NBER WORKING PAPER SERIES

OPTIMAL CONTRACTING WITH ALTRUISTIC AGENTS: A STRUCTURAL MODEL OF MEDICARE PAYMENTS FOR DIALYSIS DRUGS

Martin Gaynor
Nirav Mehta
Seth Richards-Shubik

Working Paper 27172
http://www.nber.org/papers/w27172

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
May 2020

We are grateful for helpful comments to David Dranove, George-Levi Gayle, Josh Gottlieb, Kate Ho, Amanda Kowalski, Albert Ma, Ryan McDevitt, Bob Miller, Ariel Pakes, Steven Stern, Lowell Taylor, and to audiences at presentations at Boston University, Carnegie Mellon, Georgia, Johns Hopkins, Michigan, NYU, North Carolina, Northwestern, Penn, Princeton, University of British Columbia, Wisconsin, the 2018 International Industrial Organization Conference, the 2018 Annual Meeting of the Society of Labor Economists, the 2018 Cowles Structural Microeconomics Conference, the 2018 Southern Economic Association Annual Meeting, the 2019 Annual Health Econometrics Workshop, and the 2020 Health Economics Research Organization meeting. We thank Ali Kamranzadeh, Martin Luccioni, and Cecilia Diaz Campo for excellent research assistance. The usual caveat applies. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Optimal Contracting with Altruistic Agents: A Structural Model of Medicare Payments for Dialysis Drugs
Martin Gaynor, Nirav Mehta, and Seth Richards-Shubik
NBER Working Paper No. 27172
May 2020
JEL No. D86,H51,I11,I13,I18,L14

ABSTRACT

We study physician agency and optimal payment policy in the context of an expensive medication used in dialysis care. Using Medicare claims data we estimate a structural model of treatment decisions, in which physicians differ in their altruism and marginal costs, and this heterogeneity is unobservable to the government. In a novel application of nonlinear pricing methods, we theoretically characterize the optimal unrestricted contract in this screening environment with multidimensional heterogeneity. We combine these results with the estimated model to construct the optimal contract and simulate counterfactual outcomes. The optimal contract is a flexible fee-for-service contract, which pays for reported treatments but uses variable marginal payments instead of constant reimbursement rates, resulting in substantial health improvements and reductions in costs. Our structural approach also yields important qualitative findings, such as rejecting the optimality of any linear contract, and may be employed more broadly to analyze a variety of applications.

Martin Gaynor
Heinz College
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA 15213-3890
and NBER
mgaynor@cmu.edu

Nirav Mehta
Department of Economics
University of Western Ontario
Social Science Centre 4063
London, ON N6A 5C2
Canada
nirav.mehta@uwo.ca

Seth Richards-Shubik
Department of Economics
College of Business and Economics
Lehigh University
Rausch Business Center, Room 465
621 Taylor St
Bethlehem, PA 18015
and NBER
sethrs@lehigh.edu

1 Introduction

A central problem in economics is how to design contracts to incentivize agents in the presence of asymmetric information (Tirole, 2014). This is particularly salient in health care, a very large and important part of the economy. Asymmetric information is pervasive in health care because physicians and other providers often have substantial information that is not observed by payers. Therefore, payers have to decide how to contract with providers to deliver care, while recognizing that providers possess relevant information that they do not. Consequently, the effects of better or worse incentives can have profound impacts on both expenditures and health.

The rich theoretical literature on contracting in asymmetric information environments provides a great deal of guidance on the design of optimal contracts. However, as noted by Chiappori and Salanié (2003) and by Einav et al. (2010), while there is a large empirical literature devoted to testing for the presence of asymmetric information in various contexts, fully developed applications of contract theory that are capable of generating normative conclusions are quite rare. Specifically, little empirical work leverages the power of contract theory to derive optimal payment contracts.[1] Notably, in spite of the pervasiveness and importance of contracting in health care, there is no empirical work deriving optimal contracts in this area.[2]

In this paper, we apply results from the theoretical literature on screening models (e.g., Myerson, 1981; Maskin and Riley, 1984; Goldman et al., 1984; Wilson, 1993), to empirically characterize optimal payment contracts for a particular treatment in dialysis care. The physician's choice about the treatment, a medication to boost red blood cell production, is well described by a screening environment. The medication is provided to nearly all patients in our setting, so the relevant choice is about a quantity. This quantity is observed because exact dosages are reported on insurance claims submitted by the provider to the payer (in this case, Medicare). At the same time, there is likely to be hidden information about physician characteristics that affect treatment choice. Given all of the above, a screening model is a natural fit. Thus, we can combine the theoretical results with our estimated screening model to derive an optimal payment contract, and thereby make normative comparisons between observed outcomes and those simulated under an optimal contract.

Our model describes a physician (the agent) who cares about patient health and their own compensation, providing a quantity of treatment to a patient and being paid for this service by a third party. The government (the principal) is the payer, which seeks to maximize patient health, net of the payment to the physician. Physicians differ in two unobservable dimensions: how altruistic they are towards patients (i.e., how much they care about patient health versus their own compensation) and their marginal costs of providing the treatment. This asymmetric information results in a suboptimal allocation, but the presence of altruism attenuates this distortion because it makes physicians willing to provide greater quantities, as if their marginal costs were lower. While

[1] The handful of empirical papers that consider fully optimal contracts include Wolak (1994), Gagnepain and Ivaldi (2002), and Abito (2019), which we discuss below.

[2] An important study on long-term care hospitals by Einav et al. (2018) simulates outcomes under alternative contracts and finds substantial improvements are possible, but the analysis is not intended to find the theoretically optimal contract.

the multidimensional nature of the unobserved heterogeneity of the agents may be a natural fit for our setting, it is not the typical setup in a screening model. This led us to adopt a solution method from the nonlinear pricing literature, as we discuss further below.

Our application concerns the provision of an expensive and controversial medication, epoetin alfa (or "EPO"), used to treat anemia (a lack of red blood cells) in patients with end-stage renal disease (ESRD, also known as kidney failure). Medicare is the dominant payer for the treatment of ESRD in the United States, and notably, the program spent more on EPO than any other single medication for several years in the early 2000s. We use data from Medicare insurance claims in 2008 and 2009, a period when the payment policy was stable and when there were no major informational shocks about EPO. As with most insurance claims, the treatments are observed: in this case the dosages of EPO given to each patient. Furthermore, quite uniquely, a highly specific quantitative measure of the patient's condition is available in the claims data. Providers were required to report the patient's red blood cell level (i.e., the severity of their anemia) in order to be paid for the EPO, and these blood levels are recorded on the claims. Thus, effectively, we study a monopsonist (Medicare) paying physicians for an observed quantity of treatment (dosage of EPO) to patients whose condition (red blood cell level) is observed by the principal, the agent, and the econometrician.

We specify an econometric model directly from the theoretical model, and use the patient-level claims data to estimate the structural parameters of the model. We choose a specification that yields tractable linear reduced forms while having sufficient flexibility to fit the data. The reduced forms provide simple closed-form estimates of the structural parameters, including the productivity of treatment and the joint distribution of physician altruism and marginal costs. We then use our estimates to fully characterize the unconstrained optimal contract, which results in the second-best allocation, and the constrained optimal linear contract. Both are flexible "fee-for-service" contracts, which pay for the amount of treatment provided, although they may also depend on the patient's blood level and other observable characteristics (e.g., age and the presence of comorbidities).[3] Finally, we contrast these payment contracts with the actual fee-for-service contract used by Medicare, and we simulate counterfactual outcomes under the optimal contracts.

As a general point, we note that fee-for-service contracts, which are ubiquitous in health care, may not be restrictive in settings like ours where the most relevant form of information asymmetry is about hidden types (i.e., fixed characteristics). Although these contracts do not condition on additional variables such as health outcomes, allowing the payer to do so would not increase its objective when there are no significant hidden actions (e.g., unmonitorable physician effort). In our case, because the treatment is relatively straightforward and is fully observed, the optimal (nonlinear) contract we derive is completely unrestricted, both with respect to the functional form and the arguments that enter it. For simplicity, we typically refer to this unconstrained optimal contract as the optimal "nonlinear" contract.

[3] "Linear contracts" are sometimes referred to as "two-part tariffs" (constant per-unit payment plus a fixed amount), and are distinct from "linear pricing", which has only a constant per-unit payment, and no fixed amount.

Our novel approach yields several qualitative and quantitative findings. Qualitatively, we document the presence of multidimensional unobserved heterogeneity among physicians, which implies that the optimal unrestricted contract will be nonlinear, to make use of variable marginal incentives to mitigate the effects of asymmetric information. We find that physicians are finitely altruistic, which means that they respond to monetary incentives, but also that they are substantially altruistic, which means that the optimal contract will leverage physician altruism to improve patient health at a lower cost. We also find that the observed contract used by Medicare at the time, a standard fee-for-service contract with a constant reimbursement rate, is not rationalizable even when considering only linear contracts, because the reimbursement rate was too high. This highlights a key strength of our data, which is that we can test (and end up rejecting) even constrained optimality of the observed contract.[4] To summarize, payment incentives matter, heterogeneity in both altruism and treatment cost is important in our setting, and our results indicate that moving to even the constrained optimal linear contract would substantially improve health and save money.

Quantitatively, the optimal nonlinear contract improves outcomes by reducing seemingly unjustified variation in treatment intensities while also decreasing total expenditures. Relative to the observed payment contract, the optimal nonlinear contract would increase the government's objective by an amount equal to $1,500 per patient per year. It reduces the standard deviation of dosages by 27 percent (conditional on the red blood cell level and other relevant patient characteristics), and also reduces the mean payment by 27 percent (the matching values are coincidental). We also find that, with an optimal linear contract, the per-unit payment rate would be 15% lower than the average reimbursement rate used by Medicare in 2008 and 2009. Finally, to quantify the losses due to asymmetric information, we measure the additional gains that would result if physician types were no longer private information, which would result in the first-best, full-information, allocation. These are on the order of over $2,300 per patient per month.[5]

The multidimensional nature of the unobserved heterogeneity in our model distinguishes it from the models typically used in empirical applications of screening, or hidden type, models. This may be because addressing multidimensional unobserved heterogeneity is a difficult problem in contract theory, which has rendered general analytical solutions elusive (see Mirrlees, 1986). We use the "demand profile" approach, introduced by Goldman et al. (1984) and Wilson (1993), to derive the optimal nonlinear contract. Intuitively, this approach projects the (multidimensional) distribution of unobserved types onto the (one-dimensional) quantity of treatment supplied, and then splits each level of quantity supplied into separate "markets", which may be treated as separate problems by the government (similar to Ramsey taxation, Ramsey, 1927). Although the demand profile approach is not always applicable in the presence of multidimensional heterogeneity (Deneckere and Severinov,

[4]For example, in Wolak (1994), which develops and estimates a model in which a principal seeks to regulate public utilities of (potentially) hidden types, data limitations mean the distribution of types cannot be estimated without assuming optimality of the observed contract. Paarsch and Shearer (2000) and Gayle and Miller (2009) also assume optimality of observed regime, but focus on hidden actions, as opposed to hidden types.

[5]There is other work, studying contexts different from ours, finding massive losses stemming from the presence of asymmetric information (e.g., Gayle and Miller, 2009; Abito, 2019).

2015), the conditions for this method to yield a correct solution seem to naturally be satisfied in our environment—even though we find there is multidimensional heterogeneity—because we study a supply problem not a demand problem.[6] This leads us to believe that the demand profile approach may also be more broadly applicable to certain types of screening problems.

To the best of our knowledge, there is no work that structurally estimates a model of physician treatment choices in an asymmetric information framework and uses this to characterize optimal contracts.[7] A handful of papers estimate asymmetric information models in other settings. Einav et al. (2010) discuss the small literature doing this for insurance contracts. Paarsch and Shearer (2000) characterize the optimal linear contract in a hidden action environment. Gayle and Miller (2009) also study hidden action models, quantifying the welfare loss from moral hazard. In contrast, we study a screening, or hidden type, model and flexibly characterize the optimal wage schedule. Perhaps most closely related is the fairly rich literature on optimal regulation, which considers screening models in institutional contexts that differ from ours in important ways (e.g., Wolak, 1994; Gagnepain and Ivaldi, 2002; Abito, 2019). As in the work by Gagnepain and Ivaldi and by Abito, our setting and data allow us to estimate structural parameters without imposing optimality of the observed contract. In contrast to the optimal regulation literature, we allow for multidimensional heterogeneity, which requires a different approach to characterize the optimal contract.

Our model and empirical analysis also relate to many papers in the health economics literature (see McGuire, 2000; Chalkley and Malcomson, 2000, for overviews). The basic model of physician utility is very similar to that in Gaynor et al. (2004), but allows for cost heterogeneity as well as heterogeneous altruism. De Fraja (2000) and Jack (2005) study heterogeneity across physicians, although there are various distinctions between their models and ours.[8] Like Clemens and Gottlieb (2014), we examine the impact of Medicare payment incentives, although they look at payment incentives broadly, as opposed to our focus on a specific medical context.[9] In contrast to the empirical health economics literature, we not only estimate a model of physician treatment choices and recover physician altruism, but we also make normative contributions by empirically characterizing the optimal unrestricted payment contract. Our approach allows us to characterize the full-information allocation and our novel application of the demand profile allows us to characterize the second-best allocation. Thus, we can provide the first quantitative assessment of the

[6] Specifically, the intersection determining the physician's optimal quantity is between the (typically) downward-sloping marginal transfer schedule (i.e., the marginal payment for treatment) and the upward-sloping dosage supply curve. In contrast, in the canonical screening problem, which considers an agent's demand for a good sold by a principal, the marginal price schedule and agent demand curve will both typically be downward-sloping in quantity. This would potentially lead to multiple intersections of these curves, rendering the demand profile approach inapplicable.

[7] The most similar paper on a health care application may be Einav et al. (2018), which estimates a structural model to study how dynamic incentives in payments to long-term care hospitals affect the timing of discharges. The paper conducts insightful simulations of outcomes under alternative contracts, but the environment does not feature asymmetric information and the alternative contracts are not specifically intended to be optimal.

[8] For example, Jack (2005) uses a model with unobserved effort, while in our setting the most relevant aspect of the treatment is observed (i.e., the dosage of the drug). Choné and Ma (2011) also study how physician altruism may affect the design of optimal payment contracts.

[9] Grieco and McDevitt (2017) similarly uses the specific context of dialysis care to examine an issue of broad importance in health care, the tradeoff between quantity and quality, and Eliason et al. (2019) examine the effects of corporate ownership on treatment decisions and patient outcomes.

importance of asymmetric information in a health application.

In what follows, we first provide institutional background (Section 2). In Section 3, we introduce the model and then derive the optimal nonlinear contract. Section 4 presents the data we use for our empirical analysis, and Section 5 describes the empirical implementation, including specification, identification, and estimation. Our main results comparing the optimal contracts with the observed contract are presented in Section 6. A summary and conclusions are in Section 7.

2 Institutional Background

End-stage renal disease (ESRD), or kidney failure, is a chronic and life-threatening condition that affects over half a million individuals in the United States at a given point in time. Since 1973, the Medicare program has provided universal coverage for the treatment of ESRD, regardless of age. In 2009, at the end of our study period, Medicare spent $28 billion on health care for individuals with ESRD, and of that amount, $1.74 billion was paid specifically for EPO.[10] The drug is used to treat anemia, a lack of red blood cells, which often accompanies chronic kidney disease.[11] EPO stimulates red blood cell production, and it is administered at regular intervals to try to maintain a certain target level of red blood cells.[12] The level is commonly measured in terms of the *hematocrit*, which is the volume percentage of red blood cells in the blood.

For patients on dialysis, EPO is typically administered intravenously during dialysis (specifically, hemodialysis), which occurs multiple times per week at specialized facilities called dialysis centers. The staff of these facilities consists of one medical director (a physician), with additional physicians at larger facilities, and multiple nurses and medical technicians who provide most of the direct treatment.[13] Payments are primarily made to the facilities, not the individual physician(s) or nurses, which is partly why we treat each dialysis center as a single "provider" in the empirical analysis. The main cost of providing EPO is acquiring the drug from the manufacturer (via a distributor), because its production involves an expensive biological process, and the manufacturer had a monopoly over this class of medications at the time. This motivates the assumption of constant marginal costs in our model, as the pricing was largely per-unit. Administering the drug to patients also involves non-trivial costs of staff time to prepare the dosages and monitor the injections (see Section 5.2), which is an additional source of heterogeneity across facilities.

Medicare's payment policy for EPO was debated throughout the 1990s and 2000s, largely because of concerns that the reimbursement rates were too generous and encouraged overprovision.[14]

[10] USRDS *2016 Annual Data Report*, volume 2, chapter 11; available at https://www.usrds.org/2016/view/Default.aspx. Amounts are for Medicare fee-for-service payments, and the amount for EPO includes a related drug, darbepoetin alfa, made by the same manufacturer. The total social expenditures on ESRD and these drugs were even higher because many beneficiaries also make a 20% copayment.

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

[12] A relevant medical point is that the half-life of EPO is under 12 hours, although there are longer-term effects on red blood cell levels and other health outcomes (Elliott et al., 2008).

[13] See *NEJM Catalyst*, https://catalyst.nejm.org/the-big-business-of-dialysis-care/, for an overview of how dialysis centers are run.

[14] There were concerns both that dosages were suboptimally high (i.e., marginal benefit less than marginal cost),

While dialysis itself was reimbursed with a prospective payment system known as the "composite rate," which paid a fixed amount of roughly $135 per session, EPO was a separately billable drug with its own per-unit reimbursement rate. Prior to 2005, that rate was held fixed at $10.00 per 1000 units. In 2006, Medicare adopted a new policy where the reimbursement rate was based on average sales prices calculated from data reported by the drug manufacturer. This policy, which was in effect through 2010, set a limit on the reimbursement rate each quarter, equal to 106 percent of the national average sales price from roughly six months earlier (GAO, 2006).[15] This provides the variation we need to estimate the model parameters governing how physicians respond to the marginal payment rate for EPO.

Because of the concerns about overprovision, Medicare also required dialysis centers to report a patient's hematocrit level on their insurance claims. The facilities typically filed monthly claims for each patient, which included separate lines for each dialysis session and each injection of EPO over the month. To be reimbursed for the EPO, the claims were required to report a hematocrit level taken just prior to the monthly billing cycle. Having a lab result like this in claims data is highly unusual, and it provides us with a highly specific quantitative measure of the patient's condition, in this case the severity of their anemia. Thus a key determinant of the medically appropriate treatment amount is observable, which facilitates a relatively simple approach for estimation.

Alongside the concerns about overprovision, there was substantial uncertainty about the benefits and risks of EPO (e.g., Foley, 2006). Many clinicians and medical researchers felt it was important to counteract severe anemia, to improve quality of life and address other specific risks associated with the condition. The National Kidney Foundation considered whether to recommend higher targets for the hematocrit level (The National Kidney Foundation-Kidney Disease Outcomes Quality Initiative, 2006). However, the risks associated with high dosages of EPO became clear by the mid 2000s. A major clinical trial found that patients who were given more EPO to achieve a higher target level of hematocrit suffered a higher risk of serious cardiovascular events and death (Singh et al., 2006). This study was published in November 2006, and strong warnings ("black box warnings") were added to the drug's labels in 2007.[16] As a result of this and other studies, the recommended range for hematocrit in ESRD patients remained at lower levels. For example, the National Kidney Foundation recommended the use of hemoglobin targets from 11 to 12 g/dl, corresponding to hematocrit levels of 33–36% (The National Kidney Foundation-Kidney Disease Outcomes Quality Initiative, 2007), and the FDA maintained its suggested range of hematocrit targets at 30–36%. Broadly, it seems that clinicians felt there were health benefits from providing EPO to patients with low levels of hematocrit, as well as serious risks from administering high dosages of EPO. In line with this, we specify the health function in our model as having a maximum where hematocrit equals a medical target, and decreasing on both sides from that target.

The dialysis industry was also undergoing rapid consolidation during this time period. By 2009,

and, as we discuss below, that dosages may have been high enough to harm patient health.

[15] In 2011, Medicare adopted a comprehensive "bundled" PPS for dialysis that included EPO, so the payment policy for the drug effectively switched from fee-for-service to prospective payment.

[16] During our study period of 2008 and 2009, there were no similar informational shocks of these magnitudes.

two large chains treated a combined 60 percent of dialysis patients in the US.[17] However, while the facilities within these chains had common ownership, federal regulations explicitly stated that each medical director was "responsible for the delivery of patient care and outcomes in the facility" (42 CFR §494.150, 2008). These regulations, along with the professional norms in medicine, suggest that the medical directors in chain organizations maintained clinical independence. Additionally, while the chains had common purchasing agreements for EPO, annual facility-level cost reports submitted to Medicare show variation in per-unit prices and other costs within chains (see Section 4).[18] In the empirical analysis, we treat each dialysis center as an independent entity, with its own marginal cost and degree of altruism. This allows for heterogeneity both within and between chains, and fits naturally with our theoretical framework.

3 Model

Our model uses a static screening framework, describing a one-time interaction between a principal and an agent. The government (the principal) pays a physician (the agent) to treat a patient.[19] The government seeks to maximize the benefit from patient health minus the cost of a payment to the physician. Thus, the government can be thought of as acting on behalf of patients, who receive benefits from treatment but have to fund public health insurance. The physician's utility also depends on patient health, weighted by the physician's degree of altruism, along with the cost of providing the treatment and the compensation received.

The patient arrives at the physician with a baseline health status, b_0. The physician then chooses a treatment amount, a. As is common in the literature on physician behavior (e.g., Ellis and McGuire, 1986), we assume the patient accepts the treatment exactly as prescribed by the physician.[20] In our application, b_0 is the hematocrit level from the prior month and a is the total units of EPO administered over the current month; both a and b_0 are observed by the government and the econometrician because they are reported in the monthly claims. Given the patient's baseline health status, the treatment produces health according to the function $h(a; b_0)$, which is twice differentiable in a. This function summarizes the benefits and risks of EPO for a dialysis patient with anemia. Providing amount a to a patient with baseline hematocrit level b_0 yields a resulting hematocrit level, $b_1(a; b_0)$, which is increasing in a and b_0. Health is increasing in a when the resulting level is below a medical target level, τ, but is decreasing in a when the resulting level is above this target: i.e., $\partial h/\partial a > 0$ if $b_1(a; b_0) < \tau$ and $\partial h/\partial a < 0$ if $b_1(a; b_0) > \tau$. This captures the impacts of EPO dosages on patient health—too little fails to relieve health problems due to chronic anemia, while too much increases the risk of bad health outcomes. We refer to a treatment

[17] USRDS *2011 Annual Data Report*, volume 2, chapter 10; https://www.usrds.org/atlas11.aspx.

[18] One source of variation in per-unit prices was the year-end rebates given to each facility from the manufacturer, based on the volume of EPO they purchased.

[19] Assuming the government's and physician's objectives are additively separable across patients, this model extends naturally to the empirical application with multiple physicians treating multiple patients, as we note in Section 5.

[20] As noted in Section 2, the medication is administered intravenously, while the patient is undergoing dialysis, so there is no issue with patient compliance, as opposed to patient adherence to oral medications or diet and exercise.

amount resulting in $b_1(a; b_0) > \tau$ (i.e., with negative marginal product) as *medically excessive*.[21] In the empirical implementation, we allow τ to depend on observed patient characteristics that are thought to affect the health benefits and risks of receiving EPO. Last, the function h is assumed to be strictly concave ($\partial^2 h/\partial a^2 < 0$), because patients with more severe anemia (i.e., lower hematocrit) benefit more from EPO, while the serious risks increase with higher dosages.

The degree of physician altruism, α, gives the physician's marginal rate of substitution between patient health and the physician's own income. The physician also has a constant marginal cost of treatment, z. These two attributes are unobserved by the government, and we refer to (α, z) as the physician's *type*. Heterogeneity in altruism captures differences between physicians' preferences. The treatment costs reflect the costs of acquiring and administering EPO, both of which might most naturally be expected to be heterogeneous. The types are jointly distributed according to a distribution, $F(\alpha, z)$, with the associated density $f(\alpha, z)$ that is strictly positive and differentiable over a compact set $[\underline{\alpha}, \overline{\alpha}] \times [\underline{z}, \overline{z}] \subset \mathbb{R}^2_+$, where $\underline{\alpha}$ and $\underline{z}$ are strictly positive and $\overline{\alpha}$ and $\overline{z}$ are finite.

The government sets a *payment policy*, which specifies the payment that is made to the physician based on the treatment amount and baseline hematocrit reported on the claim. The policy consists of a set of potentially nonlinear payment contracts for the treatment amount, $P(a; b_0)$, one for each possible value of b_0. In the empirical implementation, we allow the contracts to also depend on observed patient characteristics that affect τ. The payment policy, $\{P(a; b_0)\}$, is established before the physician sees the patient, and so the timing in the model can be summarized as follows:

1. Government sets the payment policy ($\{P(a; b_0)\}$)
2. Physician's type is realized (α, z)
3. Patient's baseline hematocrit level is realized (b_0)
4. Physician decides whether to participate
5. Physician chooses a treatment amount (a)
6. Outcomes occur: patient health ($h(a; b_0)$), payment to physician ($P(a; b_0)$), cost of treatment (za).

The physician's utility is a function of the patient's resulting health, weighted by the physician's degree of altruism, minus the cost of treatment, za, plus the payment from the government, $P(a; b_0)$:

$$u(a; \alpha, z, b_0, P) \equiv \alpha h(a; b_0) - za + P(a; b_0). \tag{1}$$

That is, the physician has quasilinear preferences, a standard assumption in screening models (Rochet and Stole, 2003). The physician's reservation utility is $\underline{u}$.[22]

[21] Note that "medically excessive" is a statement about the production technology h and is distinct from a normative economic concept. We will use "overprovision" to refer to economically excessive amounts. However, as will be made clear shortly, these concepts are related because a medically excessive amount will imply a suboptimally high amount.

[22] Note that $P(a; b_0) - za$, which corresponds to profits (insofar as z represents a monetary marginal cost), may be negative, which has precedent in models of motivated agents (e.g., Besley and Ghatak, 2005; Jack, 2005). See Choné

The government's objective is also a function of the patient's resulting health, weighted by a parameter, α_g, minus the payment to the physician.[23] The government's weight on patient health generically differs from the physician's weight because of the heterogeneity in α, and furthermore because the government represents the patient, α_g may be larger than the median of α, for example. The government's valuation of the outcome, where the patient has baseline hematocrit b_0 and receives treatment amount a, is as follows:

$$u_g(a; b_0, P) \equiv \alpha_g h(a; b_0) - P(a; b_0). \tag{2}$$

Because the physician's type is not observed, the government considers the expectation of this valuation over the distribution of amounts that will be chosen by different types, given the patient's baseline hematocrit b_0.[24]

We use subgame perfect Nash equilibrium to define behavior. The physician chooses a treatment amount to maximize utility function (1) given their type, the patient's baseline health, and the payment policy (this is the incentive compatibility constraint). The physician also decides whether to participate, and does not participate if the maximum possible utility would be below the reservation utility (this is the voluntary participation constraint).[25] The government sets the payment contract for each b_0, knowing how each physician type would respond. Thus, given b_0, the government's problem is to maximize the expected value of (2), subject to the physician's incentive compatibility (IC) and voluntary participation (VP) constraints, which must hold for each type:

$$\max_{P} \int_{\alpha,z} \left[u_g(a^*(\alpha, z; b_0, P); b_0, P) - P\left(a^*(\alpha, z; b_0, P); b_0\right)\right] f(\alpha, z) d\alpha dz$$

s.t.

$$a^*(\alpha, z; b_0, P) = \arg\max_{a \geq 0} \; u(a; \alpha, z, b_0, P), \quad \forall \alpha, z \qquad \text{IC}$$

$$u(a^*(\alpha, z; b_0, P); \alpha, z, b_0, P) \geq \underline{u}, \quad \forall \alpha, z \qquad \text{VP.}$$

Now we turn to the solution of the model. First, we characterize the first-best allocation, which would occur under full information. We then solve the model under asymmetric information via backward induction, starting with the physician's behavior, and then presenting our approach to derive the optimal contract, which results in the second-best allocation. This analysis is presented for a single value of the baseline hematocrit, and so b_0 is suppressed from the health function h and

and Ma (2011) for an example of a paper studying contracting in health care that constrains profits to be nonnegative. In our application, the dialysis centers provide many services so it may be reasonable to allow for negative profits from the provision of EPO.

[23] As is also standard in screening models, the principal's objective does not include the agent's objective, meaning it does not represent social welfare. This is different from the optimal regulation literature, where distortions are introduced via a distortionary cost of raising funds for the regulation program (see, e.g., Baron and Myerson, 1982).

[24] When there are multiple patients, additive separability across patients implies that the government's problem can be solved separately for each b_0.

[25] We assume the treatment amount is zero if the physician does not participate; this only affects off-equilibrium behavior.

the payment contract P. Also, we focus on interior solutions here to clarify the exposition. When solving the model for the empirical analysis, we allow for corner solutions where some physician types provide zero amounts (see Appendix D). We follow the literature in referring to this as *exclusion* (see, e.g., Armstrong, 1996), which is distinct from non-participation.

3.1 Full-Information First Best

The full-information allocation provides a benchmark against which we can measure losses due to asymmetric information. With full information, the government can effectively choose the treatment amount for each physician type, denoted $a^{*\text{FI}}(\alpha, z)$. The interior optimality condition is

$$\underbrace{\alpha_g h'(a^{*\text{FI}}(\alpha, z))}_{\text{Principal's MB}} = \underbrace{z - \alpha h'(a^{*\text{FI}}(\alpha, z))}_{\text{Agent's net MC}}. \tag{3}$$

The efficient allocation equates the principal's marginal benefit from consuming each infinitesimal unit with the agent's marginal cost of producing it, as is standard, but in this case the agent's effective, or net, marginal cost (from the principal's perspective) includes the effect of altruism. Unlike typical asymmetric information models with non-altruistic agents, here the agent derives utility from the same outcome as the principal does, and so the agent's marginal benefit from that outcome appears in the condition because it reduces the total marginal cost experienced by the agent. That is, rather than just the marginal cost of production, z, the *net marginal cost* to the agent's utility is $z - \alpha h'(a)$. Because the physician's altruism weight is positive, and h is strictly concave, this implies that treatment amounts in the efficient allocation are higher with altruism than without, which should be unsurprising. Also, we note that the efficient allocation would never have medically excessive amounts, where $h' < 0$.

3.2 Physician Behavior

Next we characterize the physician's behavior under an arbitrary differentiable payment contract P. The interior first-order condition is

$$\underbrace{z - \alpha h'(a^*)}_{nc(a^*;\alpha,z)} = \underbrace{\frac{\partial P(a^*)}{\partial a}}_{p(a^*)}. \tag{4}$$

As explained above, $z - \alpha h'(a)$ is the net marginal cost to a physician of type (α, z) for providing amount a. It will be useful to denote the net marginal cost function as $nc(a; \alpha, z) \equiv z - \alpha h'(a)$, and the marginal payment function as $p(a) \equiv \frac{\partial P(a)}{\partial a}$. The physician chooses an amount a^* that equates the net marginal cost with the marginal payment; thus $nc(a; \alpha, z)$ defines the supply curve for type (α, z). The solution is unique so long as the net marginal cost curve intersects the marginal payment curve once, from below (as discussed later in Section 3.3). Then, if $h'(a^*) > 0$, as we show will be the case under the optimal nonlinear contract, a^* is increasing in α and decreasing in z.

To see how the payment contract affects behavior by different types of physicians, it helps to start with a linear contract. Let $P^L(a) \equiv p_0 + p_1 a$ denote an arbitrary linear contract, where p_0 is a lump-sum payment, and p_1 is a constant marginal payment (i.e., the per-unit payment rate). Then rearranging (4) to $\alpha h'(a^*) = z - p_1$, it is apparent that all physician types with marginal costs below p_1 would provide amounts such that $h' < 0$, i.e., that are medically excessive, while all those with marginal costs above p_1 would not. In either case, for a given marginal cost, having a higher degree of altruism makes the physician provide a treatment amount closer to the health maximizing amount, due to the strict concavity of h.

Figure 1 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.[26] 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., $p(a_1) > p(a_2)$).[27] 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).

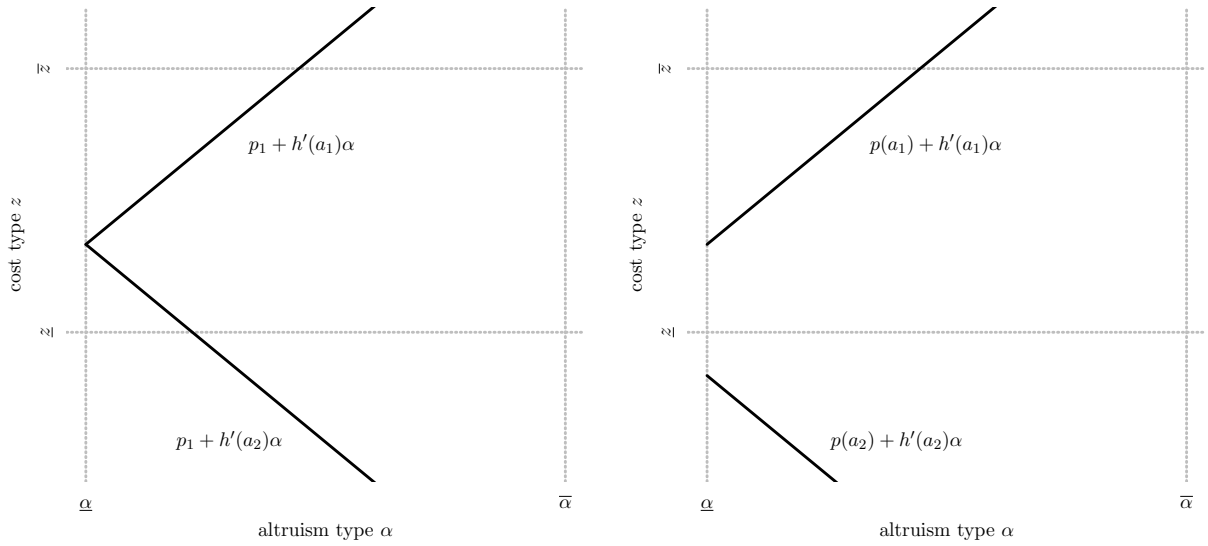
3.3 Optimal Contract

We now present our approach to solve the government's problem and thereby characterize the optimal nonlinear contract. Because agent heterogeneity in our model is multidimensional, we cannot use more common methods based on the Revelation Principle. Those methods rely on a strict ordering of agent types, so that the relevant (i.e., binding) incentive compatibility constraints can be reduced to those between adjacent types in the ordering (e.g., Myerson, 1981; Maskin and Riley, 1984). As illustrated with the isoquants above, no similar reduction of incentive compatibility constraints can be obtained under multidimensional heterogeneity, because multiple agent types may choose the same amount.

Instead, we use an analog of the "demand profile" approach (Goldman et al., 1984; Wilson, 1993), which reformulates the principal's problem in terms of finding the marginal payments for each possible quantity. The power of this approach is that it projects a multidimensional distribution

[26] 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.

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



(a) Linear contract

(b) Nonlinear contract

Figure 1: 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 $h'(a_2) < 0$.

of agent types onto a one-dimensional distribution of quantities, and the solution for each quantity can be found separately when certain conditions are satisfied.

The government's optimization problem is accordingly to set the marginal payment for each treatment amount to maximize its marginal valuation of that amount, multiplied by the probability the amount will be provided. Specifically, the government chooses the marginal payment, $p(a)$, for each potential treatment amount, $a \in A$, to maximize

$$\int_A S(p, a)[\alpha_g h'(a) - p(a)]da. \tag{5}$$

In essence, this is the infinite sum of the government's marginal valuation of each amount (i.e., the derivative of (2) with respect to a, which is inside the square brackets), weighted by the probability of receiving that marginal valuation. The function S is the analog of the demand profile in Wilson (1993), but in our case it gives a distribution of quantities supplied rather than quantities demanded. Specifically, $S(p, a)$ is the probability that the physician is a type that will provide a treatment amount of at least a, given the payment contract. In that case, the government will receive its marginal valuation at amount a, which is $\alpha_g h'(a) - p(a)$.

The set of potential treatment amounts, A, is an interval spanning zero, which corresponds to the amounts of excluded types, to $\bar{a}^{*\text{FI}} \equiv a^{*\text{FI}}(\overline{\alpha}, \underline{z})$, the amount that would be provided by the

"best" type (lowest cost, highest altruism) under full information.[28] As is standard, the voluntary participation constraint must be satisfied for any physician type. This rules out a "forcing contract" that would only pay for the maximum full-information treatment amount, for example.[29]

Assuming that the net marginal cost curve for each agent type intersects the marginal payment curve at most once, from below, which is an important regularity condition (discussed in detail below), S has a simple form:

$$S(p,a) \equiv \Pr\{p(a) \geq \underbrace{z - \alpha h'(a)}_{nc(a;\alpha,z)}\}, \tag{6}$$

where the probability is over the distribution of agent types. The single intersection of net marginal costs and marginal payments guarantees that, if the marginal payment at amount a is greater than the net marginal cost for some physician type (α, z), as expressed by the inequality in (6), then the marginal payments are greater than the net marginal costs for that type at all lower amounts as well. Hence, any type that satisfies the inequality in (6) would provide at least a, and so $S(p,a)$ as defined in (6) gives the desired probability that the marginal valuation at amount a is received.

Figure 2 provides some intuition by plotting the net marginal cost curves for two types, (α_1, z_1) and (α_2, z_2), against a marginal payment curve, $p(a)$. The net marginal cost curves are upward sloping. Their slopes are equal to $-\alpha h''(a)$, which is positive because h is strictly concave. Hence, if the marginal payment curve is downward sloping, it will intersect the net marginal cost curves once, from above, as required. Any type with a net marginal cost curve below that of type 1 at a_1^* (i.e., any (α, z) such that $z - \alpha h'(a_1^*) < z_1 - \alpha_1 h'(a_1^*)$), for example, type 2, would provide more than a_1^*.

Figure 2 suggests that this approach may be more broadly useful for solving screening problems with multidimensional heterogeneity. The demand profile approach has mainly been applied to monopoly pricing problems, but there the single-intersection condition can be more difficult to satisfy because both the consumer demand curves and the marginal price curve are typically downward sloping (see, e.g., Deneckere and Severinov, 2015, for discussion). By contrast, because marginal cost curves are typically upward sloping, the condition can be easier to satisfy in monopsony applications (i.e., purchasing goods or services).[30]

Next, using the distribution of treatment amounts generated by (6), the government's problem (5) is solved separately for each treatment amount. In addition to the single-intersection condition, this relies on the quasilinearity of the agent's preferences (i.e., no income effects), a standard

[28] We show below that the standard "no distortion at the top" result is obtained (i.e., the highest amount is undistorted) and that all other types' treatment amounts are downwards-distorted in the second-best allocation. This means that $a^{*\text{FI}}(\overline{\alpha}, \underline{z})$ is the maximum treatment amount under the optimal nonlinear contract.

[29] Even without voluntary participation constraints, the government would not choose a forcing contract. Those types for which voluntary participation is violated would provide zero, so the government could improve its objective by inducing participation from different types that would provide different amounts.

[30] To verify that the condition is satisfied in our empirical analysis, we first solve for the optimal contract and then check that no physician types have supply curves with multiple intersections with the marginal payment curve, which could be upward-sloping for some treatment amounts (indeed, this is true in our empirical results).

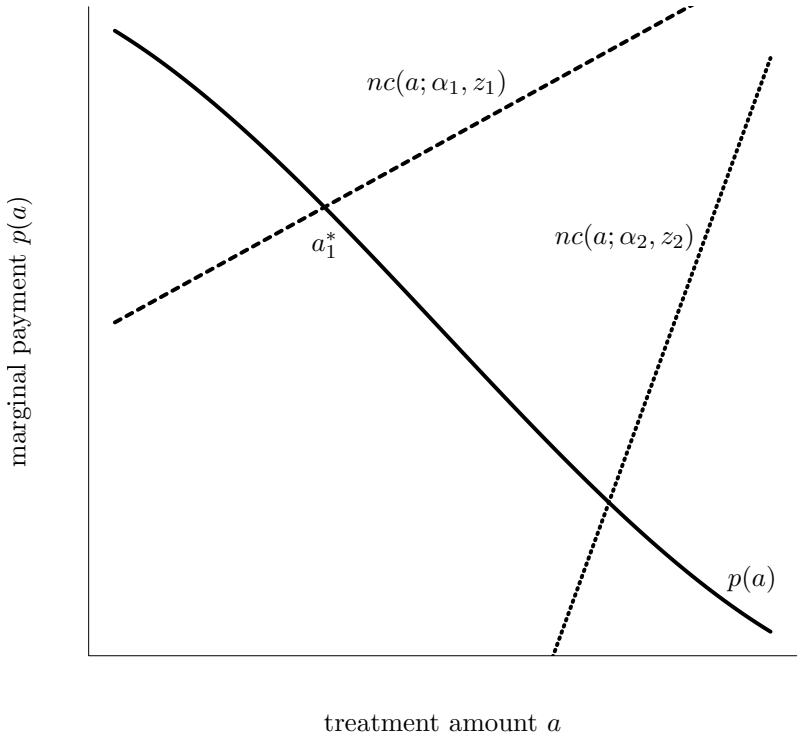


Figure 2: Example marginal payment contract and physician supply curves.

Notes: Figure plots an example marginal payment contract $p(a)$ (solid curve) and supply curves $nc(a;\alpha,z)$ for a lower altruism type (α_1, dashed line) and a higher altruism type (α_2, dotted line); both supply curves are for the same marginal cost type, i.e., $z_1 = z_2$.

assumption in screening models. Specifically, the physician's marginal utility at amount a does not depend on the marginal payment for any other amount, so the effect of $p(a)$ on the supply of amount a does not depend on the payments for other amounts.[31] The separate problems for each treatment amount are thus

$$\max_{p(a)} S(p(a),a)[\alpha_g h'(a) - p(a)], \tag{7}$$

for each $a \in A$. Splitting the principal's objective into independent problems for each quantity in this way is the central idea in the demand profile approach, which makes it tractable. It is similar to the classic idea of Ramsey (1927), which splits optimal taxation across a variety of goods into a separate problem for each good.

Finally, the optimal contract is characterized by the first-order condition of (7) for each amount,

[31] Without this separability, solving for the optimal nonlinear contract is significantly more cumbersome (Maskin et al., 1987; McAfee and McMillan, 1988). See Deneckere and Severinov (2015) for a discussion.

treating $p(a)$ as a parameter:[32]

$$\frac{\partial S(p^*(a),a)}{\partial p(a)}\left[\alpha_g h'(a) - p^*(a)\right] = S(p^*(a),a). \tag{8}$$

The contract is constructed by first solving this for $p^*(a)$, for each $a \in A$, and then integrating the marginal payments to yield P^* (see Appendix C for details). The level of P^* is set by making the participation constraint bind for the type with the lowest utility: i.e., the lowest utility in equilibrium equals the physician's reservation utility, $\underline{u}$.

Remarks

The government's first-order condition (8) is fairly intuitive. The left side of (8) represents the marginal net benefit to the government from increasing $p(a)$; i.e., the increase in the probability the physician provides amount a, $\frac{\partial S(p^*(a),a)}{\partial p(a)}$, multiplied by the government's marginal valuation at that amount, $[\alpha_g h'(a) - p^*(a)]$. The right side is the government's marginal cost of increasing $p(a)$, which is paid to all types providing amount a, given by S.

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

$$\frac{\alpha_g h'(a) - p^*(a)}{p^*(a)} = \frac{1}{\eta(a)}, \tag{9}$$

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 (9) 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:

$$\underbrace{\alpha_g h'(a)}_{\text{Principal's MB}} = \underbrace{z_i - \alpha_i h'(a)}_{\text{Agent's net MC}} + \underbrace{\frac{S(p^*(a),a)}{\frac{\partial S(p^*(a),a)}{\partial p}}}_{\text{distortion}}. \tag{10}$$

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 (10) to show that the allocation under the optimal nonlinear contract will be

[32]The optimal contract is assumed to be differentiable almost everywhere. This does not seem restrictive in our setting because we assume that the joint density function $f(\alpha, z)$ is differentiable, along with the other primitives.

downward-distorted from the first-best for all but the highest-amount type, $(\overline{\alpha}, \underline{z})$.[33] Equivalently, for any amount $a < \overline{a}^{*\text{FI}}$, fewer types choose a in the second-best because they 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 (10) 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.

4 Data

We now turn to the empirical analysis. Our primary data come from Medicare outpatient claims from renal dialysis centers (freestanding or hospital-based) in 2008 and 2009, for the treatment of patients with ESRD. The raw sample (20% of patients) contains a total of 1.4 million ESRD claims, which are typically filed monthly. Almost 90% of the claims (1.25 million) bill for at least one injection of EPO or a related medication. All claims with an injection include a baseline hematocrit level from the previous month (or a comparable hemoglobin level), but claims without an injection do not report a red blood cell level. As a consequence, we exclude claims without any injections of EPO.[34] Also, in order to avoid extreme outliers, which often reflect data entry errors, we remove observations where the amount of EPO is above its 99th percentile. Finally, we restrict to observations where the baseline hematocrit is within a broadly recommended range for using EPO, which is between 30 and 39 percent.[35] The final sample has 919,749 claims, for 74,262 unique patients, from 5,148 unique providers.

[33] 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.

[34] EPO appears on the vast majority of the claims with an injection of this class of medication (93%). The alternative drug was darbepoetin alfa. We restrict to EPO because dosages and reimbursements differ between the two drugs.

[35] The FDA-approved labeling for EPO stated a suggested target range for hematocrit of 30 to 36 percent (https://www.accessdata.fda.gov/scripts/cder/daf/index.cfm), and guidelines issued by the National Kidney Foundation recommended the use of hemoglobin targets from 11 to 12 g/dl, and not greater than 13 g/dl (The National Kidney Foundation-Kidney Disease Outcomes Quality Initiative, 2007), which is comparable to hematocrit targets from 33 to 36 percent, and not greater than 39 percent. Also, Medicare reduced the reimbursement rate by half for EPO provided to patients whose hematocrit exceeded 39 percent for three consecutive months (https://www.cms.gov/medicare-coverage-database/details/medicare-coverage-document-details.aspx?MCDId=11).

The unit of observation is the monthly claim, which reports the services given by provider i to patient j in period t. We use the dialysis centers, not individual physicians, as the providers because within each center the doctor(s), nurses, and technicians jointly provide treatment to their patients, and as noted in Section 2 the insurance payments went to the facilities. The treatment amount, a_{ijt}, is total amount of EPO administered over the claim period, and the baseline hematocrit, b_{0jt}, is the prior hematocrit level reported on the claim.[36] The payment rate, p_{1t}, is the national payment rate per 1,000 units of EPO for the quarter in which the claim was filed. These rates are listed in publicly available Medicare Part B Average Sales Price Drug Pricing Files.[37] Additionally, our empirical model includes basic patient demographics and a measure of comorbidities, because these may affect the target level of hematocrit. We use patient age, sex, and a standard measure of patient comorbidities, the Charlson Comorbidity Index, collected into a vector x_{jt}.[38]

Table 1 provides summary statistics of these variables. The average monthly dosage of EPO is 63 thousand units, with a relatively large standard deviation of 61.7 thousand units. The average baseline hematocrit is 34.8 percent, with a standard deviation of 2.2 percent. The Charlson Index, which is a count of patient comorbid conditions such as a prior heart attack (where some conditions have weights greater than one) has a mean of 1.4. Most patients have no comorbidities, as indicated by the median of zero, while those in the top quarter of the distribution have multiple comorbidities. Addendum 2 in the table lists the national payment rate for EPO for each quarter during our study period, which ranged from a low of \$8.96 in 2008Q1 to a high of \$9.62 in 2009Q3. The average payment rate across the observations in our sample is \$9.26.

Table 1 also shows the distribution of dialysis centers' acquisition costs for the drug from a separate source, the publicly available Renal Dialysis Facilities Cost Report Data from the Centers for Medicare & Medicaid Services.[39] The percentiles indicate potentially important differences across dialysis centers in their acquisition costs, even though the drug was produced by a single manufacturer.[40] (The mean and standard deviation are not presented, due to extreme outliers in the data.) As we discuss in Section 5.2, there are also nontrivial costs of administering EPO (another component of the marginal cost of treatment), which are likely to vary across dialysis

[36] For claims that report hemoglobin rather than hematocrit, we use the standard rule of thumb of multiplying by three to convert the levels (WHO, 1968).

[37] See https://www.cms.gov/Medicare/Medicare-Fee-for-Service-Part-B-Drugs/McrPartBDrugAvgSalesPrice/index.html. The national payment rates are technically limits on the allowable reimbursement rates, which may be modified for example to reflect overall healthcare costs in a local area ("geographic adjustment factors"). However, the actual reimbursement rates that can be computed from the claims are highly correlated with the national payment limits: in our sample the time-series correlation within providers is 0.98.

[38] Patient age and sex are taken from the Medicare Beneficiary Summary File. For the Charlson index, we apply the implementation from Quan et al. (2005) to Medicare inpatient claims (MEDPAR) for the patients in our sample. The Charlson index has been validated for dialysis patients (Beddhu et al., 2000).

[39] See https://www.cms.gov/Research-Statistics-Data-and-Systems/Downloadable-Public-Use-Files/Cost-Reports/Renal-Facility-265-1994-form. CMS requires dialysis centers to submit detailed annual cost reports, which include their total expenditures on EPO and the total number of units provided. From the total expenditures (less any rebates) and total units, we compute the average acquisition cost per 1,000 units of EPO for each center in the cost report data from 2008.

[40] These data also show meaningful differences in acquisition costs across dialysis centers within the same chain. For example, the interquartile ranges are \$0.22 for DaVita and \$0.41 for Fresenius, which are smaller but certainly not trivial relative to the interquartile range of \$0.92 ($= 8.15 - 7.23$) across all centers shown in Table 1.

Table 1: Summary Statistics

Variable	Mean	SD	Percentiles 10th	25th	50th	75th	90th
Monthly EPO dosage (1,000u)	63.0	61.7	8.8	20.0	42.9	84.0	143.0
Prior hematocrit level (%)	34.8	2.2	31.7	33.0	34.8	36.6	37.8
Charlson Comorbidity Index (0-16)	1.4	1.9	0	0	0	2	4
Payment rate ($/1000u)	9.26	0.24	8.96	9.07	9.20	9.58	9.62
Percentiles of EPO acquisition costs from annual cost reports for 2008							
Acquisition cost ($/1000u)			7.13	7.23	7.53	8.15	9.11
Medicare payment rate for EPO in each quarter							
Reimb. rate ($/1000u)	8.96 (2008Q1)	9.07 (Q2)	9.07 (Q3)	9.10 (Q4)	9.20 (2009Q1)	9.40 (Q2)	9.62 (Q3)	9.58 (Q4)

Notes: The EPO dosage and hematocrit level, and Charlson index come from Medicare outpatient claims data. The payment rate comes from quarterly Medicare Part B ASP Drug Pricing Files for 2008 and 2009. The distribution of EPO acquisition costs is computed from Renal Dialysis Facilities Cost Report Data for 2008. We do not present the mean or standard deviation because extreme outliers in the cost report data make those statistics unreliable, compared to the percentiles.

centers, but which are not well observed in the cost report data.

5 Empirical Implementation

We now describe how we adapt the model from Section 3 to the empirical application, and how we recover the parameters of the empirical specification from the data. The model extends to an environment with many physicians, each treating many patients, under the natural assumptions that the physicians' utility functions and the government's objective function are additively separable across patients.[41] Therefore our earlier results can be used to characterize optimal contracts in this setting, and the empirical specification is similarly presented for a treatment provided to one patient at a particular point in time. Below, we first develop the empirical specification, then discuss identification and explain the approach used for estimation, and finally present the estimates of the reduced-form and structural parameters.

[41] The static framework can be applied to multiple time periods if there are no dynamic effects of EPO (as noted in Section 2), and if the government does not consider patient histories when setting payments. This has always been the case when patient hematocrit levels are within the recommended range (i.e., not above 39%), and our analysis restricts to observations in this range.

5.1 Empirical Specification

For the empirical analysis, we assume a quadratic specification of the health function, h. This yields simple, closed-form expressions for the treatment amounts and facilitates a relatively straightforward approach to estimation.[42] The quadratic specification of h is as follows:

$$h(a; b_0) \equiv H - \frac{1}{2}[\delta a + b_0 - \tau]^2. \tag{11}$$

Here δ is a linear technology that converts the amount of EPO, a, into an increase in hematocrit (i.e., $b_1(a; b_0) = \delta a + b_0$). As with the general version of h, the maximum health is achieved when the resulting hematocrit equals the medical target level, τ, which now occurs when $a = [\tau - b_0]/\delta$. The health function also includes a positive constant $H \gg 0$, so that patient health enters positively into physician utility.[43]

With a quadratic specification of h, and with a constant marginal payment rate (p_1) as in the linear contracts that were in place during our study period, the physician's first-order condition (4) yields a linear function for the chosen treatment amounts:

$$a^*(\alpha, z; b_0, P^L) = \frac{\tau - b_0}{\delta} + \frac{p_1 - z}{\alpha\delta^2}. \tag{12}$$

We assume interior solutions apply when estimating the model because, as seen in Section 4, nearly all patients were given some amount of EPO. However, we allow for corner solutions (i.e., $a^* = 0$, which is the notion of exclusion) in the construction of the optimal contracts and in the simulations presented in Section 6.

Equation (12) implies a globally linear relationship between the patient's baseline hematocrit and the amount of EPO provided. To examine this, Figure 3 plots average dosages of EPO as a function of baseline hematocrit, separately for the first and last quarters in our data (when the national payment rates were respectively \$8.96 and \$9.58 per 1,000 units). Average dosages are monotonically decreasing in b_0, which is consistent with our model, but the relationship appears to be somewhat nonlinear, with a steeper slope at lower hematocrit levels. When the payment rate was higher (2009Q4), average dosages are larger for patients with low and medium hematocrit levels, which is also consistent with (12). However, the average dosages decrease more rapidly, and are even slightly lower for patients with high hematocrit levels, in contrast to the level shift that (12) would predict. While these aggregate plots do not provide ceteris paribus comparisons, they

[42] This specification is not crucial for the analysis. A more flexible specification could be used because, as we show in Appendix F, the health function is semiparametrically identified from the within-physician variation in baseline hematocrit (b_0) and marginal payments (p) that are observed in our data. Furthermore, the optimal nonlinear contract can be constructed for any function h that satisfies the assumptions given in Section 3. While the health function is in principle semiparametrically identified, we use the quadratic specification because it yields linear reduced forms that are highly tractable and make efficient use of the variation in the data.

[43] Specifically, we assume that H is sufficiently large such that $h(0; b_0) > 0$. This implies that the orderings of the levels of u with respect to type parameters are the same as those of derivatives of u with respect to type parameters. This kind of assumption is standard in screening models because it implies that only the participation constraint of the lowest-amount type will be binding, which simplifies characterization of the optimal nonlinear contract.

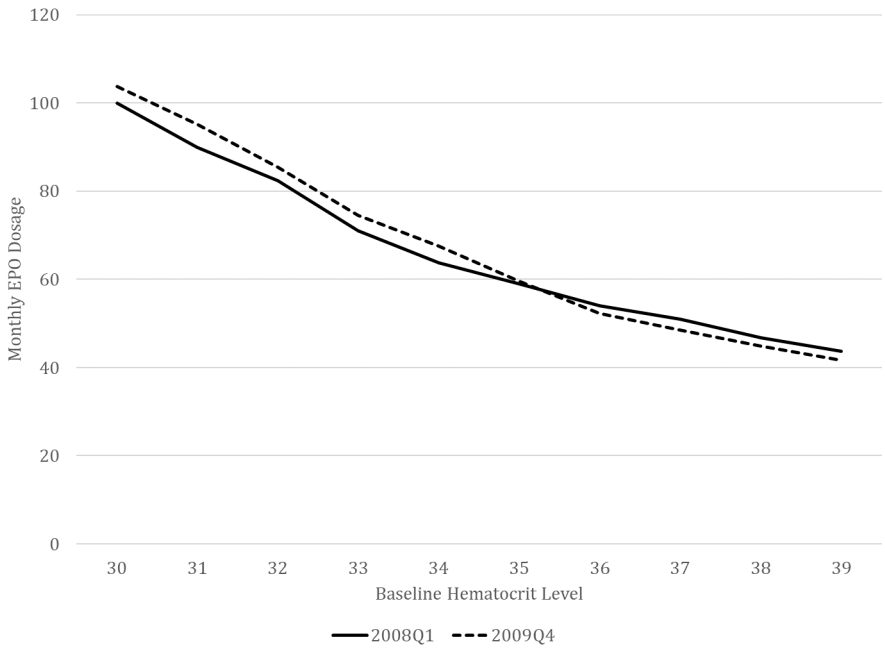


Figure 3: Mean monthly dosages of EPO in relation to baseline level of hematocrit.

suggest that certain nonlinearities absent from (12) may be empirically relevant.

To capture the potential nonlinearities suggested by Figure 3, our empirical specification adds flexibility in relation to the patient's baseline hematocrit. Specifically, we allow the model parameters to take different values when b_0 is in different intervals: i.e., when b_0 is in interval k, the parameters of the health function (11) are δ_k, τ_k, and H_k, and the distribution of (α, z) is F_k. As a consequence, each interval of baseline hematocrit can be treated separately in the estimation of the model. This approach maintains the linear, closed-form solution (12), while having sufficient flexibility to fit the possible nonlinearities in the treatment amounts suggested by Figure 3.

To provide some interpretation, the flexibility in δ means that the productivity of EPO may depend on the baseline level of hematocrit,[44] and similarly the flexibility in τ means that the target level of hematocrit for a patient may depend on their baseline level. The different distributions of (α, z) allow physicians to have potentially different altruism weights and marginal costs depending on the severity of a patient's anemia (i.e., the baseline hematocrit). On this last point, it may seem natural for physicians to have greater concern for sicker patients—our empirical results are in fact consistent with this (see Section 5.3).

Finally, as described previously, in addition to the flexibility in the parameters, the empirical model includes a vector of patient-level observable characteristics, x, which affect the target hematocrit. The parameter τ_k is accordingly extended to a vector, so that the target hematocrit for a patient with characteristics x and baseline hematocrit in interval k is $\tau_k' x$. To allow for unex-

[44] Because patients with lower baseline hematocrit are given higher dosages on average, this could approximate diminishing returns, for example.

plained variation from the econometrician's perspective, we add an independent, mean-zero shock, η. Additionally, as we make clear below, it is useful to decompose marginal cost as $z_{ik} = \mu_z + \zeta_{ik}$.

With the above extensions to (12), the observed treatment amount provided by physician i to patient j in period t is

$$a_{ijt} = \frac{\tau_k' x_{jt} - b_{0jt}}{\delta_k} + \frac{p_{1t} - [\mu_z + \zeta_{ik}]}{\alpha_{ik}\delta_k^2} + \eta_{ijtk},$$

given the patient's baseline hematocrit is in interval k. This is the empirical reduced form for the observed treatment amounts. It can be rearranged to yield reduced-form parameters and disturbances (structural parameters are in the body of the equation, reduced-form parameters are below the brackets):

$$a_{ijt} = \underbrace{\left[\frac{-1}{\delta_k}\right]}_{\beta_1^k} b_{0jt} + \underbrace{\left[\frac{1}{\alpha_{ik}\delta_k^2}\right]}_{\beta_{2i}^k} \underbrace{[p_{1t} - \mu_z]}_{\tilde{p}_t} + \underbrace{\frac{\tau_k'}{\delta_k}}_{\beta_3^{k\prime}} x_{jt} + \underbrace{\left[\frac{-\zeta_{ik}}{\alpha_{ik}\delta_k^2}\right]}_{\nu_i^k} + \underbrace{\eta_{ijtk}}_{\epsilon_{ijt}^k}. \tag{13}$$

Thus, in each hematocrit interval, our reduced form is a linear regression model with a random coefficient, β_{2i}^k, and a random effect, ν_i^k. Globally, the reduced form would be a piecewise linear regression model, but it can be estimated separately within each interval.

5.2 Identification and Estimation

In this section, we explain the approach we implement to identify and estimate the empirical model. The structural parameters to be recovered are the scalars δ_k and vectors τ_k, $k = 1 \dots K$, and the joint distributions $F_k(\alpha, z)$, $k = 1 \dots K$. One parameter, the mean of the marginal cost, is assumed to be the same across the intervals of baseline hematocrit. To identify that parameter, μ_z, we use external information on average per-unit costs of acquisition and administration of EPO, described later in this section.[45] The other parameters are identified from the reduced form (13). The values of δ_k and τ_k follow immediately from the coefficients β_1^k and β_3^k, given a value of μ_z. The joint distribution of α and z in each interval, F_k, is identified from the distribution of the random coefficient and random effect, β_2^k and ν^k.

Multiple approaches to recover the joint distributions F_k are possible. For efficiency and computational tractability we use a parametric assumption, but nonparametric methods could be employed as well. Our approach is summarized as follows (details are in Appendix E). We specify $\ln \alpha$ and z to have a joint normal distribution, so that α has a lognormal distribution with strictly positive support. Then in each hematocrit interval k, there are four unknown parameters of the joint distribution, $\mu_{\alpha,k}$, $\sigma_{\alpha,k}^2$, $\sigma_{\alpha z,k}$, and $\sigma_{z,k}^2$ (while $\mu_{z,k} = \mu_z$ is treated as known from our external information on costs).[46] Using Stein's lemma (Stein, 1981) and properties of the lognormal distribution, these parameters are identified by, and can be recovered analytically from, the first and

[45] Note that μ_z and the intercept of τ_k are not separately identified in the reduced form.

[46] Note that we follow the convention of using μ_α and σ_α (instead of $\mu_{\ln \alpha}$ and $\sigma_{\ln \alpha}$) to respectively denote the mean and standard deviation of $\ln \alpha$.

second moments of the random coefficient (β_2^k) and random effect (ν^k) in the reduced form (13). Those moments are estimated (semiparametrically) via an auxiliary regression of the residuals of (13), which is derived specifically for this purpose and takes advantage of the panel structure of the data (see Appendix E).

We note that, while this parametric approach is tractable and efficient, F_k is nonparametrically identified under the assumption that the shocks η_{ijtk} (equivalently, ϵ_{ijt}^k) are mean-independent of b_0, p_1, and x. To provide some intuition, an alternative approach to recover the joint distribution of α and z would be to estimate (13) separately for each provider (within each interval), using multiple observations per dialysis center. Consistent estimates of β_{2i}^k and ν_i^k for each provider would then yield consistent estimates of α_{ik} and ζ_{ik}, and so the empirical joint distribution of α and z could be recovered for each interval k using standard nonparametric methods (see Appendix F). However, this approach would be computationally intensive due to the large number of dialysis centers, and the estimates would be much noisier.

The identification of the structural parameters also depends on the consistency of the reduced-form estimates. We use OLS to estimate the reduced form (13), so this requires that the unobservables $(\beta_{2i}^k, \nu_i^k, \epsilon_{ijt}^k)$ are uncorrelated with the observables $(b_{0jt}, p_{1t}, x_{jt})$.[47] One possible concern is selection of patients to providers, which could make b_0 and x correlated with the provider-level unobservables, β_2^k and ν^k. We assess this concern by comparing OLS and fixed effects estimates of (13), which would indicate biases due to selection on time-invariant provider attributes (i.e., α and z). The coefficient estimates are quite similar (see Appendix Table A3, columns 4-6), which suggests that the provider-level unobservables (β_{2i}^k and ν_i^k) are not noticeably correlated with the patient characteristics (b_{0jt} and x_{jt}). We note that dialysis patients typically undergo dialysis three times a week for two to five hours at a time, and are quite debilitated after, hence travel is very costly and they do not travel far for care. Dialysis providers have a substantial amount of market power due to this (Eliason, 2017), so selection may not be a significant issue for this application.

Another possible concern is endogeneity of the national payment rate (p_{1t}). As described in Section 2, this was set each quarter based on the national average price of EPO roughly six months earlier. An individual dialysis center could not affect the national average price, but if unmodeled demand shocks were substantially correlated across centers and over time, there could be a bias because ϵ_{ijt}^k could be correlated with p_{1t}. We accordingly include a year dummy for 2009 and month dummies for each calendar month, which would address both secular and cyclical trends in demand. Assuming any effects of systematic demand shocks from dialysis centers are thereby absorbed, it is worth noting that there are other potential sources of variation in lagged prices of EPO that could generate exogenous variation in the payment rate: supply shocks from the drug manufacturer, and demand shocks from other purchasers of EPO.[48]

For the mean per-unit cost, μ_z, as noted earlier, we use external information on costs to determine the value of the parameter. Given the high price of EPO, most of the cost is for acquisition

[47] The auxiliary regression additionally requires that the second moments of the unobservables are independent of (b_0, p_1, x) within each hematocrit interval.

[48] EPO is also used extensively for chemotherapy patients and for surgery patients.

(i.e., purchasing the drug from a distributor). The Renal Dialysis Facility Cost Report Data noted in Section 4 allows us to compute per-unit acquisition costs by facility and year, and we use the median reported in Table 1, equal to \$7.53 per 1,000 units, as the acquisition component of μ_z.[49] The cost of administering EPO is also non-trivial. Several time-and-motion studies have been published to assess the cost of administering EPO, and we use estimates from Schiller et al. (2008), which is the most thorough and relevant for our time period. The results from that study imply an average cost of staff time and non-drug supplies for administering EPO equal to \$1.05 per 1,000 units.[50] Adding this to the acquisition cost, we set the value of μ_z equal to \$8.58 per 1,000 units.[51]

To estimate the reduced form (13), we use OLS separately in each hematocrit interval k. This yields estimates of β_1^k, β_3^k, and the mean of β_2^k, denoted $\bar{\beta}_2^k$. The auxiliary regression of the residuals is also estimated by OLS separately in each interval, which yields estimates of the variances and covariance of β_2^k and ν^k (see Appendix E). A cluster bootstrap is used to the compute standard errors of the structural parameters, where the clusters are the dialysis centers.

The hematocrit intervals used for estimation are three percentage points wide (e.g., $30 < b_{0jt} \leq 33$), which provides a good balance between the flexibility of the specification and the precision of the estimates from each interval. The range of hematocrit values in our estimation sample goes from 30 to 39 percent, based on treatment guidelines and payment policies in place at the time (see footnote 35). The 3-point intervals are thus convenient because they divide this range evenly, and the estimation results below indicate that this width allows sufficient power within each interval while maintaining flexibility globally.

5.3 Estimation Results

The OLS estimates of the reduced-form coefficients on the baseline hematocrit (β_1^k) and the payment rate ($\bar{\beta}_2^k$) are shown in Table 2, and the full set of estimates for all variables are provided in Appendix Table A3.[52] To interpret the coefficients, for example in the middle interval, a patient with one unit higher baseline hematocrit (say 35 vs. 34) receives 6,320 less units of EPO per month on average. Also in that interval, a one dollar increase in the payment rate (per 1,000 units) would induce providers to increase dosages by 6,390 units per month on average. Across the three intervals the magnitude of $\hat{\beta}_1$ is decreasing, which matches the decreasing magnitude of the slopes seen in Figure 3.

[49] We use the median rather than the mean because it is less sensitive to extreme outliers in the cost report data, which likely reflect data entry errors.

[50] We compute this 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 data, the median dosage per month is 45,000 units (Table 1). We use the median because it 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. Thus, we arrive at an average administration cost of \$1.05 per 1,000 units.

[51] We note that while there are not incentives to misreport acquisition costs to the government under the observed contract, conditioning on reported acquisition costs via an alternative payment contract would generate an incentive for providers to misreport acquisition costs. This makes it most natural to treat providers as possessing private information about the cost of treatment.

[52] The patient characteristics, x_{jt}, are age in years, an indicator for female sex, and indicators for each possible value of the Charlson Comorbidity Index. The regressions also include year and month dummies.

Table 2: Reduced-Form Coefficient Estimates

Variable (Coefficient)	Interval of Baseline Hematocrit > 30 to 33,	> 33 to 36,	> 36 to 39
Baseline hematocrit (β_1^k)	-9.29 (0.25)	-6.32 (0.15)	-3.56 (0.12)
Reimbursement rate ($\bar{\beta}_2^k$)	9.53 (3.02)	6.39 (2.13)	3.92 (1.97)
Obs. in interval	231,702	405,019	283,024

Notes: Estimates are from separate regressions in each interval, estimated via OLS. Regressions also include: age, sex, indicators for each value of the Charlson comorbidity index, month and year dummies. Standard errors in parentheses, computed via cluster bootstrap (clustered on dialysis center) with 250 replications.

Table 3 presents the structural parameter estimates. The parameters of the health function can be compared with certain information about EPO from the medical literature. For example, the estimate of δ_k in the middle interval implies that 1,000 units of EPO raises hematocrit by 0.158 percentage points. This and the estimates of δ_k in the other intervals are remarkably consistent with estimates of the average productivity of EPO that can be derived from results in clinical trials, which suggests the model is reasonably well specified.[53] Also, the larger values of δ_k in intervals with higher baseline hematocrit are consistent with diminishing marginal productivity of the drug, because patients with higher baseline hematocrit are given less EPO on average (Figure 3). With the estimates of τ_k, for the hematocrit target, the implied values of the individual targets ($\tau_k' x_{ijt}$) fall within the defined range for hematocrit (i.e., 0 to 100), and the averages reported in Table 3 are reasonably close to what might be expected based on clinical and policy guidelines.[54]

Next, in the distribution of physician types, the parameters $\mu_{\alpha,k}$ represent the means (and medians) of the normal distributions of $\ln \alpha$ for each interval of baseline hematocrit. The value of these parameters decreases across the intervals, which could be interpreted as a lower concern for the health of patients with less severe anemia. The median of α in interval k is $\exp(\mu_{\alpha,k})$, so for example the median in the middle interval is 18.4. This implies that if the payment rate were one dollar above the marginal cost for a physician with the median degree of altruism, that physician

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

[54] For example, guidelines issued by the National Kidney Foundation in 2007 recommended the use of hemoglobin targets from 11 to 12 g/dl, and not greater than 13 g/dl (The National Kidney Foundation-Kidney Disease Outcomes Quality Initiative, 2007), which is comparable to hematocrit targets from 33 to 36 percent, and not greater than 39 percent. These could be interpreted as possible values for the average target in our model ($\tau_k' \bar{x}_k$), assuming the guidelines ignored the cost of providing EPO. The averages reported in Table 3 are clearly larger than these values, but we consider them to be reasonably close, given that no constraints were placed on the estimates of τ_k.

Table 3: Structural Parameter Estimates

	Interval of Baseline Hematocrit		
Parameter	>30 to 33	>33 to 36	>36 to 39
Increase in hematocrit from 1000u EPO			
δ_k	0.108	0.158	0.281
	(0.003)	(0.004)	(0.009)
Mean implied hematocrit target			
$\tau_k'\bar{x}$	40.2	43.7	50.2
	(0.3)	(0.3)	(0.6)
Distribution of altruism and marginal cost types			
$\mu_{\alpha,k}$	3.54	2.91	2.99
	(0.73)	(0.83)	(1.42)
$\sigma^2_{\alpha,k}$	2.68	2.15	3.64
	(0.80)	(0.94)	(1.43)
$\sigma_{\alpha z,k}$	-0.343	-0.436	-0.371
	(0.014)	(0.062)	(0.011)
$\sigma^2_{z,k}$	0.472	0.858	0.332
	(0.162)	(0.396)	(0.073)
Obs.	231,702	405,019	283,024

Notes: Standard errors in parentheses, computed via cluster bootstrap (clustered on dialysis center) with 250 replications. Mean marginal cost, μ_z, is set at \$8.58/1000u EPO.

would provide a medically excessive dosage which raises the patient's hematocrit 0.345 percentage points beyond the medical target level, $\tau_k' x_{jt}$.[55] The marginal cost, z, is denominated in dollars, so the estimates of $\sigma^2_{z,k}$ imply standard deviations of marginal costs equal to \$0.69, \$0.93, and \$0.58 in the three intervals. For comparison, the interquartile range of acquisition costs reported in Table 1 is \$0.92. Last, these estimates indicate that heterogeneity in altruism, not marginal costs, accounts for most of the variation in dosages. For example in the middle interval, if we fix α at its mean and only allow z to vary, the standard deviation of simulated dosages is reduced by over 80%; in contrast, if we fixed z at its mean and only allowed α to vary, the standard deviation of dosages would be reduced by less than half that amount.[56]

[55] From (12), $\delta a^* + b_0 - \tau = (p_1 - z)(\alpha\delta)^{-1}$ and so we use \$1 $\times(18.4 \times 0.158)^{-1} = 0.345$.

[56] Computed for a patient with the median b_0 and mean x in this interval.

6 Quantitative Results: Optimal Contracts

This section presents our key empirical results: optimal contracts obtained using the estimated model parameters, and simulated outcomes under those contracts. We compare the allocations and surplus that would occur under the optimal constrained (i.e., linear) and unrestricted (i.e., nonlinear) contracts against those generated under the observed contract used by Medicare. The results indicate the potential value of replacing the constant per-unit payment rates in traditional fee-for-service systems with variable marginal payment rates.

Two final steps are required to compute the optimal contracts.[57] First, we must truncate the estimated type distributions to render them compact, as they are in the model. We accordingly remove the bottom and top 0.5 percent from the symmetric marginal distributions of z_k, and we remove the bottom 0.5 percent from the asymmetric marginal distributions of α_k and truncate from the top so as to maintain the estimated values of $\bar{\beta}_2^k$ (the mean of $\delta_k^{-2}\alpha_k^{-1}$).[58] Second, we must fix a value for α_g, the weight placed by the government on health relative to money. We do not assume the observed contract is optimal—moreover, we prove that at the estimated parameter values it is not possible to rationalize the observed payment rate with any value of α_g (Appendix B)—so we do not attempt to recover this parameter from the observed payment rates. Instead, we calibrate a value for α_g based on the value of a statistical life year (VSLY) and information on the relationship between hematocrit levels and mortality risk that is available from clinical trials on EPO (see Appendix G). The resulting value, $\alpha_g = 52.6$, is above the median value of α for the providers, meaning that the principal places more weight on patient health than do most agents, as might be expected.[59]

Here we present the results for a particular value of the baseline hematocrit (the median) and patient characteristics (the interval-specific mean), to illustrate in detail the differences between the optimal nonlinear, optimal linear, and observed contracts.[60] (Recall that contracts may be defined for each b_0 and x.) Contracts for values of the baseline hematocrit and patient characteristics from the other two intervals, which have different values of the estimated parameters, are presented in Appendix I. The results are qualitatively similar.

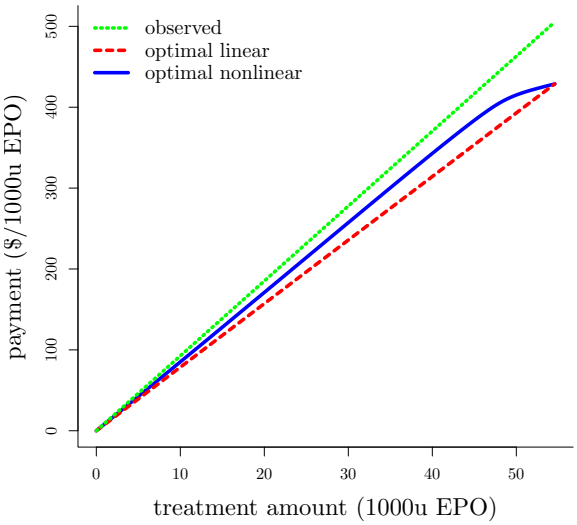
Figure 4 plots the total payments (panel a), the marginal payments (panel b), and the distributions of treatment amounts (panel c), with the nonlinear contract in blue solid lines, the linear contract in red dashed lines, and the observed contract in green dotted lines. All three contracts

[57] See Appendix D for computational details. Also, we assess the regularity condition, that no physician types' supply curves intersect the marginal payment curve under the optimal nonlinear contract more than once, and find that it is not violated (Appendix H).

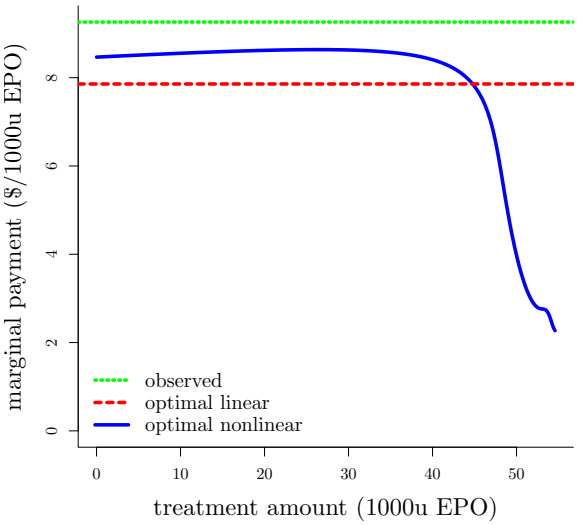
[58] Our results are not very sensitive to the choice of truncation points. Also note that by truncating the type distributions we reduce the importance of asymmetric information, which will tend to understate gains from moving to the optimal nonlinear contract.

[59] We have also computed results for a different value of α_g, which equates the mean health produced under the optimal nonlinear contract with the mean health produced under the observed contract. This shows the extent to which expenditures could be reduced, while maintaining average health outcomes. The implied value of α_g is 62.1, and the results are qualitatively and quantitatively similar to those presented here, and are available upon request.

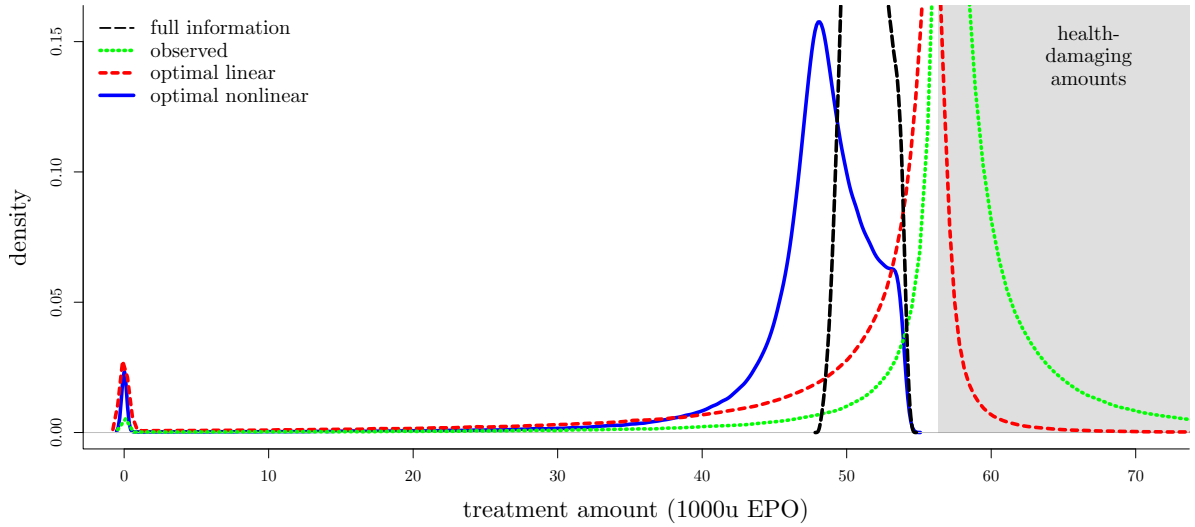
[60] The observed contract sets p_1 equal to the sample mean of the payment rates in the data, \$9.26. While there is variation in payment rates, we illustrate our results using the average to minimize clutter.



(a) Payment as a function of treatment amount



(b) Marginal payment as function of treatment amount



(c) Distribution of treatment amounts

Figure 4: Treatment and payment amounts under the optimal contract, for patients with median severity of anemia.

Notes: Figure plots treatment and payment amounts under the optimal nonlinear contract (blue, solid lines) for patients with median baseline hematocrit ($b_0 = 34.8$) and mean target hematocrit ($\tau_k' \bar{x}_k = 43.7$). Results with the optimal linear contract (red, dashed lines) and observed contract (green, dotted lines) are shown for comparison. Panel (a) plots the payment amounts, panel (b) plots marginal payments, and panel (c) plots the distribution of treatment amounts.

pay \$0 for zero provision, which occurs because the optimal contracts exclude some physicians (i.e., some types provide zero dosages in equilibrium), and so they use the same intercept of \$0 as the observed contract.[61] For positive treatment amounts, the total payments (panel a) from the optimal nonlinear contract are lower than from the observed contract, and may be higher or lower than the total payments from the optimal linear contract, depending on the treatment amount. The differences in these total payments can be non-trivial: for 45 thousand units, for example, the nonlinear contract would pay \$383.77, the linear contract would pay \$353.60, and the observed contract would pay \$416.70, per month. The marginal payment (panel b) in the nonlinear contract is roughly constant below 40 thousand units, where it lies between the fixed marginal rates of the observed and linear contracts. However, most dosages induced by the nonlinear contract are between 40 and 55 thousand units, where the marginal payment changes substantially, falling from above \$8 to about \$2 per 1,000 units.

The gray shaded area in panel c denotes medically excessive dosages, i.e., treatment amounts that reduce patient health. This plot also includes the distribution of treatment amounts in the full-information solution for comparison (black, dashed line). It is readily apparent that the treatment amounts under the observed contract are typically too high, exceeding even the health-maximizing amount ($[\tau_k'\bar{x}_k - b_0]/\delta = 56,329$ units) for all but the lowest-treatment providers. This accords with concerns that were raised about high payment rates encouraging medically excessive (not just economically excessive) provision of EPO. The optimal linear contract offers a lower payment rate, so the treatment amounts under this contract are less than those under the observed contract. However, it does not eliminate medically excessive amounts, which still occur with 19 percent of providers (see Table 4), because, as shown in Section 3, any providers with marginal costs below the constant payment rate will be induced to provide dosages that yield a negative marginal product of health, regardless of their degrees of altruism. Because the linear contract has only a single marginal payment, the government accepts these excessive dosages in order to avoid further underprovision by other types of providers (e.g., those with higher costs).

Next, to directly examine over- and underprovision, in the economic sense, Figure 5 plots the distributions (across provider types) of the deviations of the treatment amounts provided under each contract from their full information amounts.[62] Overprovision is nearly universal with the observed contract (95.3% of provider types), and it remains very common with the optimal linear contract (73.5% of provider types). In other words, under the optimal linear contract, most providers still administer dosages where the marginal benefit to the principal is below the net marginal cost for the agent. By contrast, there is no overprovision with the optimal nonlinear contract. As we discussed in Section 3.3, as is standard with optimal unrestricted contracts, the highest treatment amount equals the maximum in the full-information allocation, and all other treatment amounts

[61] The reservation utility $\underline{u}$ is set equal to the lowest utility obtained under the observed contract. A very small share of physicians (0.2%) are excluded in the simulation of the observed contract, which fixes $\underline{u}$ at the utility of a treatment amount of zero and zero payment, for the type with the lowest degree of altruism (see Appendix A.2).

[62] For example, the deviations under the optimal nonlinear contract are $a^{*\mathrm{SB}}(\alpha, z) - a^{*\mathrm{FI}}(\alpha, z)$, where $a^{*\mathrm{SB}}(\alpha, z)$ is the equilibrium treatment amount provided by type (α, z) under the second-best and $a^{*\mathrm{FI}}(\alpha, z)$ is defined in (3).

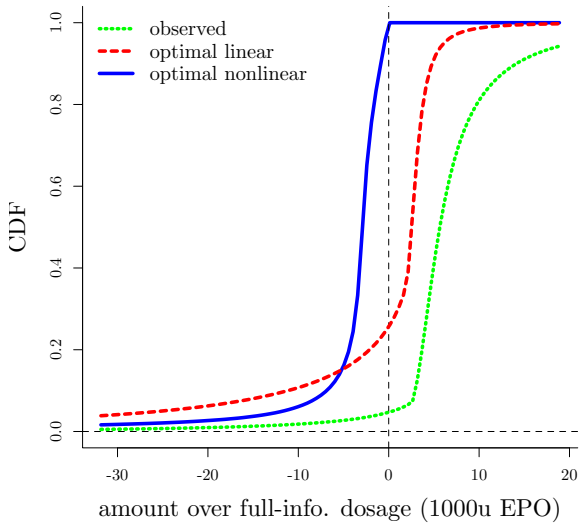


Figure 5: CDFs of deviations from full-information treatment amounts, for patients with median severity of anemia.

Notes: Figure plots the CDFs of the deviations from full-information treatment amounts under the optimal nonlinear contract (blue, solid line), optimal linear contract (red, dashed line), and baseline contract (green, dotted line), for patients with median baseline hematocrit ($b_0 = 34.8$) and mean target hematocrit ($\tau_k' \bar{x}_k = 43.7$).

are distorted downward. This further indicates the value of having flexible marginal incentives, because any amount of overprovision is dominated by a corresponding amount of underprovision that yields the same health but costs less.

Table 4 summarizes the outcomes under the three contracts. The mean dosage is smallest under the optimal nonlinear contract, and so is the mean payment. The variation in dosages, measured by the standard deviation, indicates the extent to which these contracts address the unobserved heterogeneity across providers. (Recall that b_0 and x are held constant in this example, so the medical need for treatment is the same across patients.) Compared to the observed contract, the optimal nonlinear contract reduces the standard deviation of dosages by 27%. By contrast, the optimal linear contract does not reduce the variation in dosages, because it provides a constant marginal incentive, just like the observed contract. In fact, there is *greater* variation than under the observed contract because some types are optimally excluded, which puts a non-negligible mass at zero, far from the mean. For comparison, with full information the standard deviation would be only 1.3 thousand units, about one seventh of the standard deviation under the observed contract. The variation that remains with full information reflects only the variation in altruism and costs, without any distortions due to informational frictions.

We now show in more detail how the flexible marginal incentives in a nonlinear contract allow the government to improve its objective in the presence of multidimensional unobserved heterogeneity. Figure 6a plots isoquants under the full-information (dashed lines) and second-best allocations

Table 4: Summary of Outcomes under Alternative Contracts

(patients with median severity of anemia and mean characteristics)

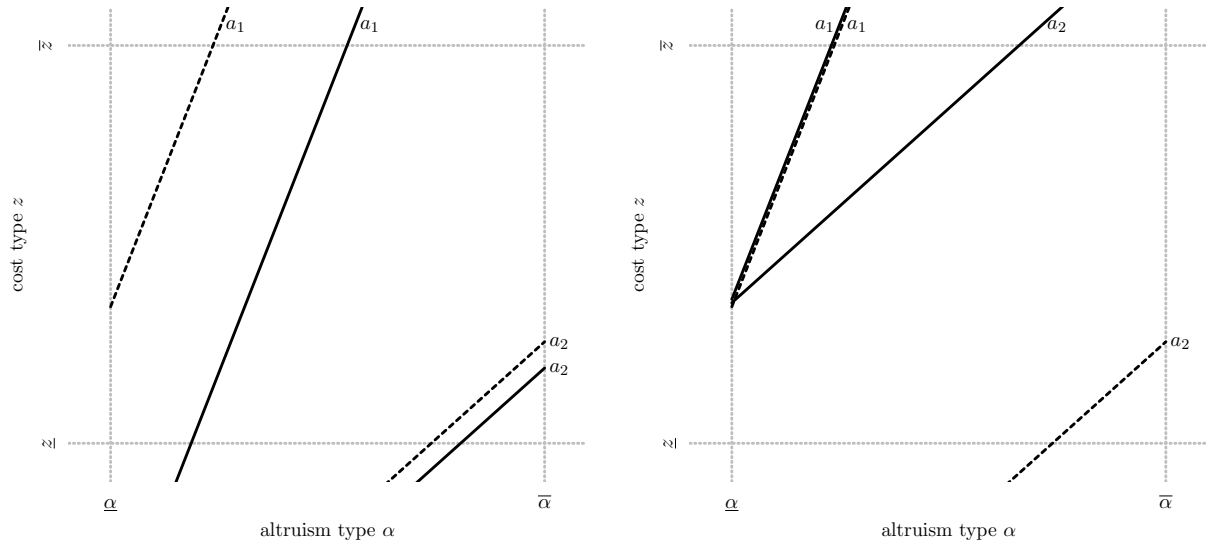
Contract	Mean Payment	Mean Dosage	Std. Dev. Dosage	Share above τ	Gain in Govt. Obj.
Observed	542	58.6	9.8	75%	
Optimal Linear	396	50.4	11.8	19%	\$98
Optimal Nonlinear	393	47.1	7.2	0%	\$125

Note: Table shows summary statistics of outcomes corresponding to contracts plotted in Figure 4. Mean and SD of dosage are in 1,000 units/month. The gain in the government objective is computed relative to the observed payment contract.

(solid lines) for the 75th and 99.99th percentile treatment amounts under the second best, which we respectively denote a_1 and a_2. The higher amount (a_2) is close to the full-information maximum ($\bar{a}^{*\text{FI}}$) because there is no distortion at the top. The physician types that provide at least a_1 or a_2 lie below (i.e., lower z and higher α) the corresponding isoquants. The isoquants under the optimal nonlinear contract are below the corresponding isoquants under full information because of the downward distortion, which is larger at the lower amount (a_1). This is in contrast to the optimal linear contract (panel b), which can result in overprovision. In particular, the isoquant for a_2 lies far above the full-information isoquant, indicating that many types provide more than the full-information maximum ($\bar{a}^{*\text{FI}}$) under the linear contract. (The coincidence of the isoquants for a_1 under full information and the optimal linear contract is itself coincidental.)

With this figure, we can see how the nonlinear and linear contracts discriminate between the two dimensions of unobserved heterogeneity, by projecting the set of physician types providing (at least) a treatment amount onto each axis. This gives the range of values from each dimension which could provide (at least) that amount. First, under full information, there exist combinations of (α, z) such that all altruism types and all cost types could provide a_1, while only strict subsets of both dimensions could provide a_2. The optimal nonlinear contract discriminates more between altruism types than cost types for a_1 (indeed, every cost type would provide a_1), and it gets fairly close to the full-information allocation under a_2. The optimal linear contract does a much poorer job of discriminating between types along either dimension—the sets of altruism types and cost types that could provide each of the pictured treatment amounts equal the full ranges in each dimension. Thus the flexible incentives provided by the nonlinear contract allow the government to better address the multidimensional heterogeneity, and we learn that the government focuses more on targeting physician altruism.

Finally, in the last column of Table 4, we show the government's gains from using optimal contracts, by calculating the increases in the government's maximized objective relative to that under the observed contract. This provides a summary measure, in dollars per patient per month, of the potential benefit to the government (and by extension, the patients represented by the



(a) Optimal nonlinear contract

(b) Optimal linear contract

Figure 6: Isoquants for the 75th percentile (a_1) and 99.99th percentile (a_2) treatment amounts under the optimal nonlinear contract.

Notes: Figure plots isoquants in the type space for two fixed amounts: the 75th percentile (a_1) and 99.99th percentile (a_2) provided under the optimal nonlinear contract. The solid lines are the isoquants for these amounts under the optimal nonlinear contract (panel a) under the optimal linear contract (panel b). For comparison, the isoquants for these amounts under full information are shown with dashed lines in each panel.

government) from the changes in outcomes discussed above.[63] There are substantial gains from using the optimal nonlinear contract, equal to \$125 per month or \$1,500 per year. By comparison, the mean monthly payment under the observed contact is \$542, so the gains are equal to 23 percent of that payment. The optimal linear contract yields almost 80 percent of the gain from optimal nonlinear contract, but the difference between the two is non-trivial, amounting to \$324 per patient per year.

7 Summary and Conclusions

In this paper, we develop an empirical analysis of optimal payment contracts, in an environment where agents have multidimensional hidden types. Our application pertains to a large sector of the economy, health care, where asymmetric information is pervasive and agents' responses to incentives can have important impacts on both health and costs. In health care, as in many other areas, allowing for multidimensional heterogeneity among producers is important for developing

[63] Aside from the fact that we consider the government's objective, not social welfare, this is analogous to standard measures of welfare changes, equivalent and compensating variation, which are equal here due to the quasilinearity of the government's objective. We do not discuss the levels of the maximized objective because the health function includes the constant H, which is unidentified and drops out from the differences shown here.

realistic empirical models that can inform payment policies.

We specify a parsimonious screening model, where the agents (dialysis providers) are heterogeneous in their altruism and marginal costs, and this heterogeneity is unobserved to the principal (the government, which pays for most dialysis care via Medicare). The unique features of our setting enable us to estimate the structural parameters of the model using simple linear reduced forms. We then construct the optimal unconstrained (i.e., nonlinear) contract by combining these estimates with an approach for solving the model based on nonlinear pricing methods.

We find that there is substantial heterogeneity among providers in our setting, in both their marginal costs and especially their degrees of altruism. This implies that an optimal contract will be nonlinear—i.e., optimal marginal incentives will vary to address the effects of asymmetric information. Moreover, because the providers are found to be substantially altruistic on average, an optimal contract will incorporate this altruism to achieve better patient health at lower cost. Our results also indicate that the payment policy in place during our study period in the early 2000s was substantially suboptimal. Moving to an optimal nonlinear contract, or even an optimal linear contract (i.e., a lower payment rate for EPO), could reduce medically excessive treatment and save the government money. However, the optimal linear contract would not eliminate medically excessive treatment, because its single marginal incentive must balance lower-cost providers (who give more treatment) against higher-cost providers (who give less treatment). This makes clear the potential value of using nonlinear pricing strategies in payment contracts for health care and other complex services.

This analysis demonstrates that it is possible to empirically examine optimal contracts in the presence of multidimensional heterogeneity, using relatively straightforward modeling strategies. While our results are specific to the institutions and data in our application, the approach we developed here may be employed more broadly to understand and construct payment contracts in a variety of real-world settings.

References

Abito, J. M., "Measuring the Welfare Gains from Optimal Pollution Regulation," *Review of Economic Studies*, 2019, forthcoming, http://www.restud.com/wp-content/uploads/2019/08/MS24699manuscript.pdf.

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.

Armstrong, M., "Multiproduct Nonlinear Pricing," *Econometrica*, 64(1):51–75, 1996.

Baron, D. P. and R. B. Myerson, "Regulating a Monopolist with Unknown Costs," *Econometrica*, 50(4):911–930, 1982.

Beddhu, S., F. J. Bruns, M. Saul, P. Seddon and M. L. Zeidel, "A Simple Comorbidity Scale Predicts Clinical Outcomes and Costs in Dialysis Patients," *American Journal of Medicine*, 108(8):609 – 613, 2000.

Besley, T. and M. Ghatak, "Competition and Incentives with Motivated Agents," *American Economic Review*, 95(3):616–636, 2005.

Chalkley, M. and J. M. Malcomson, "Government Purchasing of Health Services," in A. J. Culyer and J. P. Newhouse, eds., "Handbook of Health Economics," vol. 1, Part A of *Handbook of Health Economics*, pp. 847 – 890, Elsevier, 2000.

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

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

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

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

Deneckere, R. and S. Severinov, "Multi-dimensional Screening: A Solution to a Class of Problems," 2015, unpublished manuscript, University of California, Santa Barbara.

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

Einav, L., A. Finkelstein and N. Mahoney, "Provider Incentives and Healthcare Costs: Evidence From Long-Term Care Hospitals," *Econometrica*, 86(6):2161–2219, 2018.

Eliason, P., "Market Power and Quality: Congestion and Spatial Competition in the Dialysis Industry," 2017, unpublished manuscript, Brigham Young University.

Eliason, P. J., B. Heebsh, R. C. McDevitt and J. W. Roberts, "How Acquisitions Affect Firm Behavior and Performance: Evidence from the Dialysis Industry," *The Quarterly Journal of Economics*, 135(1):221–267, 2019.

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

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

Foley, R. N., "Do We Know the Correct Hemoglobin Target for Anemic Patients with Chronic Kidney Disease?" *Clinical Journal of the American Society of Nephrology*, 1(4):678–684, 2006.

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

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

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

Gaynor, M., J. Rebitzer and L. Taylor, "Physician Incentives in Health Maintenance Organizations," *Journal of Political Economy*, 112(4):915–931, 2004.

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

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

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

Johnson, S. G., "The NLopt nonlinear-optimization package," 2018, http://ab-initio.mit.edu/nlopt.

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

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

McAfee, R. P. and J. McMillan, "Multidimensional Incentive Compatibility and Mechanism Design," *Journal of Economic Theory*, 46(2):335–354, 1988.

McGuire, T. G., "Physician Agency," in A. J. Culyer and J. P. Newhouse, eds., "Handbook of Health Economics," vol. 1, Part A of *Handbook of Health Economics*, pp. 461–536, Elsevier, 2000.

Mirrlees, J. A., "The Theory of Optimal Taxation," in K. J. Arrow and M. D. Intriligator, eds., "Handbook of Mathematical Economics," vol. 3, pp. 1197–1249, Elsevier, 1986.

Myerson, R. B., "Optimal Auction Design," *Mathematics of Operations Research*, 6(1):58–73, 1981.

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

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.

Quan, H., V. Sundararajan, P. Halfon, A. Fong, B. Burnand, J.-C. Luthi, L. D. Saunders, C. A. Beck, T. E. Feasby and W. A. Ghali, "Coding Algorithms for Defining Comorbidities in ICD-9-CM and ICD-10 Administrative Data," *Medical Care*, 43(11):1130–1139, 2005.

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.

Rochet, J.-C. and L. A. Stole, "The Economics of Multidimensional Screening," in M. Dewatripont, L. P. Hansen and S. J. Turnovsky, eds., "Advances in Economics and Econometrics: Theory and Applications, Eighth World Congress," vol. 1 of *Econometric Society Monographs*, pp. 150–197, Cambridge University Press, 2003.

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, pMID: 17108343.

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

The National Kidney Foundation-Kidney Disease Outcomes Quality Initiative, "KDOQI Clinical Practice Guidelines and Clinical Practice Recommendations for Anemia in Chronic Kidney Disease," *American Journal of Kidney Diseases*, 47(S3):S1–S146, 2006.

—, "KDOQI Clinical Practice Guideline and Clinical Practice Recommendations for Anemia in Chronic Kidney Disease: 2007 Update of Hemoglobin Target," *American Journal of Kidney Diseases*, 50(3):471 – 530, 2007.

Tirole, J., "Market Failures and Public Policy," Nobel Prize Lecture, 2014, https://www.nobelprize.org/uploads/2018/06/tirole-lecture.pdf.

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

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.

WHO, *Nutritional Anaemias: Report of a WHO Scientific Group*, World Health Organization, Geneva, 1968.

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

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

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 6, 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 5.

Using interior physician's treatment choice functions (12), the government's problem can be written as

$$\max_{\{(p_0,p_1)\in\mathbb{R}^2\}} \int_{\underline{\alpha}}^{\overline{\alpha}} \int_{\underline{z}}^{\overline{z}} \left[\alpha_g h(a) - p_0 - p_1 a^*(\alpha, z; p_1)\right] f(\alpha, z) dz d\alpha \tag{14}$$

s.t.

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

$$a^*(\alpha, z; p_1) = \frac{\tau - b_0}{\delta} + \frac{p_1 - z}{\delta^2 \alpha}, \quad \forall(\alpha, z) \qquad \text{IC.}$$

We can eliminate the participation constraints for all types but $(\ddot{\alpha}, \ddot{z}) \equiv \arg\min_{(\alpha,z)} u(a^*(\alpha, z; p_1); \alpha, z, p_0, p_1)$, i.e., the lowest-utility type given linear contract (p_0, p_1).[64] Setting up the Lagrangian based on the remaining participation constraint, we have

$$\begin{aligned}\mathcal{L} = &\int_{\underline{\alpha}}^{\overline{\alpha}} \int_{\underline{z}}^{\overline{z}} \left[\alpha_g \left[H - \frac{[p_1 - z]^2}{2\delta^2\alpha^2}\right] - p_0 - p_1 \left[\frac{[\tau - b_0]}{\delta} + \frac{p_1 - z}{\delta^2\alpha}\right]\right] f(\alpha, z) dz d\alpha \\ &+ \mu \left[\ddot{\alpha} H + \frac{[p_1 - \ddot{z}]^2}{2\delta^2\ddot{\alpha}} + \frac{[\tau - b_0][p_1 - \ddot{z}]}{\delta} + p_0 - \underline{u}\right].\end{aligned}$$

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

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

$$\frac{\partial \mathcal{L}}{\partial p_1} = \int_{\underline{\alpha}}^{\overline{\alpha}} \int_{\underline{z}}^{\overline{z}} \left[-\alpha_g \left[\frac{p_1^* - z}{\delta^2\alpha^2}\right] - \left[\frac{[\tau - b_0]}{\delta} + \frac{p_1^* - z}{\delta^2\alpha}\right] - \frac{p_1^*}{\delta^2\alpha}\right] f(\alpha, z) dz d\alpha + \mu^* \left[\frac{p_1^* - \ddot{z}}{\delta^2\ddot{\alpha}} + \frac{\tau - b_0}{\delta}\right] = 0.$$

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

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

p_1^*:

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

$$\Rightarrow p_1^* = \frac{\alpha_g \, \mathrm{E}\left[\frac{z}{\alpha^2}\right] + \mathrm{E}\left[\frac{z}{\alpha}\right] - \frac{\ddot{z}}{\ddot{\alpha}}}{\alpha_g \, \mathrm{E}\left[\frac{1}{\alpha^2}\right] + 2\,\mathrm{E}\left[\frac{1}{\alpha}\right] - \frac{1}{\ddot{\alpha}}}. \tag{15}$$

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 $\tilde{z}^0(\alpha; p_1) \equiv \alpha\delta[\tau - b_0] + p_1$ denote the cost type indifferent between providing treatment and not, given altruism type α and payment rate p_1.[65] The government's problem, allowing for exclusion, is:

$$\max_{\{(p_0,p_1)\in\mathbb{R}^2\}} \mathrm{E}\left[u_g(a(\alpha, z; p_1); p_0, p_1)\right] = \int_{\underline{\alpha}}^{\overline{\alpha}} \int_{\underline{z}}^{\tilde{z}^0(\alpha,p_1)} \left[\alpha_g h(a^*(\alpha, z; p_1)) - p_0 - p_1 a^*(\alpha, z; p_1)\right] f(\alpha, z) dz d\alpha$$
$$+ \int_{\underline{\alpha}}^{\overline{\alpha}} \int_{\tilde{z}^0(\alpha,p_1)}^{\overline{z}} \left[\alpha_g h(0) - p_0\right] f(\alpha, z) dz d\alpha \tag{16}$$

s.t.

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

$$a^*(\alpha, z; p_1) = \begin{cases} \frac{\tau - b_0}{\delta} + \frac{p_1 - z}{\delta^2 \alpha}, & \forall\{(\alpha, z) : z < \tilde{z}^0(\alpha, p_1)\} \\ 0, & \forall\{(\alpha, z) : z \geq \tilde{z}^0(\alpha, p_1)\} \end{cases} \qquad \text{IC.}$$

(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 5.)

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

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

[65] Note that $\tilde{z}^0 \equiv \tilde{z}(\alpha; p_1, a = 0)$, where $\tilde{z}$ is defined in equation (19), in Appendix C.

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[66], we obtain the following implicit expression for p_1^*:

$$\int_{\underline{\alpha}}^{\overline{\alpha}} \int_{\underline{z}}^{\tilde{z}^0(\alpha,p_1^*)} \left[\frac{z[\alpha_g+\alpha]}{\alpha^2}\right] f(\alpha,z)dzd\alpha - \delta[\tau-b_0] \int_{\underline{\alpha}}^{\overline{\alpha}} \int_{\underline{z}}^{\tilde{z}^0(\alpha,p_1^*)} f(\alpha,z)dzd\alpha = p_1^* \int_{\underline{\alpha}}^{\overline{\alpha}} \int_{\underline{z}}^{\tilde{z}^0(\alpha,p_1^*)} \left[\frac{\alpha_g+2\alpha}{\alpha^2}\right] f(\alpha,z)dzd\alpha. \tag{17}$$

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)).

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 5.)

Let $\tilde{z}^0(\alpha;p_1) \equiv \alpha\delta[\tau-b_0]+p_1$ denote the cost type indifferent between providing treatment and not, given altruism type α and payment rate p_1.[67] Let $p_1^*(\alpha_g;\tilde{z}^0(\cdot,p_1^*))$ denote the solution to (17), 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

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

[67] Note that $\tilde{z}^0(\alpha;p_1) \equiv \tilde{z}(\alpha;p_1,a=0)$, where $\tilde{z}$ is defined in equation (19), in Appendix C. This is the same definition as in Appendix A.2, and is reproduced here for convenience.

of (15) 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*

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

which is positive because the first-order condition of the government's problem with respect 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 (17) as $\alpha_g \to \infty$, and after some manipulation and dropping the vanishing terms, we have*

$$\int_{\underline{\alpha}}^{\overline{\alpha}} \int_{\underline{z}}^{\tilde{z}^0(\alpha,p_1^*)} \frac{z}{\alpha^2} f(\alpha,z)dzd\alpha = p_1^* \int_{\underline{\alpha}}^{\overline{\alpha}} \int_{\underline{z}}^{\tilde{z}^0(\alpha,p_1^*)} \frac{1}{\alpha^2} f(\alpha,z)dzd\alpha. \tag{18}$$

Treating $\tilde{z}^0$ as a parameter, consider how an increase in $\tilde{z}^0$ (towards $\overline{z}$) would affect p_1^* defined in (18). 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 $\tilde{z}^0(\cdot,\cdot) \geq \underline{z} > 1$,[68] *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 A1 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, $p_1^*(\alpha_g; \tilde{z}^0(\cdot, p_1^*)) \leq p_1^*(\alpha_g = \infty; \tilde{z}^0(\cdot, p_1^*)) \leq p_1^*(\alpha_g = \infty; \tilde{z}^0(\cdot, p_1^*) = \overline{z}) = \mathrm{E}\left[\frac{z}{\alpha^2}\right] / \mathrm{E}\left[\frac{1}{\alpha^2}\right] < 9.26$.

Table A1: 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	8.95

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

C 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):

$$\tilde{z}(\alpha; p, a) \equiv p + \alpha h'(a). \tag{19}$$

Note that $\tilde{z}$ has intercept p and slope of $h'(a)$, both of which must be non-negative at an optimal solution $p^*(a)$.[69] 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 A1 shows, there are two cases, corresponding to $\tilde{\alpha}$. If $\tilde{\alpha} \geq \overline{\alpha}$, as depicted on the left, then

$$S(p, a) = \Pr\{\underbrace{\alpha h'(a) + p}_{\tilde{z}(\alpha; p, a)} \geq z\} = \int_{\underline{\alpha}}^{\overline{\alpha}} \int_{\underline{z}}^{\tilde{z}(\alpha; p, a)} f(\alpha, z) dz d\alpha, \tag{20}$$

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 $\tilde{\alpha}$ will choose at least the level of treatment under consideration.[70] Thus, we have

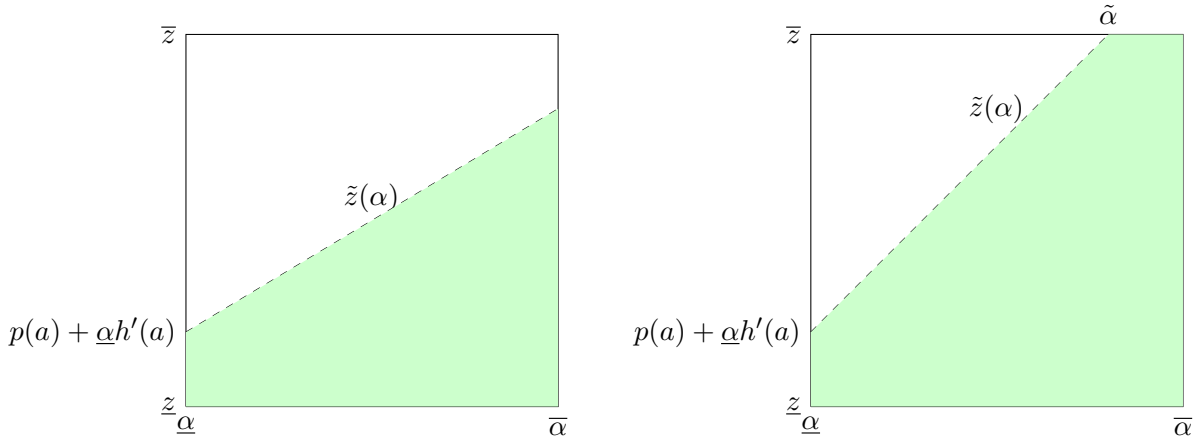
$$S(p, a) = \int_{\underline{\alpha}}^{\tilde{\alpha}(p,a)} \int_{\underline{z}}^{\tilde{z}(\alpha; p, a)} f(\alpha, z) dz d\alpha + [1 - F_\alpha(\tilde{\alpha})], \tag{21}$$

[68] 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.

[69] 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.

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

Figure A1: $\tilde{\alpha}$ cases



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

$$\frac{\partial S(p,a)}{\partial p} = \int_{\underline{\alpha}}^{\overline{\alpha}} f(\alpha, \tilde{z}(\alpha; p, a)) \underbrace{\frac{\partial \tilde{z}(\alpha; p, a)}{\partial p}}_{1} d\alpha. \tag{22}$$

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

$$\frac{\partial S(p,a)}{\partial p} = \int_{\underline{\alpha}}^{\tilde{\alpha}} f(\alpha, \tilde{z}(\alpha; p, a)) d\alpha. \tag{23}$$

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 $a \in A$.[71]

[71] Although not depicted in Figure A1, 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 $\check{\alpha}(p,a) \equiv \frac{\underline{z}-p}{h'(a)}$ denote the altruism type satisfying $\tilde{z}(\check{\alpha}) = \underline{z}$. (Note that the condition $\tilde{z}(\underline{\alpha}) < \underline{z}$ is equivalent to $\check{\alpha}(p,a) > \underline{\alpha}$.) There are two subcases. First, if $\check{\alpha}(p,a) > \overline{\alpha}$, 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, if $\check{\alpha}(p,a) \in (\underline{\alpha}, \overline{\alpha}]$ then, if $\tilde{\alpha} \geq \overline{\alpha}$ then (20) becomes

$$S(p,a) = \int_{\check{\alpha}}^{\overline{\alpha}} \int_{\underline{z}}^{\tilde{z}(\alpha;p,a)} f(\alpha, z) dz d\alpha, \tag{24}$$

and if, instead, $\tilde{\alpha} \in [\underline{\alpha}, \overline{\alpha})$, then (21) becomes

$$S(p,a) = \int_{\check{\alpha}}^{\tilde{\alpha}(p,a)} \int_{\underline{z}}^{\tilde{z}(\alpha;p,a)} f(\alpha, z) dz d\alpha + [1 - F_\alpha(\tilde{\alpha})]. \tag{25}$$

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:

$$\max_{\{(p_0,p_1)\in\mathbb{R}^2\}} \mathrm{E}\left[u_g(a(\alpha,z;p_1);p_0,p_1)\right] = \int_{\underline{\alpha}}^{\overline{\alpha}}\int_{\underline{z}}^{\overline{z}} \left[\alpha_g h(a^*(\alpha,z;p_1)) - p_0 - p_1 a^*(\alpha,z;p_1)\right] f(\alpha,z)dzd\alpha \tag{26}$$

s.t.

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

$$a^*(\alpha,z;p_1) = \max\left\{0, \frac{\tau-b_0}{\delta} + \frac{p_1-z}{\delta^2\alpha}\right\}, \quad \forall(\alpha,z) \qquad \text{IC.}$$

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

As noted in Section 5.2, 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 coefficient (β_2^k) and random effect (ν^k) in the reduced form (13). First we present an auxiliary regression of the residuals of

(13) that yields the second moments of β_2^k and ν^k (while the mean of β_2^k comes directly from (13), 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 (13) can be rearranged as

$$a_{ijt} = \beta_0^k + \beta_1^k b_{0jt} + \bar{\beta}_2^k \tilde{p}_t + \beta_3^{k\prime} x_{jt} + \underbrace{\tilde{\beta}_i^k \tilde{p}_t + \nu_i^k + \epsilon_{ijt}^k}_{r_{ijt}^k}$$

(for b_{0jt} 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 5.2. 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.

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

$$\begin{aligned} r_{ijt}^k \bar{r}_i^k &= (\tilde{\beta}_i^k \tilde{p}_t + \nu_i^k + \epsilon_{ijt}^k) \left(\frac{1}{n_i^k} \sum_{l,s:b_{0ls} \in k} \tilde{\beta}_i^k \tilde{p}_s + \nu_i^k + \epsilon_{ils}^k \right) \\ &= (\tilde{\beta}_i^k \tilde{p}_t)\tilde{\beta}_i^k \bar{\tilde{p}}_i^k + (\tilde{\beta}_i^k \tilde{p}_t)\nu_i^k + (\tilde{\beta}_i^k \tilde{p}_t)\bar{\epsilon}_i^k \\ &+ \nu_i^k \tilde{\beta}_i^k \bar{\tilde{p}}_i^k + \nu_i^k \nu_i^k + \nu_i^k \bar{\epsilon}_i^k \\ &+ \epsilon_{ijt}^k \tilde{\beta}_i^k \bar{\tilde{p}}_i^k + \epsilon_{ijt}^k \nu_i^k + \epsilon_{ijt}^k \bar{\epsilon}_i^k. \end{aligned}$$

(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:

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

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 $r_{ijt}^k \bar{r}_i$ on $\tilde{p}_t \bar{\tilde{p}}_i$, $\tilde{p}_t + \bar{\tilde{p}}$, a constant, and $\frac{1}{n_i}$.

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:

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

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: $\mu_{\alpha,k}$, $\sigma^2_{\alpha,k}$, $\sigma_{\alpha z,k}$, and $\sigma^2_{z,k}$. 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^2_α 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

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

(Also recall that δ comes directly from β_1 in (13).)

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.[72] 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, 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,[73]

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

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

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

[72] 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 $g'(X_2) = 2\alpha^{-2}$.

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

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:

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

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 (13) and the auxiliary regression above yield consistent estimates of the structural parameters.

F Identification

Here we discuss the identification of the joint density, F, and the health function, h. The data contain $(a_{ijt}, b_{0jt}, x_{jt}, p_t)$ for patients $j = 1 \ldots n_i$ at providers $i = 1 \ldots n$ in time periods $t = 1 \ldots T$. The number of time periods is fixed, but both the number of providers and the number of patients per provider go to infinity. We first show the nonparametric identification of F, given the quadratic specification of h, which requires only mean-independence of the error term η_{ijtk}. We then show the semiparametric identification of h, specifically features of the shape of the function, if its arguments enter via a known index specification.

F.1 Identification of F

Let $n_i \to \infty$, and further suppose that the number observations within each interval of baseline hematocrit (k) goes to infinity for each provider. Assume that η_{ijtk} is mean-independent of (b_{0jt}, x_{jt}, p_t): $\mathrm{E}(\eta_{ijtk}|b_{0jt}, x_{jt}, p_t) = 0$. Then OLS estimation of the reduced form (13), separately within each interval for each provider, yields consistent estimates of β_1^k, β_{2i}^k, β_3^k, and ν_i^k, for $i = 1 \ldots n$ and $k = 1 \ldots K$. The structural parameters and provider types are continuous functions of reduced-form parameters and variables, as follows:

$$\begin{aligned}
\delta_k &= -(\beta_1^k)^{-1} \\
\tau_k &= -(\beta_1^k)^{-1}\beta_3^k \\
\alpha_{ik} &= (\beta_1^k)^2(\beta_{2i}^k)^{-1} \\
z_{ik} &= \mu_z - \nu_i^k(\beta_{2i}^k)^{-1}
\end{aligned}$$

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 distributions F_k are identified from the consistent estimates of (α_{ik}, z_{ik}) for each $i = 1 \ldots n$ and $k = 1 \ldots K$.

F.2 Identification of h

We now show how a single-index assumption makes it possible to identify the shape of h, specifically its second derivative up to scale. The scale of h is not separately identified from the scale of α because they enter the physician's utility function (1) multiplicatively, but our interest is in how the slope of h changes over its domain, not its absolute magnitude. To state the single-index assumption, with some abuse of notation, let $h(a; b_0, x) = h(\delta a + b_0 - \tau' x)$, where the values of δ and τ are unknown (and we consider a particular baseline hematocrit interval so the subscript k is omitted.)

The physician's first-order condition (4) yields a moment equality,

$$\mathrm{E}[\alpha h'(\delta a + b_0 - \tau' x)\delta - z + p \mid b_0, x, p] = 0.$$

Variation in b_0 and x within the same time period and the same provider identifies the parameters δ and τ, because the marginal net cost $(z_i - p_t)$ is constant, hence the marginal health benefit $(\alpha_i h'(\cdot))$ must be constant. So, given the strict concavity of h, the index inside h must take the same value for all patients receiving treatment from that provider in that time period. This identifies the index up to scale and location. To fix the scale, the coefficient on b_0 is set to one, which gives the index a natural interpretation in the units of the hematocrit level. To fix the location, the intercept of τ may be set to zero. (This contrasts with our parametric specification of h, where the intercept of τ is also identified.)

Then given δ and τ, which determine the value of the index inside h, the shape of h is identified from variation in the payment rate across time periods. Let $y_{ijt} \equiv \delta a_{ijt} + b_{0jt} - \tau' x_{jt}$ denote the value of the index for a particular observation. The expectation of the difference between the first-order conditions (4) in two periods t and s for some provider i is then

$$\left(\mathrm{E}[\alpha_i h'(y_{ijt}) \mid p_t] - \mathrm{E}[\alpha_i h'(y_{ijs}) \mid p_s]\right)\delta = p_s - p_t.$$

Taking the ratio of these differences for two pairs of time periods, q, r and s, t, we have

$$\frac{\mathrm{E}[h'(y_{ijr}) \mid p_r] - \mathrm{E}[h'(y_{ijq}) \mid p_q]}{\mathrm{E}[h'(y_{ijt}) \mid p_t] - \mathrm{E}[h'(y_{ijs}) \mid p_s]} = \frac{p_q - p_r}{p_s - p_t}.$$

Hence the ratio of the change in the derivative of h between two points (y_{ijq} to y_{ijr}) versus two other points (y_{ijs} to y_{ijt}) is known. This essentially identifies the second derivative up to scale. For each provider there are T points of support (where T is the number of periods with different payment rates) because, as noted above, the index has the same value for all patients receiving treatment from a given provider in a given time period. However the values of the index may be different for different providers, because of the heterogeneity in α and z. Therefore this finite-difference approximation to the second derivative of h can be traced out at many points of support in the domain of h.

G 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 squared loss from a target level of hematocrit—and dollars. The clinical trial gives estimates of the mortality risk associated with different hematocrit levels, so under certain assumptions (described

below), we can find a value of α_g that equates a function of the squared difference in 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 τ (i.e., the correct medical target, where health is maximized), we can compute values of our specification of health, i.e., the squared loss from τ. We assume that the lower target used in the trial is equal to τ, so the difference in health between the two targets is equal to $\frac{1}{2}(37.8 - 33.9)^2 = 7.6$. Multiplying this by α_g gives the government's value of this difference in hematocrit levels, in terms of dollars.

If we further assume that the government's value of this difference in hematocrit levels 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. Aldy and Viscusi (2008) provides VSLY estimates of approximately \$300,000 (p. 580), so the annual value of the difference in risks would be $0.016 \times \$300,000 = \$4,800$. Because the time periods in our model are months, this would equal the government's value of the above difference in hematocrit levels over twelve periods. To summarize, we have

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

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

H Check of Regularity Condition

Figure A2 plots the supply curves (dashed, grey lines) of physician types providing each treatment amount for a patient with the median baseline hematocrit level, and shows that none intersect the marginal payment curve (solid, black line) more than once.[74]

[74] We have also verified that this regularity condition is satisfied in the other baseline hematocrit intervals.

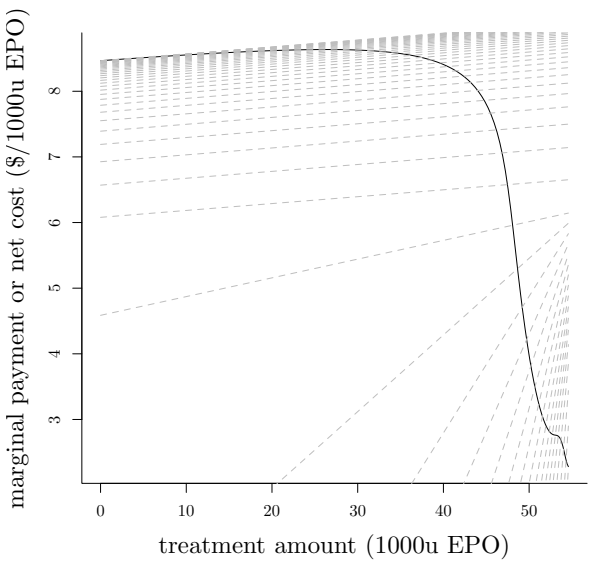


Figure A2: Regularity condition check, for patients with median severity of anemia.

Notes: Figure plots marginal payment curve (solid, black line) and physician supply curves (dashed, grey lines) for patients with median baseline hematocrit ($b_0 = 34.8$) and mean target hematocrit ($\tau_k' \bar{x}_k = 43.7$).

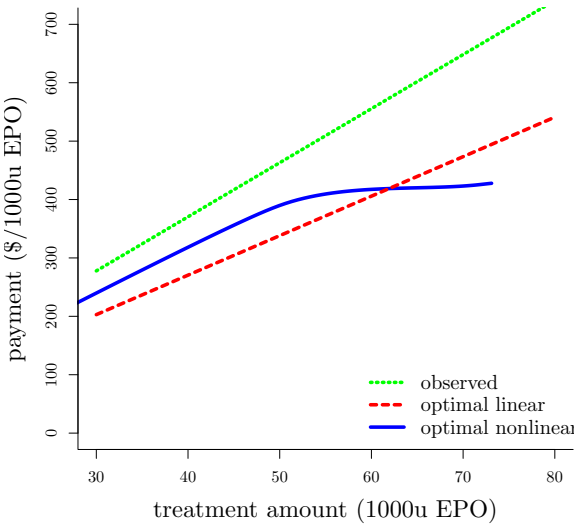
I Results for All Three Intervals of Baseline Hematocrit

This section presents the optimal contracts and outcomes under those contracts for the median baseline hematocrit and mean patient characteristics from each of the three intervals 30–33, 33–36, and 36–39, using the government's valuation of health α_g calibrated using information on the VSLY and the relationship between hematocrit levels and mortality risk.[75] Figure A3 shows the contracts; i.e., the treatment amounts and total payments. They have similar patterns to those in the median baseline hematocrit level, as discussed in the main text, with the optimal nonlinear below the observed contract and intersecting the optimal linear contract. Again, all contracts start at zero. The change in slope is more gradual at the lower baseline hematocrit, and it occurs at a higher dosage. On the other hand, the payment rate in the optimal linear contract is lower for 30-33, where patients have greater need for larger dosages. This indicates the importance of altruism in our environment: because physicians value the outcome of their patients, they can potentially be paid less to treat those who need treatment more.

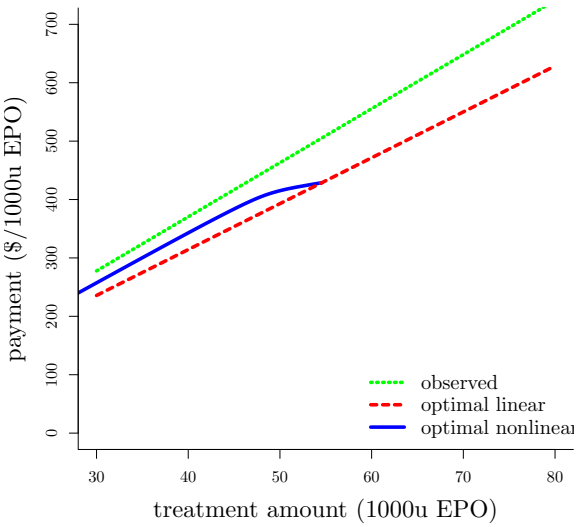
Table A2 summarizes the outcomes under these contracts. Mean dosages are lower under the optimal contracts, and accordingly so are mean payments. This reduction is beneficial to patients because under the observed contract around 80 percent of providers would give medically excessive dosages (i.e., negative marginal product) to patients with these baseline hematocrit levels. The optimal linear contract does not necessarily eliminate this obvious inefficiency: to patients

[75] The values for the median baseline hematocrit level are 32, 34.8, and 37.4 for the lower, middle, and upper intervals, respectively.

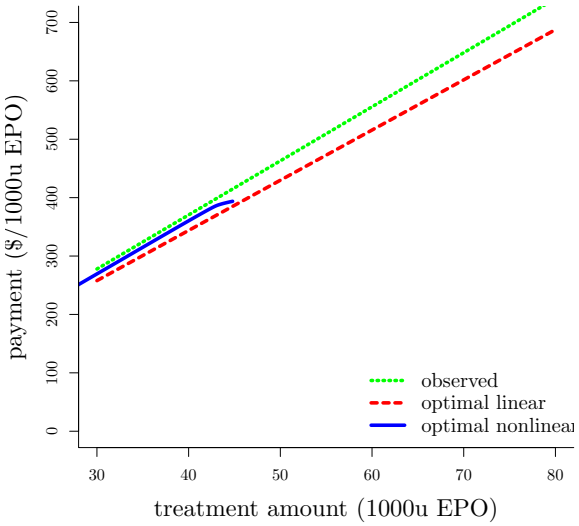
Figure A3: Optimal Nonlinear Contracts for median of each the three hematocrit intervals



(a) Payment as a function of the treatment amount, baseline HCT 30-33



(b) Payment as a function of the treatment amount, baseline HCT 33-36



(c) Payment as a function of the treatment amount, baseline HCT 36-39

Table A2: Summary of Outcomes under Optimal Contracts for median of each the three hematocrit intervals

Baseline HCT 30-33

	Mean Pmt	Mean Dosage	SD Dosage	Share above τ (%)
Observed	740	79.9	12.9	82
Optimal Linear	409	60.5	20.6	0
Optimal Nonlinear	387	54.6	12.9	0

Note: Table shows summary statistics of outcomes corresponding to contracts plotted in Figure A3a. Mean and SD of dosage are in 1,000 units/month. Treatment choice to get to hematocrit target: 75.9

Baseline HCT 33-36

	Mean Pmt	Mean Dosage	SD Dosage	Share above τ (%)
Observed	542	58.6	9.8	75
Optimal Linear	396	50.4	11.8	19
Optimal Nonlinear	393	47.1	7.2	0

Note: Table shows summary statistics of outcomes corresponding to contracts plotted in Figure A3b. Mean and SD of dosage are in 1,000 units/month. Treatment choice to get to hematocrit target: 56.3

Baseline HCT 36-39

	Mean Pmt	Mean Dosage	SD Dosage	Share above τ (%)
Observed	437	47.2	5.3	86
Optimal Linear	383	44.6	5.1	46
Optimal Nonlinear	384	42.9	2.5	0

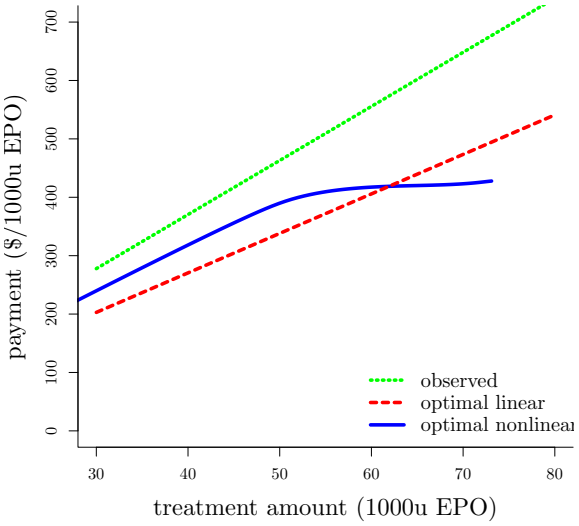
Note: Table shows summary statistics of outcomes corresponding to contracts plotted in Figure A3c. Mean and SD of dosage are in 1,000 units/month. Treatment choice to get to hematocrit target: 45.6

with the median hematocrit in the middle and upper intervals, respectively 19 and 46 percent of providers would give medically excessive dosages under it. This inefficiency does not occur with the optimal nonlinear contract because, as seen in Figure A5c, treatment amounts are below their full-information, first-best, values, all of which are strictly below what would be medically excessive (due to positive marginal costs of treatment and positive, finite, altruism).

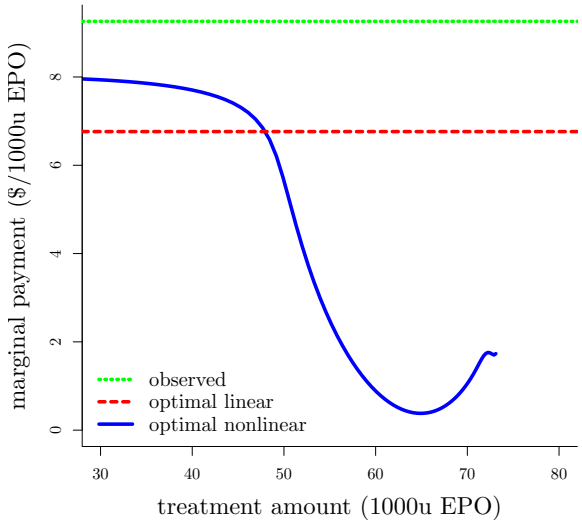
The variation in dosages, measured by the standard deviation, indicates the extent to which these contracts address the unobserved heterogeneity across providers (recall that patients have identical need for treatment in each example). The optimal nonlinear contract reduces the variation in dosages, compared to the observed contract, by 27% and 53% at the medium and high baseline hematocrit levels, respectively.[76] By contrast the optimal linear contract typically increases the variation, because it provides a constant marginal incentive, just like the observed contract, and a nontrivial share of types are (optimally) excluded, putting a non-negligible mass at zero. In contrast, under the full information scenario the standard deviations are substantially smaller (3.2, 1.3, and 0.4 thousand units per month for the low, middle, and upper intervals, respectively) but some variation remains, which reflects the variation in altruism and marginal costs.

[76]The optimal nonlinear contract does not reduce the standard deviation of dosages for the low baseline hematocrit interval (it excludes a nontrivial share of types). However, the optimal nonlinear contract reduces the standard deviation of *strictly positive dosages*, compared to the observed contract, by 16% in this interval.

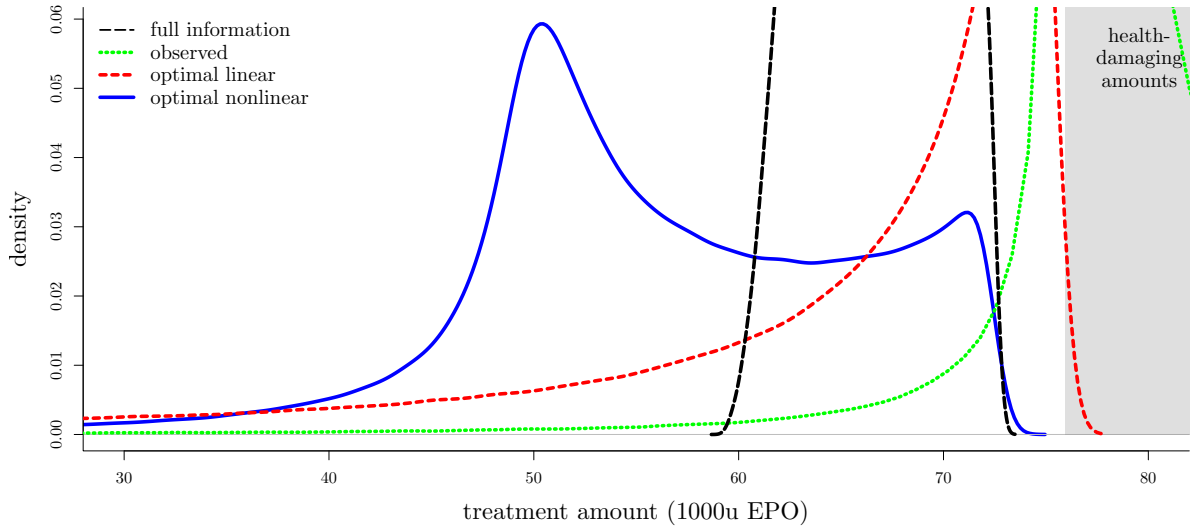
Figure A4: Optimal Nonlinear Contract Treatment Amounts and Payments, baseline HCT 30-33



(a) Payment as a function of the treatment amount

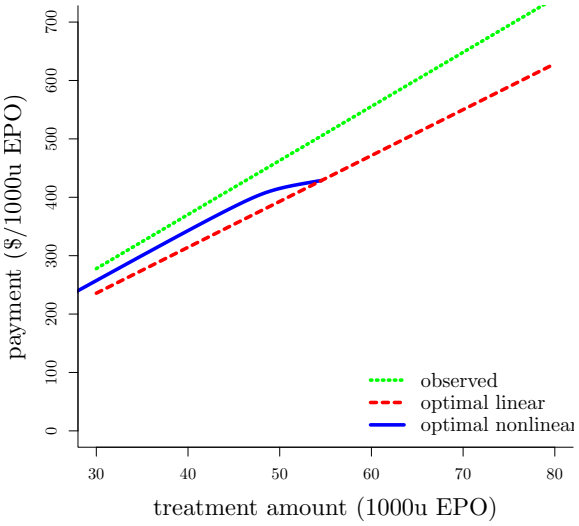


(b) Marginal payment as function of treatment amount

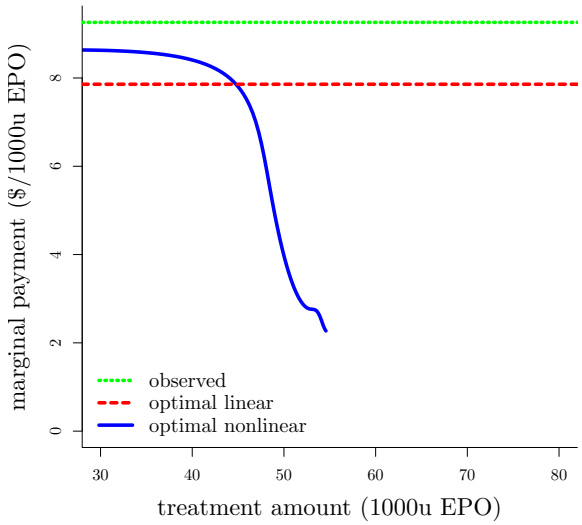


(c) Distribution of treatment amounts

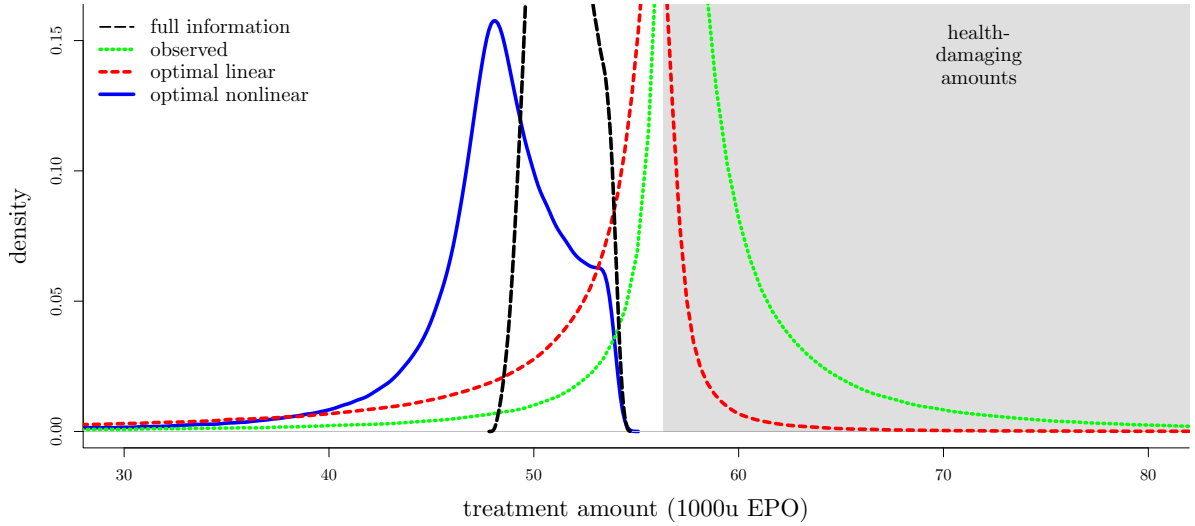
Figure A5: Optimal Nonlinear Contract Treatment Amounts and Payments, baseline HCT 33-36



(a) Payment as a function of the treatment amount

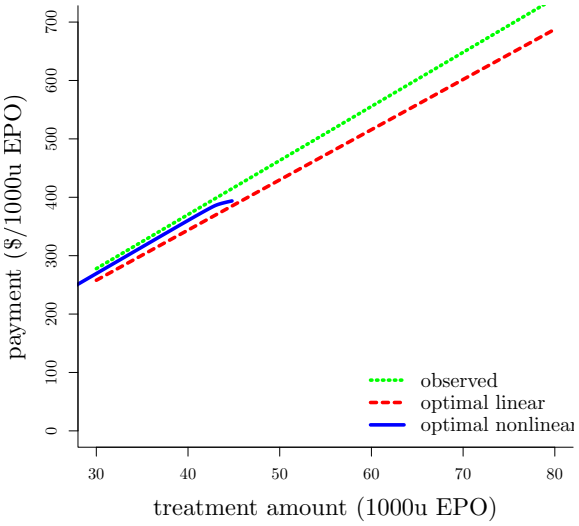


(b) Marginal payment as function of treatment amount

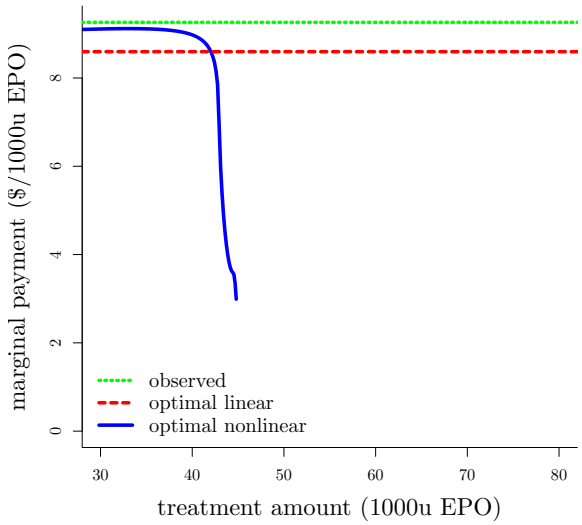


(c) Distribution of treatment amounts

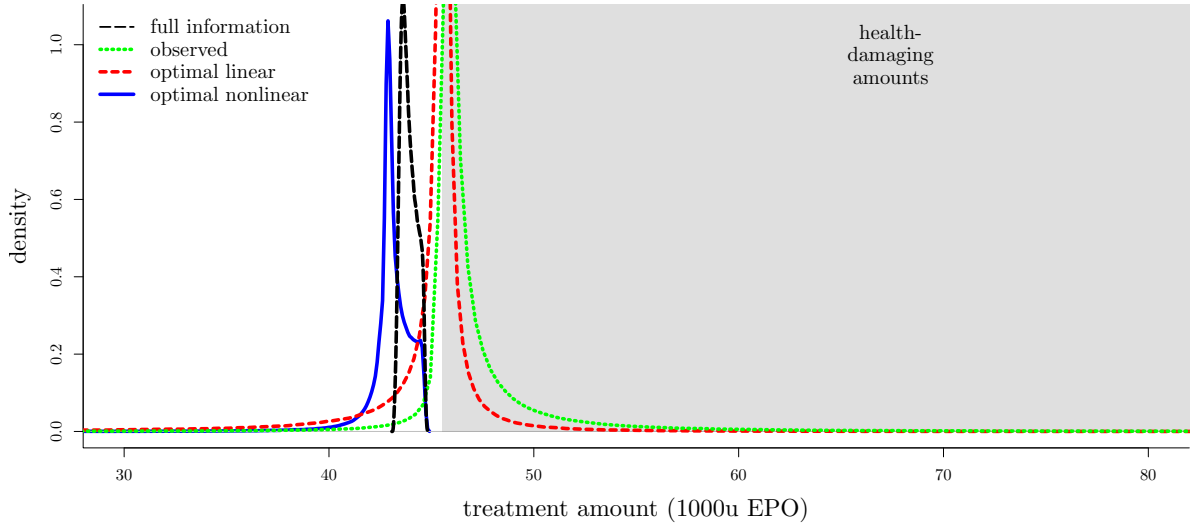
Figure A6: Optimal Nonlinear Contract Treatment Amounts and Payments, baseline HCT 36-39



(a) Payment as a function of the treatment amount



(b) Marginal payment as function of treatment amount



(c) Distribution of treatment amounts

Table A3: OLS and Fixed Effects Estimates of the Reduced Form

	OLS			Fixed Effects		
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	-9.22	-6.51	-4.00
	(0.24)	(0.15)	(0.13)	(0.19)	(0.13)	(0.12)
Reimb. rate	9.53	6.39	3.92	9.42	5.99	4.67
	(3.19)	(2.03)	(1.91)	(3.00)	(1.95)	(1.85)
Age in years	-0.41	-0.37	-0.26	-0.37	-0.33	-0.24
	(0.02)	(0.02)	(0.01)	(0.02)	(0.01)	(0.01)
Female sex	-0.89	1.54	2.89	-1.53	1.21	2.38
	(0.55)	(0.40)	(0.34)	(0.49)	(0.38)	(0.33)
Charlson=1	9.06	8.05	7.37	7.97	7.08	6.50
	(0.96)	(0.69)	(0.60)	(0.86)	(0.65)	(0.59)
Charlson=2	10.76	10.25	8.21	10.30	9.72	7.93
	(0.90)	(0.67)	(0.59)	(0.81)	(0.63)	(0.57)
Charlson=3	13.87	11.87	8.58	12.60	11.09	8.73
	(0.94)	(0.72)	(0.60)	(0.88)	(0.70)	(0.58)
Charlson=4	15.55	13.93	10.83	15.06	13.77	10.64
	(1.22)	(0.86)	(0.73)	(1.05)	(0.82)	(0.70)
Charlson=5	16.56	15.03	11.89	16.20	14.53	11.27
	(1.40)	(1.08)	(0.93)	(1.26)	(1.01)	(0.89)
Charlson=6	18.63	18.52	13.84	17.82	18.05	13.44
	(1.87)	(1.48)	(1.21)	(1.61)	(1.35)	(1.14)
Charlson=7	26.23	26.02	20.39	23.46	24.37	19.95
	(3.02)	(2.48)	(2.19)	(2.61)	(2.30)	(2.12)
Charlson=8	23.96	24.27	14.52	23.02	22.00	15.70
	(3.94)	(3.06)	(2.51)	(3.56)	(3.09)	(2.50)
Charlson=9	32.00	32.43	22.86	31.54	32.96	23.44
	(4.98)	(4.17)	(3.81)	(4.97)	(4.08)	(3.98)
Charlson=10	23.91	28.48	32.24	22.57	27.65	29.77
	(7.02)	(6.71)	(6.96)	(6.16)	(6.46)	(6.76)
Charlson=11	39.13	43.64	39.81	40.92	40.83	39.65
	(11.01)	(8.79)	(7.31)	(8.45)	(8.04)	(7.07)
Charlson=12	38.42	33.52	25.67	27.82	27.22	16.10
	(12.51)	(8.06)	(9.82)	(10.18)	(7.17)	(10.17)
Constant	392.18	294.37	192.16	388.18	299.35	207.52
	(7.93)	(5.29)	(4.98)	(6.18)	(4.60)	(4.58)
Observations	231,702	405,019	283,024	231,702	405,019	283,024
R-squared	0.029	0.028	0.021	0.029	0.027	0.021
RMSE	71.43	58.46	49.01	65.78	55.05	46.29

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