Optimal Contracting with Altruistic Agents:

A Structural Model of Medicare Reimbursements for Dialysis Drugs*

Martin Gaynor*, Nirav Mehta**, and Seth Richards-Shubik***
Carnegie Mellon and NBER*, University of Western Ontario**, Lehigh and NBER***

June 20, 2018

PRELIMINARY—PLEASE DO NOT CITE

Abstract

We study physician agency and optimal payment policy in the context of an expensive medication (epoetin alfa) used with dialysis. Using Medicare claims data we estimate a model of treatment decisions, in which physicians are partially altruistic and value both their own compensation and their patients' health. We then use the recovered parameters of the model in combination with contract theory to derive and simulate optimal linear and nonlinear reimbursement schedules. Physicians differ in their marginal costs of treatment, and this heterogeneity is unobservable to the government, which affects payment policy, along with physician altruism and the effectiveness of treatment. Comparing outcomes under these optimal contracts against those observed under the actual contracts suggests that substantial improvements in payment policy can be achieved within a fee-for-service framework.

1 Introduction

A central problem in health economics, and in health care organizations, is how to compensate physicians for their services. Physicians are typically viewed as imperfect agents for their patients, deriving utility from both their private benefits and costs and from the impact of their services on patient health. A substantial theoretical literature in health economics considers how these partially altruistic agents might behave under various payment systems, and a related empirical literature examines how physicians respond to financial incentives.

*We are grateful for helpful comments to Josh Gottlieb, Albert Ma, Lowell Taylor, and to audiences at talks at Carnegie Mellon, Georgia, Johns Hopkins, UBC, the 2018 International Industrial Organization Conference, and the 2018 Annual Meeting of the Society of Labor Economists. We thank Ali Kamranzadeh for excellent research assistance. The usual caveat applies.

We extend this work by specifying and estimating a physician utility function that depends on patient health and physician income, thereby quantifying the level of altruism among physicians in our setting. Physicians differ in their marginal costs of treatment, and this heterogeneity is unobservable to the government (the principal who chooses the payment contract). We estimate a simple linear reduced form that allows us to recover the structural parameters of the model, including physician altruism, the productivity of treatment, and the distribution of marginal costs. We then use these recovered parameters in combination with results from contract theory to derive and simulate optimal payments to physicians for service provision. Comparing the optimal contract with actual contracts suggests tractable improvements in payment policy.

Our empirical analysis is on the provision of an expensive and controversial medication used by dialysis centers to treat anemia in patients with end stage renal disease (ESRD). The medication, epoetin alpha (or "EPO"), was the largest single drug expenditure in Medicare for several years. Medicare is the predominant payer for the treatment of ESRD in the United States (at any age), and we use Medicare claims data from 2008 and 2009 to estimate our model. In the model, physicians choose the quantity of treatment to provide to each patient, based on the predicted health impact, the cost of treatment, and the payment contract. Uniquely in our setting, a quantitative measure of patient need is available in the claims records because providers were required to report a blood measurement in order to be reimbursed for EPO. The revealed weight placed on improvements in patient health relative to the physician's private marginal benefit is our measure of altruism. This parameter and the other parameters of our model are recovered in a simple manner from the coefficients of a linear fixed effects regression.

We then address the agency problem that arises when physicians are heterogeneous in their cost of providing treatment.[1] Using the estimated physician utility function, we derive both optimal (constrained) linear and unconstrained (potentially nonlinear) payment contracts. These contracts are fairly straightforward to construct from the output of the fixed effects regression. We then compare the reimbursement schedules and induced treatments under these contracts against the observed outcomes. Our results suggest that the histor-

[1] There are substantial differences across providers in the acquisition cost of the drug, as documented in Medicare renal dialysis facility cost reports (see Section 3). Heterogeneity in the level of altruism is also natural to consider here, and in work currently underway we extend the model to allow for heterogeneity in both cost and altruism. The optimal contract in the case of multidimensional heterogeneity is more difficult to characterize than that for the case of one-dimensional heterogeneity because physician "types" can no longer be ordered (see Maskin et al., 1987)). However, in our application the "demand profile" approach proposed by Wilson (1993) appears more promising than it has in the typical application where a monopolist sells a good to an agent (Deneckere and Severinov, 2015). Intuitively, we study supply, not demand, which makes it more likely that the agent's (i.e., physician's) objective will be quasiconcave for an optimal contract.

ical reimbursement rates for EPO were supra-optimally high, and by a wide margin. This is intuitive, as physician altruism and cost heterogeneity, neither of which may have been accounted for by the government, both reduce the per-unit payment in the optimal linear contract.

We derive optimal reimbursement rates in the linear contract that are substantially lower than the actual rates used by Medicare in 2008 and 2009. The optimal nonlinear contract improves outcomes further, notably by reducing seemingly unjustified variation in treatment intensities while also decreasing total expenditures. A simulation for patients with anemia at the median level of severity finds that optimal nonlinear contract that maintains the same average health as that under the actual payment contract used by Medicare reduces the standard deviation of dosages by 41 percent while the mean payment decreases by 36 percent.

In what follows, we first review the related literature (Section 1.1), then provide institutional background (Section 1.2). We introduce the model in Section 2, then derive the optimal linear contract (Section 2.1), and the optimal unconstrained contract (Section 2.2). Section 3 contains a description of the data we use for our empirical analysis, while Section 4 describes the empirical implementation, including specification, identification (Section 4.1), and estimation (Section 4.2). Quantitative results on the optimal linear and unconstrained contracts are in Section 5. Section 6 concludes.

1.1 Related literature

We view our paper as being closely related to two literatures described below.

Health economics: There is a rich theoretical literature on physician agency; much of this literature is discussed in McGuire (2000) and Chalkley and Malcomson (2000). Ellis and McGuire (1986) provide a seminal contribution. They model physicians as partially altruistic (imperfect agents) and show that partial cost reimbursement can improve outcomes when physicians are imperfect agents for their patients and inputs are noncontractible.

By contrast, there are relatively few theoretical papers on optimal contracting in health care. Chalkley and Malcomson (1998) characterize optimal contracts when patient demand does not reflect quality, and show that the optimal contract differs depending on the degree of physician altruism. De Fraja (2000) studies optimal contracts when there is heterogeneity in physician costs. Jack (2005) allows for heterogeneity in physician altruism and solves for the optimal contract in an environment where quality is noncontractible. Malcomson (2005) examines optimal contracts when providers are better informed than purchasers, with no provider altruism. Choné and Ma (2011) also study how physician altruism may affect the design of optimal payment schemes.

There is also an empirical literature studying how physicians respond to incentives and other changes in the environment. Gaynor and Pauly (1990) is an early paper showing that physicians respond strongly to financial incentive. Chandra et al. (2012) review the literature studying determinants of physician treatment choices. One determinant they focus on is physician altruism. Gaynor et al. (2004) specify and estimate a structural model of physician treatment choice where physicians are partially altruistic. Godager and Wiesen (2013) use data obtained from a laboratory experiment to document the existence of physician altruism, which they find to be heterogeneous. Clemens and Gottlieb (2014) examine the impact of financial incentives in Medicare payment for physicians and find substantial effects on supply, technology adoption, and patient outcomes.

Our model contains many components of the above literature. The basic model of physician utility is very similar to that in Gaynor et al. (2004), but allows for cost heterogeneity as well as altruism. De Fraja (2000) and Jack (2005), noted above, address heterogeneity across physicians, although there are various distinctions between their models and ours.[2] Like Clemens and Gottlieb (2014), we examine the impact of Medicare payment incentives, although they look at payment incentives broadly, as opposed to our focus on a very specific program and treatment decision.[3] Last, in contrast to the existing empirical literature, we not only estimate a model of physician treatment choices and recover physician altruism, we also empirically characterize the optimal contract and compare it to the actual contracts used in this context.

Empirical contracts: As Chiappori and Salanié (2003) discuss, there is empirical work testing for the existence of salient features for the design of optimal contracts (e.g., Chiappori et al. (2006) test for asymmetric information), but there is little work specifying and estimating structural models and using them to derive optimal contracts. This matters because the insights from the literature on contracts are most useful when applied in designing optimal policies that could be implemented in reality.

To the best of our knowledge, there is no work that structurally estimates a model of physician treatment choices in a principal-agent, or asymmetric information, framework and uses this to characterize optimal contracts. A handful of papers estimate asymmetric information models in other settings. For example, Einav et al. (2010) discuss the small

[2] For example, Jack (2005) uses a model with unobserved effort while in our setting the most relevant aspect of the treatment is observed (i.e., the dosage of the drug).

[3] Grieco et al. (2017) similarly uses the specific context of dialysis care to examine an issue of broad importance in health care, the tradeoff between quantity and quality.

literature doing this for insurance contracts. Paarsch and Shearer (2000) characterize the optimal linear contract in a hidden action environment.

Gayle and Miller (2009) also study hidden action models, quantifying the welfare loss from moral hazard. In contrast, we study a screening, or hidden information, model and flexibly characterize the optimal wage schedule. Screening models, with their focus on unobserved heterogeneity, are clearly policy relevant.

There is a fairly rich literature on optimal regulation, which considers screening models in institutional contexts that differ from ours in important ways. Wolak (1994) develops and estimates a model in which a principal seeks to regulate public utilities of (potentially) hidden types. Data limitations, including a lack of variation in regulatory regime, mean the distribution of types cannot be estimated without imposing optimality of the observed contract. Gagnepain and Ivaldi (2002), study a similar environment, but exploit variation in the regulatory regime to estimate a parametric distribution of types without having to assume optimality of the observed contract. This allows them to test whether the observed contract is optimal. Abito (2017) extends this approach to study optimal pollution regulation. As in the latter two articles, our setting and data allow us to estimate structural parameters, including agent types, without imposing optimality of the observed contract. We allow for a fully flexible type distribution, which is made possible by variation in the observed regime (i.e., reimbursement contract) and a large number of repeated measures of physicians, as each physician chooses a treatment choice for each patient they see under a variety of observed reimbursement rates.

1.2 Institutional background

ESRD, or kidney failure, is a chronic and life-threatening condition that affects over half a million individuals in the United States at a given point in time. Since 1973, the Medicare program has provided universal coverage for the treatment of ESRD, regardless of age. In 2009 Medicare spent $28 billion on health care for individuals with ESRD, and of that amount, $1.74 billion went to payments for the drug EPO.[4] The drug is used to treat anemia, a lack of red blood cells, which often accompanies chronic kidney disease.[5] It is similarly used to treat anemia in chemotherapy patients. EPO stimulates red blood cell production, and dialysis providers administer it to their patients to try to maintain a certain level of red blood cells. The level is commonly measured as *hematocrit*, which is the volume percentage

[4] USRDS *2017 Annual Data Report*, available at `https://www.usrds.org/adr.aspx`. The amount given for EPO includes a related drug darbepoetin alpha. The total social expenditures on ESRD and these drugs were even higher because many beneficiaries also make a 20% copayment.

[5] EPO is a biological product, or "biologic," but we will typically refer to it as a drug.

of red blood cells in the blood.

Medicare's payment policy for EPO was debated throughout the 1990s and 2000s, largely because of concerns that the reimbursement rates were too generous and encouraged overprovision. While dialysis itself was reimbursed with a prospective payment system (PPS) known as the "composite rate," EPO was a separately billable drug with its own per-unit reimbursement rate. Prior to 2005, the rate was held fixed at $10.00 per 1000 units. In 2006, Medicare adopted a new policy where the reimbursement rate was based on average sales prices calculated from data reported by the drug manufacturer. This policy, which was in effect through 2010 (including the years we use to estimate provider behavior), set a reimbursement rate each quarter equal to 106 percent of the national average sales price from roughly six months earlier (GAO, 2006). Later, in 2011, Medicare adopted a comprehensive "bundled" PPS for dialysis that included EPO, so the payment policy for the drug effectively switched from fee-for-service to prospective payment.[6]

Important safety concerns about EPO emerged by the mid 2000s. A major clinical trial found that patients who were given more EPO to achieve a higher target level of hematocrit suffered a higher risk of serious cardiovascular events and death (Singh et al., 2006). This study was published in November 2006, and strong warnings ("black box warnings") were added to the drug's labels in 2007. As a result of these findings, the recommended target level for hematocrit remained at a lower range, specifically 30–36%, which was the existing standard at the time (e.g., the range for which the FDA had approved the drug).

Medicare claims for dialysis care, which are the source of our data, are typically filed monthly and include separate lines for each administration of EPO. The drug is most commonly administered intravenously during dialysis, which occurs multiple times per week at specialized facilities called dialysis centers. The staff of these facilities typically consists of one medical director (a physician) and multiple nurses and medical technicians, and payments are primarily made to the dialysis centers, not the physician(s).[7] To be reimbursed for EPO, the centers are required to report a hematocrit level taken just prior to the monthly billing cycle. This is an unusual feature in claims data and provides a quantitative measure of patient need, in this case the severity of the anemia, which is a key component of our model. Last, a relevant medical point is that the half-life of EPO is under 12 hours (Elliott et al., 2008), so there is no direct stock effect of the drug from one month to the next. This partly supports our use of a static model applied separately to each month of treatment, although

[6] In future work we will evaluate the decreases in dosages that occurred when the bundled PPS was adopted, and compare these to the simulated dosages under our optimal contracts.

[7] See *NEJM Catalyst*, `https://catalyst.nejm.org/the-big-business-of-dialysis-care/`, for an overview. Some dialysis centers have multiple physicians on staff, but in the empirical analysis we treat each facility as a single provider.

there are longer-term effects on red blood cell production and other health outcomes.

2 Model

In our theoretical model there is one time period, one principal, and one agent. The government (the principal) hires a physician (the agent) to treat a patient.[8] The government seeks to maximize patient health, net of the cost of transfers to the physician. Thus the government can be thought of as acting on behalf of patients, who receive benefits from treatment but have to fund public health insurance through taxes.[9] Physician utility depends on patient health, the cost of the treatment provided, and the compensation received.

The patient arrives at the physician with a baseline health status, e_0, which in our application is the hematocrit level from the prior month. The physician then chooses an amount of treatment (the action), a, which is the total units of EPO administered over the month. As is standard in this literature, we assume the patient passively accepts the treatment prescribed by their physician (e.g., Ellis and McGuire (1986)). Both e_0 and a are observed by the government since they are reported in the monthly claims. Given the patient's baseline health status, the treatment produces health according to the function $h(a, e_0)$. Health is increasing in the amount of treatment when the resulting hematocrit level is below the target, e_τ, is decreasing when the resulting level is above the target, and is concave in treatment (i.e., $\partial h/\partial a > 0$ if $h(a, e_0) < e_\tau$, $\partial h/\partial a < 0$ if $h(a, e_0) > e_\tau$, and $\partial^2 h/\partial a^2 < 0$). This reflects the fact that patients with more severe anemia (i.e., lower hematocrit) benefit more from EPO, while there are serious risks from over-provision.

The physician is of a "type," indexed by $i \in I$, which is unobserved by the government. Here the type represents the physician's cost of providing treatment. We will refer to physicians by their type i when it is convenient and not confusing to do so. The cost type determines the per-unit cost of treatment and is denoted by z_i. We order cost types such that $z_i < z_{i+1}$; i.e., lower indices correspond to lower treatment costs. Treatment cost has support on a closed interval $[\underline{z}, \overline{z}]$ and a distribution F_z, which is known to the government, with density f_z and mean μ_z. Finally, let $\underline{i} \equiv \{i : z_i = \underline{z}\}$ and $\bar{i} \equiv \{i : z_i = \overline{z}\}$ respectively denote the indices corresponding to the lowest and highest cost types.

The government sets a reimbursement policy (the payment contract, or "wage" schedule), $w(a, e_0)$, that may depend on both the treatment amount and the baseline health. This can be understood as a set of (potentially) nonlinear contracts, one for each possible value of e_0.

[8] Section 4 discusses how this model extends naturally to multiple physicians with multiple patients.

[9] The government's objective does not include the physician's surplus, so it does not represent social welfare. This is of course natural in a principal-agent model.

The reimbursement policy is established before the physician sees the patient. The timing of the model is summarized below.

Timing:

1. Government sets the wage schedule ($w(a, e_0)$)
2. Physician's type is realized (z_i)
3. Patient's baseline hematocrit level is realized (e_0)
4. Physician decides whether to participate
5. Physician chooses an amount of treatment (a)
6. Outcomes occur: patient health ($h(a, e_0)$), government payment to physician ($w(a, e_0)$), cost of treatment ($c(a, z_i)$)

The utility function for a physician of cost type i is

$$u_i(a; e_0, w) \equiv \alpha_p h(a, e_0) - c(a, z_i) + w(a, e_0), \tag{1}$$

where $c(\cdot)$ is the cost function, which gives the total cost of providing treatment amount a for a physician with cost factor z_i. As described above, $h(a, e_0)$ gives the resulting health of the patient, $w(a, e_0)$ gives the reimbursement amount from the government, and the altruism parameter, α_p, is the weight placed on patient health.

The government values patient health minus the payment to the physician.[10] The weight that the government places on patient health, α_g, may be different than the weight placed by physicians (e.g., this weight may be larger because the government represents the patients). The government's net value of the outcome, given the amount of treatment provided, is thus

$$u_g(a, e_0) \equiv \alpha_g h(a, e_0) - w(a, e_0). \tag{2}$$

The government's expected value of the outcome integrates this over the actions that would be taken by different physician types, given the reimbursement policy.

We use subgame perfect Nash equilibrium to define behavior. The physician chooses a treatment amount to maximize utility function (1) given the wage schedule (and the patient's baseline health). The government sets the wage schedule knowing how the physician will respond. The government's problem is therefore to maximize the expected value of (2) subject to the physician's incentive compatibility and voluntary participation constraints, which we assume must hold for all physicians.

[10] As noted earlier, the government's objective is not intended to represent social welfare.

We next provide a specification that yields intuitive, closed-form solutions, and which is the basis of our empirical specification. The functions c and h are, respectively, linear and quadratic, and the government is restricted to offer linear contracts. In addition to providing clear intuition, a linear contract was the actual payment policy during the period we use to estimate the model, creating a tight link between our model, estimation strategy, and the institutional context. In Section 2.2, we summarize the solution for the optimal unrestricted contract (i.e., the "second-best"). Although the solution for the optimal unrestricted contract would, with minor adaptation, work for more general specifications of c and h (see Appendix A.2 for details), we maintain the linear c and quadratic h because they are sufficiently flexible for our empirical work.

2.1 Optimal Linear Contract

The specifications here, which are also applied in the empirical analysis, are as follows. The health production function is a quadratic loss in the distance from the target level of hematocrit:

$$h(a, e_0) \equiv -\frac{1}{2}[\delta a + e_0 - e_\tau]^2,$$

where δ is a linear technology that converts the dosage a of EPO to an increase in hematocrit.[11] The bliss point of the health function is achieved when baseline hematocrit plus this output from the drug equals the target level (i.e., $\delta a + e_0 = e_\tau$). The cost function is $c(a; z_i) = az_i$; in other words, the cost factor z_i is the marginal cost of providing the drug, and there is no fixed cost.[12] Finally, the government offers a linear contract for each e_0:

$$w(a, e_0) \equiv w_0(e_0) + w_1(e_0) \cdot a.$$

This means that there is a fixed payment (w_0) and a per-unit payment (w_1), and these amounts may depend on the baseline hematocrit.[13] For convenience, however, we suppress the dependence of $w(\cdot)$ and other functions on e_0 when it is not confusing to do so.

We solve by backward induction. Given the patient's baseline hematocrit and the corre-

[11] To some extent this equates health with the amount of red blood cells in the blood, which has been questioned in the medical literature (Jacques et al., 2011). Interpreted more broadly as health, this specification with a bliss point represents the tradeoff between the benefits of increasing very low blood counts against the serious risks associated with EPO, and the fact that those risks outweigh the benefits when the blood count is above the target level.

[12] Adding a fixed cost that is constant across types would not affect the results.

[13] In fact, the Medicare payment policy during our analysis period included a reduced reimbursement rate for claims where the prior hematocrit level was above 39%.

sponding linear contract, the physician of cost type i solves

$$\max_{a \geq 0} \ -\alpha_p \frac{1}{2}[\delta a + e_0 - e_\tau]^2 - a z_i + w_0 + w_1 a.$$

At an interior solution, the optimal treatment amount is

$$a_i^* = a^*(z_i; e_0, w) \equiv \frac{e_\tau - e_0}{\delta} + \frac{w_1 - z_i}{\delta^2 \alpha_p}, \tag{3}$$

which is unique due to the weak concavity of c and w, which, when added to the strictly concave h, produces a strictly concave physician objective. Here, we assume the conditions are such that an interior solution always applies. Hence, under the equilibrium contract, a physician of any cost type will provide a positive amount of the drug to a patient with any baseline hematocrit level that the government would like to have treated (i.e., any $e_0 < e_\tau$). We relax the interior solution requirement in Section 2.2, so that high-cost physicians may choose not to provide the drug under the optimal nonlinear contract.

The physician behavior above implies that the hematocrit level resulting from the chosen treatment amount (i.e., $\delta a_i^* + e_0$) may be less than, equal to, or greater than the target level e_τ, depending on the wage schedule and the physician's marginal cost of EPO. If the marginal cost is greater than the reimbursement rate ($z_i > w_1$) then the resulting hematocrit will be below the target. If the marginal cost of acquiring and administering EPO is less than the reimbursement rate, then the resulting hematocrit will be above the target. This matches concerns that were raised about high reimbursement rates encouraging over-provision of EPO.

The government's problem is to set a slope w_1 and an intercept w_0 for the payment contract (for each e_0) that maximizes the expectation of (2), while ensuring participation and treatment. This problem is expressed as follows:

$$\max_{(w_0, w_1) \in \mathbb{R}^2} \int_{\underline{z}}^{\overline{z}} \left[\alpha_g h(a, e_0) - w_0 - w_1 a\right] f_z(z) dz \tag{4}$$

$$\begin{aligned} \text{s.t.} \quad & \\ & a = a^*(z_i; e_0, w), \forall i \in I \qquad && \text{IC} \\ & u_i(a^*(z_i; e_0, w); e_0, w) \geq 0, \forall i \in I \qquad && \text{VP}, \end{aligned}$$

where the functions u_i and a^* are defined in equations (1) and (3). The incentive compatibility (IC) constraints recognize that the physician implements her optimal treatment amount given the contract, and the voluntary participation constraints (VP) require the government to offer payments such that action a_i^* provides nonnegative utility for any cost type.[14] Since lower cost types have strictly greater utility for any actions and payments, this simplifies to a Lagrangian with one constraint because all VP constraints are slack except the one for the highest cost type $\overline{z}$.

Solving the Lagrangian yields the optimal intercept (i.e., fixed payment) and slope (i.e., per-unit rate) given below (details are in Appendix A.1):

$$\begin{aligned} w_0^* &= \frac{\left[2+\frac{\alpha_g}{\alpha_p}\right]\overline{z}-\left[1+\frac{\alpha_g}{\alpha_p}\right]\mu_z}{\delta\left[1+\frac{\alpha_g}{\alpha_p}\right]}\left[[e_\tau - e_0] + \frac{\left[1+\frac{\alpha_g}{\alpha_p}\right]\mu_z-\left[2+\frac{\alpha_g}{\alpha_p}\right]\overline{z}}{2\delta[\alpha_p+\alpha_g]}\right] \\ w_1^* &= \mu_z - \frac{\alpha_p}{\alpha_p+\alpha_g}\overline{z}. \end{aligned} \tag{5}$$

The optimal per-unit rate w_1^* is decreasing in both physician altruism and the upper limit of the cost type distribution, which is related to the "spread" of this distribution. The maximum cost type thus affects the strength of the incentives given to physicians of all types, because the participation constraint for this type is the one that binds. The higher is $\overline{z}$, the more expensive it is for the government to induce a positive action: the intercept w_0^* is higher while the per-unit rate w_1^* is lower (see Appendix A.1.1 for details). Also, we note that the patient's baseline hematocrit (e_0) only appears in the intercept, which keeps this contract fairly simple. The fixed payment (w_0^*) would vary with the patient's measured severity, but there would be a common per-unit rate (w_1^*) as in traditional fee-for-service systems.

Substituting the optimal per-unit rate into the physician's treatment choice function (3) yields the equilibrium action under the optimal linear contract:

$$a_i^{*\text{linear}} = \frac{e_\tau - e_0}{\delta} - \frac{z_i}{\delta^2\alpha_p} + \frac{\mu_z}{\delta^2\alpha_p} - \frac{\overline{z}}{\delta^2[\alpha_g+\alpha_p]}. \tag{6}$$

This is decreasing in the physician's own marginal cost (z_i) but increasing in the average marginal cost (μ_z).

It is useful to compare these equilibrium treatment amounts with the first-best amounts that would occur in a full-information scenario, where physician cost types are observable and contractible.[15] Given the assumed functional forms, the treatment amounts in the full-

[14] The utility of the outside option is normalized to zero.

[15] Under full information, the government chooses a desired treatment for each z_i, which has the interior optimality condition for each z_i of $[\alpha_g + \alpha_p]\, h'(a_i^{*\text{full info}}) = z_i$. Payments in the full-information scenario would take the physician's cost type as an argument and would extract all the surplus from the physician. The payment to a physician of cost type z_i would be $w_i^{*\text{full info}} = a_i^{*\text{full info}} z_i - \alpha_p h(a_i^{*\text{full info}})$.

information scenario would be

$$a_i^{*\text{full info}} = \frac{e_\tau - e_0}{\delta} - \frac{z_i}{\delta^2[\alpha_g + \alpha_p]}. \tag{7}$$

Unlike the equilibrium amount under the linear contract (6), the full-information amount for a physician of cost type z_i does not depend on the other cost types in the economy (i.e., μ_z and $\overline{z}$ do not appear in 7). The mean treatment amount is smaller under the optimal linear contract than in the full information benchmark, but there is a threshold in z such that cost types below the threshold provide more under the linear contract.[16] Thus, although the full-information solution features a higher average treatment level, the optimal linear contract results in over-provision by low-cost types and under-provision by high-cost types, because all types are given the same marginal incentive. The unrestricted (potentially nonlinear) contract presented next will not result in this kind of pooling, as it will allow the government to separate cost types by providing variable marginal incentives.

2.2 Optimal Unrestricted Contract

This section presents and discusses the optimal unrestricted contract for a given baseline hematocrit level e_0. (As we did earlier, for convenience we will suppress the dependence of the wage contract and other functions on e_0 when it is not confusing to do so.) Our solution approach draws on the price discrimination literature (Maskin and Riley, 1984), where we invoke the revelation principle to derive an optimal contract specifying treatments and payments. For details please see Appendix A.2.

Using our functional form assumptions, the optimal (i.e., "second-best") induced actions are

$$a_i = \frac{e_\tau - e_0}{\delta} - \frac{z_i + \frac{F_z(z_i)}{f_z(z_i)}}{\delta^2[\alpha_g + \alpha_p]}. \tag{8}$$

This is distorted away from the full-information solution in (7) due to the $\frac{F_z(z_i)}{f_z(z_i)}$ term, which is positive, hence the treatment amounts under the optimal unrestricted contract are lower than those in the full-information solution (except for the lowest-cost type, where $F_z(z_i) = 0$). The treatment amounts are decreasing in cost type if the hazard $\lambda_z \equiv \frac{f_z(z_i)}{F_z(z_i)}$ is decreasing in z; as shown in Appendix A.2, this allows the optimal contract to be (indirectly) implemented with a single, nonlinear payment schedule. As with the full-information solution, the optimal treatment amount is increasing in patient need (i.e., decreasing in e_0) and in government and

[16] The mean of (6) is $\frac{e_\tau - e_0}{\delta} - \frac{\overline{z}}{\delta^2[\alpha_g+\alpha_p]}$ while the mean of (7) is $\frac{e_\tau - e_0}{\delta} - \frac{\mu_z}{\delta^2[\alpha_g+\alpha_p]}$, and $\overline{z} \geq \mu_z$. Subtracting (7) from (6), we have $a_i^{*\text{linear}} > a_i^{*\text{full info}}$ iff $z_i > \mu_z + \frac{\alpha_p}{\alpha_g}[\mu_z - \overline{z}]$

physician valuations of patient health, and is decreasing in the physician's cost of treatment. Additionally, we note that higher degrees of physician altruism can blunt the distortion away from the first-best that is caused by the unobserved heterogeneity. This highlights the potential importance of allowing for both physician heterogeneity and physician altruism when characterizing optimal payment contracts.

The payments in the unrestricted contract can be derived from the optimal treatment amounts and the specification of physician utility (see Appendix A.2). The result is

$$w(a_i) = \left[\int_{i+1}^{\bar{i}} a_j dj \right] - [\alpha_p h(a_i) - a_i z_i]. \tag{9}$$

This expression involves the integral of actions by higher cost types (z_{i+i} to $\bar{z}$) because the additive separability in our utility specification allows the IC constraints to simplify in a way that yields this integral as the utility of cost type z_i in equilibrium (see Appendix A.3). Notably, as in Ellis and McGuire (1986), the payments in the optimal unrestricted contract result in partial cost sharing if $\alpha_p > 0$; i.e., if physicians are at all altruistic.[17]

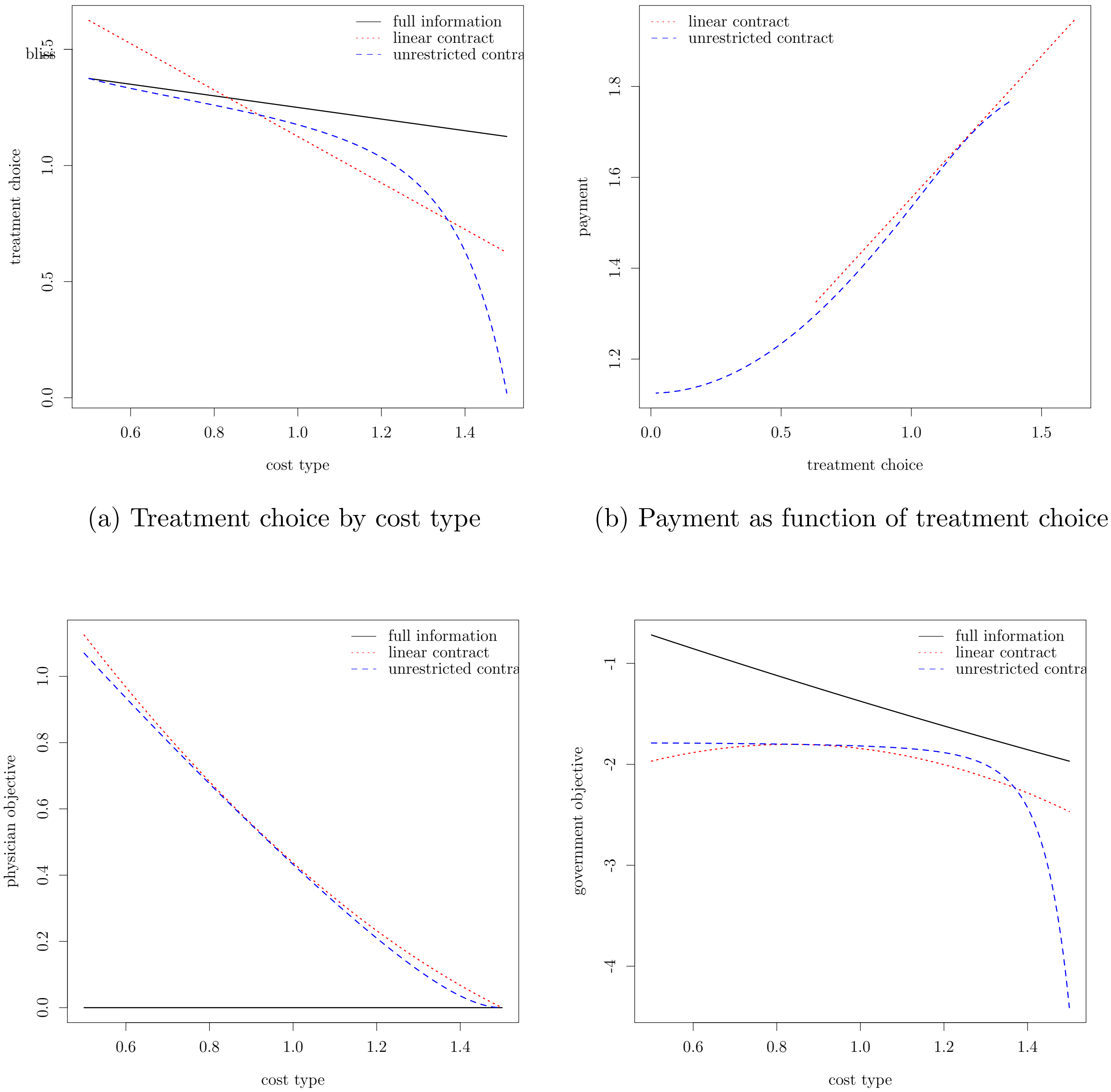
Figure 1 plots outcomes for full-information (black, solid line), linear contract (red, dotted line), and unrestricted contract (blue, dashed line) scenarios for a specific parameterization of the model, where physician cost types are distributed according to a truncated normal distribution.[18] Figure 1a shows the induced action (treatment amount) for each cost type under each scenario. Under full information (the "first-best"), the treatment amount decreases as the cost factor z increases. This is also true in both asymmetric information scenarios. However the linear contract, which offers the same marginal incentive to all types, leads to over-provision for low-cost types and under-provision for high-cost types. In contrast, the unrestricted contract, which separates types by providing variable marginal incentives, leads to under-provision for all but the lowest-cost type, $\underline{z}$. Incentives are optimally weakened for higher cost types, resulting in treatment choices that are lower than in the first-best, and the amount of this distortion relative to the full-information solution is increasing in cost type.

Figure 1b shows how the payment depends on the treatment amount under the linear and unrestricted contracts; in other words, this figure plots the payment schedules. Both schedules are upward-sloping, meaning that pure prospective payment is not optimal in this example, similar to Ellis and McGuire (1986). Also similar to Ellis and McGuire (1986), neither contract implements full cost sharing. The range of payments under the unrestricted contract differs from the range under the linear contract, because the nonlinear schedule provides lower marginal rates and induces lower treatment amounts among low-cost physicians,

[17] This is because $\frac{\partial w(a_i)}{\partial a_i} = z_i - \alpha_p h'(a_i)$, which is strictly less than the marginal cost z_i for $e_0 < e_\tau$.

[18] Specifically, $\alpha_p = 1$, $\alpha_g = 3$, $z \sim N(1, 1/16)$ with truncation points $1 \pm 1/2$.

Figure 1: Comparison of Outcomes in Model Simulation



(a) Treatment choice by cost type

(b) Payment as function of treatment choice

(c) Physician objective by cost type

(d) Government objective by cost type

and because it induces substantially lower treatment amounts among high-cost physicians as well. (However, as seen in Figure 1a, the unrestricted contract induces higher treatment amounts than does the linear contract among cost types in the middle of the range.)

The qualitative relationship between payment and treatment amounts depends on the asymmetric information. In a full-information scenario, the government could fully compensate high-cost types to implement the desired action, thereby paying them more than low-cost types and leading to a *negative* relationship between payment and treatment amounts.[19] However, under such a contract asymmetric information would enable low-cost types to obtain large surpluses by pretending to be high-cost types. Thus, incentive compatibility requires that payment and treatment amounts must both be lower for high-cost types, so that low-cost types enjoy a larger surplus (i.e., are paid more) by providing larger amounts. This results in a payment contract that increases with treatment amounts.

Finally, Figures 1c and 1d respectively plot the physician objective (u_i) and the government objective (u_g) against the cost type for each scenario. Since the government can extract all of the surplus under full information, all cost types receive zero utility in that scenario (Figure 1c). For the optimal linear and nonlinear contracts under asymmetric information, only the highest cost type receives zero utility (because the participation constraint only binds for this type), while all lower cost types receive positive surpluses. The physician surpluses are in fact quite similar under the linear and nonlinear contracts in this example. As can be seen from Figure 1d, the government naturally fares best under full information and second-best with the unrestricted contract.[20]

3 Data

Our data come primarily from Medicare outpatient claims from renal dialysis centers (freestanding or hospital-based) in 2008 and 2009, for the treatment of patients with ESRD. As noted earlier, these facilities provide dialysis treatment to patients multiple times per week, and claims are typically filed monthly. EPO can be administered at each visit, and each injection is individually listed as a separate line on the claim.

The raw 20% sample for 2008 and 2009 contains 1.4 million ESRD claims with 11.1 million claim lines for EPO or related medications.[21] Almost 90% of the claims bill for at

[19] This can be seen by implicitly differentiating the binding participation constraint, resulting in $\frac{\partial w_i^{*\text{full info}}}{\partial a_i^{*\text{full info}}} = -\left[z_i - \alpha_p h'(a_i^{*\text{full info}})\right]$, where we know the term in the brackets is weakly positive because the full-information treatment choice solves $[\alpha_g + \alpha_p]\, h'(a_i^{*\text{full info}}) = z_i$, and $\alpha_g \geq 0$.

[20] The substantial downward distortion of the treatment amounts from high-cost physicians results in lower values of the government's objective from these types, but this is in the far upper tail of the distribution.

[21] Epoetin alpha constitutes 97.6% of the claim lines for this class of medication in our sample. The

least one injection of these drugs (1.25 million claims). All claims with an injection include a baseline hematocrit level from the previous month (or a related hemoglobin level), but claims without an injection typically do not report a red blood cell level. As a consequence, for the present analysis we exclude claims without any injections of EPO.

The unit of analysis is the (typically) monthly claim, which reports the treatment given by provider i to patient j in period t.[22] The total amount of EPO injected over the month is the action, a_{ijt}, and the prior hematocrit level reported in the claim is the baseline hematocrit, $e_{0,jt}$.[23] The reimbursement rate, w_{1t}, is the national payment rate per 1,000 units of EPO for the quarter in which the claim was filed. These rates are listed in Medicare Part B Average Sales Price Drug Pricing Files available on the CMS website.[24] The claims also list the actual payments for each injection of EPO, so a claim-specific reimbursement rate can be computed. These actual reimbursement rates are highly correlated with the national payment rates.[25]

In order to avoid extreme outliers, which often reflect data entry errors, we remove observations where the baseline hematocrit ($e_{0,jt}$) is above its 99th percentile or below its 1st percentile, or where the amount of EPO (a_{ijt}) is above its 99th percentile. The final analytic sample has 1.1 million claims for 76,985 unique patients from 5,150 unique providers. The providers are defined as facilities, not individual physicians, because within each facility there are multiple doctors and nurses who jointly treat their patients.[26]

Table 1 provides summary statistics on the three main variables in our analytic sample. The average monthly dosage of EPO is 67.0 thousand units, with a relatively large standard deviation of 66.4 thousand units, and the average baseline hematocrit is 34.5 percent (volume percentage of red blood cells in the blood). The national reimbursement rate has a mean of \$9.26 per 1,000 units of EPO and ranges from a low of \$8.96 in 2008Q1 to a high of \$9.62 in 2009Q3. Table 1 also presents information on the distribution of acquisition costs for EPO from a separate source, the Renal Dialysis Facility Cost Reports. The Centers for Medicare and Medicaid Services (CMS) requires dialysis facilities to submit detailed annual cost reports, which include their total expenditures on EPO and the total number

alternative drug was darbepoetin alpha. For the present analysis we restrict to epoetin alpha because dosages and reimbursements differ between the two drugs.

[22] Note that we now use i to index providers instead of unobserved types.

[23] For claims that report hemoglobin rather than hematocrit, we use the standard rule of thumb of multiplying by three to convert the levels.

[24] See `https://www.cms.gov/Medicare/Medicare-Fee-for-Service-Part-B-Drugs/McrPartBDrugAvgSalesPrice/index.html`.

[25] In our analytic sample the correlation is 0.92, net of provider fixed effects.

[26] Also, many facilities belong to large, national chains (DaVita and Fresenius), but we treat the individual facilities separately.

Table 1: Summary Statistics

				Percentiles				
Variable		Mean	SD	10th	25th	50th	75th	90th
Monthly EPO dosage (1,000u)		67.0	66.4	8.8	20.0	45.0	90.0	156.2
Prior hematocrit level (%)		34.5	3.4	30	32.4	34.8	36.9	38.7
Reimbursement rate (\$/1000u)		9.26	0.24					
Addendum 1 – Percentiles of EPO acquisition costs from annual cost reports:								
Acquisition cost (\$/1000u)				7.13	7.23	7.53	8.15	9.11
Addendum 2 – Medicare reimbursement rate for EPO in each quarter:								
Reimb. rate (\$/1000u)	8.96 (2008Q1)	9.07 (Q2)	9.07 (Q3)	9.10 (Q4)	9.20 (2009Q1)	9.40 (Q2)	9.62 (Q3)	9.58 (Q4)

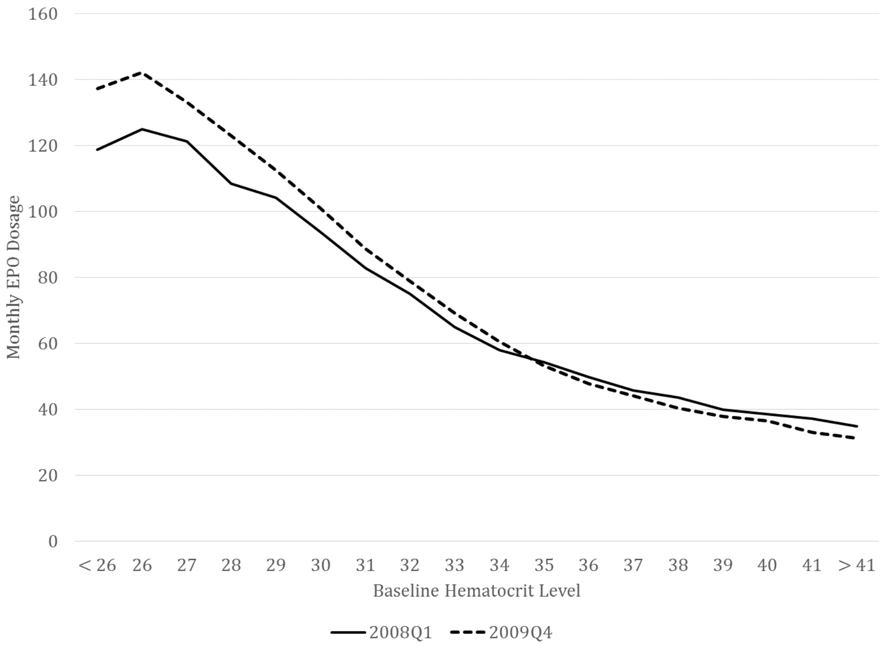
Notes: The EPO dosage and hematocrit come from Medicare outpatient claims data. The reimbursement rate comes from quarterly Medicare Part B ASP Drug Pricing Files for 2008 and 2009. This latter variable takes one of eight values depending on the quarter, as listed in Addendum 2, so we do not present its percentiles. The distribution of EPO acquisition costs shown in Addendum 1 is computed from Renal Dialysis Facilities Cost Report Data for 2008. We do not present the mean or standard deviation because extreme outliers in the cost report data make those statistics unreliable, compared to the percentiles.

of units provided. These data are publicly available on the CMS website.[27] From the total expenditures (less any rebates) and total units, we compute the average acquisition cost per 1,000 units for each facility in the cost report data for 2008. The percentiles listed in the table show non-trivial differences across providers in the acquisition cost, even though the drug was produced by a single manufacturer.

To present some preliminary evidence on the relationships between patient need, payment rates, and drug provision, Figure 2 plots average dosages of EPO as a function of baseline hematocrit in the first and last quarters of our analytic sample. In both periods the dosages decline with higher hematocrit levels, as would be expected, and the decline is steeper at lower levels. A notable difference between the two periods is that in 2009Q4, when the reimbursement rate was higher, average dosages are higher for patients with low hematocrit levels, and these dosages decrease more rapidly in relation to hematocrit. Our empirical model, discussed next, will use specifications designed to fit these qualitative features.

[27] `https://www.cms.gov/Research-Statistics-Data-and-Systems/Downloadable-Public-Use-Files/Cost-Reports/RenalFacility.html`

Figure 2: Mean Monthly Dosages of EPO in Relation to Baseline Level of Hematocrit



4 Empirical Implementation

We now describe how we adapt the model from Section 2 to our empirical context. The model naturally extends to an environment with many agents, each treating many patients, under the assumptions that the providers' utility functions and the government's objective function are additive across patients.[28]

The empirical specification extends the functional forms introduced in Section 2.1 by adding flexibility in relation to the patient's baseline hematocrit. Specifically we add $\delta(e_0)$ to the parameter δ, which allows the productivity of EPO to depend on the baseline level of hematocrit.[29] Also we add $\alpha(e_0)$ to the altruism parameter, so that the total weight on the patient's health in the physician's utility is $\alpha_p + \alpha(e_0)$. This means that physicians may have greater or lesser concern for the health of patients with different baseline levels of

[28] The static framework can be applied to to multiple time periods if there are no dynamic effects of EPO (as discussed in Section 1.2), and if the government does not use treatment histories in setting reimbursement rates. This has always been the case when patient hematocrit levels are within the recommended range. However, the Medicare payment policy during our analysis period paid reduced rates for EPO given to patients who had high hematocrit levels for three consecutive months (over 39%).

[29] This allows us to approximate diminishing returns in a simple fashion that maintains our closed-form solutions.

hematocrit.[30] The physician's utility function is thus

$$-[\alpha_p + \alpha(e_0)]\frac{1}{2}\left[[\delta + \delta(e_0)]a + e_0 - e_\tau\right]^2 - az_i + w_0 + w_1 a.$$

The parameters $\alpha(e_0)$ and $\delta(e_0)$ are specified take different values over certain intervals of e_0, noted below, which makes this a piecewise quadratic function of a and e_0.

Given this utility function, the optimal treatment amount (at an interior solution) is

$$a_i^* = \frac{[e_\tau - e_0]}{\delta + \delta(e_0)} + \frac{w_1 - z_i}{[\alpha_p + \alpha(e_0)][\delta + \delta(e_0)]^2}. \tag{10}$$

The treatment policy function is thus piecewise linear, with the segments corresponding to the intervals in e_0. As we will show, this specification has the advantages of maintaining closed-form solutions while having sufficient flexibility to fit the qualitative features seen in Figure 2.

To allow for unexplained variation from the econometrician's perspective we add an independent, mean-zero shock η. We also decompose the individual cost type as $z_i = \mu_z + \zeta_i$. The observed treatment amount provided by physician i to patient j at time t is then

$$a_{ijt} = \frac{[e_\tau - e_{0jt}]}{\delta + \delta(e_{0jt})} + \frac{w_{1t} - [\mu_z + \zeta_i]}{[\alpha_p + \alpha(e_{0jt})][\delta + \delta(e_{0jt})]^2} + \eta_{ijt}, \tag{11}$$

where e_{0jt} is patient j's baseline hematocrit in period t and w_{1t} is the national reimbursement rate for EPO in period t. Within each interval of baseline hematocrit, in which the parameters $\alpha(e_0)$ and $\delta(e_0)$ are constant, (11) is linear in e_0 and w_1.

4.1 Identification

To derive an optimal contract we need values for the following structural parameters: α_p and $\alpha(e_0)$ (altruism weights), δ and $\delta(e_0)$ (productivity of EPO), and e_τ (target hematocrit level). We also need the distribution of z, because the mean and maximum (μ_z and $\overline{z}$) appear in the optimal linear contract and the inverse hazard ratio ($F_z(z)/f_z(z)$) appears in the optimal unrestricted contract. These parameters and the distribution of z can be identified from the reduced form (11), except that we require additional data on costs to establish the mean value of z (described below).

[30] We still refer to physicians when discussing the model, but to be clear, we use facilities as the providers in the empirical analysis.

The reduced form can be re-expressed as follows:

$$a_{ijt} = \sum_k 1\{e_{0jt} \in E^k\} \left[\beta_0^k + \beta_1^k e_{0jt} + \beta_2^k w_{1t}\right] + \nu_i + \epsilon_{ijt}, \tag{12}$$

where $1\{e_{0jt} \in E^k\}$ indicates that the baseline hematocrit is in interval E^k. This is a piecewise linear regression model with provider fixed effects. The parameters $\delta + \delta(E^k)$ and $\alpha_p + \alpha(E^k)$ are recovered from the coefficients of this regression as follows:

$$\delta + \delta(E^k) = -\frac{1}{\beta_1^k} \quad \text{and} \quad \alpha_p + \alpha(E^k) = \frac{[\beta_1^k]^2}{\beta_2^k},$$

with the $\delta(E^k)$ and $\alpha(E^k)$ in one interval normalized to zero.[31] The result for the altruism parameters has intuitive appeal: it is the ratio of the responsiveness to a measure of patient need (squared) to the responsiveness to remuneration. The intercept in each interval, β_0^k, identifies a linear combination of e_τ and μ_z, and the distribution of the provider fixed effects identifies the distribution of deviations from the mean cost. We use external data, from the Renal Dialysis Facility Cost Reports, to identify the mean per-unit cost, μ_z. Specifically we take the median value reported in Table 1 as the value of μ_z, which is equal to \$7.53 per 1,000 units.[32] This then completes the identification of the distribution of z (from the fixed effects) and the target level of hematocrit e_τ (from the intercepts).

The identification of the structural parameters also depends on the consistency of the reduced-form coefficient estimates. There are two variables in the reduced form (12): the baseline hematocrit (e_{0jt}) and the national payment rate (w_{1t}). These must be uncorrelated with the error term (ϵ_{ijt}) net of the provider fixed effects (ν_i). The baseline hematocrit satisfies this so long as any selection of patients to providers did not change over the two years of our estimation period. That makes the within-provider variation in e_{0jt} across patients and over time exogenous.[33] The national payment rate, as described in Section 1.2, was determined by the national average price of EPO roughly six months earlier. An

[31] The coefficients of the reduced form are the following combinations of the structural parameters:

$$\beta_0^k = \frac{e_\tau}{\delta+\delta(E^k)} - \frac{\mu_z}{[\alpha_p+\alpha(E^k)][\delta+\delta(E^k)]^2} \qquad \beta_1^k = -\frac{1}{\delta+\delta(E^k)} \qquad \beta_2^k = \frac{1}{[\alpha_p+\alpha(E^k)][\delta+\delta(E^k)]^2}.$$

[32] We use the median rather than the mean because it is less sensitive to extreme outliers in the cost report data. Also to be clear, the cost reports reflect the acquisition cost but not any additional costs of administering the drug. The labor and equipment costs for injecting EPO are small relative to its acquisition cost but may nevertheless be non-trivial. In that case μ_z represents a lower bound on the full per-unit cost.

[33] A more subtle concern is the fact that the hematocrit level from the prior month reflects the treatment given in the prior month. So long as each provider's treatment protocols are stable during our analysis period, the provider fixed effects would address this as well.

Table 2: Reduced-Form Coefficient Estimates

	Interval of Baseline Hematocrit					
Coefficient	Up to 27,	> 27 to 30,	> 30 to 33,	> 33 to 36,	> 36 to 39,	> 39
β_0^k	352.1	173.7	281.1	229.3	153.8	28.3
Constant term	(46.0)	(51.3)	(28.2)	(18.7)	(17.9)	(25.1)
β_1^k	-6.20	-7.56	-9.48	-6.70	-4.00	-2.06
Baseline HCT	(0.14)	(0.37)	(0.20)	(0.13)	(0.13)	(0.16)
β_2^k	-7.20	17.09	11.45	7.22	4.97	10.40
Reimb. rate	(5.21)	(5.64)	(3.10)	(2.03)	(1.94)	(2.76)
Obs. in interval	97,983	82,640	216,391	379,530	267,245	83,344

Notes: Estimates are from a single fixed-effects regression where variables are interacted with indicators for the listed ranges in baseline hematocrit (HCT). Standard errors in parentheses, clustered by provider.

individual facility could not affect the national average price, but if demand shocks were substantially correlated across facilities and over time, there could be a bias. However we include a year dummy for 2009 and month dummies for each calendar month. These would address both secular and cyclical trends in demand. Also, dialysis facilities were not the only purchasers of EPO because the drug was also widely used in chemotherapy. Thus there are two potential sources of exogenous variation in the lagged prices that determined w_{1t}: demand shocks from chemotherapy providers and supply shocks from the drug manufacturer.

4.2 Estimation

We estimate (12) via standard fixed effects estimation with provider fixed effects. As noted above, the regression also includes separate effects for year and month (not year x month). The year effect(s) capture possible secular changes, while the month effects account for billing behavior that may change throughout the year (e.g., a spike in claims dated December 31). Table 2 presents the estimates of the main coefficients. These come from a single regression with separate coefficients for each of the listed intervals of baseline hematocrit. For example, in the interval from 30 to 33, a patient with one unit higher hematocrit (say 32 vs. 31) receives 9,480 less units of EPO per month on average. Also for patients in the interval from 30 to 33, a one dollar increase in the reimbursement rate (per 1,000 units) would induce providers to increase dosages by 11,450 units per month on average.

The intervals of baseline hematocrit were chosen to reflect treatment guidelines and

Table 3: Structural Parameter Estimates

	Interval of Baseline Hematocrit					
Parameter	Up to 27,	> 27 to 30,	> 30 to 33,	> 33 to 36,	> 36 to 39,	> 39
Providers' value of health in dollars						
$\alpha_p + \alpha(E^k)$	-5.34	3.34	7.84	6.22	3.22	0.41
	(3.89)	(1.15)	(2.15)	(1.77)	(1.28)	(0.12)
Increase in HCT from 1000u EPO						
$\delta + \delta(E^k)$	0.161	0.132	0.105	0.149	0.250	0.486
	(0.004)	(0.007)	(0.002)	(0.003)	(0.008)	(0.037)
Implied HCT target using $\mu_z = \$7.53$						
e_τ	48.0	40.0	38.8	42.3	47.8	51.9
	(1.3)	(1.2)	(0.5)	(0.4)	(0.7)	(2.1)

Notes: Values are recovered from reduced form coefficients, using the value of $\mu_z = \$7.53$, as described in Section 4.1. Delta method standard errors in parentheses.

policies in place at the time, and to balance the flexibility of the specification with the precision of the estimates for each interval. The FDA approved EPO for use in patients with hematocrit between 30 and 36, and Medicare reduced the reimbursement rate for EPO provided to patients with hematocrit above 39.[34] Using 3-point intervals divides this range from 30 to 39 evenly, and the estimation results indicate that this width allows sufficient power within each interval while maintaining flexibility globally. We consider the estimates in the intervals from 30 up to 39 to be the most reliable, because these are the most common clinically (over 80% of the observations are in these intervals) and because this range has at least implicit approval from the FDA or Medicare policy. Within this range there is decreasing responsiveness to both baseline hematocrit and the reimbursement rate, going from lower to higher intervals. This matches the decreasing magnitude of the slopes seen in Figure 2.

Table 3 presents the structural parameters, recovered as described above. Across the intervals from 30 to 39, the altruism weight decreases while the productivity of EPO increases. The former could be interpreted as a lower concern for the health of patients with less severe anemia. The latter is consistent with diminishing marginal productivity of EPO.[35] The magnitudes of the altruism parameters imply that, for example in the interval from 30 to

[34] Medicare did not use 36% as the cutoff for payment reductions because of the acknowledged difficulty in maintaining a target level exactly over time.

[35] Patients with lower baseline hematocrit are given more EPO on average, and the estimates of $\delta + \delta(E^k)$ reflect the average productivity of EPO for patients with baseline hematocrit in each range.

33, physicians would require \$7.84 to compensate for providing an amount of EPO intended to achieve a hematocrit above or below the target level by 1 percentage point. The estimate of $\delta + \delta(E^k)$ in this interval implies that 1,000 units of EPO raises hematocrit by 0.105 percentage points. This and the other estimates of the average productivity of EPO across the intervals from 30 to 39 are consistent with estimates obtained from clinical trials.[36] The values of the implied hematocrit target are also reasonably close to what would be expected based on clinical and policy guidelines.[37]

5 Results: Optimal Contracts

We now present our main empirical results: optimal contracts obtained with the parameters recovered above, and simulated outcomes under those contracts. First we compute the optimal linear contract given by (5) and use it to simulate counterfactual dosage amounts for the entire sample. We then construct the optimal unrestricted contract for a particular value of baseline hematocrit (the median), and compare this contract and its induced outcomes with the linear contract and the observed contract (i.e., the actual reimbursement rates).

Generating the optimal contracts requires fixing a value for α_g, the weight placed by the government on health relative to money. However this parameter could not be recovered without making further assumptions, for example that the observed payment policy is optimal, an assumption we do not want to make. Instead, for the linear contract, we consider two possible values for α_g that span a wide range: specifically, $\alpha_g = \alpha_p$ as the low value and $\alpha_g = 10\alpha_p$ as the high value. For the unrestricted contract, we set the value of α_g so that the mean level of health produced under the contract matches the mean level under the observed contract (the implied value is $\alpha_g = 4.4\alpha_p$).[38] Additionally, for the distribution of z, because the extremes of a nonparametrically estimated distribution can be highly sensitive to outliers, we trim the recovered distribution at the 5th and 95th percentiles.[39]

Table 4 presents the results for the optimal linear contract. The observed reimbursement

[36] The average dosages and the average increases from initial hemoglobin levels reported in Singh et al. (2006) imply average productivities of 0.143 and 0.167 for the two treatment groups in the study (our calculations). Also, Tonelli et al. (2003) construct a dose-response curve based on results from five other clinical trials, which indicates average marginal productivities ranging from 0.135 to 0.241 for hematocrit levels in this range.

[37] These implied targets are slightly higher than one would expect based on the FDA and Medicare policies (e.g., 36 or 39). This indicates a misspecification in the model, for example in the quadratic loss around the target, but also that the misspecification is mild because the implied targets are reasonably close to the expected amounts.

[38] A similar exercise with the optimal linear contract would be less meaningful because different values of w_1^* implied by different values of α_g shift all dosages up or down by a uniform amount—see (3).

[39] This results in a value of $\bar{z} = 12.03$ (\$/1,000 units of EPO), rather than 24.47.

Table 4: Simulation of Average Monthly Dosages of EPO under Linear Contracts

Scenario	Interval of Baseline Hematocrit					
(reimb. rate)	Up to 27,	> 27 to 30,	> 30 to 33,	> 33 to 36,	> 36 to 39,	> 39
Observed reimbursement rates						
(\$8.96–\$9.62)	79.6	114.2	85.5	62.1	48.6	39.7
Optimal linear contract, $\alpha_g = 10\alpha_p$						
(\$6.44)	99.8	65.9	53.2	42.0	35.0	13.8
Optimal linear contract, $\alpha_g = \alpha_p$						
(\$1.54)	135.0	2.4	6.7	11.4	13.6	0.2

Notes: Predicted monthly dosages are generated for each observation in the sample using the estimated version of (12), with the payment rate set to the observed amounts, or \$6.44, or \$1.54, as indicated.

rates (w_{1t}) range from \$8.96 to \$9.62 over time. The optimal reimbursement rate (w_1^*) in the linear contract with $\alpha_g = 10\alpha_p$ is \$6.44, and with $\alpha_g = \alpha_p$ is \$1.54.[40] The fixed payment (w_0^*) varies with the patient's baseline hematocrit (e.g., for the median level, $w_0^* =$ \$168), but this does not affect the simulated dosages because only the marginal payment affects the physician's optimal amount (at an interior solution).[41] We then use the estimated reduced form (12) to predict a dosage for each observation (i.e., for each patient in each month) under each contract.[42] With the observed reimbursement rates the average predicted monthly dosages for patients within each interval of baseline hematocrit match the average observed dosages (by construction). With the higher counterfactual rate (\$6.44 per 1,000 units) dosages are reduced by about 30–40%, and with the lower rate they are reduced even further. Interestingly, with the higher rate the proportional reductions are largest in the top interval of baseline hematocrit, where Medicare sought to discourage the use of EPO.

Next we derive an optimal unrestricted contract that maintains the mean level of health produced under the observed contract while having lower total payments and less variation in dosages.[43] This shows the extent to which expenditures can be reduced with a nonlinear contract, which provides better incentives for a heterogeneous population of agents. The contract is computed for the median level of baseline hematocrit ($e_0 = 34.8$), and is compared

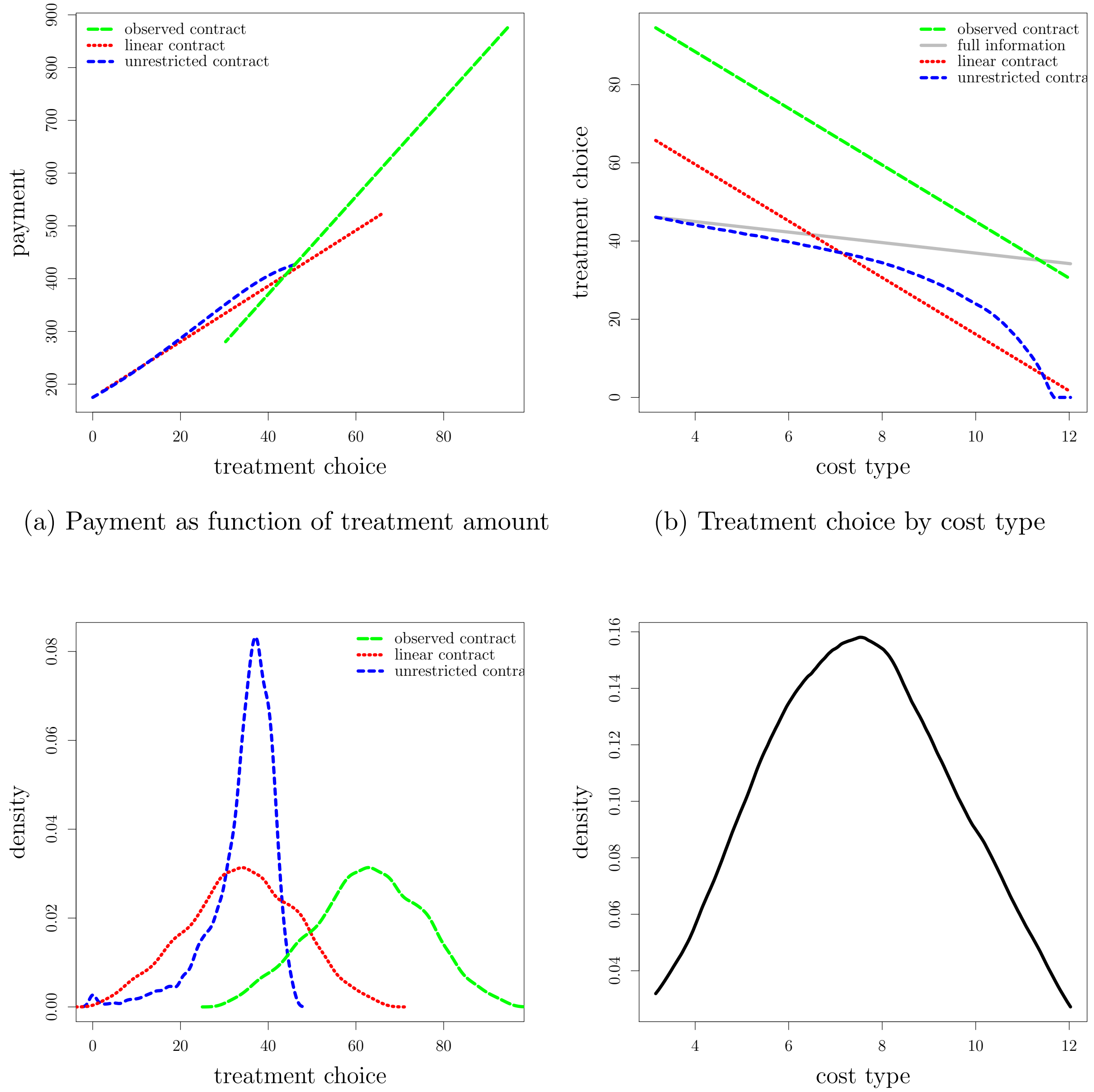
[40] Although α_p varies across the intervals of baseline hematocrit, the reimbursement rate does not vary because we have set α_g to be proportional to α_p, which simplifies the formula for w_1^* in (5).

[41] As noted earlier, our analysis of the linear contract assumes interiority to provide relatively simple results, but the analysis of the unrestricted contract does not impose this.

[42] Negative predicted dosages are replaced with 0.

[43] As noted above, this implies a value for the government's weight on health ($\alpha_g = 4.4\alpha_p$).

Figure 3: Optimal Nonlinear Contract to Match Average Level of Health



(a) Payment as function of treatment amount

(b) Treatment choice by cost type

(c) Distribution of treatment amounts

(d) Distribution of cost types

Notes: Figure plots treatment and payment amounts under an optimal nonlinear contract that results in the same average health as does the observed contract. An optimal linear contract and the observed contract are shown for comparison. The contracts and outcomes are computed for a patient with the median level of baseline hematocrit (34.8). The value of α_g was calibrated to match the average equilibrium value of health (h) under the nonlinear contract to the average value of health under the observed contract. The observed contract uses the sample mean of the payment rate ($9.26). Panel (a) shows the payment contract (i.e., total payment as a function of the treatment amount) for the observed contract and the optimal linear and unrestricted contracts derived from the estimated model. Panel (b) shows the treatment amounts chosen under these contracts as a function of the physician cost type. Panel (c) shows the distribution of treatment amounts and panel (d) shows the distribution of cost types.

Table 5: Summary of Outcomes under Optimal Contracts

Contract	Mean Pmt	Mean Dosage	SD Dosage
Observed	\$582	62.9	12.5
Linear	\$355	34.0	12.5
Unrestricted	\$373	34.3	7.3
Full Information	\$306	40.2	2.3

Note: Table shows summary statistics of outcomes plotted in Figure 3. Mean and SD of dosage are in 1,000 units/month.

with the corresponding optimal linear contract and with the observed contract.[44]

Figure 3 presents the results, with the unrestricted contract in blue (dashed lines), the linear contract in red (dotted lines), and the observed contract in green (long-dashed lines). Figure 3a plots the contracts themselves; i.e., how the total payment depends on the treatment amount. Figure 3c underneath shows the distribution of treatment amounts provided under each contract (to patients with $e_0 = 34.8$). The optimal unrestricted contract has a roughly similar average payment rate as the optimal linear contract, but the ability to provide different marginal incentives for different dosages results in a much more compressed distribution of treatment amounts under the nonlinear contract. The observed contract is steeper than the optimal linear contract because it has a higher per-unit rate, but the latter includes a nonzero intercept and pays more for dosages below a certain amount.

Figure 3b shows how treatment amounts are related to physician cost types under each contract (panel d underneath plots the distribution of cost types). The full-information solution is included for comparison (gray, solid line). The treatment amounts under the observed contract are typically too high, exceeding first-best, full-information dosages for all but the highest cost types. The optimal linear contract offers a uniformly lower payment rate, so dosages under this contract lie below those under the observed contract and the two move in parallel. Treatment amounts under the optimal unrestricted contract, which better separates physician types, are much closer to the full-information solution over most of the distribution of cost types. This indicates the importance of the subtle differences between the optimal linear and nonlinear contracts seen in Figure 3a. Last, in contrast to both the observed and optimal linear contracts, the optimal unrestricted contract leads to mild under-provision for most cost types, and the highest cost types choose not to provide the drug.

Table 5 summarizes the outcomes under these contracts (again, for patients with the median level of hematocrit). With the observed contract, the predicted mean payment is \$582

[44] The linear contract uses the same value of α_g; it does not hold mean health constant. The observed contract has $w_0 = 0$ and sets w_1 equal to the sample mean of the actual reimbursement rates, \$9.26.

and the predicted mean dosage is 62,900 units of EPO per month. The optimal linear contract reduces the mean payment by 39% and the mean dosage by 46%. The optimal linear contract does not change the variation in dosages, however, because it provides a constant marginal incentive just like the observed contract. The optimal unrestricted contract, by contrast, is nonlinear and induces a substantial reduction in the variation in treatment amounts across physicians. The standard deviation of the dosages falls by 41%, while the mean dosage is essentially the same as with the optimal linear contract. The optimal unrestricted contract is thereby able to achieve a substantial reduction in expenditures while maintaining the same average health as under the observed contract. With the optimal unrestricted contract, the average payment falls from \$582 to \$373, a 36% decrease.[45] Finally, for comparison, the full information scenario indicates what could be obtained in the absence of asymmetric information about physician costs. The mean dosage is slightly higher than under the optimal contracts, while the mean payment is lower (because all physicians receive zero surplus). Some variation in treatment amounts remains, which reflects the physicians' actual marginal costs.

6 Summary and Conclusions

In this paper we analyze optimal contracts for agents with partial altruism and hidden types. We do this for a particularly important sector of the economy, health care, where agents' responses to incentives can have important impacts on both health and costs. We are able to do this using a simple, parsimonious principal-agent model where doctors care about their patients' health and about their own profits, and the government cares about patients' health and program cost.

We specify and estimate a simple reduced form linear model that allows us to recover the underlying structural parameters. We find that doctors are imperfectly altruistic, and that there is substantial cost heterogeneity among them (unobservable to the government). We then derive the optimal linear and the optimal unconstrained (nonlinear) contracts, using the structural parameter estimates.

We calculate the impacts that using the optimal nonlinear contract would have on spending and on health. We find that Medicare could realize large (36%) cost savings while maintaining patient health at the same mean level as under the current payment system by using the optimal unrestricted contract, and drastically reduce variation in treatment across

[45] The mean payment is larger under the nonlinear contract compared to the linear contract, but this relatively small additional expenditure yields better patient health because of the reduction in treatment variation; details are available upon request.

patients (a 41% decrease in the standard deviation of drug dosages).

Importantly, we also find that optimal contracts are not fully prospective—physicians are optimally paid partially on a retrospective basis, and the optimal contract involves partial cost sharing.

These findings provide new empirical evidence on the structure of optimal contracts under asymmetric information, given agents' actual behavior. They can also inform policy decisions about physician payment. As we have shown, contract form can have very consequential impacts on both treatment and costs.

References

Abito, J. M., "Welfare Gains From Optimal Pollution Regulation," *Working Paper*, 2017.

Chalkley, M. and J. M. Malcomson, "Contracting for Health Services when Patient Demand does not Reflect Quality," *Journal of Health Economics*, 17(1):1 – 19, 1998, ISSN 0167-6296, doi: https://doi.org/10.1016/S0167-6296(97)00019-2.

—, "Government Purchasing of Health Services," in A. J. Culyer and J. P. Newhouse, eds., "Handbook of Health Economics," vol. 1 of *Handbook of Health Economics*, pp. 847 – 890, Elsevier, 2000, doi: https://doi.org/10.1016/S1574-0064(00)80174-2.

Chandra, A., D. Cutler and Z. Song, "Who Ordered That? The Economics of Treatment Choices in Medical Care," *Handbook of Health Economics*, 2:397–432, 2012.

Chiappori, P.-A., B. Jullien, B. Salanié and F. Salanié, "Asymmetric Information in Insurance: General Testable Implications," *RAND Journal of Economics*, 37(4):783–798, 2006.

Chiappori, P.-A. and B. Salanié, "Testing Contract Theory: A Survey of Some Recent Work," in M. Dewatripont, L. P. Hansen and S. Turnovky, eds., "Advances in Economics and Econometrics: Eighth World Congress," vol. 1, pp. 115–149, Cambridge University Press, 2003.

Choné, P. and C.-t. A. Ma, "Optimal Health Care Contract Under Physician Agency," *Annals of Economics and Statistics/Annales d'Économie et de Statistique*, pp. 229–256, 2011.

Clemens, J. and J. D. Gottlieb, "Do Physicians' Financial Incentives Affect Medical Treatment and Patient Health?" *American Economic Review*, 104(4):1320–49, 2014, doi: 10.1257/aer.104.4.1320.

De Fraja, G., "Contracts for Health Care and Asymmetric Information," *Journal of Health Economics*, 19(5):663–677, 2000.

Deneckere, R. and S. Severinov, "Multi-dimensional Screening: A Solution to a Class of Problems," Tech. rep., mimeo, econ. ucsb. edu, 2015.

Einav, L., A. Finkelstein and J. Levin, "Beyond Testing: Empirical Models of Insurance Markets," *Annual Review of Economics*, 2(1):311–336, 2010.

Elliott, S., E. Pham and I. C. Macdougall, "Erythropoietins: A Common Mechanism of Action," *Experimental Hematology*, 36(12):1573–1584, 2008.

Ellis, R. P. and T. G. McGuire, "Provider Behavior under Prospective Reimbursement: Cost Sharing and Supply," *Journal of Health Economics*, 5(2):129–151, 1986.

Gagnepain, P. and M. Ivaldi, "Incentive Regulatory Policies: The Case of Public Transit Systems in France," *RAND Journal of Economics*, pp. 605–629, 2002.

GAO, "End-Stage Renal Disease: Bundling Medicare's Payment for Drugs with Payment for All ESRD Services Would Promote Efficiency and Clinical Flexibility," Tech. Rep. GAO-07-77, U.S. Government Accountability Office, Washington, DC, 2006.

Gayle, G.-L. and R. A. Miller, "Has Moral Hazard Become a More Important Factor in Managerial Compensation?" *American Economic Review*, 99(5):1740–1769, 2009.

Gaynor, M. and M. Pauly, "Compensation and Productive Efficiency in Partnerships: Evidence From Medical Groups Practice," *Journal of Political Economy*, pp. 544–573, 1990.

Gaynor, M., J. Rebitzer and L. Taylor, "Physician Incentives in Health Maintenance Organizations," *Journal of Political Economy*, pp. 915–931, 2004.

Godager, G. and D. Wiesen, "Profit Or Patients' Health Benefit? Exploring The Heterogeneity In Physician Altruism," *Journal of Health Economics*, 32(6):1105–1116, 2013.

Grieco, P. L., R. C. McDevitt et al., "Productivity and Quality in Health Care: Evidence from the Dialysis Industry," *Review of Economic Studies*, 84(3):1071–1105, 2017.

Jack, W., "Purchasing Health Care Services From Providers With Unknown Altruism," *Journal of Health Economics*, 24(1):73–93, 2005.

Jacques, L. B., T. S. Jensen, J. Rollins, K. Long and E. Koller, "Final Decision Memorandum for CAG # 00413N: Erythropoiesis Stimulating Agents (ESA) for Treatment of Anemia in Adults with CKD Including Patients on Dialysis and Patients not on Dialysis," Tech. Rep. CAG # 00413N, Centers for Medicare and Medicaid Services, Washington, DC, 2011.

Malcomson, J. M., "Supplier Discretion Over Provision: Theory and an Application to Medical Care," *RAND Journal of Economics*, 36(2):412–432, 2005.

Maskin, E., J. J. Laffont, J. Rochet, T. Groves, R. Radner and S. Reiter, *Optimal Nonlinear Pricing with Two-Dimensional Characteristics*, pp. 256–266, University of Minnesota Press, Minneapolis, 1987.

Maskin, E. and J. Riley, “Monopoly with Incomplete Information,” *RAND Journal of Economics*, 15(2):171–196, 1984.

McGuire, T. G., “Physician Agency,” *Handbook of Health Economics*, 1:461–536, 2000.

Paarsch, H. J. and B. Shearer, “Piece Rates, Fixed Wages, and Incentive Effects: Statistical Evidence from Payroll Records,” *International Economic Review*, 41(1):59–92, 2000.

Singh, A. K., L. Szczech, K. L. Tang, H. Barnhart, S. Sapp, M. Wolfson and D. Reddan, “Correction of Anemia with Epoetin Alfa in Chronic Kidney Disease,” *New England Journal of Medicine*, 355(20):2085–2098, 2006, doi: 10.1056/NEJMoa065485, pMID: 17108343.

Tonelli, M., W. C. Winkelmayer, K. K. Jindal, W. F. Owen and B. J. Manns, “The Cost-effectiveness of Maintaining Higher Hemoglobin Targets with Erythropoietin in Hemodialysis Patients,” *Kidney International*, 64:295–304, 2003.

Wilson, R. B., *Nonlinear Pricing*, Oxford University Press on Demand, 1993.

Wolak, F. A., “An Econometric Analysis of the Asymmetric Information, Regulator-Utility Interaction,” *Annales d’Economie et de Statistique*, pp. 13–69, 1994.

A Model Details

A.1 Linear Contract with Cost Heterogeneity

The government's problem in (4) maximizes patient health, net the cost of transfers, subject to using a linear contract. Note that we can remove all participation constraints except for that for physician cost type $\overline{z}$ because $u_p(a, z)$ is decreasing in z; i.e., the only participation constraint that binds is that of the highest cost type. Using interior physician's policy functions, government's problem (4) can be rewritten as

$$\max_{\{(w_0,w_1)\in\mathbb{R}^2\}} \int_{\underline{z}}^{\overline{z}} \left[u_g(a) - w_0 - w_1 a(z, w)\right] f(z)dz \tag{13}$$

$$\begin{aligned} &\text{s.t.} \\ &u_{\overline{z}}(a^*(\overline{z}, w), w) \geq 0, && \text{VP} \\ &a^*(z, w) = \frac{e_\tau - e_0}{\delta} + \frac{w_1 - z}{\delta^2 \alpha_p}, \quad \forall z && \text{IC.} \end{aligned}$$

Setting up the Lagrangian based on this constraint and using interior physician's policy functions we obtain

$$\begin{aligned} \mathcal{L} = &\int_{\underline{z}}^{\overline{z}} \left[\alpha_g(\frac{-(w_1 - z)^2}{2\delta^2\alpha_p^2}) - w_0 - w_1(\frac{(e_\tau - e_0)}{\delta} + \frac{w_1 - z}{\delta^2\alpha_p})\right] f(z)dz \\ &+ \mu \left[\frac{(w_1 - \overline{z})^2}{2\delta^2\alpha_p} + \frac{(e_\tau - e_0)(w_1 - \overline{z})}{\delta} + w_0\right]. \end{aligned}$$

First-order conditions with respect to w_0 and w_1 yield the following system of equations:

$$\int_{\underline{z}}^{\overline{z}} [-f(z)dz] + \mu = 0 \Rightarrow \mu = F(\overline{z}) = 1$$

$$\int_{\underline{z}}^{\overline{z}} \left[-\alpha_g \left[\frac{w_1 - z}{\delta^2\alpha_p^2}\right] - \left[\frac{(e_\tau - e_0)}{\delta} + \frac{w_1 - z}{\delta^2\alpha_p}\right] - \frac{w_1}{\delta^2\alpha_p}\right] f(z)dz + \mu \left[\frac{w_1 - \overline{z}}{\delta^2\alpha_p} + \frac{e_\tau - e_0}{\delta}\right] = 0.$$

The first part of the second equation has to do with the average of the cost type distribution while the last term of the second equation, following the multiplier μ, pertains to the binding participation constraint of the highest-cost-type. Using $\mu = 1$, from the first equation, the second equation can be simplified further to

$$
\begin{aligned}
&-\frac{[\alpha_g+\alpha_p]w_1}{\delta^2\alpha_p^2}-\frac{\overline{z}}{\delta^2\alpha_p}+\int_{\underline{z}}^{\overline{z}}\left[\frac{[\alpha_g+\alpha_p]z}{\delta^2\alpha_p^2}f(z)dz\right]=0\\
&\Rightarrow w_1^*=\int_{\underline{z}}^{\overline{z}}zf(z)dz-\frac{\alpha_p\overline{z}}{\alpha_p+\alpha_g}=\mu_z-\frac{\alpha_p}{\alpha_p+\alpha_g}\overline{z}
\end{aligned}
$$

The next step is to characterize w_0^*, which we do using the binding participation constraint of $\overline{z}$:

$$
\begin{aligned}
u_{\overline{z}}(a^*(\overline{z},w),w)&=\frac{(w_1-\overline{z})^2}{2\delta^2\alpha_p}+\frac{(e_\tau-e_0)(w_1-\overline{z})}{\delta}+w_0=0\\
\Rightarrow w_0^*&=-\frac{w_1-\overline{z}}{\delta}\left[(e_\tau-e_0)+\frac{w_1-\overline{z}}{2\delta\alpha_p}\right]
\end{aligned}
$$

Plugging in $w_1^*=\mu_z-\frac{\alpha_p}{\alpha_p+\alpha_g}\overline{z}$ one can characterize w_0 as

$$
w_0=\frac{\left[2+\frac{\alpha_g}{\alpha_p}\right]\overline{z}-\left[1+\frac{\alpha_g}{\alpha_p}\right]\mu_z}{\delta\left[1+\frac{\alpha_g}{\alpha_p}\right]}\left[\frac{e_\tau-e_0}{\delta}+\frac{\mu_z\left[1+\frac{\alpha_g}{\alpha_p}\right]-\left[2+\frac{\alpha_g}{\alpha_p}\right]\overline{z}}{2\delta\left[\alpha_p+\alpha_g\right]}\right]
$$

A.1.1 Comparative Statics

Comparative statics for w_1^* Based on equation (5) it is clear that

$$
\begin{aligned}
&\frac{\partial w_1^*}{\partial\overline{z}}<0\\
&\frac{\partial w_1^*}{\partial\alpha_g}>0\\
&\frac{\partial w_1^*}{\partial\alpha_p}=-\frac{\mu\alpha_g}{[\alpha_p+\alpha_g]^2}<0.
\end{aligned}
$$

Comparative statics for w_0^* To do comparative statics for w_0^*, we first define the indirect utility function for a physician of type z, $v_p(z;w)$, which is obtained by plugging the interior solution of the physician's action into the utility function. Note that this constraint for the highest cost type is binding under the optimal linear contract, i.e., they get no surplus. Thus,

$$
v_p(\overline{z};w_0^*,w_1^*)=\frac{[w_1^*-\overline{z}]^2}{2\delta^2\alpha_p}+\frac{[e_\tau-e_0][w_1^*-\overline{z}]}{\delta}+w_0^*=0.
$$

We use the binding constraint on the indirect utility function for the highest cost type for the comparative static analysis.

$$\frac{\partial v_p(\overline{z}; w_0^*, w_1^*)}{\partial \overline{z}} = 0 \iff \frac{\partial w_0^*}{\partial \overline{z}} - 2\frac{w_1^* - \overline{z}}{2\delta^2 \alpha_p} + 2\frac{w_1^* - \overline{z}}{2\delta^2 \alpha_p}\frac{\partial w_1^*}{\partial \overline{z}} - \frac{[e_\tau - e_0]}{\delta} + \frac{[e_\tau - e_0]}{\delta}\frac{\partial w_1^*}{\partial \overline{z}} = 0$$

$$\iff \frac{\partial w_0^*}{\partial \overline{z}} = \left[1 - \frac{\partial w_1^*}{\partial \overline{z}}\right]\left[\frac{w_1^* - \overline{z}}{\delta^2 \alpha_p} + \frac{[e_\tau - e_0]}{\delta}\right] = a_{\overline{z}}^{*linear}\left[1 - \frac{\partial w_1^*}{\partial \overline{z}}\right] > 0.$$

$$\frac{\partial v_p(\overline{z}; w_0^*, w_1^*)}{\partial \overline{z}} = 0 \iff \frac{\partial w_0^*}{\partial \alpha_g} + 2\frac{w_1^* - \overline{z}}{2\delta^2 \alpha_p}\frac{\partial w_1^*}{\partial \alpha_g} + \frac{[e_\tau - e_0]}{\delta}\frac{\partial w_1^*}{\partial \alpha_g} = 0$$

$$\iff \frac{\partial w_0^*}{\partial \alpha_g} = -\frac{\partial w_1^*}{\partial \alpha_g}\left[\frac{w_1^* - \overline{z}}{\delta^2 \alpha_p} + \frac{[e_\tau - e_0]}{\delta}\right] = -a_{\overline{z}}^{*linear}\frac{\partial w_1^*}{\partial \alpha_g} < 0.$$

$$\frac{\partial v_p(\overline{z}; w_0^*, w_1^*)}{\partial \overline{z}} = 0 \iff \frac{\partial w_0^*}{\partial \alpha_g} - \frac{[w_1^* - \overline{z}]^2}{2\delta^2 \alpha_p^2} + \frac{w_1^* - \overline{z}}{\delta^2 \alpha_p}\frac{\partial w_1^*}{\partial \alpha_p} + \frac{[e_\tau - e_0]}{\delta}\frac{\partial w_1^*}{\partial \alpha_p} = 0$$

$$\iff \frac{\partial w_0^*}{\partial \alpha_p} = \frac{[w_1^* - \overline{z}]^2}{2\delta^2 \alpha_p^2} - \frac{\partial w_1^*}{\partial \alpha_p}\left[\frac{w_1^* - \overline{z}}{\delta^2 \alpha_p} + \frac{[e_\tau - e_0]}{\delta}\right]$$

$$= -h(a_{\overline{z}}^{*linear}, e_0) - \frac{\partial w_1^*}{\partial \alpha_p} a_{\overline{z}}^{*linear} > 0.$$

A.2 Unrestricted Contract with Cost Heterogeneity

We now solve for the optimal unrestricted contract for a given baseline hematocrit level e_0. (As we did earlier, for convenience we will suppress the dependence of the wage schedule and other functions on e_0 when it is not confusing to do so.)

First, we show how two useful properties for the agent's utility function can be obtained from reasonable assumptions on the primitives. The properties are monotonicity and single-crossing with regard to the unobserved heterogeneity. They arise from the assumptions on health production function described earlier and from monotonicity and complementarity assumptions in the cost function. These assumptions are formally stated below.

Assumption 1. *Technology and Costs*

We assume the technology and cost functions have the following properties:

1. *Health production function:*

   (i) $\frac{\partial h}{\partial a} > 0$ *if* $e_0 < e_\tau$ *and* $\frac{\partial h}{\partial a} < 0$ *if* $e_0 \geq e_\tau$

   (ii) $\frac{\partial^2 h}{\partial a^2} < 0$

2. *Cost function:*

   (i) $c(0; z) = 0$ *and*

   (ii) $\frac{\partial c}{\partial a} > 0$ *and* $\frac{\partial^2 c}{\partial^2 a} \geq 0$

   (iii) $\frac{\partial c}{\partial z} > 0$ *and* $\frac{\partial^2 c}{\partial z \partial a} > 0.$

A1(1) reflects the aforementioned fact that patients with low hematocrit benefit more from EPO, while there are serious risks from above-target provision. A1(2.ii) is the standard assumption that marginal costs are positive and nondecreasing and A1(2.iii) says that lower cost types also face lower marginal costs of increasing their action; such an assumption is standard in the theoretical literature studying screening models. Note that the assumed quadratic h and linear c satisfy A1. Further note that, although the optimal contracts derived in this section are for the assumed functional forms of h and c, the results characterizing the optimal contract and treatment choices can be adapted without much trouble for other h and c that satisfy A1.

Given these properties in A1, the utility function has the following useful properties. It will be convenient to express the utility of physician of type i at their prescribed action a_i as

$$u_i \equiv \underbrace{\alpha_p h(a_i) - c(a_i; z_i)}_{\equiv \hat{u}_p(a_i; z_i)} + t_i, \tag{14}$$

i.e., it can be decomposed into a component depending on patient health and cost of treatment, $\hat{u}_p(a_i; z_i)$, and a component representing the transfer from the government to the physician.

Proposition 1. *Given the above assumptions, the agent's utility function has the following properties:*

1. *Monotonicity:* $z_i < z_j$ *implies* $\hat{u}_p(a; z_i) \geq \hat{u}_p(a; z_j)$*; this inequality is strict if* $a > 0$

   Proof. Eq. (1) shows that the only difference in utilities comes from $c(a, z)$. A1(2.i,iii) together imply that $c(a; z_i) \leq c(a; z_j)$ (with a strict inequality if $a > 0$). □

2. *Single crossing:* $z_i < z_j$ *implies* $\frac{\partial \hat{u}_p(a; z_i)}{\partial a} \geq \frac{\partial \hat{u}_p(a; z_j)}{\partial a}$*; this inequality is strict if* $a > 0$

   Proof. This follows directly from eq. (1) and A1(2.iii). □

Now we turn to the physician's problem and the government's problem, which we restate here for convenience.

Physician's problem: Given w and e_0, physician i chooses a drug treatment level a_i to solve

$$\max_{a \geq 0} u_i(a; e_0, w).$$

The solution $a^*(z; e_0, w)$ will satisfy

$$\alpha_p h'(a^*) + w'(a^*) \leq c'(a^*; z_i).$$

Government's problem: The government solves

$$\max_{w \in \mathbb{W}} \int_{\underline{z}}^{\overline{z}} \left[\alpha_g h(a, e_0) - w(a)\right] f_z(z) dz \tag{15}$$

$$\begin{aligned} &\text{s.t.} \\ &a = a^*(z_i; e_0, w), && \text{IC} \\ &u_i(a^*(z_i; e_0, w); e_0, w) \geq 0, \forall i \in I && \text{VP,} \end{aligned}$$

where $\mathbb{W}$ is the set of functions containing all possible contracts.

Characterizing the optimal unrestricted wage schedule is complicated, due to the considerable flexibility allowed by $\mathbb{W}$. Therefore, we take the standard approach of considering the equivalent direct revelation mechanism, in which the government offers a menu of pairs $\{(a_i, t_i)\}_{i \in I}$, where each $(a_i, t_i) \in \mathbb{R}_+ \times \mathbb{R}$, to the physician, who truthfully reveals their type i.[46] In such a mechanism, the physician tells the government their type, after which the government implements their "prescribed" action a_i. We can then determine the optimal $w(a_i)$ by identifying the relevant t_i for action a_i.[47]

We show how to arrive at the government's transformed problem in Appendix A.3, which allows for a direct solution of treatment levels for a physician of each cost type. For convenience, we reproduce the government's transformed, or "relaxed", problem here:

$$\max_{\{a_i\}_{i \in I}} \int_{\underline{z}}^{\overline{z}} \left[\left[\hat{u}_g(a_i) + \hat{u}_p(a_i; z_i)\right] f_z(z_i) - a_i F(z_i)\right] dz_i. \tag{16}$$

Note that the relaxed problem drops the monotonicity constraint required to implement the direct revelation mechanism. As is standard, we proceed by first solving for the set of actions

[46] See the literature on price discrimination, e.g., Maskin and Riley (1984).

[47] For each baseline hematocrit e_0, the resulting $w(a_i)$ could be thought of as either a menu of linear schedules, one for each cost type $i \in I$, or a single, potentially nonlinear, schedule. We treat the optimal contract as the latter, as we find it a more intuitive representation. Note that the relevant equilibrium concept when invoking the revelation principle is Bayesian Nash Equilibrium, a refinement of subgame perfect Nash Equilibrium.

and then verifying that the actions satisfy monotonicity.

Solution: By differentiating (16) with respect to the action of cost-type i, we obtain

$$\frac{\partial \hat{u}_g(a_i)}{\partial a_i} + \frac{\partial \hat{u}_p(a_i; z_i)}{\partial a_i} = \frac{F_z(z_i)}{f_z(z_i)},$$

where $\frac{F_z(z_i)}{f_z(z_i)}$ is the reciprocal of the reverse hazard function, $\lambda_z(z_i)$; i.e., $\lambda_z(z_i)$ represents the likelihood of being cost type z_i conditional on not being *above* that cost type. Because the right side is strictly positive, the action of a_i will be distorted for all but the lowest-cost type; i.e., we obtain the standard "no distortion at the top" result (where the "top" is the best type with the lowest cost). If z has the monotone decreasing hazard property, then $\frac{F_z(z_i)}{f_z(z_i)}$ is increasing in i, meaning actions become increasingly distorted as we consider higher and higher cost types.[48]

To provide more intuition, we show the optimal treatment choice, using our functional form assumptions:

$$a_i = \frac{e_\tau - e_0}{\delta} - \frac{z_i + \frac{F_z(z_i)}{f_z(z_i)}}{\delta^2[\alpha_g + \alpha_p]}. \tag{17}$$

As discussed above, treatment choice will be decreasing in cost type, i.e., the required monotonicity condition (20) will be satisfied, if λ_z is decreasing in z. Optimal treatment choice is increasing in need (i.e., decreasing in baseline hematocrit e_0) and government and physician valuation of health, and is decreasing in the cost of treatment. Note that the aforementioned distortions mean that actions characterized in (17) are lower than those under full information, (7), for all but the lowest-cost type. However, increases in physician altruism α_p can blunt the distortions induced by unobserved heterogeneity. This highlights the usefulness of allowing for physician heterogeneity and physician altruism when characterizing optimal remuneration for physicians.

Based on the optimal treatment choice the nonlinear transfer, or wage, function can be obtained by substituting for surplus

$$t_i = w(a_i) = u_i - \hat{u}_p(a_i; z_i) = \left[\int_{z_{i+1}}^{\overline{z}} a_j dj\right] + a_i z_i - \alpha_p h(a_i),$$

which, as in Ellis and McGuire (1986), results in partial cost sharing if $\alpha_p > 0$, i.e., physicians are altruistic.[49] Using our functional form assumptions to use (17) to substitute for treatment

[48] This is analogous to the standard monotone increasing hazard condition, because as we decrease z we encounter "better" cost types.

[49] This is because $\frac{\partial w(a_i)}{\partial a_i} = z_i - \alpha_p h'(a_i)$, which is strictly less than the marginal cost z_i for $e_0 < e_\tau$.

choice results in the following expression for the equilibrium transfer

$$t_i = \frac{\alpha_p}{2}\left[\frac{z_i + \frac{F_z(z_i)}{f_z(z_i)}}{\delta^2[\alpha_g+\alpha_p]}\right]^2 + a_i z_i + \int_{z_{i+1}}^{\overline{z}} \left[\frac{e_{\mathcal{T}} - e_0}{\delta} - \frac{z_j + \frac{F_z(z_j)}{f_z(z_j)}}{\delta^2[\alpha_g+\alpha_p]}\right] dz_j. \tag{18}$$

We now address *indirect* implementation, which is relevant for applying the above solution to the real world using a schedule depending on a, instead of *direct* implementation via revelation. Assuming the hazard is monotone decreasing, a_i is strictly monotonic in type. If t_i is also strictly monotonic in type, then we could indirectly implement the direct revelation mechanism via a single nonlinear schedule $t(a_i) = w(a_i) = t_i$, since each cost type would have a unique treatment choice and each transfer would correspond to a unique action and, hence, cost type. Due to the single-crossing condition, transfer t_i increasing in action a_i would be equivalent to t_i being decreasing in cost type z_i. Differentiating (18), the transfer function will be decreasing in z when

$$\frac{\partial t_i}{\partial z_i} = \frac{\partial a_i}{\partial z_i}\underbrace{\left[z_i - \alpha_p \frac{\partial h(a_i)}{\partial a_i}\right]}_{-\frac{\partial \hat{u}_p(a;z_i)}{\partial a}} < 0.$$

The monotone decreasing hazard property implies that $\frac{\partial a_i}{\partial z_i} < 0$. The fact that the government seeks to induce physicians to implement higher actions than they would under autarky (which would satisfy $\frac{\partial \hat{u}_p(a;z_i)}{\partial a}|_{a=a^{*,\text{autarky}}} = 0$) implies that $\frac{\partial \hat{u}_p(a;z_i)}{\partial a} < 0$ for the equilibrium action, meaning $\frac{\partial t_i}{\partial z_i} < 0$, i.e., $\frac{\partial t_i}{\partial a_i} > 0$. Intuitively, transfers serve to compensate physicians for treatment costs. If physicians have a high enough valuation of patient health, these transfers become less important. As with the optimal linear contract, to understand how physician altruism affects incentive strength we now examine the slope of the transfer function is affected by an increase in altruism. First, note that $\frac{\partial \hat{u}_p(a;z_i)}{\partial a} < 0 \Leftrightarrow \frac{F_z(z_i)}{f_z(z_i)} < \frac{\alpha_g}{\alpha_p} z_i$.[50] Differentiating $\frac{\partial t_i}{\partial z_i}$ with respect to α_p, we find $\frac{\partial^2 t_i}{\partial z_i \partial \alpha_p} = \frac{\partial^2 a_i}{\partial z_i \partial \alpha_p}\left[z_i - \alpha_p \frac{\partial h(a_i)}{\partial a_i}\right] + \frac{\partial a_i}{\partial z_i}\left[-\frac{\partial h(a_i)}{\partial a_i} - \alpha_p \frac{\partial^2 h(a_i)}{\partial a_i^2}\frac{\partial a_i}{\partial \alpha_p}\right]$, which is positive if $\frac{F_z(z_i)}{f_z(z_i)} < \frac{\alpha_g}{\alpha_p} z_i$, i.e., if the condition guaranteeing a monotonic wage function holds. Since this condition ensures that $\frac{\partial t_i}{\partial z_i} < 0$, increasing physician altruism dampens the relationship between physician treatment choices and reimbursement. Intuitively, higher

[50] Plugging in for the optimal treatment choice $a = \frac{e_{\mathcal{T}} - e_0}{\delta} - \frac{z + \frac{F_z(z)}{f_z(z)}}{\delta^2[\alpha_g + \alpha_p]}$, we obtain

$$\frac{\partial \hat{u}_p}{\partial a} = -\left[z - \alpha_p\left[\frac{z + \frac{F_z(z)}{f_z(z)}}{\alpha_p + \alpha_g}\right]\right] = -\left[\frac{\alpha_g z - \alpha_p \frac{F_z(z)}{f_z(z)}}{\alpha_p + \alpha_g}\right].$$

Therefore, $\frac{\partial \hat{u}_p}{\partial a} \leq 0 \iff \alpha_g z > \alpha_p \frac{F_z(z)}{f_z(z)}$.

altruism reduces the importance of cost heterogeneity, the driving force behind the existence of an increasing wage function.

A.3 Further Details for Unrestricted Contract with Cost Heterogeneity

Using our functional form assumptions, the government's problem can be written as

$$\max_{\{(a_i,t_i)\}_{i\in I}} \int_{\underline{z}}^{\overline{z}} \left[\hat{u}_g(a_i) - t_i\right] f_z(z_i) dz_i \tag{19}$$

$$\begin{aligned} &\text{s.t.} \\ &\qquad u_i \geq 0, \forall i && \text{VP} \\ &\qquad \hat{u}_p(a_i; z_i) + t_i \geq \hat{u}_p(a_j; z_i) + t_j, \forall i, j && \text{IC,} \end{aligned}$$

i.e., the government maximizes its objective, where $\hat{u}_g(a) = \alpha_g h(a)$, given the constraints that all physicians must participate (VP) and that no type of physician has an incentive to mimic a physician of another type (IC).

To solve for the optimal set of actions, we first use the definition of u_i (from (14)) eliminate t_i. Using the expression for a physician's utility, the IC constraints imply that $u_i \geq u_{i+1} + a_{i+1}[z_{i+1} - z_i]$ and $u_{i+1} \geq u_i + a_i[z_i - z_{i+1}]$, which can be combined to produce $a_i \geq a_{i+1}$; i.e., treatment choice is monotonic, and, in particular, nonincreasing, in z.[51]

We next examine which of the constraints in (19) will not bind. First, as was also the case in the linear schedule, we can remove all participation constraints except for that for the physician with cost type $\overline{z}$, because $\hat{u}_p(a; z)$ is decreasing in z. Second, the single-crossing property means that the IC constraints will be "upwards" binding, because "better" (i.e., lower-cost) types could always at least obtain "worse" (i.e., higher-cost) types' utility by mimicking them (i.e., lower-cost types could take higher actions). Therefore, the IC constraints for all but the physician with the highest-cost type, $\overline{z}$, will bind. We will sometimes abuse notation and refer to types using their cost level as their index, e.g., $a_{\overline{z}} = a(\overline{z})$.

Dropping the slack constraints and imposing monotonicity, the government's problem

[51] Using i's surplus, the IC constraint in eq. (19) can be re-written as $u_i \geq \hat{u}_p(a_j; z_i) + t_j$; substituting for t_j and re-arranging, we obtain $u_i \geq u_j + a_j\left[z_j - z_i\right]$.

becomes

$$\max_{\{a_i\}_{i\in I}} \int_{\underline{z}}^{\overline{z}} \left[\hat{u}_g(a_i) + \hat{u}_p(a_i; z_i) - u_i\right] f_z(z_i) dz_i \tag{20}$$

s.t.

$$\begin{aligned} &u_{\overline{z}} = 0 && \text{VP} \\ &u_i = u_{i+1} + a_{i+1}[z_{i+1} - z_i], \text{ for } \{i : z_i < \overline{z}\} && \text{IC} \\ &a_i \geq a_{i+1}, \forall i && \text{M.} \end{aligned}$$

The remaining VP constraint says the government will extract all the surplus from the highest-cost type, $\overline{z}$. Because $a_i \geq 0$ and $z_i < z_{i+1}$, $\forall i$, the IC, or truth-telling, constraint implies that physician surplus weakly increases as we decrease cost types, and strictly increases when the cost type immediately above has a positive action (as is assumed here). Intuitively, a lower-cost type captures surplus because she can pretend to be the higher-cost type, by taking a lower action; this surplus disappears when the higher-cost type does not choose a positive treatment level. Using the binding VP constraint and recursively substituting for u_i in the binding IC constraints, we obtain the following expression for type i's surplus:

$$u_i = \int_{i+1}^{\bar{i}} a_j dj. \tag{21}$$

Therefore, the government chooses treatments to solve

$$\max_{\{a_i\}_{i\in I}} \int_{\underline{z}}^{\overline{z}} \left[\hat{u}_g(a_i) + \hat{u}_p(a_i; z_i) - \left[\int_{i+1}^{\bar{i}} a_j dj\right]\right] f_z(z_i) dz_i$$

$$\Leftrightarrow$$

$$\max_{\{a_i\}_{i\in I}} \int_{\underline{z}}^{\overline{z}} \left[\left[\hat{u}_g(a_i) + \hat{u}_p(a_i; z_i)\right] f_z(z_i) - a_i F(z_i)\right] dz_i \tag{22}$$

Note that we have temporarily dropped the monotonicity constraint. As is standard, we proceed by first solving for the set of actions and then verifying that the actions satisfy monotonicity.