

\[
\pi(x_j) \geq \pi(x_N) + \mu \sum_{i=j}^{N-1} h(x_{i+1}, \theta_i) - h(x_i, \theta_i) \\
\ldots \\
\pi(x_1) \geq \pi(x_N) + \mu \sum_{i=1}^{N-1} h(x_{i+1}, \theta_i) - h(x_i, \theta_i).
\]

**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 \( \pi(x_N) - \epsilon > 0 \). 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 \( \epsilon \) will strictly increase the objective function.

\[
SWF = h(x_N, \theta_N) - c(x_N) - (1 - \eta)\pi(x_N) + \sum_{i=1}^{N-1} h(x_i, \theta_i) - c(x_i) - (1 - \eta)\pi(x_i)
\]

\[
< h(x_N, \theta_N) - (1 - \eta)\pi(x_N) + (1 - \eta)\epsilon - c(x_N) + \sum_{i=1}^{N-1} h(x_i, \theta_i) - \pi(x_i) - c(x_i).
\]

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

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

For an arbitrary \( j \), suppose \((IC\; j \rightarrow j + 1)\) 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 \((IC_j \rightarrow j+1)\) is still satisfied, and reducing \(\pi(x_j)\) by \(\epsilon\) strictly increases the objective function.

\[
SWF = h(x_j, \theta_j) - c(x_j) - (1 - \eta)\pi(x_j) + \sum_{i \neq j} h(x_i, \theta_i) - c(x_i) - (1 - \eta)\pi(x_i)
\]

\[
< h(x_j, \theta_j) - (1 - \eta)\pi(x_j) + (1 - \eta)\epsilon - c(x_j) + \sum_{i \neq j} h(x_i, \theta_i) - \pi(x_i) - c(x_i).
\]

Therefore, it is optimal to continue reducing \(\pi(x_N)\) until \((IC_j \rightarrow j + 1)\) is binding.

**Step 7: Characterize the solution.**

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

\[
\pi(x_j) = \mu \sum_{i=j}^{N-1} h(x_{i+1}, \theta_i) - h(x_i, \theta_i) \quad \text{, } j \in \{1, \ldots, N - 1\}
\]

\[
\pi(x_N) = 0.
\]

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

\[
SWF = \sum_{j=1}^{N} h(x_j, \theta_j) - c(x_j) - (1 - \eta)\pi(x_j)
\]

\[
= \sum_{j=1}^{N} h(x_j, \theta_j) - c(x_j) - (1 - \eta) \left( \mu \sum_{i=j}^{N-1} h(x_{i+1}, \theta_i) - h(x_i, \theta_i) \right)
\]

\[
= \sum_{j=1}^{N} h(x_j, \theta_j) - c(x_j) - (1 - \eta) \cdot \mu \sum_{j=1}^{N-1} j \cdot (h(x_{j+1}, \theta_j) - h(x_j, \theta_j))
\]

It follows that

\[
\max_{x_j} h(x_N, \theta_N) - c(x_N) + \sum_{j=1}^{N-1} h(x_j, \theta_j) - c(x_j) - (1 - \eta)\mu \cdot j \cdot (h(x_{j+1}, \theta_j) - h(x_j, \theta_j))
\]
\[ \frac{\partial SWF}{\partial x_N} = h_x(x_N, \theta_N) - c_x(x_N) - (1 - \eta) \mu(N - 1) h_x(x_N, \theta_{N-1}) = 0 \]

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

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

**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—

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

Monotonicity for all types below—

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

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

\[ \pi(x_j) = \mu \sum_{i=j}^{N-1} h(x_{i+1}, \theta_i) - h(x_i, \theta_i), \quad j \in \{1, \ldots, T - 1\} \]

\[ \pi(x_T) = 0. \]
\[ SWF = \frac{1}{N} \sum_{j=1}^{T} h(x_j, \theta_j) - c(x_j) + \frac{1}{N} \sum_{j=T}^{N} (h(x_T, \theta_j) - c(x_T)) - (1 - \eta) \cdot \mu \cdot \frac{1}{N} \sum_{j=1}^{T-1} j \cdot (h(x_{j+1}, \theta_j) - h(x_j, \theta_j)) \]

\[ \implies \]

\[ \frac{\partial SWF}{\partial x_T} = \sum_{j=T}^{N} \frac{h(x_T, \theta_j) - c(x_T) - (1 - \eta) \cdot \mu(T - 1)h_x(x_T, \theta_{T-1}) = 0}{h(x_T, \theta_j) - c(x_j) + \mu(1 - \eta)[j \cdot h_x(x_j, \theta_j) - (j - 1) \cdot h_x(x_j, \theta_{j-1})]} \]

\[ \implies \frac{\partial SWF}{\partial x_j} = h_x(x_j, \theta_j) - c_x(x_j) + \mu(1 - \eta)h_x(x_1, \theta_1) = 0 \]

The monotonicity condition for the top becomes–

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

**Treatment level for the threshold type. \( \theta_T \) is above his first best level:**

\[ h_x(x_T, \theta_T) - c_x(x_T) + \sum_{j=T+1}^{N} h_x(x_T, \theta_j) \geq (N - 1 - T)h_x(x_T, \theta_{T+1}) \]

\[ \implies h_x(x_T, \theta_T) - c_x(x_T) \leq (1 - \eta) \mu \left( (T - 1)h_x(x_T, \theta_{T-1}) - (N - 1 - T)h_x(x_T, \theta_{T+1}) \right) \]

\[ \leq 0 \]

**Treatment level for the highest type. \( \theta_N \), is below first best level:**

\[ \sum_{j=T}^{N} (h_x(x_T, \theta_j) - c_x(x_T) - (1 - \eta) \cdot \mu(T - 1)h_x(x_T, \theta_{T-1}) = 0 \]

\[ \implies h_x(x_T, \theta_N) - c_x(x_T) = (1 - \eta) \cdot \mu(T - 1) \frac{h_x(x_T, \theta_{T-1})}{\geq 0} + \left( h_x(x_T, \theta_N) - \sum_{j=T}^{N} h_x(x_T, \theta_j) \right) \]

\[ \geq 0 \]
Step 10: Checking that the solution to this problem satisfies the other constraints.

Revisiting \((IC \ j+1 \rightarrow \ j)\), and plugging in the solution of the modified problem, it follows that,

\[
\begin{align*}
\mu h(x_{j+1}, \theta_{j+1}) + \pi(x_{j+1}) &\geq \mu h(x_j, \theta_{j+1}) + \pi(x_j) \\
\mu h(x_{j+1}, \theta_{j+1}) - \mu h(x_j, \theta_{j+1}) &\geq \pi(x_j) - \pi(x_{j+1}) \\
&= \mu \left( \sum_{i=j}^{N-1} h(x_{i+1}, \theta_i) - h(x_i, \theta_i) - \sum_{i=j+1}^{N-1} h(x_{i+1}, \theta_i) - h(x_i, \theta_i) \right) \\
&= h(x_{j+1}, \theta_{j+1}) - h(x_j, \theta_{j+1}) \geq h(x_{j+1}, \theta_j) - h(x_j, \theta_j)
\end{align*}
\]

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.

### 6.2 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, \(\phi\), is

\[
\frac{dSWF}{d\phi} = \left( \frac{1}{\mu} - \frac{1}{1-\phi} \right) \bar{\varepsilon}_\phi + (1 - \eta)(\bar{\varepsilon}_\phi + c(x_N) - \bar{c}) - (1 - \eta)c(x_N)\bar{c}_\phi,
\]

where \(\bar{\varepsilon}_\phi \equiv \sum_{i=1}^{N} \gamma_i \varepsilon^i_\phi\) 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 \(\phi\) via the first order condition of the SWF with respect to \(\phi\), 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,

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

s.t.

\[
\begin{align*}
\mu h_x(x_\theta, \theta) - c_x(x_\theta)(1 - \phi) &= 0 \quad \phi \leq 1 \\
t + \phi c(x_\theta) - c(x_\theta) &\geq 0 \forall x_\theta
\end{align*}
\]

where (LIC) stands for the local incentive constraint of the physician. Since the set of incentive compatible \(x_\theta\)'s are monotonically increasing in \(\theta\), and costs are non-decreasing in treatment level
$x$, it follows that the set of (PC) constraints for all types below $\theta_N$ 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 $\theta_N$.

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 $x_\theta$ is increasing in $\theta$, the only binding (PC) will be that of type $N$.

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

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

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

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

6.3 Data Appendix

The service categorization was done according to the following ranges.

- Administrative, Miscellaneous, & Investigational: HCPCS A9150-A9999
- Blood Services: CPT 99195; HCPC P9010-P9100, P2028-P2038, G0460, C9137, C9139
- Cardiovascular: CPT 92920-93799; HCPCS G9680
- Dialysis: CPT 90935-90999; HCPCS G0257, G0365
- Drugs and Administration: CPT 96360-96549; HCPCS J0120-J9999, Q2043, Q2049, Q2050, Q5101, G0498, G0260
- Emergency: HCPCS G0378-G0390
- Emerging Technologies (Level 3): All level 3 codes ending with T or F; HCPCS C9478
- Evaluation & Management: CPT 99201-99499; HCPCS G0245-G0249, G0463
- Gastroenterology: CPT 91010-91299; HCPCS G9684
• Imaging & Radiology: CPT 70010-79999; HCPCS C8900-C8937, G0297, R0070-R0076

• Medical Devices: HCPCS C1713-C2615

• Medical and Surgical Supplies: HCPCS A4206-A8004, C8957-C9488

• Neurological: CPT 95700-96020

• Opthamology: CPT 92002-92499

• Outpatient Therapy: CPT 97151-97804, 98960-98962; HCPCS G0237-G0239, G0424

• Pathology and Laboratory: CPT 80047-89398

• Preventative & Screening: CPT 76706, 77052, 77057, 77063, 77067, 77078-77085, 76977, 80061, 82465, 83718, 84478, 82947, 82950-82951, 97802-97804, 96127; HCPCS G0389, G0402-G0405, G0108-G0109, G0270-G0271, G0123-G0124, G0130, G0141-G0148, P3000-P3001, Q0091, G0451, G0101-G0199, G0431-G0449, G0396-G0399

• Psychiatry and Psychology: CPT 90785-90899; HCPCS G0296, G0410, G0411, G0473

• Pulmonary: CPT 94002-94799; HCPCS G0302-G0307, G0277, G9679, G9681

• Surgery: CPT 10004-69990; HCPCS G0416, C5271, C5273, C5275, C5277

• Vaccine: CPT 90281-90756; HCPCS G0008-G0010, Q2034-Q2039