Optimal Contracting with Altruistic Agents: Medicare Payments for Dialysis Drugs

ONLINE APPENDIX

Martin Gaynor
Carnegie Mellon University
and NBER

Nirav Mehta
University of Western Ontario

Seth Richards-Shubik
Lehigh University
and NBER

January 27, 2023

Contents

A Optimal Linear Contract

A.1 Optimal Linear Contract when there is No Exclusion

In this section we solve for the optimal linear contract for the case where no physician types are excluded in equilibrium, i.e., all physicians would choose strictly positive treatment amounts. Although we allow for corner solutions for treatment amounts in our quantitative results, in Section V, the current exercise is useful because our proof that the observed payment rate cannot be rationalized draws on this result (see Appendix B). Note that, while we use the more general h notation for the health production function when it simplifies expressions, results here were obtained using the quadratic-loss parameterization of h, specified in Section IV.

Using interior physician's treatment choice functions (10), the government's problem can be written as

$$\max_{\{(p_0,p_1)\in\mathbb{R}^2\}} \int_{\underline{\alpha}}^{\overline{\alpha}} \int_{\underline{z}}^{\overline{z}} \left[\alpha_g h(a) - p_0 - p_1 a^*(\alpha, z; p_1)\right] f(\alpha, z) dz d\alpha \tag{O1}$$

s.t.

$$u(a^*(\alpha, z; p_1); \alpha, z, p_0, p_1) \geq \underline{u}, \quad \forall(\alpha, z) \qquad \text{VP}$$

$$a^*(\alpha, z; p_1) = \frac{\tau - b}{\delta} + \frac{p_1 - z}{\delta^2 \alpha}, \quad \forall(\alpha, z) \qquad \text{IC.}$$

We can eliminate the participation constraints for all types but

$$(\ddot{\alpha}, \ddot{z}) \equiv \underset{(\alpha,z)}{\arg\min}\ u(a^*(\alpha, z; p_1); \alpha, z, p_0, p_1),$$

i.e., the lowest-utility type given linear contract (p_0, p_1).[1] Setting up the Lagrangian based on the remaining participation constraint, we have

$$\mathcal{L} = \int_{\underline{\alpha}}^{\overline{\alpha}} \int_{\underline{z}}^{\overline{z}} \left[\alpha_g \left[H - \frac{[p_1 - z]^2}{2\delta^2\alpha^2}\right] - p_0 - p_1 \left[\frac{[\tau - b]}{\delta} + \frac{p_1 - z}{\delta^2 \alpha}\right]\right] f(\alpha, z) dz d\alpha$$
$$+ \mu \left[\ddot{\alpha} H + \frac{[p_1 - \ddot{z}]^2}{2\delta^2 \ddot{\alpha}} + \frac{[\tau - b][p_1 - \ddot{z}]}{\delta} + p_0 - \underline{u}\right].$$

[1] If $h > 0$ then $(\ddot{\alpha}, \ddot{z}) = (\underline{\alpha}, \overline{z})$, by the envelope condition.

First-order conditions with respect to p_0 and p_1 yield the following system of equations:

$$\frac{\partial \mathcal{L}}{\partial p_0} = \int\limits_{\underline{\alpha}}^{\overline{\alpha}} \int\limits_{\underline{z}}^{\overline{z}} \left[-f(\alpha, z) dz d\alpha\right] + \mu^* = 0 \Rightarrow \mu^* = 1$$

$$\frac{\partial \mathcal{L}}{\partial p_1} = \int\limits_{\underline{\alpha}}^{\overline{\alpha}} \int\limits_{\underline{z}}^{\overline{z}} \left[-\alpha_g \left[\frac{p_1^* - z}{\delta^2 \alpha^2}\right] - \left[\frac{[\tau - b]}{\delta} + \frac{p_1^* - z}{\delta^2 \alpha}\right] - \frac{p_1^*}{\delta^2 \alpha}\right] f(\alpha, z) dz d\alpha + \mu^* \left[\frac{p_1^* - \ddot{z}}{\delta^2 \ddot{\alpha}} + \frac{\tau - b}{\delta}\right] = 0.$$

Using $\mu^* = 1$, from the first equation, the second equation can be simplified further to solve for p_1^*:

$$\int\limits_{\underline{\alpha}}^{\overline{\alpha}} \int\limits_{\underline{z}}^{\overline{z}} \left[\frac{\alpha_g [p_1^* - z]}{\delta^2 \alpha^2} + \frac{2 p_1^*}{\delta^2 \alpha} - \frac{z}{\delta^2 \alpha}\right] f(\alpha, z) dz d\alpha = \frac{p_1^* - \ddot{z}}{\delta^2 \ddot{\alpha}}$$

$$\Rightarrow p_1^* = \frac{\alpha_g \, \mathrm{E}\left[\frac{z}{\alpha^2}\right] + \mathrm{E}\left[\frac{z}{\alpha}\right] - \frac{\ddot{z}}{\ddot{\alpha}}}{\alpha_g \, \mathrm{E}\left[\frac{1}{\alpha^2}\right] + 2 \, \mathrm{E}\left[\frac{1}{\alpha}\right] - \frac{1}{\ddot{\alpha}}}. \tag{O2}$$

If desired, one could then characterize p_0^* in terms of p_1^*, using the binding participation constraint of $(\ddot{\alpha}, \ddot{z})$.

A.2 Optimal Linear Contract when there is Exclusion

Let $\tilde{z}^0(\alpha; p_1) \equiv \alpha\delta[\tau - b] + p_1$ denote the cost type indifferent between providing treatment and not, given altruism type α and payment rate p_1.[2] The government's problem, allowing for exclusion, is:

$$\max_{\{(p_0, p_1) \in \mathbb{R}^2\}} \mathrm{E}\left[u_g(a(\alpha, z; p_1); p_0, p_1)\right] = \int\limits_{\underline{\alpha}}^{\overline{\alpha}} \int\limits_{\underline{z}}^{\tilde{z}^0(\alpha, p_1)} \left[\alpha_g h(a^*(\alpha, z; p_1)) - p_0 - p_1 a^*(\alpha, z; p_1)\right] f(\alpha, z) dz d\alpha$$
$$+ \int\limits_{\underline{\alpha}}^{\overline{\alpha}} \int\limits_{\tilde{z}^0(\alpha, p_1)}^{\overline{z}} \left[\alpha_g h(0) - p_0\right] f(\alpha, z) dz d\alpha \tag{O3}$$

s.t.

$$u(a^*(\alpha, z; p_1); \alpha, z, p_0, p_1) \geq \underline{u}, \quad \forall (\alpha, z) \qquad \text{VP}$$

$$a^*(\alpha, z; p_1) = \begin{cases} \frac{\tau - b}{\delta} + \frac{p_1 - z}{\delta^2 \alpha}, & \forall \{(\alpha, z) : z < \tilde{z}^0(\alpha, p_1)\} \\ 0, & \forall \{(\alpha, z) : z \geq \tilde{z}^0(\alpha, p_1)\} \end{cases} \qquad \text{IC.}$$

[2] Note that $\tilde{z}^0 \equiv \tilde{z}(\alpha; p_1, a = 0)$, where $\tilde{z}$ is defined in equation (O6), in Appendix C.2.

(Note that, while we use the more general h notation for the health production function when it simplifies expressions, results here were obtained using the quadratic-loss parameterization of h, specified in Section IV.)

Note that the equilibrium utility of excluded type (α, z) is $u(0; \alpha, z, p_0, p_1) = \alpha h(0) + p_0$, i.e., it does not depend on z and is increasing in α; this, combined with the fact that the treatment amount is increasing in α when $h'(a) > 0$ (which is satisfied at $a = 0$), implies that only the participation constraint for the lowest-altruism type will bind. Setting up the Lagrangian based on the lowest-altruism-type's participation constraint, we have

$$\mathcal{L} = \int\limits_{\underline{\alpha}}^{\overline{\alpha}} \int\limits_{\underline{z}}^{\tilde{z}^0(\alpha,p_1)} \left[\alpha_g h(a^*(\alpha, z; p_1)) - p_0 - p_1 a^*(\alpha, z; p_1)\right] f(\alpha, z) dz d\alpha + \int\limits_{\underline{\alpha}}^{\overline{\alpha}} \int\limits_{\tilde{z}^0(\alpha,p_1)}^{\overline{z}} \left[\alpha_g h(0) - p_0\right] f(\alpha, z) dz d\alpha$$
$$+ \mu \left[\underline{\alpha} h(0) + p_0 - \underline{u}\right].$$

Differentiating with respect to p_0, we obtain $\mu^* = 1$ and $p_0^* = \underline{u} - \underline{\alpha} h(0)$. Differentiating with respect to p_1, and simplifying a good bit,[3] we obtain the following implicit expression for p_1^*:

$$\int\limits_{\underline{\alpha}}^{\overline{\alpha}} \int\limits_{\underline{z}}^{\tilde{z}^0(\alpha,p_1^*)} \left[\frac{z[\alpha_g + \alpha]}{\alpha^2}\right] f(\alpha, z) dz d\alpha - \delta[\tau - b] \int\limits_{\underline{\alpha}}^{\overline{\alpha}} \int\limits_{\underline{z}}^{\tilde{z}^0(\alpha,p_1^*)} f(\alpha, z) dz d\alpha = p_1^* \int\limits_{\underline{\alpha}}^{\overline{\alpha}} \int\limits_{\underline{z}}^{\tilde{z}^0(\alpha,p_1^*)} \left[\frac{\alpha_g + 2\alpha}{\alpha^2}\right] f(\alpha, z) dz d\alpha. \tag{O4}$$

B Rationalizability of Observed Payment Rate

The model parameters governing physician behavior are identified without assuming optimality of the observed payment contract. Given our use of physicians' revealed preference to identify these parameters, it is natural to consider whether a revealed preference approach could also inform our value for α_g. In this section, we show that there does not exist a value of α_g such that the optimal linear contract equals the sample mean payment rate, \$9.26/1000u at any of the baseline hematocrit levels considered in our results section, given the estimated parameters. Put differently, the fact that we cannot use the observed payment contract to back out a value of α_g implies that we reject optimality of the observed payment contract; this is in contrast to early work in the empirical contracts literature, which needed to assume optimality of the observed regime to identify model parameters (e.g., Wolak (1994)) but similar to more recent work (e.g., Abito (2019)).

[3] The details are tedious, and are available upon request.

Unlike the case where there is no equilibrium exclusion under the optimal linear contract (see Appendix A.1), the payment rate under the optimal linear contract when there are excluded types is only characterized via a cumbersome implicit expression (see Appendix A.2), which is not ideal because, without further guidance, one would have to exhaustively search through all possible values of α_g to prove the assertion that there did not exist a value of α_g that could rationalize the observed payment rate. Therefore, we adopt an alternative approach, which is to obtain a tractable expression for an upper bound of the optimal linear payment rate, which we then show is below that in the data. (Note that, while we use the more general h notation for the health production function when it simplifies expressions, results here were obtained using the quadratic-loss parameterization of h, specified in Section IV.)

Let $\tilde{z}^0(\alpha; p_1) \equiv \alpha\delta[\tau - b] + p_1$ denote the cost type indifferent between providing treatment and not, given altruism type α and payment rate p_1.[4] Let $p_1^*(\alpha_g; \tilde{z}^0(\cdot, p_1^*))$ denote the solution to (O4), where we assume $p_1^*(\alpha_g; \tilde{z}^0(\cdot, p_1^*)) > 0$. The second argument indicates that the correct cost type, which depends on p_1^*, is used as the upper limit of integration for the inner integral.

We first show in Proposition 1 that $p_1^*(\alpha_g; \tilde{z}^0(\cdot, p_1^*))$ is increasing in α_g. We then show in Proposition 2 that $p_1^*(\infty; \overline{z})$, i.e., the optimal linear payment rate with no exclusion and infinite value of α_g, bounds $p_1^*(\infty; \tilde{z}^0(\cdot, p_1^*))$ from above. This is particularly useful because, taking the limit of (O2) as $\alpha_g \to \infty$, we have $p_1^*(\infty; \overline{z}) = \mathrm{E}\left[\frac{z}{\alpha^2}\right] / \mathrm{E}\left[\frac{1}{\alpha^2}\right]$, which is a very simple explicit expression that can be evaluated using only model primitives.

Proposition 1 **($p_1^*(\alpha_g; \tilde{z}^0(\cdot, p_1^*))$ increasing in α_g).** *The government's choice of p_1^* will be increasing in α_g if $p_1^* > 0$ and the government's objective exhibits complementarity between α_g and p_1 (Vives, 2001, Theorem 2.3). Intuitively, if the government finds it worthwhile to pay physicians to increase their treatment amounts, it does so due to the health benefit. Increasing its valuation of this benefit, α_g, would naturally increase the government's "input" choice, p_1. Because the government's objective is smooth, this complementarity takes the form of a positive cross-partial derivative. We have*

$$\frac{\partial^2 \mathrm{E}\left[u_g(\alpha, z, p_0, p_1)\right]}{\partial \alpha_g \partial p_1} = \int\limits_{\underline{\alpha}}^{\overline{\alpha}} \int\limits_{\underline{z}}^{\tilde{z}^0(\alpha, p_1)} \left[\frac{\partial h(a^*(\alpha, z, p_1))}{\partial a^*} \frac{\partial a^*}{\partial p_1}\right] f(\alpha, z) dz d\alpha,$$

which is positive because the first-order condition of the government's problem with respect

[4] Note that $\tilde{z}^0(\alpha; p_1) \equiv \tilde{z}(\alpha; p_1, a = 0)$, where $\tilde{z}$ is defined in equation (O6), in Appendix C.2. This is the same definition as in Appendix A.2, and is reproduced here for convenience.

to p_1 returns (for $p_1^* > 0$)

$$\alpha_g \int_{\underline{\alpha}}^{\overline{\alpha}} \int_{\underline{z}}^{\tilde{z}^0(\alpha,p_1)} \left[\frac{\partial h(a^*(\alpha,z,p_1))}{\partial a^*}\frac{\partial a^*}{\partial p_1}\right] f(\alpha,z)dzd\alpha - \int_{\underline{\alpha}}^{\overline{\alpha}} \int_{\underline{z}}^{\tilde{z}^0(\alpha,p_1)} \left[a^*(\alpha,z,p_1) + p_1^*\frac{\partial a^*}{\partial p_1}\right] f(\alpha,z)dzd\alpha = 0$$

$$\Rightarrow \alpha_g \int_{\underline{\alpha}}^{\overline{\alpha}} \int_{\underline{z}}^{\tilde{z}^0(\alpha,p_1)} \left[\frac{\partial h(a^*(\alpha,z,p_1))}{\partial a^*}\frac{\partial a^*}{\partial p_1}\right] f(\alpha,z)dzd\alpha > 0$$

$$\Rightarrow \int_{\underline{\alpha}}^{\overline{\alpha}} \int_{\underline{z}}^{\tilde{z}^0(\alpha,p_1)} \left[\frac{\partial h(a^*(\alpha,z,p_1))}{\partial a^*}\frac{\partial a^*}{\partial p_1}\right] f(\alpha,z)dzd\alpha > 0,$$

where the second line obtains if $p_1^* > 0$ (as was assumed) and there is a positive measure of non-excluded types. □

Proposition 2 ($p_1^*(\infty;\tilde{z}^0(\cdot,p_1^*)) < p_1^*(\infty;\overline{z})$)**.** *Taking the limit of (O4) as $\alpha_g \to \infty$, and after some manipulation and dropping the vanishing terms, we have*

$$\int_{\underline{\alpha}}^{\overline{\alpha}} \int_{\underline{z}}^{\tilde{z}^0(\alpha,p_1^*)} \frac{z}{\alpha^2} f(\alpha,z)dzd\alpha = p_1^* \int_{\underline{\alpha}}^{\overline{\alpha}} \int_{\underline{z}}^{\tilde{z}^0(\alpha,p_1^*)} \frac{1}{\alpha^2} f(\alpha,z)dzd\alpha. \tag{O5}$$

Treating $\tilde{z}^0$ as a parameter, consider how an increase in $\tilde{z}^0$ (towards $\overline{z}$) would affect p_1^* defined in (O5). The derivative of the left side with respect to $\tilde{z}^0$ is $\int_{\underline{\alpha}}^{\overline{\alpha}} \frac{\tilde{z}^0(\alpha,p_1^*)}{\alpha^2} f(\alpha,\tilde{z}^0(\alpha,p_1^*))d\alpha$. The derivative of the double-integral expression on the right side with respect to $\tilde{z}^0$ is $\int_{\underline{\alpha}}^{\overline{\alpha}} \frac{1}{\alpha^2} f(\alpha,\tilde{z}^0(\alpha,p_1^*))d\alpha$. Because we have $\tilde{z}^0(\cdot,\cdot) \geq \underline{z} > 1$,[5] the left side will increase more than the double integral on the right side, meaning $\frac{\partial p_1^*}{\partial \tilde{z}^0} > 0$ and, therefore, $p_1^*(\infty;\tilde{z}^0(\cdot,p_1^*)) < p_1^*(\infty;\overline{z})$. □

Table O1 shows that the upper bound derived above for the optimal linear payment rate is lower than the observed payment rate, 9.26, for the median baseline HCT level in each of the three baseline HCT intervals. Combining this with Propositions 1-2, there cannot exist a value of α_g that rationalizes the observed payment rate for any of these baseline HCT levels. That is, $p_1^*(\alpha_g;\tilde{z}^0(\cdot,p_1^*)) \leq p_1^*(\alpha_g = \infty;\tilde{z}^0(\cdot,p_1^*)) \leq p_1^*(\alpha_g = \infty;\tilde{z}^0(\cdot,p_1^*) = \overline{z}) = \mathrm{E}\left[\frac{z}{\alpha^2}\right] / \mathrm{E}\left[\frac{1}{\alpha^2}\right] < 9.26$.

[5] The lower bounds of the marginal cost type distribution for the low, medium, and high baseline HCT intervals are, respectively, 6.81, 6.19, and 7.10 \$/1000u EPO.

Table O1: Upper bound for optimal linear payment rate

	Baseline HCT interval		
	30-33	33-36	36-39
$p_1^*(\infty; \overline{z})$	8.96	9.10	8.96

Note: $p_1^*(\infty; \overline{z}) = \mathrm{E}\left[\frac{z}{\alpha^2}\right] / \mathrm{E}\left[\frac{1}{\alpha^2}\right]$.

C Model Details

C.1 Restrictiveness of Linear Contracts

Figure O1 illustrates how the two-dimensional physician types map into treatment amounts, under an arbitrary linear contract and an arbitrary nonlinear contract. With either contract, the set of types that will provide the treatment amount a is a line in the support of (α, z): see that (4) rearranges to $z = p(a) + h'(a)\alpha$. The figure plots two such isoquants for amounts a_1 and a_2, where a_2 is medically excessive.[6] The immediately apparent difference between the linear and nonlinear contracts is that with a linear contract (panel a), the intercept of the isoquants is fixed at p_1, while it can change with the nonlinear contract (panel b) because the marginal payment can vary (e.g., $p(a_1) > p(a_2)$).[7] This suggests the difficulty of designing a linear contract that induces appropriate treatment amounts. For example, a linear contract would have difficulty avoiding medically excessive amounts because the payment rate (p_1) would have to be below the marginal cost of the lowest-cost type ($\underline{z}$) to avoid downward slopes, which would likely exclude a nontrivial share of higher-cost types. Nonlinear contracts can avoid this particular tension because, as illustrated by the isoquant for a_2 in the right panel, the marginal payments for medically excessive amounts (e.g., $p(a_2)$) can be set below the marginal cost of the lowest-cost type ($\underline{z}$), which places such isoquants entirely outside the support of (α, z).

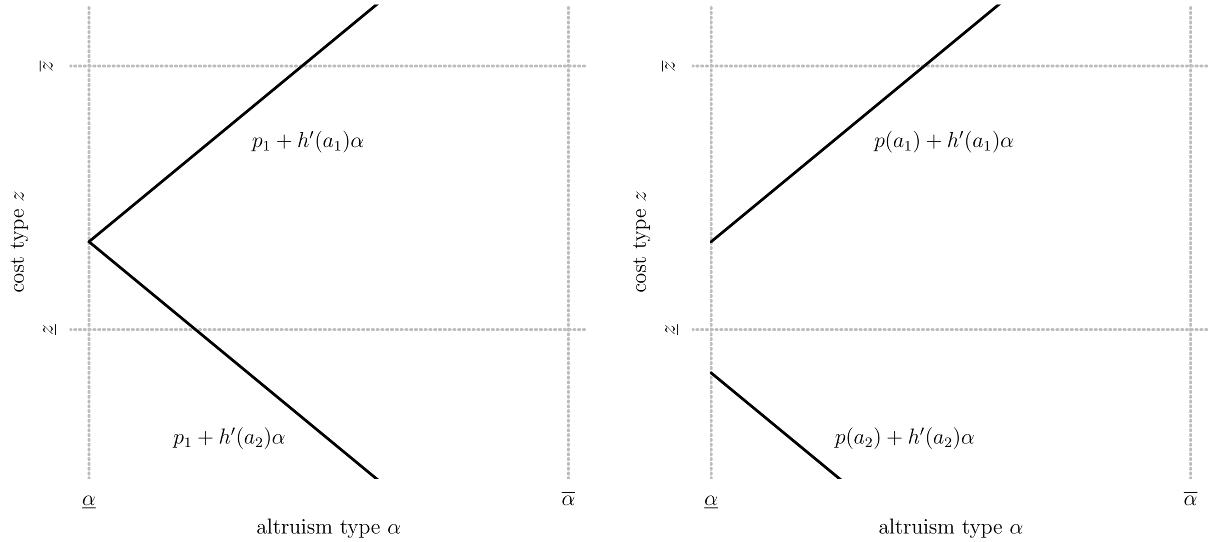
C.2 Details for Solution of Optimal Nonlinear Contract

We now show how to express S in terms of the joint density $f(\alpha, z)$. It will be convenient to define the cost type indifferent about choosing treatment a (given p):

$$\tilde{z}(\alpha; p, a) \equiv p + \alpha h'(a). \tag{O6}$$

[6] That is, $h'(a_2) < 0$. Also note that the slope of the isoquants is $h'(a)$, so downward slopes correspond to medically excessive amounts.

[7] We set $\underline{\alpha} = 0$ only for this illustration, to show the intercept on the plot.



(a) Linear contract

(b) Nonlinear contract

Figure O1: Isoquants for example contracts.

Notes: Figure plots isoquant curves in the type space for an example linear contract (left), which has a constant payment rate of p_1, and an example nonlinear contract (right), which has a variable marginal payment, given by the function p, where $p_1 = p(a_1) > p(a_2)$. The treatment amounts are such that $h'(a_1) > 0$ and $h'(a_2) < 0$.

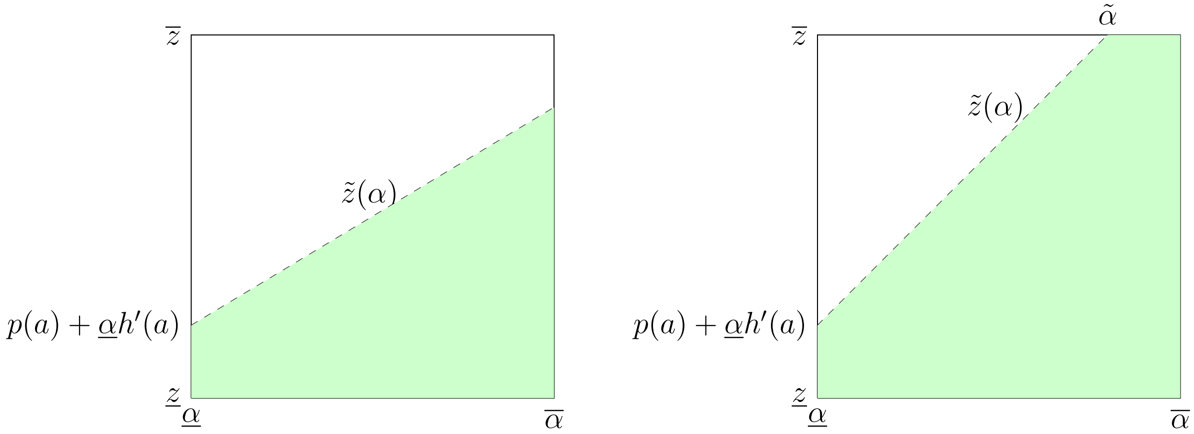
Note that $\tilde{z}$ has intercept p and slope of $h'(a)$, both of which must be non-negative at an optimal solution $p^*(a)$.[8] We also define $\tilde{\alpha}(p,a) = \frac{\overline{z}-p(a)}{h'(a)}$ as the altruism type satisfying $\tilde{z}(\tilde{\alpha}) = \overline{z}$. Suppose that $\tilde{z}(\underline{\alpha}) \geq \underline{z}$. As Figure O2 shows, there are two cases, corresponding to $\tilde{\alpha}$. If $\tilde{\alpha} \geq \overline{\alpha}$, as depicted on the left, then

$$S(p,a) = \Pr\{\underbrace{\alpha h'(a) + p}_{\tilde{z}(\alpha;p,a)} \geq z\} = \int_{\underline{\alpha}}^{\overline{\alpha}} \int_{\underline{z}}^{\tilde{z}(\alpha;p,a)} f(\alpha,z)dzd\alpha, \qquad \text{(O7)}$$

where the types choosing at least a are in the green region. Otherwise, as depicted on the right, we have $\tilde{\alpha} \in [\underline{\alpha}, \overline{\alpha})$, which means that all cost types with altruism types of at least $\tilde{\alpha}$

[8] If $p^* < 0$ then the government would not seek to induce the physician to increase their treatment amount from autarky. If $h' < 0$ at the optimum, the government could save money and improve health by paying for a lower amount.

Figure O2: $\tilde{\alpha}$ cases



will choose at least the level of treatment under consideration.[9] Thus, we have

$$S(p,a) = \int_{\underline{\alpha}}^{\tilde{\alpha}(p,a)} \int_{\underline{z}}^{\tilde{z}(\alpha;p,a)} f(\alpha,z)dzd\alpha + [1 - F_\alpha(\tilde{\alpha})], \tag{O8}$$

where F_α denotes the marginal CDF of α.

To solve for p^* using (8), we also need to differentiate S above with respect to (the parameter) p. If $\tilde{\alpha} \geq \overline{\alpha}$, we have

$$\frac{\partial S(p,a)}{\partial p} = \int_{\underline{\alpha}}^{\overline{\alpha}} f(\alpha, \tilde{z}(\alpha;p,a)) \underbrace{\frac{\partial \tilde{z}(\alpha;p,a)}{\partial p}}_{1} d\alpha. \tag{O9}$$

If $\tilde{\alpha} < \overline{\alpha}$, we have

$$\frac{\partial S(p,a)}{\partial p} = \int_{\underline{\alpha}}^{\tilde{\alpha}} f(\alpha, \tilde{z}(\alpha;p,a))d\alpha. \tag{O10}$$

Note that both $S(p,a)$ and $\frac{\partial S(p,a)}{\partial p}$ are continuous at $\alpha = \tilde{\alpha}(p,a)$. The solution p^* is then obtained by solving (8) for p^* for each $a \in A$.[10]

[9] There is a trivial third case, where $\tilde{\alpha}(p,a) < \underline{\alpha}$; in this case, $S(p,a) = 1$ and $\frac{\partial S(p,a)}{\partial p} = 0$.

[10] Although not depicted in Figure O2, when $\tilde{\alpha}(p,a) \geq \underline{\alpha}$, it is possible that $\tilde{z}(\underline{\alpha}) < \underline{z}$. Here, the integration limits for α must be adapted to account for $\tilde{z}(\alpha)$ crossing the α axis from below. Let $\breve{\alpha}(p,a) \equiv \frac{\underline{z}-p}{h'(a)}$ denote the altruism type satisfying $\tilde{z}(\breve{\alpha}) = \underline{z}$. (Note that the condition $\tilde{z}(\underline{\alpha}) < \underline{z}$ is equivalent to $\breve{\alpha}(p,a) > \underline{\alpha}$.) There are two subcases. First, if $\breve{\alpha}(p,a) > \overline{\alpha}$, then even the most altruistic physician type would not provide the level of treatment under consideration at marginal transfer p, meaning $S(p,a) = 0$ and $\frac{\partial S(p,a)}{\partial p} = 0$. Second,

C.3 Intuition and Normative Aspects of the Optimal Contract

We can divide both sides of (8) by $p^*(a)$ and $\frac{\partial S(p^*(a),a)}{\partial p}$ to obtain the expression

$$\frac{\alpha_g h'(a) - p^*(a)}{p^*(a)} = \frac{1}{\eta(a)}, \tag{O13}$$

where $\eta(a) \equiv \frac{\partial S(p^*(a),a)}{\partial p} \frac{p^*(a)}{S(p^*(a),a)}$ is the elasticity of supply at a. Note the similarity of the expression in (O13) to the Lerner Index for monopoly pricing, i.e., $\frac{p-c'}{p} = \frac{1}{\eta}$, where p and c' are, respectively, the marginal price and marginal cost and η is the elasticity of demand. Our expression differs from that because the government is a monopsonist and, instead of a marginal cost of production c', the government has a marginal valuation of treatment, $\alpha_g h'$. Intuitively, the principal's objective is lower (i.e., it extracts less surplus) where supply is more responsive to price changes (i.e., the elasticity of supply is larger).

We now turn to the normative properties of the second-best allocation. To analyze this, let i index a type that is marginal at a, i.e., $\alpha_i h'(a) - z_i + p^*(a) = 0$. Using this type's first order condition to eliminate $p^*(a)$ from (8) and rearranging, we obtain

$$\underbrace{\alpha_g h'(a)}_{\text{Principal's MB}} = \underbrace{z_i - \alpha_i h'(a)}_{\text{Agent's net MC}} + \underbrace{\frac{S(p^*(a),a)}{\frac{\partial S(p^*(a),a)}{\partial p}}}_{\text{distortion}}, \tag{O14}$$

i.e., at the second-best equilibrium allocation, the principal's marginal benefit of providing a equals the agent's marginal net cost plus a term representing the distortion from the first-best.

We can use (O14) to show that the allocation under the optimal nonlinear contract will be downward-distorted from the first-best for all but the highest-amount type, $(\overline{\alpha}, \underline{z})$.[11] Equivalently, for any amount $a < \overline{a}^{*\text{FI}}$, fewer types choose a in the second-best because they

if $\breve{\alpha}(p,a) \in (\underline{\alpha}, \overline{\alpha}]$ then, if $\tilde{\alpha} \geq \overline{\alpha}$ then (O7) becomes

$$S(p,a) = \int_{\breve{\alpha}}^{\overline{\alpha}} \int_{\underline{z}}^{\tilde{z}(\alpha;p,a)} f(\alpha,z)dzd\alpha, \tag{O11}$$

and if, instead, $\tilde{\alpha} \in [\underline{\alpha}, \overline{\alpha})$, then (O8) becomes

$$S(p,a) = \int_{\breve{\alpha}}^{\tilde{\alpha}(p,a)} \int_{\underline{z}}^{\tilde{z}(\alpha;p,a)} f(\alpha,z)dzd\alpha + [1 - F_\alpha(\tilde{\alpha})]. \tag{O12}$$

[11] Recall that at an interior solution under the optimal linear contract a^* is increasing in α and decreasing in z when the regularity condition holds.

are being distorted downwards. To see this, first recall that $S(p(a), a)$ is the probability the physician would choose at least a. Hence, the numerator of the distortion, $S(p^*(a), a)$, is strictly positive for all but the maximum treatment amount, which is only provided by the highest-amount type (which has a measure of zero). Also the denominator of the distortion, $\frac{\partial S(p^*(a),a)}{\partial p(a)}$, is positive because the probability in (6) increases with $p(a)$. Hence the right side of (O14) is larger than the right side of (3) for all but the highest-amount type. Because h is strictly concave, the second-best treatment amount is therefore below the first-best amount for all but the maximum treatment amount. $S(p^*(a), a)$ increases as we consider lower dosages, and the distortion typically increases, as well.

As noted by Goldman et al. (1984), this result is very similar to that of Ramsey (1927), who studies a government tasked with raising a certain amount of revenue via distortionary taxation of a variety of commodities. As is well known, the optimal second-best tax rates are set in proportion to the inverse of the elasticity of demand, and the lower the elasticity of demand, the closer to the first-best allocation for that commodity. Analogously here, the lower the elasticity of supply, the smaller the distortion.

D Computational Details

D.1 Computation of Optimal Linear Contract

In practice, we numerically compute (p_0^*, p_1^*) by using the COBYLA algorithm in the R implementation of the NLopt library (Powell, 1994; Johnson, 2018; R Core Team, 2019), which allows for constrained optimization computation of the government's problem under a linear contract, where we embed exclusion into the physician's choice of treatment amount to solve:

$$\max_{\{(p_0,p_1)\in\mathbb{R}^2\}} \mathrm{E}\left[u_g(a(\alpha, z; p_1); p_0, p_1)\right] = \int_{\underline{\alpha}}^{\overline{\alpha}} \int_{\underline{z}}^{\overline{z}} \left[\alpha_g h(a^*(\alpha, z; p_1)) - p_0 - p_1 a^*(\alpha, z; p_1)\right] f(\alpha, z) dz d\alpha \tag{O15}$$

s.t.

$$u(a^*(\alpha, z; p_0, p_1); \alpha, z, p_0, p_1) \geq \underline{u}, \quad \forall(\alpha, z) \qquad \text{VP}$$

$$a^*(\alpha, z; p_1) = \max\left\{0, \frac{\tau - b}{\delta} + \frac{p_1 - z}{\delta^2 \alpha}\right\}, \quad \forall(\alpha, z) \qquad \text{IC.}$$

(Note that, while we use the more general h notation when it simplifies expressions, these results were obtained using the quadratic-loss parameterization of h, in Section IV.) We evaluate the participation constraints on a grid of (α, z), where there are 700 points of support for α, spanning $[\underline{\alpha}, \overline{\alpha}]$, and 400 points of support for z, spanning $[\underline{z}, \overline{z}]$.

D.2 Computation of Optimal Nonlinear Contract

We compute the optimal nonlinear contract by solving (8), the details of the constituent parts of which are described in Appendix C.2, using the BBoptim subroutine contained in the BB package in R (Varadhan and Gilbert, 2009). We solve (8) for a grid of 100 amounts. The lowest value of the grid is zero because we allow for optimal exclusion via the nonlinear contract. The maximum value of the grid is 0.01 below the full-information amount for the highest-treatment-choice type; we use this as the maximum point due to the numerical issues incumbent in evaluating derivatives at the upper corner of the treatment amount space (which is the same as the upper bound of the full-information treatment amount space, due to the downwards-distortion of equilibrium amounts under the optimal nonlinear contract). Finally, we fit a spline to the grid of treatment amounts, which is what we use for our quantitative results.

E Identification

Here we discuss the identification of the health function, h, and the joint distribution of provider altruism and marginal cost functions. Identification is done separately for each baseline hematocrit interval k; we suppress the k subscript in this appendix. The number of time periods is fixed, but both the number of providers and the number of patients per provider go to infinity. For an arbitrary provider i, there is rich variation in (b, x, p_1, a), where patient characteristics (b, x) vary between patients and over time, the (constant) marginal reimbursement rate p_1 varies over time, and observed treatment choices a are the sum of a provider's equilibrium treatment choice $a_i^*(p_1, b, x)$ and an econometric error, η, which is mean-independent of (b, x, p_1): $\mathrm{E}(\eta | b, x, p_1) = 0$.[12]

We start by studying identification of more general specifications for the health function and the provider type distribution than we use in our empirical implementation (specified in Section IV). We maintain the assumption of quasilinear utility for providers. We also allow for provider-level heterogeneity in the intercept of marginal cost functions, though here we also allow for (homogeneous) convexity in marginal cost functions, which allows for cost functions with heterogeneous convexity. We show that the marginal product of treatment on health, $h'(a; b, x)$, is identified to scale under a single-index specification for the arguments

[12] This is the same as in our empirical specification; see eq. (11).

of h',[13] and, therefore, that the sign of the marginal effect of treatment on health is identified. The scale parameter is the mean of provider altruism, μ_α.[14] We provide a test for $\mu_\alpha > 0$ and also show identification of the joint distribution of altruism (given the scale μ_α) and marginal cost functions. Finally, we also show that the choice of μ_α has no bearing on our main, normative, results (Section E.2).

In Section E.3, we show the nonparametric identification of $F(\alpha, z)$, given the quadratic specification of h and constant marginal cost function we use in our empirical specification.

E.1 Identification of h', and Joint Distribution of Costs and Altruism

Consider the following general model of utility for arbitrary provider i:

$$U_i(h(a;b), P(a;p_1), c_i(a)),$$

where $\frac{\partial U_i}{\partial h} \geq 0$, $\frac{\partial U_i}{\partial P} > 0$, and $\frac{\partial U_i}{\partial c} \leq 0$, i.e., utility is weakly increasing in health, increasing in money, and weakly decreasing in cost.[15] The production function h and reimbursement function P are common across providers, but the other functions may differ across providers. The observed contract is linear in a, so we have $P(a;p_1) = p_0 + p_1 a$.[16] We also assume that h is strictly concave in a and that c_i is weakly convex in a. Based on our application, we make two further assumptions. We assume that $h'(0;b) > 0$; if this condition did not hold there would be no reason for the government to incentivize any provision of treatment at b. We also assume $\frac{\partial^2 h(a;b)}{a} b < 0$, i.e., the marginal product of dosage is lower, the higher is the baseline hematocrit level.[17]

Our goal is to see what about utility and its argument functions (h, P, c_i) is identified from $a_i^*(p_1, b)$ (which we sometimes write as a_i^* for brevity), using the interior solution for the provider's optimal dosage and our assumptions about shape restrictions, i.e., about the signs of first and second derivatives.

[13] If not explicit, all derivatives are with respect to the dosage a, e.g., $h' = \frac{\partial h}{\partial a}$.

[14] In a slight abuse of notation, μ_α denotes the mean of α in this section. This is in contrast to when we describe our empirical specification or estimation results, where it refers to the mean of $\ln \alpha$.

[15] The identification of the effect of x on health is identical to identification of the effect of a on health, so we suppress x for the remainder of this section.

[16] The intercept, p_0 does not vary in the data so we do not specify it explicitly as an argument of $P(\cdot)$. Also note that the observed payment contract does not vary with patient characteristics (in contrast to the optimal payment contract), so we do not include those as arguments of $P(\cdot)$ here.

[17] This is consistent with our index specification, wherein both a and b affect the index entering h.

The first order condition to maximize utility is

$$\frac{\partial U_i}{\partial P}\, p_1 = -\,\frac{\partial U_i}{\partial c}\,\frac{\partial c_i(a_i^*)}{\partial a} - \frac{\partial U_i}{\partial h}\,\frac{\partial h(a_i^*;b)}{\partial a},$$

which, when we divide by $\frac{\partial U_i}{\partial P} > 0$ to get a marginal rate of substitution, becomes

$$p_1 = -\frac{\frac{\partial U_i}{\partial c}}{\frac{\partial U_i}{\partial P}}\,\frac{\partial c_i(a_i^*)}{\partial a} - \frac{\frac{\partial U_i}{\partial h}}{\frac{\partial U_i}{\partial P}}\,\frac{\partial h(a_i^*;b)}{\partial a}\,.$$

The utility level or the levels of any of the functions (h, P, c_i) are not identified from the optimality condition. Our approach will be to use a combination of assumptions and (pure) normalizations to obtain values for $\frac{\partial U_i}{\partial h}, \frac{\partial U_i}{\partial P}, \frac{\partial U_i}{\partial c}$ and then see what is identified about the derivatives of the arguments to utility. Now we add one more assumption and three pure normalizations. Assume quasilinear utility in P, which means we can normalize $\frac{\partial U_i}{\partial P} = 1$, giving

$$p_1 = -\,\frac{\partial U_i}{\partial c}\,\frac{\partial c_i(a_i^*)}{\partial a} - \frac{\partial U_i}{\partial h}\,\frac{\partial h(a_i^*;b)}{\partial a}\,.$$

Note that $\frac{\partial U_i}{\partial c}$ is not separable from $\frac{\partial c_i}{\partial a}$, so without loss of generality we can normalize $\frac{\partial U_i}{\partial c} = -1$. Similarly, $\frac{\partial U_i}{\partial h}$ is not separable from $\frac{\partial h(a_i^*;b)}{\partial a}$, so we can normalize $\frac{\partial U_i}{\partial h} = \alpha_i \geq 0$.[18]

With the normalizations, we now have

$$p_1 = \frac{\partial c_i(a_i^*)}{\partial a} - \alpha_i\,\frac{\partial h(a_i^*;b)}{\partial a},$$

which says that an interior treatment choice, $a_i^*(p_1, b)$ equates the marginal reimbursement rate p_1 with the provider's "net marginal cost of treatment", i.e., their marginal cost of treatment, net the provider's marginal benefit from treatment coming from any improvement in patient health (which may be negative if $h'(a_i^*;b) < 0$).

Polynomial approximation We show identification using a polynomial approximation to the above FOC. Specifically, we approximate the marginal cost and marginal health benefit using polynomials:

$$p_1 = \underbrace{\left[c_{0i} + c_{1i}{\cdot}a_i^* + c_{2i}{\cdot}(a_i^*)^2 + \cdots\right]}_{\approx \frac{\partial c_i(a_i^*)}{\partial a}} - \underbrace{\alpha_i\left[h_0 + h_{1a}{\cdot}a_i^* + h_{1b}{\cdot}b + h_{2a}{\cdot}(a_i^*)^2 + h_{2b}{\cdot}b^2 + h_{2ab}{\cdot}a_i^*{\cdot}b + \cdots\right]}_{\approx \alpha_i \frac{\partial h(a_i^*;b)}{\partial a}}. \tag{O16}$$

For concreteness, consider the case in which both polynomials were of degree two. Higher-

[18] Note too that b is excluded from the cost function. This a natural assumption, because without it one could not separate the cost and health functions.

degree polynomials would also be identified (as would be lower-degree ones, like those we use in our empirical implementation). We assume $c_i'(a) = c_{0i} + c_1 a + c_2 a^2$, i.e., we allow for heterogeneity in the intercept of marginal costs and also allow for non-constant marginal costs within provider.[19]

With an infinite number of patients per provider and the mean-independence of η, the equilibrium treatment choices $a_i^*(p_1, b)$ are identified for each observed value of (p_1, b). With the observed variation in (p_1, b), this directly identifies the reduced-form parameters (θ^i, γ^i) listed in braces under the rearranged version of (O16) below:

$$p_1 = \underbrace{(c_{0i} - \alpha_i h_0)}_{\theta_0^i} + \underbrace{(c_1 - \alpha_i h_{1a})}_{\theta_1^i} a_i^* + \underbrace{(c_2 - \alpha_i h_{2a})}_{\theta_2^i} (a_i^*)^2 + \underbrace{(-\alpha_i h_{1b})}_{\gamma_1^i} b + \underbrace{(-\alpha_i h_{2b})}_{\gamma_2^i} b^2 + \underbrace{(-\alpha_i h_{2ab})}_{\gamma_{2a}^i} a_i^* b \tag{O17}$$

Given our polynomial approximation, our goal is to identify the parameters $\{c_{0i}, \alpha_i\}$ for each provider, and the common parameters c_1, c_2, h_0, h_{1a}, h_{2a}, h_{1b}, h_{2b}, h_{2ab}.

First, the derivative of $\alpha_i h'$ with respect to b (which is approximated using the terms $\alpha_i h_{1b}$, $\alpha_i h_{2b}$, and $\alpha_i h_{2ab}$) is identified from γ_1^i and γ_2^i, because of the exclusion restriction that b does not affect costs c_i.

Our approach to identify the remaining parameters is to assume an index assumption on the arguments of $h(\cdot)$, which links the derivative of h with respect to a to the derivative of h with respect to b.[20]

We denote means of the reduced-form parameters taken across providers using $\bar{\cdot}$: e.g., $\overline{\gamma}_1 = -\mu_\alpha h_{1b}$ (we use μ_α to denote the mean of α and μ_{c_0} to denote the mean of c_0). Thus we have

$$\begin{aligned}
\overline{\theta}_0 &= \mu_{c_0} - \mu_\alpha h_0 \qquad \text{(O18)}\\
\overline{\theta}_1 &= c_1 - \mu_\alpha h_{1a}\\
\overline{\theta}_2 &= c_2 - \mu_\alpha h_{2a}\\
\overline{\gamma}_1 &= -\mu_\alpha h_{1b}\\
\overline{\gamma}_2 &= -\mu_\alpha h_{2b}\\
\overline{\gamma}_{2a} &= -\mu_\alpha h_{2ab}.
\end{aligned}$$

[19] Heterogeneity in the "intercept" of the marginal cost is fairly flexible. Regardless, it is not clear how to separately identify heterogeneity in higher-degree terms of the marginal cost function (e.g., provider-specific c_{1i}); intuitively, we take averages across providers and c_{1i} and a_i^* would be correlated due to the optimality of a_i^*. Note that if c_1 and other higher-order terms in the cost function are all equal to zero, then we have $c_{0i} = z_i$ (the latter being the constant marginal cost we use in our empirical implementation).

[20] There may be other sets of assumptions yielding identification; for example, identification may be obtained by restricting $c'(a)$ to be lower order. Therefore our approach should be viewed as sufficient but not necessary.

Identification then proceeds as follows:

1. Test any of the restrictions $\overline{\gamma}_1 = 0$, $\overline{\gamma}_2 = 0$, or $\overline{\gamma}_{2a} = 0$. If we reject then we can set the scale of α by choosing a positive value for μ_α.[21] We have identified (up to the scale of μ_α) the parameters

$$h_{1b} = -\overline{\gamma}_1/\mu_\alpha, \quad h_{2b} = -\overline{\gamma}_2/\mu_\alpha, \quad h_{2ab} = -\overline{\gamma}_{2a}/\mu_\alpha, \quad \alpha_i = \mu_\alpha \gamma_1^i/\overline{\gamma}_1 \text{ for } i = 1, \ldots, n.$$

2. Invoke the single-index assumption, which means we can write $h(a; b) = g(\kappa_{ab} a + b)$, where κ_{ab} is a constant to be identified. As is standard in single-index models (see, e.g., Ichimura, 1993; Härdle et al., 2004), the scale and location of the index are not identified. Setting the coefficient on b equal to one fixes the scale and gives the index a natural interpretation, in the units of the hematocrit level. We have set the location to zero; note that this nonidentification means the intercept of $\tau' x$ in our empirical specification is identified from functional form.

   The single-index assumption implies that

$$\frac{\partial^2 h(a;b)}{\partial a^2} = \kappa_{ab} \frac{\partial^2 h(a;b)}{a} b. \tag{O19}$$

   For example, in our empirical specification we have $\kappa_{ab} = \delta$.[22] With our 2nd-degree polynomial approximation, we have

$$\frac{\partial^2 h(a;b)}{\partial a^2} \approx h_{1a} + 2h_{2a} \cdot a + h_{2ab} \cdot b$$
$$\frac{\partial^2 h(a;b)}{a} b \approx h_{1b} + 2h_{2b} \cdot b + h_{2ab} \cdot a,$$

   so, with the index assumption we have at an optimum

$$[h_{1a} + 2h_{2a} \cdot a_i^* + h_{2ab} \cdot b] = \kappa_{ab} [h_{1b} + 2h_{2b} \cdot b + h_{2ab} \cdot a_i^*] \,. \tag{O20}$$

   Step 1 identified the parameters on the right of (O20), other than κ_{ab}. We then need to observe at least three vectors of $(b, p_1, a_i^*(b, p_1))$ to exactly identify $h_{1a}, h_{2a}, \kappa_{ab}$; more than three would yield overidentification and, thus, better estimates. (The same argument holds with higher-degree polynomials, but the number of points required increases.)

   We have now identified all of the parameters of h' to scale, except for h_0.

[21] Recall that α_i is non-negative.

[22] In our empirical specification, we have $h'(a; b, x) = \delta [\tau' x - b - \delta a]$, so $h_{1a} = -\delta^2$ and $h_{1b} = -\delta$.

3. Using the second and third lines of (O18), we can identify c_1 and c_2:

$$c_1 = \overline{\theta}_1 + \mu_\alpha h_{1a}$$
$$c_2 = \overline{\theta}_2 + \mu_\alpha h_{2a}.$$

   Note that these parameters are identified (and not just to scale).

4. Next, we use the average marginal cost μ_z (obtained from external data) to identify the average intercept of the marginal cost function, μ_{c_0}, which combined with the average intercept also identifies h_0 to scale. The mean marginal cost (over providers, patients, and time periods) is $\mathrm{E}\left[c_{0i} + c_1 a_i^*(b, p_1) + c_2(a_i^*(b, p_1))^2\right] = \mu_{c_0} + c_1\overline{a} + c_2\overline{a^2}$. Equating this with μ_z, we can solve for μ_{c_0} given that we have identified c_1 and c_2:

$$\mu_{c_0} = \mu_z - \left[c_1\overline{a} + c_2\overline{a^2}\right].$$

   Then, using the first line of (O18), we have

$$h_0 = \frac{\mu_{c_0} - \overline{\theta}_0}{\mu_\alpha},$$

   i.e., h_0 is identified to scale.

5. Finally, we identify c_{0i} via the provider-specific intercept:

$$c_{0i} = \theta_0^i + \alpha_i h_0 = \theta_0^i + \frac{\mu_\alpha \gamma_1^i}{\overline{\gamma}_1} \frac{\mu_z - \left[\overline{\theta}_0 + c_1\overline{a} + c_2\overline{a^2}\right]}{\mu_\alpha} = \theta_0^i + \frac{\gamma_1^i}{\overline{\gamma}_1}\left[\mu_z - \left[\overline{\theta}_0 + c_1\overline{a} + c_2\overline{a^2}\right]\right],$$

   where all the terms on the right have been identified (and not just to scale).

Identification of the sign of h' Here we note that the identification of the sign of h' (i.e., whether treatments are health improving or health damaging on the margin) does not rely on the scale normalization. Rearranging (O16), we have

$$c_i'(a_i^*(p_1, b)) - p_1 = \alpha_i h'(a_i^*(p_1, b); b),$$

where (b, p_1) are data varying within provider i and we have shown identification of c_i' and (trivially) $a_i^*(p_1, b)$. Then if (as we find), $\alpha_i > 0$, we have

$$\mathrm{sign}(c_i'(a_i^*(p_1, b)) - p_1) = \mathrm{sign}(h'(a_i^*(p_1, b); b)),$$

i.e., we have identified the sign of h' at $a_i^*(p_1, b)$. In particular, consider the triplet (i, b, p_1) such that $c_i'(a_i^*(p_1, b)) = p_1$. For any such triplet, $a_i^*(p_1, b)$ identifies the health-maximizing treatment amount (i.e., where $h' = 0$).

E.2 Invariance to the Choice of the Scale of α

We now show how the choice of μ_α does not affect the optimal contract or any of our normative results.

For simplicity suppose we have a one-degree polynomial for health (this is not necessary but makes the exposition cleaner):

$$h'(a; b) = h_0 + h_{1a}a + h_{1b}b = \frac{\pi_o}{\mu_\alpha} + \frac{\pi_{1a}}{\mu_\alpha}a + \frac{\pi_{1b}}{\mu_\alpha}b,$$

where $\pi.$ are all identified and $\mu_\alpha > 0$ is the scale of α. Our calibration of α_g uses the change in $h(a; b)$ when we increase the treatment from a lower to a higher level, respectively, a_L and a_H. We first definitely integrate our (identified-to-scale) h' to return the (identified-to-scale) health level:

$$h(a; b) = H + \frac{\pi_o}{\mu_\alpha}a + \frac{\pi_{1a}}{\mu_\alpha}\frac{a^2}{2} + \frac{\pi_{1b}}{\mu_\alpha}ab,$$

where H is the integration constant. The difference in which, given $b = b_{cal}$ can be written as

$$\Delta h_{cal} \equiv h(a_L; b_{cal}) - h(a_H; b_{cal}) = \left[\frac{\pi_o}{\mu_\alpha} + \frac{\pi_{1b}}{\mu_\alpha}b\right][a_H - a_L] + \frac{\pi_{1a}}{\mu_\alpha}\frac{a_H^2 - a_L^2}{2} = \frac{q_{cal}}{\mu_\alpha},$$

where q_{cal} is identified because $a_L = 0$, a_H is based on an experimental intervention, and $b_{cal} = -\pi_0/\pi_{1b}$, which cancels out the intercept term (and is in any case identified).

We then calibrate α_g from the expression

$$\alpha_g \Delta h_{cal} = \chi \to \alpha_g = \frac{\chi}{q_{cal}}\mu_\alpha,$$

where χ is another known constant based on the experimental intervention. Therefore, α_g perfectly scales with μ_α.

Now consider the government's problem, cast in terms of the demand profile:

$$\max_{P(a)} \int_A S(p(a), a)[\alpha_g h'(a; b) - p(a)]da$$
$$\text{s.t.}$$
$$S(p(a), a) = \Pr\{p(a) \geq c'(a; z) - \alpha h'(a; b)\},$$

where $p(a) = \frac{\partial P(a)}{\partial a}$ and $P(a)$ contains a constant that satisfies voluntary participation for all providers. We have shown the invariance of $\alpha_g h'(a;b)$ and $\alpha h'(a;b)$ to the choice of $\mu_\alpha > 0$, and have also shown that $c'(a,z)$ (where, $c'(a;z_i) = c'_i(a)$ used above) is identified independently from μ_α. This means that changing the value of μ_α does not affect the government's problem, meaning it does not affect the optimal contract (unrestricted or constrained) or any of the normative results.

E.3 Special Case: Identification of F Given Quadratic h

Here we show how the joint distribution of provider types, F, is nonparametrically identified given the quadratic specification of the health function, h.[23] Recall that η is mean-independent of (b, x, p_1) and that the number of observations per provider goes to infinity. Then OLS estimation of the reduced form (11), separately for each provider, yields consistent estimates of β_1, β_{2i}, β_3, and ν_i for each provider.

The structural parameters and provider types are continuous functions of reduced-form parameters and variables, as follows:

$$\begin{aligned}
\delta &= -(\beta_1)^{-1} \\
\tau &= -(\beta_1)^{-1}\beta_3 \\
\alpha_i &= (\beta_1)^2(\beta_{2i})^{-1} \\
z_i &= \mu_z - \nu_i(\beta_{2i})^{-1}
\end{aligned}$$

Hence the structural parameters and provider types are identified by and can be consistently estimated from the reduced-form coefficients of the provider-specific regressions. Finally, the joint distribution F is identified from the consistent estimates of (α_i, z_i) for each provider i.

Thus it is in concept possible to estimate the reduced form for each provider and then use nonparametric density estimation to recover F. However, we do not however pursue this approach because it would be computationally intensive and the resulting estimates would be much noisier.

F Recovery of $F(\alpha, z)$

As noted in Section IV.B, we recover $F_k(\alpha, z)$ under a distributional assumption, where $\ln\alpha$ and z have a joint normal distribution. Here we show how we estimate the parameters of that distribution, which are recovered from the first and second moments of the random

[23] We continue to suppress the k denoting the baseline hematocrit interval.

coefficient (β_2^k) and random effect (ν^k) in the reduced form (11). First we present an auxiliary regression of the residuals of (11) that yields the second moments of β_2^k and ν^k (while the mean of β_2^k comes directly from (11), and the mean of ν^k is zero). Then we derive closed-form expressions for the parameters of $F_k(\alpha, z)$ as functions of these moments.

To develop the auxiliary regression, let $\bar{\beta}_2^k$ denote the mean of β_2^k, and decompose the random coefficient as $\beta_2^k = \bar{\beta}_2^k + \tilde{\beta}^k$. Then (11) can be rearranged as

$$a_{ijt} = \beta_1^k b_{jt} + \bar{\beta}_2^k \tilde{p}_t + \beta_3^{k\prime} x_{jt} + \underbrace{\tilde{\beta}_i^k \tilde{p}_t + \nu_i^k + \epsilon_{ijt}^k}_{r_{ijt}^k}$$

(for b_{jt} in interval k). The OLS coefficient on $\tilde{p}_t$ is a consistent estimate of the mean of the random coefficient, $\mathrm{E}(\beta_2^k)$, under the assumptions discussed in Section IV.B. The auxiliary regression then uses the composite residual, r_{ijt}^k, times the provider-level mean residual, $\bar{r}_i^k$ (taken within interval k), as its dependent variable. This yields consistent estimates of the second moments, $\mathrm{V}(\beta_2^k)$, $\mathrm{V}(\nu^k)$, and $\mathrm{Cov}(\beta_2^k, \nu^k)$, as we show next.[24]

First expand the product of the composite residual and the provider-level mean residual as follows:

$$\begin{aligned}
r_{ijt}^k \bar{r}_i^k &= (\tilde{\beta}_i^k \tilde{p}_t + \nu_i^k + \epsilon_{ijt}^k) \left(\frac{1}{n_i^k} \sum_{l,s: b_{ls} \in k} \tilde{\beta}_i^k \tilde{p}_s + \nu_i^k + \epsilon_{ils}^k \right) \\
&= (\tilde{\beta}_i^k \tilde{p}_t) \tilde{\beta}_i^k \bar{\tilde{p}}_i^k + (\tilde{\beta}_i^k \tilde{p}_t) \nu_i^k + (\tilde{\beta}_i^k \tilde{p}_t) \bar{\epsilon}_i^k \\
&+ \nu_i^k \tilde{\beta}_i^k \bar{\tilde{p}}_i^k + \nu_i^k \nu_i^k + \nu_i^k \bar{\epsilon}_i^k \\
&+ \epsilon_{ijt}^k \tilde{\beta}_i^k \bar{\tilde{p}}_i^k + \epsilon_{ijt}^k \nu_i^k + \epsilon_{ijt}^k \bar{\epsilon}_i^k .
\end{aligned}$$

(The variables of the form $\bar{z}_i^k$ denote means taken among the observations for provider i where the patient's baseline hematocrit is in interval k, and n_i^k is the number of such observations.) The expectation of this product conditional on the payment rates and the number of observations is as follows:

$$\begin{aligned}
E[r_{ijt}^k \bar{r}_i^k | \tilde{p}_t, \bar{\tilde{p}}_i^k, n_i^k] &= V(\tilde{\beta}^k) \tilde{p}_t \bar{\tilde{p}}_i^k + \mathrm{Cov}(\tilde{\beta}^k, \nu^k) \tilde{p}_t + 0 \\
&+ \mathrm{Cov}(\tilde{\beta}^k, \nu^k) \bar{\tilde{p}}_i^k + V(\nu^k) + 0 \\
&+ 0 + 0 + E[\epsilon_{ijt}^k \bar{\epsilon}_i^k] \\
&= V(\tilde{\beta}^k) \cdot \tilde{p}_t \bar{\tilde{p}}_i^k + \mathrm{Cov}(\tilde{\beta}^k, \nu^k) \cdot [\tilde{p}_t + \bar{\tilde{p}}_i^k] + V(\nu^k) + V(\epsilon^k) \cdot \frac{1}{n_i^k} .
\end{aligned}$$

[24] This assumes that the second moments of the unobservables ($\tilde{\beta}_i^k, \nu_i^k, \epsilon_{ijt}^k$) are independent of the observables, while OLS estimation of (11) assumes their first moments are independent of the observables.

This assumes that the error terms ϵ_{ijt}^k are orthogonal to $\tilde{\beta}_i^k$ and ν_i^k and are uncorrelated across observations. Last, note that $\mathrm{V}(\beta_2^k) = \mathrm{V}(\tilde{\beta}^k)$ and $\mathrm{Cov}(\beta_2^k, \nu^k) = \mathrm{Cov}(\tilde{\beta}^k, \nu^k)$. Thus, we can consistently estimate the desired variances and covariance of β_2^k and ν^k by performing a regression of $r_{ijt}\bar{r}_i$ on $\tilde{p}_t\bar{\tilde{p}}_i$, $\tilde{p}_t + \bar{\tilde{p}}$, a constant, and $\frac{1}{n_i}$, within each interval k.

Now we show how these reduced-form moments are mapped to the parameters of $F_k(\alpha, z)$. The joint normal distribution of $\ln\alpha$ and z is specified as follows:

$$\begin{pmatrix} \ln\alpha \\ z \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_{\alpha,k} \\ \mu_z \end{pmatrix}, \begin{bmatrix} \sigma_{\alpha,k}^2 & \sigma_{\alpha z,k} \\ \sigma_{\alpha z,k} & \sigma_{z,k}^2 \end{bmatrix}\right)$$

The value of μ_z is treated as known from our external information on costs, which leaves four parameters to recover for each hematocrit interval: $\mu_{\alpha,k}$, $\sigma_{\alpha,k}^2$, $\sigma_{\alpha z,k}$, and $\sigma_{z,k}^2$. The expressions for these parameters as functions of the reduced-form moments are derived below. These parameters are recovered separately for each interval k, so we omit that index here to simplify the derivations.

a) First we obtain μ_α and σ_α^2 from $\mathrm{E}(\beta_2)$ and $\mathrm{V}(\beta_2)$, using the following properties of the log-normal distribution:

(i) If X has a log-normal distribution, where $\ln X \sim N(\mu, \sigma^2)$, then

$$\mu = \ln\left(\frac{(\mathrm{E}(X))^2}{\sqrt{\mathrm{V}(X) + (\mathrm{E}(X))^2}}\right) \qquad \text{and} \qquad \sigma^2 = \ln\left(1 + \frac{\mathrm{V}(X)}{(\mathrm{E}(X))^2}\right),$$

(ii) and if $Y = X^{-1}$, then $\ln Y \sim N(-\mu, \sigma^2)$.

Hence, because α is log-normal, and $\alpha^{-1} = \delta^2\beta_2$, we have

$$\mu_\alpha = -\ln\left(\frac{\delta^2(\mathrm{E}(\beta_2))^2}{\sqrt{\mathrm{V}(\beta_2) + (\mathrm{E}(\beta_2))^2}}\right) \qquad \text{and} \qquad \sigma_\alpha^2 = \ln\left(1 + \frac{\mathrm{V}(\beta_2)}{(\mathrm{E}(\beta_2))^2}\right).$$

(Also recall that δ comes directly from β_1 in (11).)

b) Next we obtain $\sigma_{\alpha z}$ from $\mathrm{Cov}(\beta_2, \nu)$, along with $\mathrm{E}(\beta_2)$ and $\mathrm{V}(\beta_2)$. First, we use the definitions $\beta_2 \equiv \delta^{-2}\alpha^{-1}$ and $\nu \equiv -(z - \mu_z)\beta_2$ to put the reduced-form covariance in terms of the structural parameters:

$$\mathrm{Cov}(\nu, \beta_2) = \mathrm{Cov}(-(z-\mu_z)\delta^{-2}\alpha^{-1}, \delta^{-2}\alpha^{-1}) = \delta^{-4}\mathrm{Cov}(-(z-\mu_z)\alpha^{-1}, \alpha^{-1}).$$

Then we use the definitional relationship between the covariance and expectations:

$$\delta^{-4}\mathrm{Cov}(-(z-\mu_z)\alpha^{-1},\alpha^{-1}) = \delta^{-4}\mathrm{E}[-(z-\mu_z)\alpha^{-2}] - \delta^{-4}\mathrm{E}[-(z-\mu_z)\alpha^{-1}]\cdot\mathrm{E}[\alpha^{-1}].$$

Now we apply Stein's lemma (Stein, 1981) to the terms $\mathrm{E}[-(z-\mu_z)\alpha^{-1}]$ and $\mathrm{E}[-(z-\mu_z)\alpha^{-2}]$. We use a version of the lemma for two variables, stated as follows: if X_1 and X_2 are jointly normally distributed, g is differentiable, and the relevant expectations exist, then

$$\mathrm{E}[(X_1-\mu_1)g(X_2)] = \mathrm{Cov}(X_1,X_2)\cdot\mathrm{E}[g'(X_2)].$$

Let $X_1 = -z$, $X_2 = -\ln\alpha$, and $g(X_2) = e^{X_2}$ or $g(X_2) = e^{2X_2}$ as appropriate.[25] Then we have

$$\begin{aligned}
&\mathrm{E}[-(z-\mu_z)\alpha^{-1}] = \sigma_{\alpha z}\mathrm{E}[\alpha^{-1}] = \sigma_{\alpha z}\delta^2\mathrm{E}(\beta_2);\\
&\mathrm{E}[-(z-\mu_z)\alpha^{-2}] = \sigma_{\alpha z}2\mathrm{E}[\alpha^{-2}] = \sigma_{\alpha z}2\delta^4\mathrm{E}(\beta_2^2) = \sigma_{\alpha z}2\delta^4[\mathrm{V}(\beta_2)+\mathrm{E}(\beta_2)^2].
\end{aligned}$$

The first equality in each line above applies the lemma, and the second equality uses $\alpha^{-1} = \delta^2\beta_2$ (by definition). The last equality in the second line uses the definitional relationship between the variance and expectations. Finally we insert these results into the expression for $\mathrm{Cov}(\nu,\beta_2)$:

$$\begin{aligned}
\mathrm{Cov}(\nu,\beta_2) &= \delta^{-4}\left(\sigma_{\alpha z}2\delta^4[\mathrm{V}(\beta_2)+\mathrm{E}(\beta_2)^2] - \sigma_{\alpha z}\delta^2\mathrm{E}(\beta_2)\cdot\delta^2\mathrm{E}(\beta_2)\right)\\
&= \sigma_{\alpha z}\left(2\mathrm{V}(\beta_2)+\mathrm{E}(\beta_2)^2\right).
\end{aligned}$$

Therefore,

$$\sigma_{\alpha z} = \frac{\mathrm{Cov}(\nu,\beta_2)}{2\mathrm{V}(\beta_2)+\mathrm{E}(\beta_2)^2}.$$

c) Last, we obtain σ_z^2 from $\mathrm{V}(\nu)$, and the other moments, as follows. As with the covariance in part (b), we first put the reduced-form variance in terms of the structural parameters, and then use the relationship between the variance and expectations:

$$\begin{aligned}
\mathrm{V}(\nu) = \mathrm{V}(-(z-\mu_z)\delta^{-2}\alpha^{-1}) &= \delta^{-4}\mathrm{V}(-(z-\mu_z)\alpha^{-1})\\
&= \delta^{-4}\mathrm{E}[(-(z-\mu_z))^2\alpha^{-2}] - \delta^{-4}\mathrm{E}[-(z-\mu_z)\alpha^{-1}]^2.
\end{aligned}$$

From the derivations in part (b), we have $\mathrm{E}[-(z-\mu_z)\alpha^{-1}] = \sigma_{\alpha z}\delta^2\mathrm{E}(\beta_2)$ in the second term,

[25] Note that for $g(X_2) = e^{X_2}$ then $g(X_2) = \alpha^{-1}$ and $g'(X_2) = \alpha^{-1}$, or for $g(X_2) = e^{2X_2}$ then $g(X_2) = \alpha^{-2}$ and $g'(X_2) = 2\alpha^{-2}$.

so we must now derive the result for $\mathrm{E}[(-(z-\mu_z))^2\alpha^{-2}]$ in the first term.

We start by integrating out z via the use of iterated expectations. First,

$$\mathrm{E}[(-(z-\mu_z))^2\alpha^{-2}] = \mathrm{E}[\alpha^{-2}\mathrm{E}[(-(z-\mu_z))^2|\alpha]].$$

Then, using the relationship between the variance and expectations on the inner conditional expectation,[26]

$$\mathrm{E}[(-(z-\mu_z))^2|\alpha] = \mathrm{V}[-(z-\mu_z)|\alpha] + \mathrm{E}[-(z-\mu_z)|\alpha]^2$$

Because z and $\ln\alpha$ are joint normal (as are $-z$ and $-\ln\alpha$), we have

$$\begin{aligned}\mathrm{V}[-(z-\mu_z)|\alpha] &= \mathrm{V}[-z|-\ln\alpha] = \sigma_z^2 - \frac{\sigma_{\alpha z}^2}{\sigma_\alpha^2}\\ \mathrm{E}[-(z-\mu_z)|\alpha]^2 &= (\mathrm{E}[-z|-\ln\alpha]+\mu_z)^2 = \left(\frac{\sigma_{\alpha z}}{\sigma_\alpha^2}(-\ln\alpha+\mu_\alpha)\right)^2.\end{aligned}$$

Substituting these back into the outer (unconditional) expectation, we have

$$\mathrm{E}[(-(z-\mu_z))^2\alpha^{-2}] = \left(\sigma_z^2 - \frac{\sigma_{\alpha z}^2}{\sigma_\alpha^2}\right)\mathrm{E}[\alpha^{-2}] + \left(\frac{\sigma_{\alpha z}}{\sigma_\alpha^2}\right)^2 \mathrm{E}[\alpha^{-2}(-\ln\alpha+\mu_\alpha)^2].$$

In part (b) we showed that $\mathrm{E}[\alpha^{-2}] = \delta^4[\mathrm{V}(\beta_2)+\mathrm{E}(\beta_2)^2]$, so we must now derive a result for $\mathrm{E}[\alpha^{-2}(-\ln\alpha+\mu_\alpha)^2]$ in the second term.

To do this we apply Stein's lemma to $-\ln\alpha$, although to simplify the expressions, here we write X in place of $-\ln\alpha$. In the univariate case the lemma is stated as follows: if X is normally distributed, g is differentiable, and the relevant expectations exist, then $\mathrm{E}[(X-\mu_X)g(X)] = \mathrm{V}(X)\cdot\mathrm{E}[g'(X)]$. This must be applied twice, as follows:

$$\begin{aligned}&\mathrm{E}[\alpha^{-2}(-\ln\alpha+\mu_\alpha)^2] = \mathrm{E}[e^{2X}(X-\mu_X)^2] =\\ \text{(i) }&\mathrm{E}[(X-\mu_X)\cdot\underbrace{e^{2X}(X-\mu_X)}_{g(X)}] = \sigma_X^2\mathrm{E}[\underbrace{2e^{2X}(X-\mu_X)+e^{2X}}_{g'(X)}] =\\ \text{(ii) }&\sigma_X^2\mathrm{E}[(X-\mu_X)\cdot\underbrace{2e^{2X}}_{g(X)}] + \sigma_\alpha^2\mathrm{E}[e^{2X}] = (\sigma_X^2)^2\mathrm{E}[\underbrace{4e^{2X}}_{g'(X)}] + \sigma_X^2\mathrm{E}[e^{2X}]\\ &= (4(\sigma_X^2)^2+\sigma_X^2)\mathrm{E}[e^{2X}] = (4(\sigma_\alpha^2)^2+\sigma_\alpha^2)\mathrm{E}[\alpha^{-2}]\end{aligned}$$

[26] Note this is not simply the conditional variance of z because μ_z is not the conditional mean.

Substituting this in above, we have

$$\begin{aligned}
\mathrm{E}[(-(z-\mu_z))^2\alpha^{-2}] &= \left(\sigma_z^2 - \frac{\sigma_{\alpha z}^2}{\sigma_\alpha^2}\right)\mathrm{E}[\alpha^{-2}] + \left(\frac{\sigma_{\alpha z}}{\sigma_\alpha^2}\right)^2 (4(\sigma_\alpha^2)^2 + \sigma_\alpha^2)\mathrm{E}[\alpha^{-2}] \\
&= \left(\sigma_z^2 + 4(\sigma_{\alpha z})^2\right)\mathrm{E}[\alpha^{-2}] \\
&= \left(\sigma_z^2 + 4(\sigma_{\alpha z})^2\right)\delta^4[\mathrm{V}(\beta_2) + \mathrm{E}(\beta_2)^2].
\end{aligned}$$

where the last equality uses $\mathrm{E}[\alpha^{-2}] = \delta^4[\mathrm{V}(\beta_2) + \mathrm{E}(\beta_2)^2]$ from part (b). Finally, bringing the results together, we have

$$\begin{aligned}
\mathrm{V}(\nu) &= \delta^{-4}\left((\sigma_z^2 + 4(\sigma_{\alpha z})^2)\delta^4[\mathrm{V}(\beta_2) + \mathrm{E}(\beta_2)^2] - (\sigma_{\alpha z}\delta^2\mathrm{E}[\beta_2])^2\right) \\
&= (\sigma_z^2 + 4(\sigma_{\alpha z})^2)[\mathrm{V}(\beta_2) + \mathrm{E}(\beta_2)^2] - (\sigma_{\alpha z})^2\mathrm{E}(\beta_2)^2
\end{aligned}$$

Therefore

$$\sigma_z^2 = \frac{\mathrm{V}(\nu) + (\sigma_{\alpha z})^2\mathrm{E}(\beta_2)^2}{\mathrm{V}(\beta_2) + \mathrm{E}(\beta_2)^2} - 4(\sigma_{\alpha z})^2.$$

□

Thus we have closed-form expressions for the structural parameters $\mu_{\alpha,k}$, $\sigma_{\alpha,k}^2$, $\sigma_{\alpha z,k}$, and $\sigma_{z,k}^2$ as functions of the reduced-form moments $\mathrm{E}(\beta_2^k)$, $\mathrm{V}(\beta_2^k)$, $\mathrm{V}(\nu^k)$, and $\mathrm{Cov}(\beta_2^k, \nu^k)$. This establishes that the parameters of $F_k(\alpha, z)$ are uniquely identified by these moments (along with δ and the external information on μ_z). Furthermore these expressions are continuous, so the consistent estimates of the reduced-form moments from the OLS estimation of (11) and the auxiliary regression above yield consistent estimates of the structural parameters.

G Calibrations

G.1 Calibration of μ_z

As described in the paper, we use external information on the costs of acquiring and administering EPO to calibrate the value of the mean per-unit cost, μ_z. For the acquisition cost, we use the median across facilities of the per-unit cost of purchasing the drug from a distributor (net of discounts and rebates), computed from Renal Dialysis Facility Cost Report Data, which is equal to \$7.53 per 1,000 units. For the administration cost, we compute an average per-unit cost of staff time and non-drug supplies based on results from Schiller et al. (2008), as follows. Schiller et al. (2008) reports an average cost for EPO administration of \$3.63 per dialysis session, and an average of 13.0 sessions per month, for a total cost of

\$47.19 per month. From our claims data, the median dosage per month is about 45,000 units.[27] (We refer to the median rather than the mean because the former is not sensitive to large dosages that occur with low probability, which were unlikely in the smaller sample used by the Schiller et al. (2008) study.) Dividing \$47.19 by 45,000, we arrive at an average administration cost of \$1.05 per 1,000 units. Adding this to the acquisition cost, we obtain a value of μ_z equal to \$8.58 per 1,000 units.

G.2 Calibration of α_g

We use information on the relationship between hematocrit levels and mortality risk from a large clinical trial (Singh et al., 2006) and an estimate of the value of a statistical life-year (VSLY) from Aldy and Viscusi (2008) to calibrate the value of α_g. The parameter expresses the conversion (i.e., marginal rate of substitution) in the government's objective function between health—specified as a quadratic function of the dosage of EPO—and dollars. The clinical trial gives estimates of the mortality risk associated with different hematocrit levels (which result from different dosages), so under certain assumptions (described below), we can find a value of α_g that equates the difference in a quadratic function of the hematocrit levels with the difference in mortality risks multiplied by the VSLY.

The clinical trial (Singh et al., 2006) compared outcomes between patients with chronic kidney disease who were randomly assigned to target levels of hemoglobin equal to 11.3 g/dl and 13.5 g/dl. The lower target group achieved a mean hemoglobin level of 11.3 g/dl, comparable to a 33.9% hematocrit level, while the higher target group only achieved a mean hemoglobin level of 12.6 g/dl, comparable to a 37.8% hematocrit level. The cumulative probability of death or serious cardiovascular event (e.g., heart attack, stroke) was 0.175 for the higher target group and 0.135 for the lower target group (p. 2090), over a period of about 30 months (Figure 3, p. 2093). Assuming a uniform distribution of these events over time, the difference in the probability of death or serious cardiovascular event over one year would be 0.016 between the higher and lower target groups. Thus we have a relationship between hematocrit levels and the annual risk of death or a debilitating health event, at two points in the distribution of hematocrit.

If we assume how the targets used in the trial relate to the true point where health is maximized (i.e., where $h'(a; b, x) = 0$), we can compute the difference in health from the two targets, as defined by our quadratic specification. We assume that the lower target used in the trial is equal to τ, where health is maximized, implying that the difference in health

[27] This is the median without restricting to the three main hematocrit intervals.

from the two targets is equal to 7.6, as follows:

$$\left(-\frac{1}{2}(33.9-\tau)^2\right)-\left(-\frac{1}{2}(37.8-\tau)^2\right)=\frac{1}{2}(33.9-33.9)^2+\frac{1}{2}(37.8-33.9)^2=7.6.$$

Multiplying this by α_g will give the government's value of this difference in health, in terms of dollars.

If we further assume that the government's value of this difference in health comes entirely from the difference in the risk of death or a debilitating health event, we can find the monetary value of this difference in health by multiplying a VSLY estimate by the difference in these risks from the two target levels. Aldy and Viscusi (2008) provides VSLY estimates of approximately $300,000 (p. 580), so the annual value of the difference in risks would be $0.016 \times \$300,000 = \$4,800$. Finally, because the time periods in our model are months, this would equal the government's value of the above difference in health over twelve periods. Therefore, we have

$$12 \times 7.6\,\alpha_g = 0.016 \times \$300,000,$$

which yields our calibrated value of $\alpha_g = 52.6$.

H Posterior Means of α and z

Given the estimated distributions of α and z, posterior distributions can be computed for each provider by applying Bayes' Theorem, as follows. Let $g(a|b,p,x;\alpha,z)$ denote the density function for the dosage conditional on the patient's covariates (b,p,x) and the provider's type (α,z). To fully specify this density function, a distribution for the error term η (equivalently, ϵ) in the reduced form (11) is needed (note that the reduced form shows how a is a function of η and the other variables and parameters). Accordingly, let η have a normal distribution with mean zero and variance σ_η^2, and denote its density as $\phi(\eta;\sigma_\eta^2)$.

For a provider i with a set of patient-month observations $JT(i)$, the posterior density of (α,z) is proportional to

$$\prod_{jt\in JT(i)} g(a_{ijt}|b_{jt},p_t,x_{jt};\alpha,z)\cdot f_k(\alpha,z)$$

(see, e.g., Train, 2009, Chapter 11). We use this to compute posterior means of α and z for each provider (in each hematocrit interval k) via Monte Carlo integration. First we draw values of (α,z) from the estimated distribution $F_k(\alpha,z)$. Then with each draw, $(\hat{\alpha}_{ik}^s,\hat{z}_{ik}^s)$, we calculate the value of the error term for each observation, $jt \in JT(i)$, as follows:

$$\hat{\eta}^s_{ijt} = a_{ijt} - \left[\frac{-1}{\delta_k}\right] b_{jt} + \left[\frac{1}{\hat{\alpha}^s_{ik}\delta^2_k}\right] p_{1t} + \left[\frac{\tau'_k}{\delta_k}\right] x_{jt} + \left[\frac{-\hat{z}^s_{ik}}{\hat{\alpha}^s_{ik}\delta^2_k}\right] \tag{O21}$$

(this comes from rearranging the reduced form). The conditional density of the dosage for each observation, $g(a_{ijt}|b_{jt}, p_t, x_{jt}; \hat{\alpha}^s_{ik}, \hat{z}^s_{ik})$, is equal to the density of this error term, $\phi(\hat{\eta}^s_{ijt}; \sigma^2_{\eta,k})$. Finally, the posterior mean of α for provider i (in hematocrit interval k) is equal to

$$\frac{\sum_{s=1}^{S} \hat{\alpha}^s_{ik} \prod_{jt\in JT(i)} \phi(\hat{\eta}^s_{ijt}; \sigma^2_{\eta,k})}{\sum_{s=1}^{S} \prod_{jt\in JT(i)} \phi(\hat{\eta}^s_{ijt}; \sigma^2_{\eta,k})},$$

and similarly for the posterior mean of z. To complete these computations, the estimated parameters are used in (O21), and the variance $\sigma^2_{\eta,k}$ is set equal to the variance of the reduced-form residuals in that interval.

Table O2 presents summary statistics on these provider-level posterior means, by ownership type and by chain affiliation. Among for-profit dialysis centers, for example, the median of the center-specific posterior means of α is 31.8 and the mean is 36.3, in the bottom interval of baseline hematocrit. By comparison, among non-profit and governmental centers, the median is 34.8 and the mean is 41.0 in that interval, indicating somewhat greater weight placed on patient health, on average. The posterior means of z, the marginal cost, are noticeably lower among for-profit centers, with medians and means below \$8.60 in all intervals, while the medians and means among non-profit and governmental centers are mostly above \$8.70. However the distributions of α and z also overlap substantially between these two groups of providers. In all intervals, the standard deviations of the provider-level posterior means within each group are much larger than the differences between the medians or means of the two groups.

We see similar patterns comparing providers in the two large chains against all other providers. In almost all cases, the posterior means of α are somewhat lower in DaVita and Fresenius centers, compared to all other centers. The marginal costs are also consistently lower for centers in the two large chains, compared to other centers. The variation in marginal costs is lower within the large chains as well, typically by one quarter to one third. This is broadly consistent with the variation in acquisition costs observed in Medicare cost report data (see footnote 32 in the paper).

Table O2: Distribution of Provider-Level Posterior Means

	Altruism (α)			Marginal Cost (z)		
	Interval of Baseline Hematocrit			Interval of Baseline Hematocrit		
	$>$ 30 to 33,	$>$ 33 to 36,	$>$ 36 to 39	$>$ 30 to 33,	$>$ 33 to 36,	$>$ 36 to 39
I) Ownership Type						
a) Non-profit and governmental						
Median	34.8	15.1	21.1	8.70	8.71	8.61
Mean	41.0	16.5	28.0	8.88	8.96	8.74
Std. Dev.	82.7	23.3	61.2	0.60	0.83	0.40
b) For-profit						
Median	31.8	15.3	20.5	8.55	8.50	8.58
Mean	36.3	17.3	27.3	8.55	8.55	8.59
Std. Dev.	59.3	25.3	63.6	0.39	0.64	0.28
II) Chain Affiliation						
c) DaVita						
Median	28.8	15.7	22.1	8.50	8.44	8.58
Mean	33.3	17.5	26.3	8.44	8.43	8.58
Std. Dev.	57.0	24.6	42.6	0.30	0.56	0.22
d) Fresenius						
Median	30.7	14.4	18.6	8.55	8.47	8.56
Mean	35.7	16.7	26.2	8.55	8.51	8.53
Std. Dev.	55.2	24.6	73.0	0.34	0.59	0.24
e) Other/Indep.						
Median	35.6	16.6	20.5	8.61	8.65	8.60
Mean	40.0	17.5	28.1	8.72	8.80	8.68
Std. Dev.	66.5	26.0	69.1	0.51	0.75	0.36

Posterior means computed in each interval for each facility using estimated model parameters, as described in Appendix H. Ownership type and chain affiliation of each facility taken from Medicare cost report data.

I Check of Regularity Condition

Figure O3 plots the supply curves (dashed, grey lines) of physician types providing each treatment amount for a patient with the median baseline hematocrit level in the lowest, middle, and highest baseline hematocrit intervals, and shows that none intersect the marginal payment curve (solid, black line) more than once.

J Sensitivity Analyses and Other Assessments

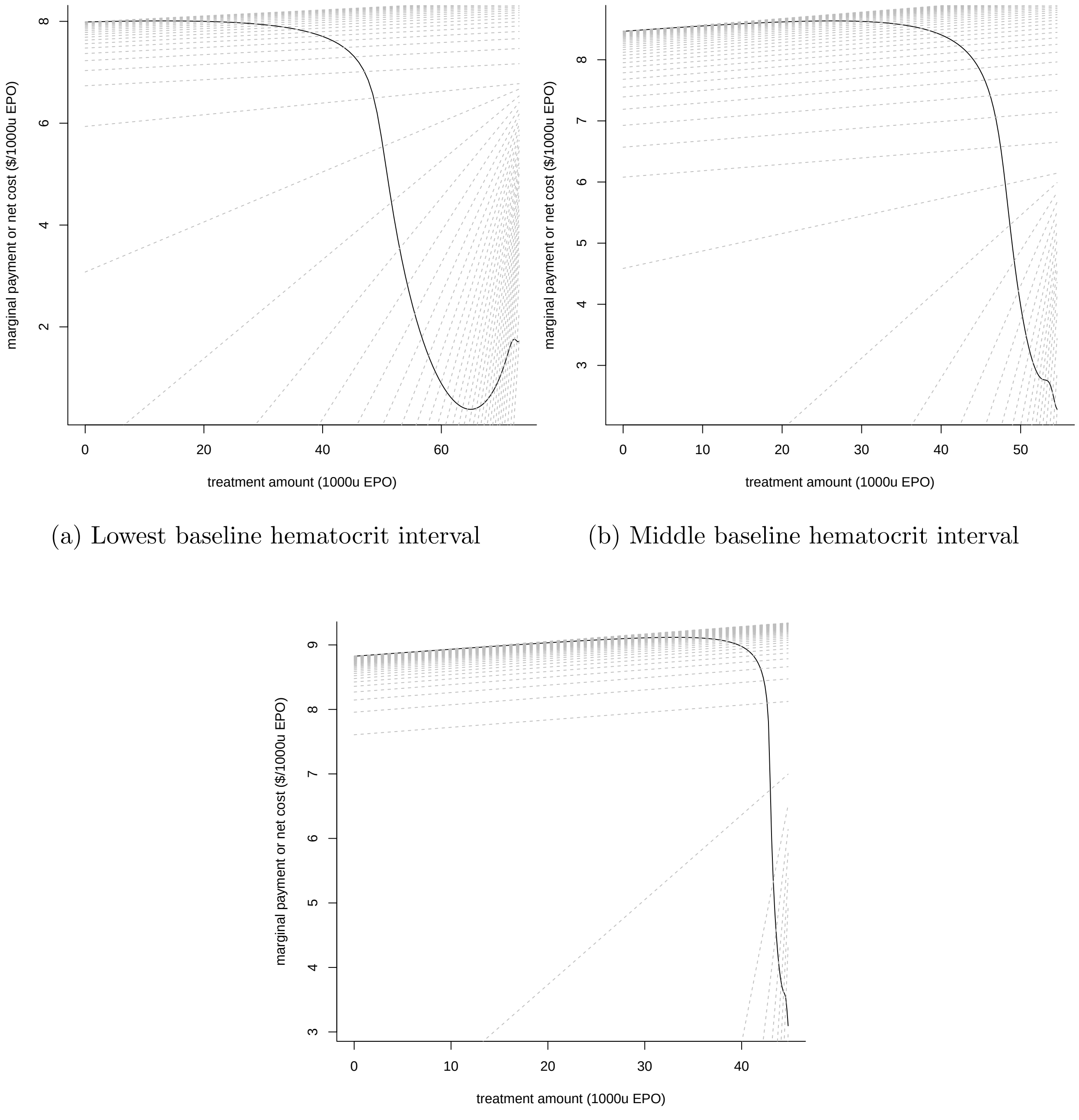
J.1 Robustness of the Reduced Form

Table O3 presents estimates of the reduced form using fixed effects for either the provider, physician, or patient. The sample sizes for the regressions with physician fixed effects are slightly smaller because some observations do not include a physician identifier. The regressions with patient fixed effects omit patient characteristics because they have limited variation within patients over time. Table O4 provides the full estimation results for the alternative specifications of the reduced form reported in Table 3, columns 4 to 9. Table O5 provides the results for the main specification with asymptotic standard errors clustered on chains rather than facilities.

J.2 Variability of Hematocrit within Patients over Time

Table O6 describes the variability of hematocrit levels within patients over time, by showing the distribution of hematocrit values reported on patients' prior monthly claims given the values on their current monthly claims. Each column shows this distribution for a one-percentage-point interval in the current hematocrit. For example, among patients with current hematocrit greater than 34 and less than or equal to 35 (the column labeled ">34 - 35"), 16.4% had hematocrit in that same interval reported on their prior monthly claim, while 11.2% and 5.5% had hematocrit levels of >33 - 34 and >31 - 32, respectively (the corresponding rows in that column).

The prior monthly claim is defined as the claim with a start date of its claim period between 25 and 34 days before the start date of the current claim period. (In rare cases where multiple such claims are found, the claim with the lowest encrypted claim ID number is used.) Such a prior monthly claim could not be found for about one-fifth of the current monthly observations, which mostly reflects new beneficiaries without prior claims.



(a) Lowest baseline hematocrit interval

(b) Middle baseline hematocrit interval

(c) Highest baseline hematocrit interval

Figure O3: Regularity condition check, for patients with different severities of anemia.

Notes: Figure plots marginal payment curve (solid, black line) and physician supply curves (dashed, grey lines) for patients with median baseline hematocrit and mean target hematocrit in the lowest (panel a), middle (panel b), and highest (panel c) baseline hematocrit intervals.

Table O3: Fixed Effects Estimates of the Reduced Form

	Provider (Facility) Fixed Effects			Physician Fixed Effects			Patient Fixed Effects		
Interval:	> 30 to 33,	> 33 to 36,	> 36 to 39	> 30 to 33,	> 33 to 36,	> 36 to 39	> 30 to 33,	> 33 to 36,	> 36 to 39
Variable	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Hematocrit	-9.22	-6.51	-4.00	-9.25	-6.44	-3.90	-6.10	-4.72	-3.90
	(0.19)	(0.13)	(0.12)	(0.20)	(0.12)	(0.12)	(0.15)	(0.08)	(0.09)
Reimb. rate	9.42	5.99	4.67	7.34	6.95	4.29	9.90	7.15	11.99
	(3.00)	(1.95)	(1.85)	(2.99)	(1.93)	(1.86)	(2.17)	(1.18)	(1.28)
Age in years	-0.37	-0.33	-0.24	-0.37	-0.33	-0.24			
	(0.02)	(0.01)	(0.01)	(0.02)	(0.01)	(0.01)			
Female sex	-1.53	1.21	2.37	-1.25	1.38	2.56			
	(0.49)	(0.38)	(0.33)	(0.51)	(0.40)	(0.34)			
Charlson=1	7.95	7.06	6.49	8.00	7.12	6.58			
	(0.86)	(0.65)	(0.59)	(0.87)	(0.65)	(0.58)			
Charlson=2	10.28	9.70	7.92	10.35	9.32	7.58			
	(0.81)	(0.63)	(0.57)	(0.82)	(0.63)	(0.56)			
Charlson=3	12.58	11.08	8.72	12.55	11.06	8.44			
	(0.88)	(0.70)	(0.58)	(0.90)	(0.71)	(0.59)			
Charlson=4	15.03	13.76	10.63	15.28	13.37	10.72			
	(1.05)	(0.82)	(0.70)	(1.06)	(0.84)	(0.72)			
Charlson=5	16.18	14.51	11.26	16.07	14.94	11.72			
	(1.26)	(1.01)	(0.89)	(1.28)	(1.05)	(0.89)			
Charlson=6	17.79	18.03	13.43	18.41	19.01	13.66			
	(1.61)	(1.35)	(1.14)	(1.62)	(1.35)	(1.13)			
Charlson=7	23.43	24.35	19.94	23.61	24.71	19.83			
	(2.61)	(2.30)	(2.12)	(2.83)	(2.42)	(2.17)			
Charlson=8	22.99	21.98	15.69	20.84	21.82	15.16			
	(3.56)	(3.09)	(2.50)	(3.44)	(2.95)	(2.53)			
Charlson=9	31.52	32.94	23.43	33.38	31.48	24.26			
	(4.97)	(4.08)	(3.98)	(4.99)	(4.19)	(3.94)			
Charlson=10	22.54	27.63	29.77	19.85	27.83	27.59			
	(6.16)	(6.46)	(6.76)	(7.08)	(6.35)	(6.82)			
Charlson=11	40.89	40.81	39.64	35.09	40.68	38.61			
	(8.45)	(8.04)	(7.07)	(8.24)	(8.08)	(7.17)			
Charlson=12	27.80	27.21	16.09	32.19	27.89	22.44			
	(10.18)	(7.17)	(10.17)	(9.61)	(7.51)	(8.46)			
Observations	231,702	405,019	283,024	230,455	402,811	281,411	231,702	405,019	283,024
R-squared	0.029	0.027	0.021	0.029	0.026	0.020	0.085	0.056	0.044
RMSE	65.78	55.05	46.29	66.05	55.15	46.34	48.45	38.40	33.61

Each column is a separate regression. Regressions also include month and year dummies.
Robust standard errors in parentheses, clustered on panel unit for facility and physician fixed effects.

Table O4: Alternative Specifications of the Reduced Form

	No Patient Observables			Comorbidity Indicators		
Interval:	> 30 to 33,	> 33 to 36,	> 36 to 39	> 30 to 33,	> 33 to 36,	> 36 to 39
Variable	(1)	(2)	(3)	(4)	(5)	(6)
Hematocrit	-9.61	-6.39	-3.46	-9.25	-6.32	-3.56
	(0.24)	(0.15)	(0.13)	(0.24)	(0.15)	(0.13)
Reimb. rate	9.81	6.13	4.26	9.39	6.07	4.07
	(3.20)	(2.04)	(1.92)	(3.20)	(2.03)	(1.91)
Age in years				-0.41	-0.37	-0.27
				(0.02)	(0.02)	(0.01)
Female sex				-0.70	1.62	2.95
				(0.55)	(0.40)	(0.34)
Myocardial inf.				-0.65	0.28	-0.74
				(1.09)	(0.88)	(0.74)
Cong. hrt. failure				9.23	9.07	7.04
				(0.80)	(0.59)	(0.50)
Periph. vasc. dis.				4.10	3.60	3.11
				(1.01)	(0.78)	(0.66)
Cerebro vasc. dis.				-2.41	-0.26	-0.44
				(1.19)	(0.98)	(0.74)
Dementia				-2.86	0.10	0.19
				(2.73)	(1.96)	(1.58)
Chron. pulm. dis.				3.51	3.04	1.97
				(0.88)	(0.65)	(0.58)
Rheumatic dis.				6.30	8.44	5.30
				(2.18)	(1.81)	(1.50)
Peptic ulcer dis.				9.50	7.24	6.33
				(2.15)	(1.71)	(1.41)
Mild liver dis.				6.57	4.05	3.30
				(2.23)	(1.62)	(1.37)
Diabetes w/out comp.				4.69	4.48	3.62
				(0.73)	(0.56)	(0.48)
Diabetes w/chron. comp.				1.47	0.83	0.72
				(0.80)	(0.59)	(0.51)
Hemi/para-plegia				3.44	2.95	0.93
				(3.25)	(2.38)	(2.03)
Any malignancy				12.62	10.73	8.29
				(1.95)	(1.57)	(1.38)
Mod/severe liver dis.				18.14	21.82	17.08
				(5.17)	(3.77)	(3.47)
Metastatic tumor				14.65	10.90	11.07
				(4.55)	(3.60)	(3.45)
AIDS/HIV				20.55	21.82	18.14
				(3.99)	(3.22)	(2.96)
Observations	231,702	405,019	283,024	231,702	405,019	283,024
R-squared	0.014	0.009	0.005	0.031	0.030	0.022
RMSE	71.98	59.01	49.40	71.37	58.40	48.98

Each column is a separate regression. Regressions also include month and year dummies.
Robust standard errors in parentheses, clustered on dialysis centers.

Table O5: Alternative Clusters for the Standard Errors

	Clustered on Dialysis Centers			Clustered on Chains		
Interval:	> 30 to 33,	> 33 to 36,	> 36 to 39	> 30 to 33,	> 33 to 36,	> 36 to 39
Variable	(1)	(2)	(3)	(4)	(5)	(6)
Hematocrit	-9.29	-6.32	-3.56	-9.29	-6.32	-3.56
	(0.24)	(0.15)	(0.13)	(0.46)	(0.98)	(0.40)
Reimb. rate	9.53	6.39	3.92	9.53	6.39	3.92
	(3.19)	(2.03)	(1.91)	(7.83)	(6.52)	(4.26)
Age in years	-0.41	-0.37	-0.26	-0.41	-0.37	-0.26
	(0.02)	(0.02)	(0.01)	(0.03)	(0.02)	(0.02)
Female sex	-0.88	1.55	2.89	-0.88	1.55	2.89
	(0.55)	(0.40)	(0.34)	(1.14)	(0.54)	(0.54)
Charlson=1	9.03	8.03	7.36	9.03	8.03	7.36
	(0.96)	(0.69)	(0.60)	(1.26)	(1.05)	(0.67)
Charlson=2	10.73	10.23	8.20	10.73	10.23	8.20
	(0.90)	(0.67)	(0.59)	(1.49)	(0.99)	(0.63)
Charlson=3	13.84	11.85	8.57	13.84	11.85	8.57
	(0.94)	(0.72)	(0.60)	(1.74)	(1.04)	(0.64)
Charlson=4	15.52	13.91	10.82	15.52	13.91	10.82
	(1.22)	(0.86)	(0.73)	(2.50)	(1.65)	(0.94)
Charlson=5	16.53	15.00	11.88	16.53	15.00	11.88
	(1.40)	(1.08)	(0.93)	(2.96)	(1.95)	(1.31)
Charlson=6	18.61	18.50	13.83	18.61	18.50	13.83
	(1.87)	(1.48)	(1.21)	(2.86)	(3.20)	(1.51)
Charlson=7	26.20	26.00	20.38	26.20	26.00	20.38
	(3.02)	(2.48)	(2.19)	(3.99)	(3.97)	(3.90)
Charlson=8	23.92	24.24	14.50	23.92	24.24	14.50
	(3.94)	(3.06)	(2.51)	(3.48)	(2.95)	(2.53)
Charlson=9	31.98	32.42	22.85	31.98	32.42	22.85
	(4.98)	(4.17)	(3.81)	(5.78)	(5.84)	(2.72)
Charlson=10	23.88	28.45	32.23	23.88	28.45	32.23
	(7.02)	(6.71)	(6.96)	(5.31)	(7.77)	(5.02)
Charlson=11	39.10	43.62	39.80	39.10	43.62	39.80
	(11.01)	(8.79)	(7.31)	(8.76)	(7.06)	(6.64)
Charlson=12	38.39	33.50	25.66	38.39	33.50	25.66
	(12.51)	(8.06)	(9.82)	(12.21)	(6.21)	(9.20)
Observations	231,702	405,019	283,024	231,702	405,019	283,024
R-squared	0.029	0.028	0.021	0.029	0.028	0.021
RMSE	71.43	58.46	49.01	71.43	58.46	49.01

Each column is a separate regression. Regressions also include month and year dummies.
Robust standard errors in parentheses, clustered on dialysis centers or chains as indicated.

Table O6: Distribution of Hematocrit on Current and Prior Month Claims

						Current HCT					
Lagged HCT	**=,< 30**	**>30 - 31**	**>31 - 32**	**>32 - 33**	**>33 - 34**	**>34 - 35**	**>35 - 36**	**>36 - 37**	**>37 - 38**	**>38 - 39**	**> 39**
=,< 30	0.363	0.210	0.155	0.109	0.076	0.056	0.041	0.035	0.031	0.029	0.029
>30 - 31	0.088	0.116	0.080	0.064	0.048	0.036	0.027	0.022	0.019	0.016	0.017
>31 - 32	0.088	0.103	0.125	0.088	0.070	0.055	0.043	0.034	0.029	0.026	0.024
>32 - 33	0.107	0.137	0.149	0.168	0.134	0.112	0.090	0.073	0.062	0.055	0.048
>33 - 34	0.081	0.106	0.120	0.133	0.157	0.129	0.108	0.089	0.076	0.068	0.056
>34 - 35	0.067	0.089	0.106	0.124	0.141	0.164	0.137	0.121	0.102	0.090	0.073
>35 - 36	0.069	0.088	0.102	0.126	0.149	0.174	0.205	0.184	0.171	0.155	0.131
>36 - 37	0.040	0.049	0.055	0.066	0.082	0.097	0.120	0.151	0.139	0.135	0.118
>37 - 38	0.031	0.035	0.039	0.046	0.056	0.069	0.090	0.111	0.145	0.136	0.128
>38 - 39	0.028	0.030	0.033	0.037	0.045	0.055	0.074	0.097	0.118	0.156	0.159
> 39	0.037	0.035	0.036	0.040	0.042	0.052	0.065	0.083	0.108	0.134	0.218
Matched	75,275	37,391	50,978	90,691	93,551	103,853	134,913	89,221	73,106	67,450	66,975
(Pct)	62.8%	73.3%	76.5%	79.5%	81.4%	81.9%	82.6%	81.9%	81.2%	80.2%	76.5%
Unmatched	44,513	13,595	15,667	23,380	21,307	22,895	28,500	19,652	16,929	16,666	20,620
(Pct)	37.2%	26.7%	23.5%	20.5%	18.6%	18.1%	17.4%	18.1%	18.8%	19.8%	23.5%
Total	119,788	50,986	66,645	114,071	114,858	126,748	163,413	108,873	90,035	84,116	87,595

Each column shows the distribution of hematocrit levels reported on the prior monthly claim, given the level on the current monthly claim. The proportions are among those claims where a prior claim could be found, defined as a claim with a start date between 25 and 34 days before the current start date. The numbers of current claims with (Matched) and without (Unmatched) prior month claims are reported at the bottom.

J.3 Distributions of Facility Residuals and Test of Unimodality

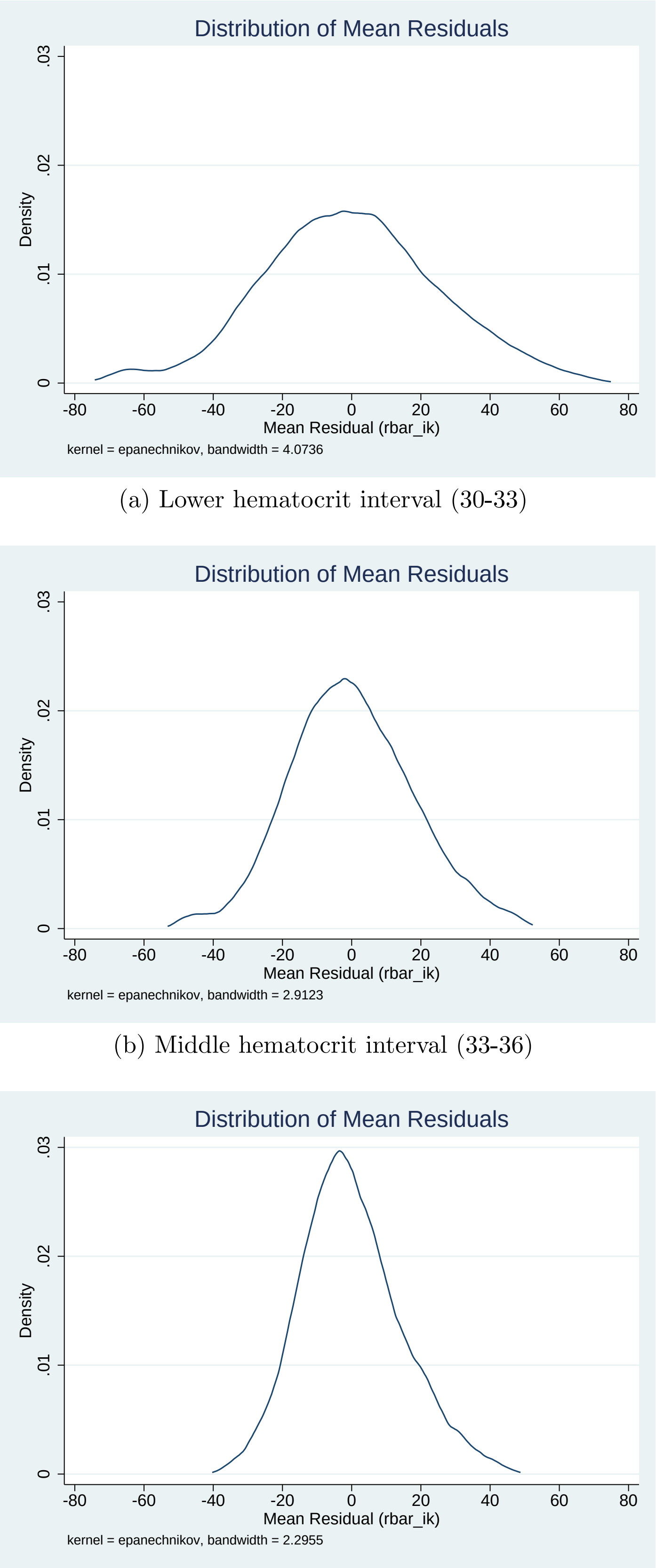
Figure O4 shows the distributions of the facility-level mean residuals ($\bar{r}_i^k$, defined in Appendix F) in each hematocrit interval. We formally test the null hypothesis of unimodality for these three distributions using a "dip test" (Hartigan and Hartigan, 1985), implemented with the user-written command `diptest` in Stata (Cox, 2009). The test statistics (p-values) in each interval are as follows: 0.0033 (0.9930), 0.0035 (0.9930), and 0.0029 (1.0000). Thus the null hypothesis of unimodality is not rejected in any interval; indeed, the test statistics are quite small with p-values quite close to one.

J.4 Downstream Medical Costs

We combine estimates of the effects of EPO on transfusions and hospitalizations from Eliason et al. (2022) with estimates of the average costs of transfusions and hospitalizations from other sources noted below, to calculate a rough estimate of the the change in downstream medical costs under the optimal nonlinear contract.

The exact sources and values are as follows:

- Effect of 1,000u of EPO on monthly transfusion rate: -0.000586 (Eliason et al. (2022), Table 7, column 4 – IV estimate of the effect on transfusions)



(a) Lower hematocrit interval (30-33)

(b) Middle hematocrit interval (33-36)

(c) Upper hematocrit interval (36-39)

Figure O4: Distribution of facility-level mean residuals ($\bar{r}_i^k$)

- Effect of 1,000u of EPO on monthly hospitalization rate: 0.000205 (Eliason et al. (2022), Table 8, column 2 – IV estimate of the effect on hospitalization for any cause)
- Mean expenditure per outpatient transfusion episode among a sample of chronic dialysis patients: \$854 (Gitlin et al., 2012, Table 2)
- Mean per-person per-year Medicare inpatient expenditures for ESRD patients in 2009: \$25,244 (United States Renal Data System, 2020, Figure 9.6)
- Mean per-person per-year number of hospitalizations for ESRD patients in 2009: 1.82 (United States Renal Data System, 2020, Figure 4.1)
- Mean Medicare inpatient expenditures per hospitalization: \$25,244 / 1.82 = \$13,870 (derived from above)

With these values, we calculate the change in downstream costs that would result from the change in the mean monthly dosage of EPO under the optimal nonlinear contract, equal to -11.4 thousand units, as follows:

$$-11.4 \times \left[(-0.000586 \times \$854) + (0.000205 \times \$13,870)\right] = -\$26.71 \text{ per patient per month.}$$

K Forcing Contract

This section describes how we compute the forcing contract implementing the maximum dosage under the full-information allocation, $\overline{a}^{*\text{FI}}$, and associated gains to the government over the observed contract for the middle hematocrit interval.

Let $P^{\text{force}}(a)$ denote the forcing payment contract, where

$$P^{\text{force}}(a) = \begin{cases} \overline{P} & \text{for } a = \overline{a}^{*\text{FI}} \\ -\infty & \text{else} \end{cases}. \tag{O22}$$

Solving the principal's problem then amounts to finding the value of $\overline{P}$ that maximizes its objective, subject to the usual voluntary participation constraint and an adapted incentive compatibility constraint that reflects the forcing nature of the contract. This is accomplished by making the participation constraint of the type $(\underline{\alpha}, \overline{z})$ bind (note that the payment amount for $a \neq \overline{a}^{*\text{FI}}$ is relevant only for off-equilibrium behavior, and, as such, doesn't matter so long as it's less than $\overline{P}$). The solution is $\overline{P}^* = \underline{u} - \underline{\alpha} h(\overline{a}^{*\text{FI}}) + \overline{z}\overline{a}^{*\text{FI}}$ and the principal's associated objective is $\alpha_g h(\overline{a}^{*\text{FI}}) - \overline{P}^*$.

The results for the middle baseline hematocrit interval are presented in the bottom row of Table O7. While there are no medically excessive treatments under this forcing contract, the payment is larger than even under the observed payment contract, leaving massive information rents to better types. Indeed, the gain in the government objective over the observed contract (presented in the last column) is a fifth of that under the optimal nonlinear contract. This makes sense, as this (and any other) forcing contract was in the set of contracts considered by the principal when solving for the optimal unrestricted contract. Intuitively, while this forcing contract does implement a desired treatment amount for one particular agent type (highest altruism, lowest cost), the cost of getting the vast majority of agents to implement this amount is larger than the principal's valuation of any associated health benefit.

Table O7: Summary of Outcomes under Forcing and other Contracts for Patients with Median Severity of Anemia

	Mean	Mean	Std. Dev.	Share	Gain in
	Payment	Dosage	Dosage	above τ	Govt. Obj.
Observed	542	58.6	9.8	75	
Optimal Linear	396	50.4	11.8	19	\$ 98
Optimal Nonlinear	393	47.1	7.2	0	\$ 125
Forcing Contract	582	54.6	0	0	\$ 24

L Importance of Both Dimensions of Heterogeneity

One of the strengths of our framework is that we are not beholden to an assumption that there is only one dimension of heterogeneity (or, for that matter, that there exists multidimensional heterogeneity). Rather, the model can recover the variation in different dimensions and we can quantify the importance of different types of unobserved heterogeneity. Given that we found altruism heterogeneity to be more substantial than heterogeneity in marginal costs, a natural question is whether the latter type of heterogeneity matters, from a normative perspective. Accordingly, we have examined the importance of heterogeneity in z by reducing the variance of z from its estimated value of 0.858 (which is different from zero at standard significance levels) to 0.10, and then solving for the optimal nonlinear contract in this counterfactual environment.[28]

The contracts are shown in Figure O5, for dosages of 40,000 units of EPO and greater (this corresponds to over 90% of treatment amounts). Figure O5 plots the marginal payment rates of the optimal nonlinear contract under our baseline parameterization (solid, blue, line) and when the variance of z is reduced (dashed, red, line). The main difference is that the marginal payment is higher for dosages up to about 48,000 units. This reflects an increase in the marginal costs of formerly low-cost providers, when z is shrunk toward the mean. (Note that very high-cost providers are not pictured here because they provide treatment amounts lower than 40,000 units.) The dosages above 48,000 units come from types with sufficiently high altruism that their behavior is not substantially affected by changes in marginal costs.

We have also computed how the optimal nonlinear contract based on the counterfactual parameterization featuring less heterogeneity in z would affect the gains to the government from better contracting. We compute that the government would on average gain \$125 per patient/month from moving to the optimal nonlinear contract from the observed contract.[29] Using instead the optimal nonlinear contract resulting from misspecifying the model with less heterogeneity in z, the government would gain \$113 per patient-month. Some of the reduction in the gain comes directly from the higher payments under the misspecified nonlinear contract, which do not outweigh the government's valuation of the resulting increases in patient health. Thus, taking into account the full extent of the variation in z would improve the government's gain by just over 10%.

[28] We retain a positive value for σ_z^2 to avoid re-writing our algorithm to solve for the optimal nonlinear contract. Because we found non-trivial heterogeneity in both altruism and marginal costs, we wrote our algorithm assuming there was nonzero variance in each dimension; this means the results we present below likely understate the importance of heterogeneity in z.

[29] We do this for the set of comparable types, i.e., those choosing treatment levels common to the baseline distribution and the distribution under reduced σ_z^2; as this set comprises 99.8% of provider types, the value presented here is virtually identical to the value presented in our baseline results in Table 5.

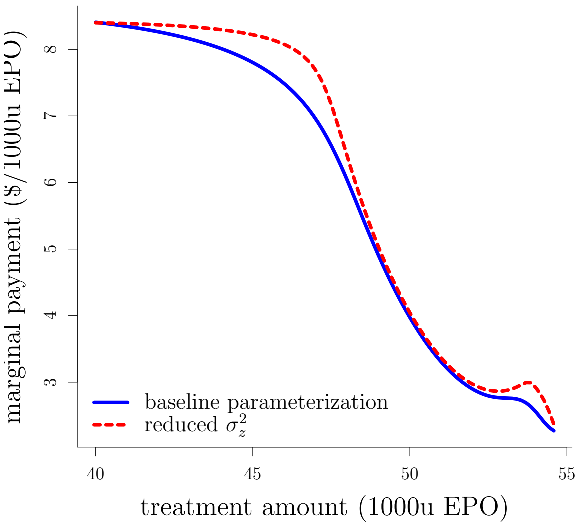


Figure O5: Comparison of marginal payments for nonlinear contracts under baseline parameterization and under parameterization with reduced σ_z^2.

References

Abito, J. M., “Measuring the Welfare Gains from Optimal Pollution Regulation,” *Review of Economic Studies*, 2019, forthcoming.

Aldy, J. E. and W. K. Viscusi, “Adjusting the Value of a Statistical Life for Age and Cohort Effects,” *Review of Economics and Statistics*, 90(3):573–581, 2008.

Cox, N. J., “DIPTEST: Stata module to compute dip statistic to test for unimodality,” Statistical Software Components, Boston College Department of Economics, 2009.

Eliason, P. J., B. Heebsh, R. J. League, R. C. McDevitt and J. W. Roberts, “The Effect of Bundled Payments on Provider Behavior and Patient Outcomes: Evidence from the Dialysis Industry,” 2022, unpublished manuscript.

Gitlin, M., J. A. Lee, D. M. Spiegel, J. L. Carson, X. Song, B. S. Custer, Z. Cao, K. A. Cappell, H. V. Varker, S. Wan and A. Ashfaq, “Outpatient red blood cell transfusion payments among patients on chronic dialysis,” *BMC Nephrology*, 13:145–153, 2012.

Goldman, M. B., H. E. Leland and D. S. Sibley, “Optimal Nonuniform Prices,” *Review of Economic Studies*, 51(2):305–319, 1984.

Härdle, W., M. Müller, S. Sperlich and A. Werwatz, *Nonparametric and Semiparametric Models*, vol. 1, Springer, 2004.

Hartigan, J. A. and P. M. Hartigan, “The Dip Test of Unimodality,” *The Annals of Statistics*, 13(1):70–84, 1985, ISSN 00905364.

Ichimura, H., “Semiparametric Least Squares (SLS) and Weighted SlS Estimation of Single-Index Models,” *Journal of Econometrics*, 58(1-2):71–120, 1993.

Johnson, S. G., “The NLopt Nonlinear-Optimization Package,” 2018, http://ab-initio.mit.edu/nlopt.

Powell, M. J., “A Direct Search Optimization Method That Models the Objective and Constraint Functions by Linear Interpolation,” in “Advances in Optimization and Numerical Analysis,” pp. 51–67, Springer, 1994.

R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2019.

Ramsey, F. P., “A Contribution to the Theory of Taxation,” *Economic Journal*, 37(145):47–61, 1927.

Schiller, B., S. Doss, E. De Cock, M. A. Del Aguila and A. R. Nissenson, “Costs of Managing Anemia with Erythropoiesis-Stimulating Agents During Hemodialysis: A Time and Motion Study,” *Hemodialysis International*, 12(4):441–449, 2008.

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.

Stein, C. M., “Estimation of the Mean of a Multivariate Normal Distribution,” *Annals of Statistics*, 9(6):1135–1151, 1981.

Train, K. E., *Discrete Choice Methods with Simulation*, Cambridge University Press, 2009.

United States Renal Data System, *2020 USRDS Annual Data Report: Epidemiology of kidney disease in the United States*, National Institutes of Health, National Institute of Diabetes and Digestive and Kidney Diseases, Bethesda, MD, 2020.

Varadhan, R. and P. Gilbert, “BB: An R Package for Solving a Large System of Nonlinear Equations and for Optimizing a High-Dimensional Nonlinear Objective Function,” *Journal of Statistical Software*, 32(4):1–26, 2009.

Vives, X., *Oligopoly Pricing: Old Ideas and New Tools*, MIT Press, 2001.

Wolak, F. A., “An Econometric Analysis of the Asymmetric Information, Regulator-Utility Interaction,” *Annales d’Economie et de Statistique*, 34:13–69, 1994.