Optimal Contracting with Altruistic Agents:

A Structural Model of Medicare Reimbursements for Dialysis Drugs*

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

December 16, 2019

PRELIMINARY—PLEASE DO NOT CIRCULATE OR CITE

Abstract

We study physician agency and optimal payment policy in the context of an expensive medication (epoetin alfa) used with dialysis. Using Medicare claims data we estimate a structural model of treatment decisions, in which physicians differ in their altruism (how much they value their patients' health versus their own compensation) and their marginal costs of treatment, and this heterogeneity is unobservable to the government. We characterize the optimal nonlinear payment contract in the presence of this two-dimensional heterogeneity, and then use the recovered parameters of the model, in combination with our theoretical characterization of the contract, to derive and simulate optimal reimbursement schedules. Comparing outcomes under the optimal contract against those observed under the actual contract suggests that substantial improvements can be achieved within a fee-for-service framework.

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

1 Introduction

A central problem in economics is how to design contracts to incentivize agents to perform optimally in the presence of asymmetric information. While this problem has been extensively addressed theoretically, the empirical literature on incentive design is much less developed. In particular, there has been little work that empirically characterizes optimal contracts. The design of contracts is an especially important problem in health economics. Physicians are imperfect agents for patients and for health care payers (government or private insurers), and asymmetric information is pervasive. Payers have to decide how to pay doctors to deliver care, while recognizing that doctors possess relevant information that they do not.

Physician payment systems offer an excellent application for contract theory. First, the institutional facts fit a classic environment considered in contract theory—screening models. A public or private insurer (in our case, Medicare, which acts as a monopsonist) typically pays unobservably heterogeneous physicians using a single offered contract, where pay is based on an observable variable (usually the quantity of a service produced). Second, there are rich, detailed data in health care, due to the nature of payment and administrative oversight. This permits empirical analysis of econometric models derived from contract theory. Third, there are important consequences of physician payment systems. Health care is a very large and important sector of the economy. Moreover, physician incentives can lead to substantial impacts on patients' health. Consequently, the effects of better or worse incentives can have profound impacts on both expenditures and health.

In this paper we characterize optimal payment contracts for a physician who cares about patient health and about money. In the typical asymmetric information setting, there is one dimension of unobservable agent heterogeneity (e.g., valuations of a good differ across agents). In our application we permit agents to be heterogeneous in two dimensions: doctors differ in how altruistic they are towards patients (how much they care about patient health versus money) and in their marginal costs of treating patients. Intuitively, altruism reduces the net marginal cost of treatment, which reduces the distortions caused by asymmetric information. We use results from the literature on multidimensional screening models to characterize the optimal contract in the presence of two-dimensional agent heterogeneity.

We then specify an econometric model directly derived from the theoretical characterization, and use data from the US Medicare program to estimate the structural parameters of the model. We estimate via simple linear reduced forms that afford closed-form estimates of the structural parameters of the model, including the productivity of treatment, and the joint distribution of physician altruism and marginal costs. We are then able to fully characterize the unconstrained optimal nonlinear contract and the constrained optimal linear contract, and contrast those with the actual payment contract used by Medicare. Our results show that payment incentives matter, that heterogeneity in both altruism and marginal costs of treatment is important, and that moving to even a constrained linear optimal contract would improve treatment and save money.

Our application considers physician decisions about the provision of an expensive and controversial medication, epoetin alfa (or "EPO"), used to treat anemia in patients with end stage renal disease (ESRD). Medicare is the predominant payer for the treatment of ESRD in the United States, and the program spent more on EPO than any other single medication for several years in the early 2000s. We use Medicare claims data from 2008 and 2009 to estimate our model, a period when the payment system was stable and when there were no major informational shocks about EPO. Uniquely in our setting, a quantitative measure of a determinant of patient need is available in the claims records because providers were required to report a blood measurement in order to be reimbursed for EPO.

Notably, we consider an environment with two-dimensional heterogeneity. Addressing multidimensional unobserved heterogeneity in agent characteristics is a difficult problem in contract theory and related areas such as auctions, and general analytical solutions have been elusive (see Mirrlees (1986)).[1] We use the "demand profile" approach, introduced by Goldman et al. (1984) and Wilson (1993), to solve the model and derive the optimal unrestricted (i.e., nonlinear) contract. This approach was developed for monopolist price discrimination problems, but, as recently noted by Deneckere and Severinov (2015), is not always applicable in the presence of multidimensional heterogeneity. We suggest that the demand-profile approach may be particularly applicable in our context, where we study the supply of a good (i.e., drug dosage), and may therefore also be more applicable to certain types of screening problems more generally. Indeed, the conditions for this method to yield a correct solution—even in the presence of multidimensional heterogeneity—seem to naturally be satisfied in our environment. Briefly, in our setting, the intersection determining the agent's (here, the physician's) optimal quantity is between the (typically) downward-sloping

[1] Our environment also has similarities those that studied in the literature on optimal taxation in hidden information environments, which was initiated by Mirrlees (1971). The empirical literature on optimal taxation has focused on one-dimensional heterogeneity, due to the well-understood complications arising when agents have types described by more than one dimension (Maskin et al. (1987), McAfee and McMillan (1988), Armstrong (1996)). Much of the empirical literature on optimal taxation adopts a "sufficient statistics" approach, which affords a relatively agnostic way of computing the welfare effects of infinitesimal changes in the contract (see, e.g., Saez (2001)), or quantitatively examines the effects of a restricted class of mechanisms, without theoretically characterizing the optimal contract (see, e.g., Blundell and Shephard (2011)). Our paper offers a tractable way to fully and analytically characterize the unconstrained optimal contract. See Judd and Su (2006) for an example computing optimal tax policy for multidimensional types using numerical methods.

marginal transfer schedule (i.e., the marginal reimbursement for treatment) and the upward-sloping dosage supply curve.[2]

We find that reimbursement rates in the optimal linear contract are substantially lower than the actual rates used by Medicare in 2008 and 2009. The optimal nonlinear contract improves outcomes further, notably by reducing seemingly unjustified variation in treatment intensities while also decreasing total expenditures. A simulation for patients with anemia at the median level of severity finds that the optimal nonlinear contract would increase Medicare's objective by an amount equal to $1440 per patient per year, over the actual payment contract used at the time. It does this by reducing the standard deviation of dosages (for a patient of the same need) by 22 percent while the mean payment decreases by 29 percent.

In what follows, we first review the related literature (Section 1.1), then provide institutional background (Section 2). In Section 3, we introduce the model and then derive the optimal nonlinear contract, which results in the second-best allocation in the presence of asymmetric information. Section 4 contains a description of the data we use for our empirical analysis, and Section 5 describes the empirical implementation, including specification, identification, and estimation. Quantitative results comparing the optimal contracts with the observed contract are presented in Section 6.

1.1 Related literature

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

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

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

[2] In contrast, in typical price discrimination problems, the marginal price schedule and agent demand curve will both typically be downward-sloping in quantity (which, in our setting, would be treatment dosage). This would potentially lead to multiple intersections of these curves, rendering the demand profile approach inapplicable.

literature doing this for insurance contracts. Paarsch and Shearer (2000) characterize the optimal linear contract in a hidden action environment. Gayle and Miller (2009) also study hidden action models, quantifying the welfare loss from moral hazard. In contrast, we study a screening, or hidden information, model and flexibly characterize the optimal wage schedule. Screening models, with their focus on unobserved heterogeneity, are clearly policy relevant.

There is also a literature on optimal regulation, which considers screening models in institutional contexts that differ from ours in important ways. Wolak (1994) develops and estimates a model in which a principal seeks to regulate public utilities of (potentially) hidden types. Data limitations, including a lack of variation in regulatory regime, mean the distribution of types cannot be estimated without imposing optimality of the observed contract. Gagnepain and Ivaldi (2002), study a similar environment, but exploit variation in the regulatory regime to estimate a parametric distribution of types without having to assume optimality of the observed contract. This allows them to test whether the observed contract is optimal. Abito (2018) extends this approach to study optimal pollution regulation. As in the latter two articles, our setting and data allow us to estimate structural parameters, including agent types, without imposing optimality of the observed contract. In principle, we could allow for a fully flexible type distribution, due to variation in the observed regime (i.e., reimbursement contract) and a large number of repeated measures of physicians, as each physician chooses a treatment choice for each patient they see under a variety of observed reimbursement rates. That being said, we adopt a parametric approach to the type distribution for efficiency and computational tractability. In contrast to the latter two articles, we allow for two-dimensional heterogeneity, which substantially complicates characterization of the optimal contract.

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

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

There is also an empirical literature studying how physicians respond to incentives and other changes in the environment. Gaynor and Pauly (1990) is an early paper showing that physicians respond strongly to financial incentives. Chandra et al. (2012) review the literature studying determinants of physician treatment choices. One determinant they focus on is physician altruism. Gaynor et al. (2004) specify and estimate a structural model of physician treatment choice, where physicians are partially altruistic. Godager and Wiesen (2013) use data obtained from a laboratory experiment to document the existence of physician altruism, which they find to be heterogeneous. Clemens and Gottlieb (2014) examine the impact of financial incentives in Medicare payment for physicians and provide evidence of effects on supply, technology adoption, and patient outcomes. Einav et al. (2018) find a strong effect of dynamic incentives in payments to long-term care hospitals on the timing of discharges, and they simulate outcomes under a variety of alternative payment schedules.

Our model and empirical analysis relate to many components of the above literature. The basic model of physician utility is very similar to that in Gaynor et al. (2004), but allows for cost heterogeneity as well as (heterogeneous) altruism. De Fraja (2000) and Jack (2005), noted above, address heterogeneity across physicians, although there are various distinctions between their models and ours.[3] 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.[4] Last, in contrast to the existing empirical literature, we not only estimate a model of physician treatment choices and recover physician altruism, we also empirically characterize the optimal contract and compare it to the actual contracts used in this context.

2 Institutional Background

ESRD, or kidney failure, is a chronic and life-threatening condition that affects over half a million individuals in the United States at a given point in time. Since 1973, the Medicare program has provided universal coverage for the treatment of ESRD, regardless of age. In 2009, at the end of our study period, Medicare spent $28 billion on health care for individuals

[3] 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).

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

with ESRD, and of that amount, \$1.74 billion went to payments for EPO.[5] The drug is used to treat anemia (a lack of red blood cells), which often accompanies chronic kidney disease.[6] 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.[7] The level is commonly measured in terms of the *hematocrit*, which is the volume percentage of red blood cells in the blood. For dialysis patients, 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 typically consists of one medical director (a physician) and multiple nurses and medical technicians, and payments are primarily made to the dialysis centers, not the physician(s).[8]

Medicare's payment policy for EPO was debated throughout the 1990s and 2000s, largely because of concerns that the reimbursement rates were too generous and encouraged overprovision. While dialysis itself was reimbursed with a prospective payment system known as the "composite rate," which paid a fixed amount of roughly \$130 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 (including the years we use to estimate the model), 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).[9] Additionally, Medicare required dialysis centers to report the hematocrit levels of their patients, and the reimbursement rate was cut in half for patients with relatively high levels (over 39%) for three consecutive months.[10]

Important safety concerns about EPO emerged by the mid 2000s. A major clinical trial found that patients who were given more EPO to achieve a higher target level of hematocrit suffered a higher risk of serious cardiovascular events and death (Singh et al., 2006). This

[5] USRDS *2016 Annual Data Report*, volume 2, chapter 11; available at `https://ppp.usrds.org/2016/viep/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.

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

[7] 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).

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

[9] 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.

[10] A Medicare Coverage Document describing this monitoring policy is available at `https://ppp.cms.gov/medicare-coverage-database/details/medicare-coverage-document-details.aspx?MCDId=11`.

study was published in November 2006, and strong warnings ("black box warnings") were added to the drug's labels in 2007. As a result of 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% (NKF, 2007), and the FDA maintained its range of hematocrit levels for which EPO was approved in ESRD patients at 30–36%.

The dialysis industry was also undergoing rapid consolidation during this time period. By 2009, two large chains treated a combined 60 percent of dialysis patients in the US.[11] However, while the facilities within these chains had common ownership, their medical staff maintained a good degree of independence. When facilities were acquired, their staff remained largely intact, and medical directors retained clinical authority as is the professional norm in medicine. In the empirical analysis, we treat each dialysis center as an independent decision maker, which allows for this heterogeneity within chains and fits more naturally with our theoretical framework.[12]

Our data come from Medicare insurance claims for dialysis care, which are filed by and paid to the individual dialysis centers, typically once per month for each patient. The claims include separate lines for each administration of EPO over the month. To be reimbursed for EPO, the dialysis centers are required to report a hematocrit level taken just prior to the monthly billing cycle. This lab result is included in the claims data—which provides a quantitative measure of a determinant of patient need, in this case the severity of the anemia. Therefore, this key component of our model is observable, which is highly unusual for a health application.

3 Model

Our model uses a static principal-agent framework, describing a one-time interaction between the principal and an agent. The government (the principal) pays a physician (the agent) to treat a patient.[13] 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.

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

[12] Direct evidence of variation in certain costs both within and between chains can be found in annual facility-level cost reports submitted to Medicare. See Section 4 for a description of those data.

[13] Section 5 discusses how this model extends to multiple physicians treating multiple patients.

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.[14] 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 $b_1(a; b_0) > \tau$. We refer to a treatment amount resulting in $b_1(a; b_0) > \tau$ (i.e., with negative marginal product) as *medically excessive*.[15] In the empirical implementation, we also 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 physician's degree of altruism, α, gives the marginal rate of substitution between patient health and own compensation in the physician's utility. The physician also has a constant marginal cost of treatment, z, which in our application reflects the costs of acquiring and administering EPO. These two attributes are unobserved by the government, and we refer to (α, z) as the physician's *type*. Heterogeneity in altruism may capture differences between physician's preferences. The treatment costs include acquisition and administration costs, both of which may naturally 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 reimbursement policy, which specifies the payment that would be made to the physician based on the treatment amount and baseline hematocrit reported on

[14] As noted in Section 2, the medication is administered intravenously, while the patient is undergoing dialysis. This is in contrast to administration of treatment in many other typical medical applications, e.g., having a patient with heart disease exercise and watch their diet.

[15] 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 clear shortly, these concepts are related because a medically excessive amount will imply a suboptimally high amount.

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.[16] In the empirical implementation, the contracts also depend on the observed patient characteristics that affect τ. The reimbursement 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 reimbursement 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, $\alpha h(a;b_0)$, minus the cost of treatment, za, plus the payment from the government, $P(a;b_0)$, as follows:

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

Thus the physician has quasilinear preferences, a standard assumption in screening models (Rochet and Stole, 2003). The physician's reservation utility is $\underline{u}$.[17]

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.[18] 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

[16] Allowing the contract to condition on b_0 is not unrealistically complicated; in fact, the Medicare payment policy specified a reduced reimbursement rate for claims where the prior hematocrit level was above 39% for three consecutive months.

[17] 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 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.

[18] As is also standard in screening models, the government's objective does not include the physician'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).

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 actions that would be taken by different types, conditional on b_0. The government's overall objective function adds these conditional expectations across b_0, because it is observable, therefore, contractible.

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 reimbursement 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).[19] The government sets the payment contract for each b_0, knowing how each physician type would respond. Thus, for each b_0, the government's problem is to maximize the expected value of (2), subject to the physician's incentive compatibility and voluntary participation constraints, which must hold for each type.

Now we turn to the solution of the model. First we characterize the first-best allocation that would occur under full information. Then we solve the model under asymmetric information via backward induction, starting with the physician's behavior, which describes how altruism affects the agent's choices, 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 health, so b_0 is suppressed from the health function h and the payment contract P. Also here we focus on interior solutions to clarify the exposition. When solving the model numerically for the empirical analysis, we allow for corner solutions—i.e., where some physician types provide zero amounts (see Appendix D). This is relatively straightforward, and 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

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

condition is

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

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. 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. (Note that the efficient allocation would never have medically excessive amounts, where $h' < 0$.)

3.2 Physician Behavior

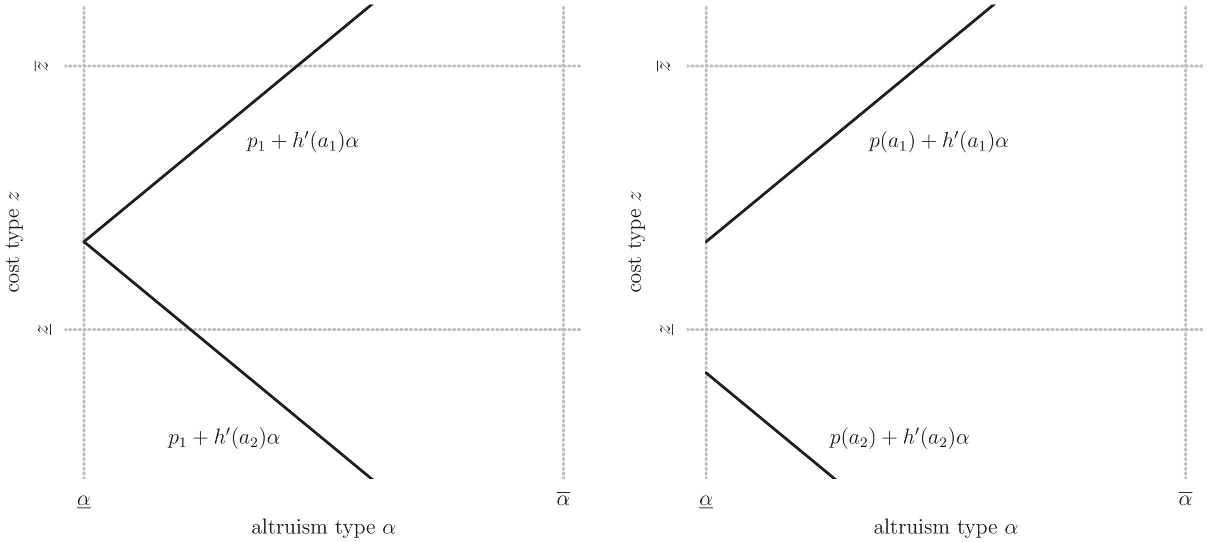
Next we characterize the physician's behavior under an arbitrary payment contract P. The interior first-order condition of the physician's utility function (1) is

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

The function $nc(a; \alpha, z) \equiv z - \alpha h'(a)$ is the net marginal cost to a physician of type (α, z) for providing amount a, and $p(a) \equiv \frac{\partial P(a)}{\partial a}$ is the marginal payment at amount a. The physician will choose 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 marginal payment curve intersects the net marginal cost curve once, from above (discussed in Section 3.3). Then, if $h'(a^*) > 0$, as we show will be the case under an 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., reimbursement 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,



(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 reimbursement rate of p_1, and an example nonlinear contract (right), which has a variable reimbursement rate, described by the function p, where $p_1 = p(a_1) > p(a_2)$. The treatment amounts are $a_1 < \frac{\tau - b_0}{\delta} < a_2$, i.e., $h'(a_1) > 0$ and $h'(a_2) < 0$.

under an arbitrary linear contract and an arbitrary nonlinear contract. With either contract, the set of types that would provide the treatment amount a^* is a line in the support of (α, z) (note that (4) rearranges to $z = p(a^*) + h'(a^*)\alpha$). The figure plots these isoquants for treatment amounts a_1 and a_2, where $a_1 < \frac{\tau - b_0}{\delta} < a_2$. The slope of the isoquants is $h'(a^*)$, so downward slopes correspond to medically excessive amounts. The most 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 changes with the nonlinear contract (panel b) because the marginal payment can vary ($p(a_1) > p(a_2)$).[20] This suggests the difficulty of designing a linear contract that induces desired treatment amounts. For example, a linear contract would have difficulty avoiding medically excessive amounts because the reimbursement rate (p_1) would have to be below the marginal cost of the lowest-cost provider ($\underline{z}$) to avoid downward slopes, which would likely exclude many higher-cost types. Nonlinear contracts can avoid this tension, as illustrated by the isoquant for a_2 in the right panel, which lies entirely outside the type space.

[20] We set $\underline{\alpha} = 0$ only for this illustration.

3.3 Optimal Contract

We now present our approach to solve the government's problem and thereby characterize the optimal nonlinear contract. Because the agent's 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 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 take the same action.

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 the marginal payments for each possible quantity. The power of this approach is that it projects a multidimensional distribution of agent types into a one-dimensional distribution of quantities, and the solution for each quantity can be found separately.

The government's optimization problem is accordingly to set the marginal payment for each treatment amount to maximize the marginal valuation of that amount, multiplied by the probability the amount would 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), 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 would provide a treatment amount of at least a, given the payment contract. In that case, the government would receive the 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 actions of excluded types, to $\overline{a}^{*\mathrm{FI}} \equiv a^{*\mathrm{FI}}(\overline{\alpha}, \underline{z})$, the amount that would be provided by the "best" type (lowest cost, highest altruism) under full information.[21] As is standard, the voluntary participation constraint must be satisfied for any physician type. This rules out a "forcing contract" that would only reimburse for the maximum amount

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

provided under full information, for example.[22]

Assuming that the net marginal cost curves for each agent type intersect the marginal payment curve at most once, from below, which is an important regularity condition (discussed more below), then 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, $p(a)$. These types respectively provide the amounts a_1^* and a_2^*, where the curves intersect. 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 $z - \alpha h'(a_1^*) < z_1 - \alpha_1 h'(a_1^*)$) would provide more than a_1^* because $nc(a; \alpha, z)$ would intersect $p(a)$ at a larger amount.

Figure 2 suggests that this approach may be more broadly useful for solving principal-agent 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). 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).

[22] 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.

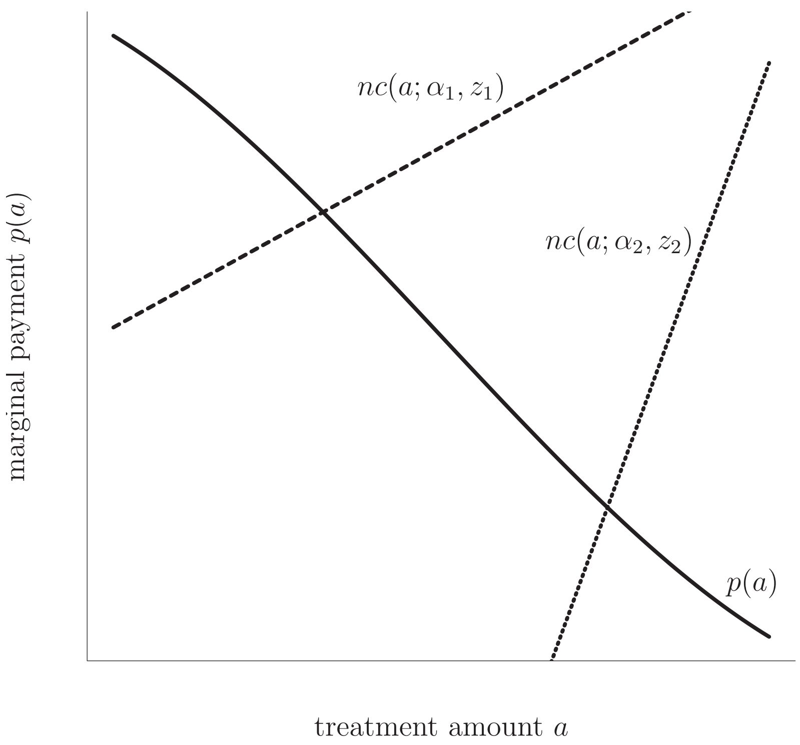


Figure 2: Example marginal payment contract and physician supply curves.

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., $z_1 = z_2$.

The government's problem can then be solved separately for each treatment amount, using the distribution of treatment amounts given by (6).[23] In addition to the single-intersection condition, this also relies on the quasilinearity of the agent's preferences, or no income effects, a standard 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 supply is independent of the payments for other amounts. The separate problems for each treatment amount are

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

for each $a \in A$. Splitting the principal's objective into independent problems for each

[23] 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.

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, treating $p(a)$ as a parameter:[24]

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

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, $\frac{\partial S(p^*(a),a)}{\partial p(a)}$, times 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.

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

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

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 "inverse elasticity rule" of 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 because the government is a monopsonist and, instead of a variable marginal cost of production c', the government has a nonconstant marginal valuation of treatment, $\alpha_g h'$. Intuitively, the government's "margin" is lower (i.e., it extracts less surplus) where supply is more responsive to price changes (i.e., the elasticity of supply is larger).

Finally, we examine the normative properties of the second-best allocation. 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

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

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 downward-distorted from the first-best for all but the highest-action type, $(\overline{\alpha}, \underline{z})$.[25] Equivalently, fewer types choose $a < \overline{a}^{*\mathrm{FI}}$ 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 $S(p^*(a), a)$ is strictly positive for all but the maximum treatment amount, which is provided by the highest-action type (which has a measure of zero). Also the partial derivative $\frac{\partial S(p^*(a),a)}{\partial p(a)}$ is positive because the probability in (6) increases with $p(a)$. Hence the RHS of (10) is larger than the RHS of (3) for all but the highest-action type. Therefore, because h is strictly concave, the second-best treatment amount is 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.[26]

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 monthly, with 11.1 million claim lines for EPO or related medications.[27] Almost 90% of the claims bill for at least one injection of these drugs (1.25 million claims). All claims with an injection include a baseline hematocrit level from the previous month (or a related hemoglobin level), but claims without an injection do not

[25] 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.

[26] This is difficult to show in a general setting (Wilson, 1993). However, with one dimension of heterogeneity (say, unobserved marginal cost types z that are distributed according to F_z, with associated density f_z), the distortion is captured by the inverse of the reverse hazard function of the type (in this example, $F_z(z)/f_z(z)$), which is monotonic for many commonly used distributions (Bagnoli and Bergstrom, 2005).

[27] Epoetin alfa constitutes 97.6% of the claim lines for this class of medication in our sample. The alternative drug was darbepoetin alfa. We restrict to epoetin alfa because dosages and reimbursements differ between the two drugs.

report a red blood cell level. As a consequence, we exclude claims without any injections of EPO. 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 or the baseline hematocrit is above its 99th percentile or below its 1st percentile. The final sample has 1.1 million claims, for 76,985 unique patients from 5,150 unique providers.

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 as the providers, not individual physicians, because within each center the doctor(s), nurses, and technicians jointly provide treatment to their patients. 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.[28] The reimbursement 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.[29] Additionally, our empirical model includes basic patient demographics and a measure of comorbidities, because these may affect the target level of hematocrit. We use age, sex, and the Charlson Comorbidity Index, collected into a vector x_{jt}.[30]

Table 1 provides summary statistics on the main variables. The average monthly dosage of EPO is 67.0 thousand units, with a relatively large standard deviation of 66.4 thousand units. The average baseline hematocrit (the volume percentage of red blood cells in the blood) is 34.5 percent, with a smaller standard deviation of 3.4 percent. The Charlson index, which is a count of 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.

The national reimbursement rate for EPO, which varies by quarter, ranged from a low of \$8.96 in 2008Q1 to a high of \$9.62 in 2009Q3 (Addendum 2 in the table), and has a mean of \$9.26 across the observations in our sample. Table 1 also shows the distribution of dialysis centers' acquisition costs for the drug from a separate source, the publicly available Renal

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

[29] See `https://ppp.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 0.92.

[30] 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).

Table 1: Summary Statistics

			Percentiles				
Variable	Mean	SD	10th	25th	50th	75th	90th
Monthly EPO dosage (1,000u)	67.0	66.4	8.8	20.0	45.0	90.0	156.2
Prior hematocrit level (%)	34.5	3.4	30	32.4	34.8	36.9	38.7
Charlson Comorbidity Index (0-16)	1.4	2.0	0	0	0	3	4
Reimbursement rate ($/1000u)	9.26	0.24	8.96	9.07	9.20	9.58	9.62
Addendum 1 – Percentiles of EPO acquisition costs from annual cost reports for 2008:							
Acquisition cost ($/1000u)			7.13	7.23	7.53	8.15	9.11
Addendum 2 – Medicare reimbursement rate for EPO in each quarter:							
Reimb. rate ($/1000u) 8.96	9.07	9.07	9.10	9.20	9.40	9.62	9.58
(2008Q1)	(Q2)	(Q3)	(Q4)	(2009Q1)	(Q2)	(Q3)	(Q4)

Notes: The EPO dosage and hematocrit level, and Charlson index come from Medicare outpatient claims data. The reimbursement rate comes from quarterly Medicare Part B ASP Drug Pricing Files for 2008 and 2009. The distribution of EPO acquisition costs shown in Addendum 1 is computed from Renal Dialysis Facilities Cost Report Data for 2008. We do not present the mean or standard deviation because extreme outliers in the cost report data make those statistics unreliable, compared to the percentiles.

Dialysis Facilities Cost Report Data from the Centers for Medicare & Medicaid Services.[31] The percentiles indicate potentially important differences across dialysis centers in their acquisition costs, even though the drug was produced by a single manufacturer. (The mean and standard deviation are not presented, due to extreme outliers in the data.) Additionally, as we discuss in Section 5.2, there are nontrivial costs of administering EPO, which are also likely to vary across dialysis centers, but this cannot be constructed in a straightforward manner from the cost report data.

[31] See `https://ppp.cms.gov/Research-Statistics-Data-and-Systems/Dopnloadable-Public-Use-Files/Cost-Reports/RenalFacility.html`. 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.

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.[32] We can therefore use our earlier results to characterize optimal contracts in the empirical model, and accordingly the empirical specification is developed for a single treatment, provided to one patient at a particular point in time. Below, we first develop the empirical specification, then discuss identification and explain the particular approach used for estimation, and finally present the estimates of the reduced-form and structural parameters.

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. However this specification is not crucial for our 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 is observed in our data. Furthermore, the optimal nonlinear contract could be constructed for any function h that satisfies the assumptions given in Section 3.

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.[33]

[32] The static framework can be applied to multiple time periods if there are no dynamic effects of EPO (discussed in Section 2), and if the government does not use treatment histories in setting reimbursement rates. This has always been the case when patient hematocrit levels are within the recommended range. The Medicare payment policy during our analysis period paid reduced rates for EPO given to patients who had relatively high hematocrit levels for three consecutive months (over 39%). Our main analysis restricts to observations below this threshold.

[33] Specifically, we assume that H is sufficiently large such that $h(0; b_0) > 0$. This implies that the orderings

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, as follows:

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

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., $a^* = 0$) in the construction of the optimal contracts and in the simulations presented in Section 6, meaning that some physician types may be excluded.

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 reimbursement 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 reimbursement 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 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 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 essentially 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,[34] and similarly the flexibility in τ means that

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 is a standard assumption in screening models because it implies that only the participation constraint of the lowest-action type will be binding, which simplifies characterization of the optimal nonlinear contract.

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

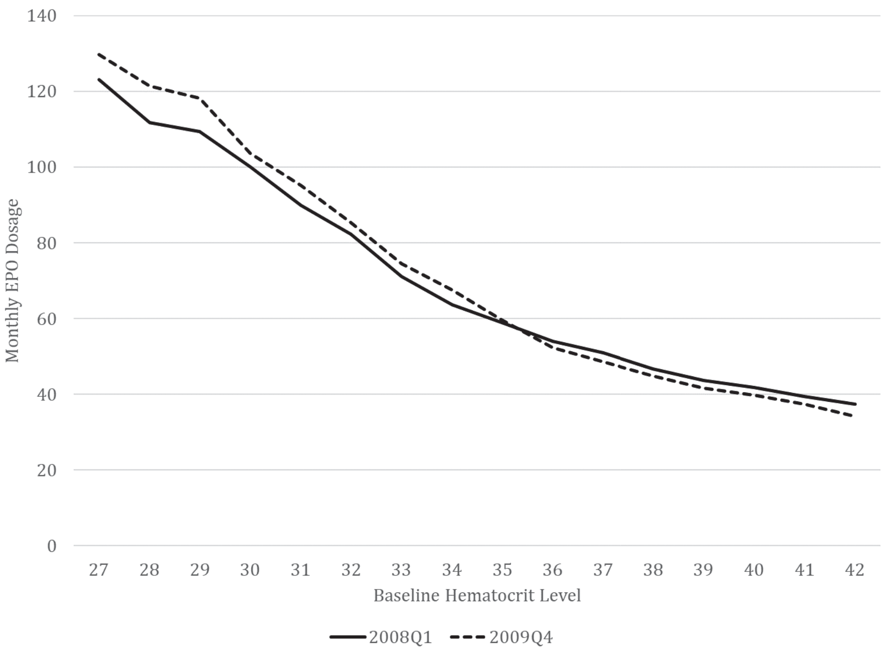


Figure 3: Mean monthly dosages of EPO in relation to baseline level of hematocrit.

the target level of hematocrit may depend on the 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, but not for marginal costs to change with severity—and our empirical results are in fact consistent with this (see Section 5.3).

Finally, 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 $\tau_k' x$. Also, to allow for unexplained variation from the econometrician's perspective, we add an independent, mean-zero shock, η. Additionally, as will be clear below, it is useful to decompose the 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 as follows to yield reduced-form parameters and disturbances:

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

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 primarily explain the approach we implement to identify and estimate the empirical model. However, we also note the nonparametric identification of the type distribution, F (discussed below). This and the semiparametric identification of h (noted earlier) are discussed further in Appendix F.

The structural parameters to be recovered in our empirical model are the scalars δ_k and vectors τ_k, $k = 1 \ldots K$, and the joint distributions $F_k(\alpha, z)$, $k = 1 \ldots K$. 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.[35] 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 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 log-normal distribution with strictly positive support. Then in each hematocrit interval k, there are four unknown parameters, $\mu_\alpha(k)$, $\sigma_\alpha^2(k)$, $\sigma_{\alpha z}(k)$, and $\sigma_z^2(k)$, since $\mu_z(k) = \mu_z$ is treated as known from our external information on costs.[36] Using Stein's lemma (Stein, 1981) and properties of the log-normal distribution, these parameters are identified by, and can be recovered analytically from, the first and 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

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

[36] As is standard, these parameters apply to the normal distribution of $\ln \alpha$ and z.

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).

This parametric approach is tractable and efficient, but it is not necessary because F_k is nonparametrically identified under the assumption that the shocks η_{ijtk} (equivalently, ϵ^k_{ijt}) 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), which is theoretically possible in our application given the large numbers of observations per dialysis center. Consistent estimates of β^k_{2i} and ν^k_i 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 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 $(\beta^k_{2i}, \nu^k_i, \epsilon^k_{ijt})$ are uncorrelated with the observables $(b_{0jt}, p_{1t}, x_{jt})$.[37] One possible concern is selectivity of patients to providers, which could make b_0 and x correlated with the provider-level unobservables, β^k_2 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 A2), which suggests that the provider-level unobservables (β^k_{2i} and ν^k_i) are not noticeably correlated with the patient characteristics (b_{0jt} and x_{jt}). 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^k_{ijt} 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 (e.g., chemotherapy providers).

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

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

acquisition (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.[38] The cost of administering EPO is also non-trivial. Several time-and-motion studies have been published to assess the administration cost, 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.[39] Adding this to the acquisition cost, we set the value of μ_z equal to \$8.58 per 1,000 units.

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. We focus on the intervals from 30 to 39, based on treatment guidelines and payment policies in place at the time. Specifically, the FDA had approved EPO for use in patients with hematocrit between 30 and 36, and Medicare reduced the reimbursement rate for EPO provided to patients with hematocrit above 39. Most of the observations in our sample fall into this range (over 80%). The 3-point intervals are thus convenient because they divide the range from 30 to 39 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 reimbursement rate ($\bar{\beta}_2^k$) are shown in Table 2, and the full set of estimates for all variables

[38] 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.

[39] 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.

Table 2: Reduced-Form Coefficient Estimates

Variable	Interval of Baseline Hematocrit		
(Coefficient)	> 30 to 33,	> 33 to 36,	> 36 to 39
Baseline hematocrit	-9.60	-6.54	-3.67
(β_1^k)	(0.24)	(0.15)	(0.13)
Reimbursement rate	11.16	6.99	3.72
($\bar{\beta}_2^k$)	(3.30)	(2.11)	(1.98)
Obs. in interval	216,390	379,527	267,245

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.

are provided in Appendix Table A1.[40] To interpret the coefficients, for example in the middle interval, a patient with one unit higher baseline hematocrit (say 35 vs. 34) receives 6,540 less units of EPO per month on average. Also in that interval, a one dollar increase in the reimbursement rate (per 1,000 units) would induce providers to increase dosages by 6,990 units per month on average. In terms of the model, providers are on average finitely altruistic. Across the three intervals the magnitude of $\hat{\beta}_1$ is decreasing, which matches the decreasing magnitude of the slopes seen in Figure 3.

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 0.153 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.[41] 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

[40] 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.

[41] 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.

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	0.104	0.153	0.273
	(0.003)	(0.004)	(0.010)
Mean implied hematocrit target			
$\tau_k' \bar{x}$	40.0	43.4	50.0
	(0.3)	(0.3)	(0.6)
Distribution of altruism and marginal cost types			
$\mu_\alpha(k)$	3.47	3.32	3.21
	(0.84)	(0.87)	(1.43)
$\sigma_\alpha^2(k)$	2.72	3.01	3.84
	(0.87)	(0.90)	(1.45)
$\sigma_{\alpha z}(k)$	-0.353	-0.385	-0.379
	(0.011)	(0.011)	(0.012)
$\sigma_z^2(k)$	0.325	0.301	0.290
	(0.078)	(0.060)	(0.068)
Obs.	216,390	379,527	267,245

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.

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.[42]

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.

[42] 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 (NKF, 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, e.g., 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.

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 of α in the middle interval is 27.7. This implies that if the reimbursement rate were one dollar above the marginal cost for some physician with the median degree of altruism, that physician would provide a medically excessive dosage which raises the patient's hematocrit 0.236 percentage points beyond the target level, $\tau_k' x_{jt}$.[43] The marginal cost, z, is denominated in dollars, so the estimates of $\sigma_z^2(k)$ imply standard deviations of marginal costs equal to \$0.57, \$0.55, and \$0.54 in the three intervals. These estimates are quite close, economically and statistically, which suggests that marginal costs may not depend on the patient's baseline hematocrit. (Also, 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 median and only allow z to vary, the standard deviation of simulated dosages is reduced by 90%.[44]

6 Quantitative Results: Optimal Contracts

This section presents our main 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. The results indicate the potential value of replacing constant payment rates in traditional fee-for-service systems with variable marginal payment rates.

Two last steps are required to compute the optimal contracts.[45] 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 $\delta_k^{-2}\alpha_k^{-1}$).[46] Additionally, 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 our parameter values it is not possible to rationalize the observed payment

[43] From (12), $\delta a^* + b_0 - \tau = (p_1 - z)(\alpha\delta)^{-1}$ and so we use $\$1 \times (27.7 \times 0.153)^{-1} = 0.236$.

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

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

[46] 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.

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.[47]

Here we present the results for a particular value of 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.[48] (Recall that contracts may be defined for each b_0 and x.) Contracts for other 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.

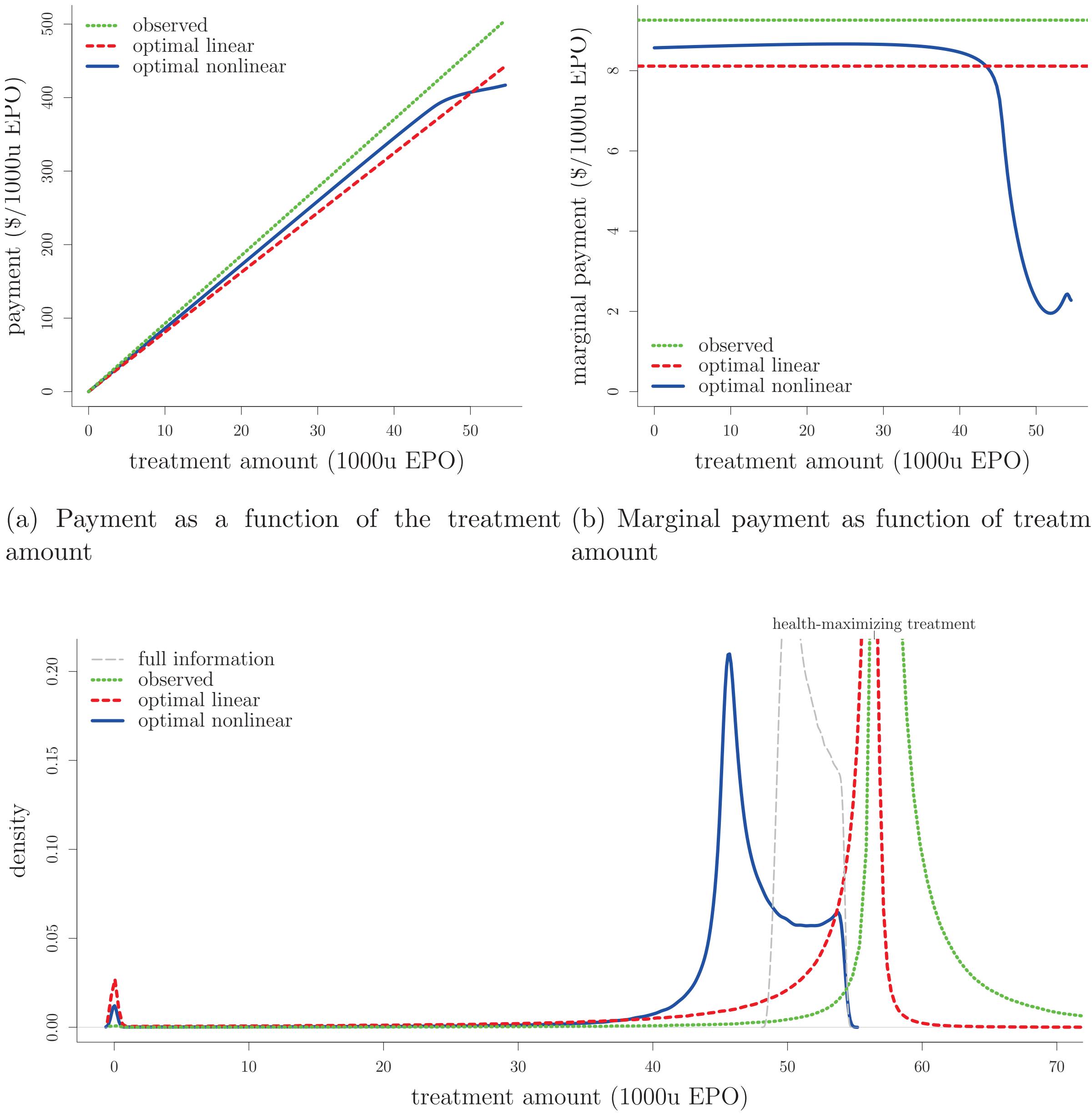
Figure 4 plots the total payments (panel a), the marginal payments (panel b), and the distributions of treatment amounts (panel c), with the nonlinear contract in blue solid lines, the linear contract in red dashed lines, and the observed contract in green dotted lines. All three contracts pay \$0 for 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.[49] For positive treatment amounts, the total payments 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 \$385.89, the linear contract would pay \$365.13, and the observed contract would pay \$416.70, per month. The marginal payment 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.[50]

[47] 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 71, and the results are qualitatively and quantitatively similar to those presented here (available upon request).

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

[49] The reservation utility $\underline{u}$ is set equal to the lowest utility obtained under the observed contract. A very small share of physicians (less than 0.01%) are excluded in the simulation of the observed contract, which fixes $\underline{u}$ at the utility of zero action and zero payment for the type with the lowest degree of altruism (see Appendix A.2).

[50] The small increase in the marginal payment for dosages above 50 thousand units is driven by the tails



(a) Payment as a function of the 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 ($\tau_k' \bar{x}_k = 43.4$). 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.

The distributions of the treatment amounts in Figure 4c include the full-information solution for comparison (gray, long-dashed line). It is readily apparent that the treatment amounts under the observed contract are typically too high, exceeding even the health-maximizing amount (i.e., $[\tau - b_0]/\delta = 56,445$ units) for all but the lowest-treatment types of providers. This accords with the concerns that were raised at the time about high reimbursement 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 15 percent of types 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 examine over- and underprovision in the economic sense, Figure 5 plots the CDFs of deviations of the treatment amounts from their full information amounts, under each contract. Overprovision is nearly universal with the observed contract (98.5% of provider types), and it remains very common with the optimal linear contract (84.4% of provider types). In other words, under the optimal linear contract, most provider types 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 is standard with optimal unrestricted contracts, the highest action equals the maximum in the full-information allocation, and all other treatment amounts 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 patients have identical need for treatment in this example.) Compared to the observed contract, the optimal nonlinear contract reduces the standard deviation of dosages by 21%. 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

of the distribution of α and z.

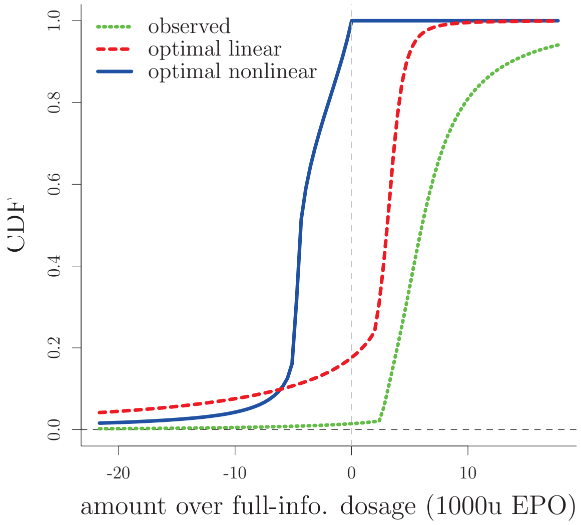


Figure 5: Deviations over full-information treatment amounts, for patients with median severity of anemia.

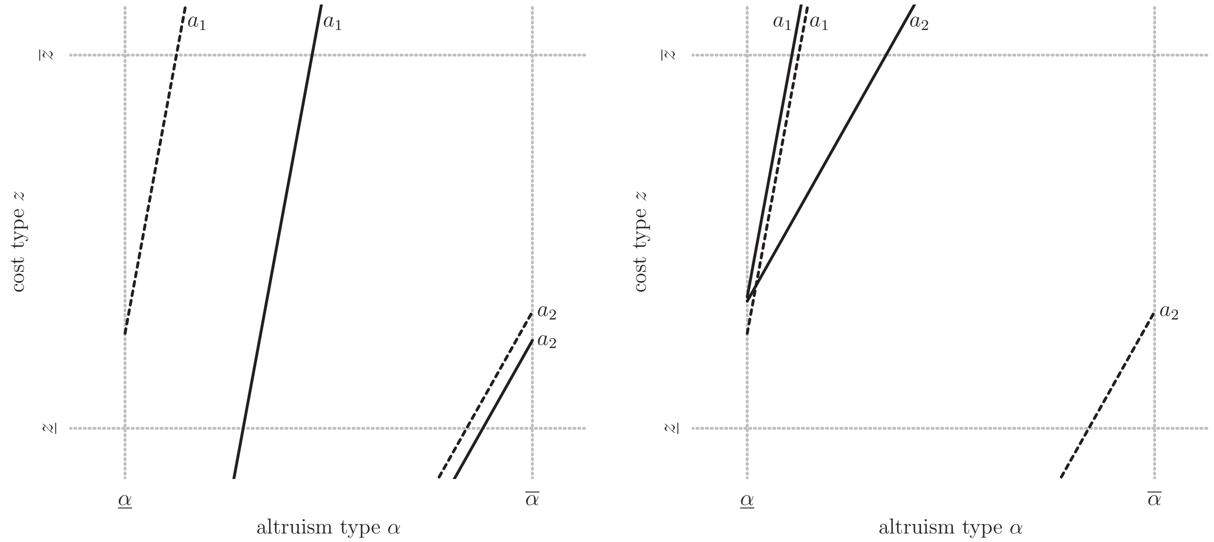
Notes: Figure plots the CDF of the deviation over full information of treatment amounts under the optimal nonlinear contract (blue, solid line), optimal linear contract (red, dashed line), and baseline contract (green, dotted line), for patients with median baseline hematocrit ($b_0 = 34.8$) and mean target hematocrit ($\tau_k' \bar{x}_k = 43.4$). For example, the line for the optimal nonlinear contract is the CDF of $[a^{*\mathrm{SB}}(\alpha, z) - a^{*\mathrm{FI}}(\alpha, z)]$, where $a^{*\mathrm{SB}}(\alpha, z)$ is the equilibrium treatment amount for type (α, z) under the second-best and $a^{*\mathrm{FI}}(\alpha, z)$ is defined in (3).

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	548	59.2	7.8	88%	
Optimal Linear	420	51.7	10.2	15%	\$92
Optimal Nonlinear	389	46.8	6.1	0%	\$120

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.



(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 isoquant curves 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. Panel (a) plots the isoquants for these amounts under the optimal nonlinear contract, and panel (b) plots the isoquants for these same amounts under the optimal linear contract. For comparison, the isoquants for these amounts under full information are plotted with dashed lines in each panel.

mass at zero. For comparison, with full information the standard deviation would be only 1.5 thousand units, about one fifth 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 6 plots isoquants for the 75th percentile (a_1) and 99.99th percentile (a_2) amounts provided under the optimal nonlinear contract (panel a, black lines), along with the the isoquants for these same amounts under full-information (dashed lines in both panels). The types that provide at least these amounts are to the southeast of the corresponding isoquants. The isoquants under the optimal nonlinear contract are below the isoquants under full information because of the downward distortion (which is greater for smaller amounts). This is in contrast to the isoquants for these amounts under the optimal linear contract (panel b, black lines). In particular, the isoquant for amount a_2 lies far above of the full information isoquant, indicating that many types provide more than what is essentially the full-information maximum ($\bar{a}^{*\mathrm{FI}}$), under the linear contract.

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 government) from the changes in outcomes presented above.[51] There are substantial gains from using the optimal nonlinear contract, equal to \$120 per month or \$1440 per year. 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 almost \$340 per patient per year. We can also calculate the gains in the government's maximized objective in the full information scenario, which provides a measure of the losses due to asymmetric information. The difference between the gains in the full information scenario and the best feasible gains under a nonlinear contract amount to almost \$2800 per patient per month, which suggests that asymmetric information is indeed a major feature of this environment.

7 Summary and Conclusions

In this paper we analyze optimal contracts for agents with multidimensional hidden types. We do this for a particularly important sector of the economy, health care, where asymmetric information is pervasive and where providers' responses to incentives have important impacts on both health and costs. Qualitatively, we document the presence of multidimensional unobserved heterogeneity among physicians, which implies that an optimal contract would mitigate the effects of asymmetric information by paying variable marginal rates. We find that physicians respond to financial incentives, but that they are also altruistic, so the optimal contract can leverage physician altruism to increase patient health at lower cost. Quantitatively, we project that the government could improve its objective by \$1440 per patient-year by switching from the observed contract to the optimal nonlinear contract. These results provide new empirical evidence on the structure of optimal contracts, and they can inform policy decisions about physician payment systems. As we have shown, the form of a payment contract may have substantial impacts on both treatment and costs.

[51] 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 but drops out from the differences shown in Table 4.

References

"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, ISSN 0272-6386, doi: https://doi.org/10.1053/j.ajkd.2007.06.008.

Abito, J. M., "Welfare Gains From Optimal Pollution Regulation," *Working Paper*, 2018.

Aldy, J. E. and W. K. Viscusi, "Adjusting the value of a statistical life for age and cohort effects," *The Review of Economics and Statistics*, 90(3):573–581, 2008.

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

Bagnoli, M. and T. Bergstrom, "Log-Concave Probability and its Applications," *Economic Theory*, 26(2):445–469, 2005.

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," *The American Journal of Medicine*, 108(8):609 – 613, 2000, ISSN 0002-9343, doi: https://doi.org/10.1016/S0002-9343(00)00371-5.

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

Blundell, R. and A. Shephard, "Employment, Hours of Work and the Optimal Taxation of Low-Income Families," *Review of Economic Studies*, 79(2):481–510, 2011.

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

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

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

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

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

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

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

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

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

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

Einav, L., A. Finkelstein and N. Mahoney, "Provider Incentives and Healthcare Costs: Evidence From Long-Term Care Hospitals," *Econometrica*, 86(6):2161–2219, 2018, doi: 10.3982/ECTA15022.

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

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

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

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

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

Gaynor, M. and M. Pauly, “Compensation and Productive Efficiency in Partnerships: Evidence From Medical Groups Practice,” *Journal of Political Economy*, 98(3):544–573, 1990.

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

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

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

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

Johnson, S. G., “The NLopt nonlinear-optimization package,” 2018.

Judd, K. and C.-L. Su, “Optimal Income Taxation with Multidimensional Taxpayer Types,” Tech. rep., Working Paper, earliest available draft April, 2006.

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

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

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

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,” *Handbook of Health Economics*, 1:461–536, 2000.

Mirrlees, J. A., “An Exploration in the Theory of Optimum Income Taxation,” *Review of Economic Studies*, 38(2):175–208, 1971.

—, “The Theory of Optimal Taxation,” *Handbook of Mathematical Economics*, 3:1197–1249, 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, ISSN 00257079.

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*, vol. 1 of *Econometric Society Monographs*, pp. 150–197, Cambridge University Press, 2003, doi: 10.1017/CBO9780511610240.006.

Saez, E., “Using Elasticities to Derive Optimal Income Tax Rates,” *Review of Economic Studies*, 68(1):205–229, 2001.

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, doi: 10.1056/NEJMoa065485, pMID: 17108343.

Stein, C. M., “Estimation of the Mean of a Multivariate Normal Distribution,” *The Annals of Statistics*, 9(6):1135–1151, 1981, ISSN 00905364.

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.

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

Wolak, F. A., "An Econometric Analysis of the Asymmetric Information, Regulator-Utility Interaction," *Annales d'Economie et de Statistique*, 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 $(\ddot{\alpha}, \ddot{z}) \equiv \text{arg min}_{\,(\alpha,z)}\, u(a^*(\alpha, z; p_1); \alpha, z, p_0, p$ i.e., the lowest-utility type given linear contract (p_0, p_1).[52] Setting up the Lagrangian based on the remaining participation constraint, we have

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

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

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

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

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

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

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

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

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 reimbursement rate p_1.[53] The government's problem, allowing for exclusion, is:

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

$$+ \int\limits_{\underline{\alpha}}^{\overline{\alpha}} \int\limits_{\tilde{z}^0(\alpha, p_1)}^{\overline{z}} \left[\alpha_g h(0) - p_0 \right] f(\alpha, z) dz d\alpha \tag{16}$$

s.t.

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

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

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

(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 action 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

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

Differentiating with respect to p_0, we obtain $\mu^* = 1$ and $p_0^* = \underline{u} - \underline{\alpha} h(0)$. Differentiating with respect to p_1, and simplifying a good bit[54], we obtain the following implicit expression for p_1^*:

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

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 (2018)).

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

Unlike the case where there is no equilibrium exclusion under the optimal linear contract (see Appendix A.1), the payment rate under the optimal linear contract when there are excluded types is only characterized via a cumbersome implicit expression (see Appendix A.2), which is not ideal because, without further guidance, one would have to exhaustively search through all possible values of α_g to prove the assertion that there did not exist a value of α_g that could rationalize the observed payment rate. Therefore, we adopt an alternative approach, which is to obtain a tractable expression for an upper bound of the optimal linear payment rate, which we then show is below that in the data. (Note that, while we use the more general h notation for the health production function when it simplifies expressions, results here were obtained using the quadratic-loss parameterization of h, specified in Section 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.[55] 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 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*

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

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

[55] 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.

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

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

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* $\tilde{z}^0(\cdot,\cdot) \geq \underline{z} > 1$*,*[56] *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 5 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, $p_1^*(\alpha_g;\tilde{z}^0(\cdot,p_1^*)) \leq p_1^*(\alpha_g = \infty;\tilde{z}^0(\cdot,p_1^*)) \leq p_1^*(\alpha_g = \infty;\tilde{z}^0(\cdot,p_1^*) = \overline{z}) = \mathrm{E}\left[\frac{z}{\alpha^2}\right] / \mathrm{E}\left[\frac{1}{\alpha^2}\right] < 9.26$.

[56] The lower bounds of the marginal cost type distribution for the low, medium, and high baseline HCT intervals are, respectively, 7.11, 7.17, and 7.19 \$/1000u EPO.

Table 5: Upper bound for optimal linear payment rate

	Baseline HCT interval		
	30-33	33-36	36-39
$p_1^*(\infty;\overline{z})$	8.974	8.995	8.956

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 $p^*(a)$.[57] 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 7 shows, there are two cases, corresponding to $\tilde{\alpha}$. If $\tilde{\alpha} \geq \overline{\alpha}$, as depicted on the left, then

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

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.[58] Thus, we have

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

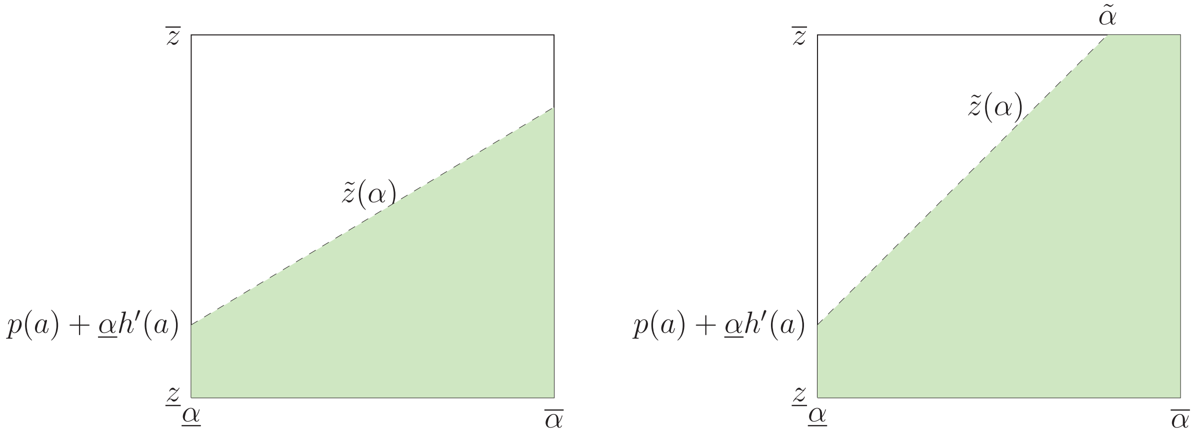
where F_α denotes the marginal CDF of α.

To solve for p^* using (8), we also need to differentiate S above with respect to (the

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

[58] 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 7: $\tilde{\alpha}$ cases



parameter) p. If $\tilde{\alpha} \geq \overline{\alpha}$, we have

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

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

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

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 $a \in A$.[59]

[59] Although not depicted in Figure 7, when $\tilde{\alpha}(p,a) \geq \underline{\alpha}$, it is possible that $\tilde{z}(\underline{\alpha}) < \underline{z}$. Here, the integration limits for α must be adapted to account for $\tilde{z}(\alpha)$ crossing the α axis from below. Let $\breve{\alpha}(p,a) \equiv \frac{\underline{z}-p}{h'(a)}$ denote the altruism type satisfying $\tilde{z}(\breve{\alpha}) = \underline{z}$. (Note that the condition $\tilde{z}(\underline{\alpha}) < \underline{z}$ is equivalent to $\breve{\alpha}(p,a) > \underline{\alpha}$.) There are two subcases. First, if $\breve{\alpha}(p,a) > \overline{\alpha}$, then even the most altruistic physician type would not provide the level of treatment under consideration at marginal transfer p, meaning $S(p,a) = 0$ and $\frac{\partial S(p,a)}{\partial p} = 0$. Second, if $\breve{\alpha}(p,a) \in (\underline{\alpha}, \overline{\alpha}]$ then, if $\tilde{\alpha} \geq \overline{\alpha}$ then (20) becomes

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

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

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

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 action:

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

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 actions. The lowest value of the grid is zero, i.e., we allow for optimal exclusion via the nonlinear contract. The maximum value of the grid is 0.01 below the full-information action 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 action space (which is the same as the upper bound of the full-information action space, due to the downwards-distortion of equilibrium actions under the optimal nonlinear contract).

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

As noted in Section 5.2, in practice we recover $F(\alpha, z)$ under a distributional assumption, and we recover the parameters of the distribution using estimates from OLS estimation of the reduced form (13) and an auxiliary regression of its residuals. Here we describe the details of this approach.

We specify $F(\alpha, z)$ such that $\ln \alpha$ and z have a joint normal distribution. The distribution is parameterized as follows (restricting to one element of $\ln \alpha$, for hematocrit interval k):

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

The value of μ_z is treated as known from our external information on costs. This leaves four parameters to recover for each hematocrit interval: $\mu_\alpha(k)$, $\sigma_\alpha^2(k)$, $\sigma_{\alpha z}(k)$, and $\sigma_z^2(k)$. As detailed below, these parameters are uniquely identified by the first and second moments of the random coefficient (β_2^k) and random effect (ν^k) in the reduced form (13).

The required moments of β_2^k and ν^k are recovered via OLS estimation of (13) and an auxiliary regression of its residuals. Let $\bar{\beta}_2^k$ denote the mean of β_2^k, and then decompose the random coefficient as $\beta_2^k = \bar{\beta}_2^k + \tilde{\beta}^k$. Then (13) can be rearranged as

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

(for b_{0jt} in interval k). The 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. An auxiliary regression of the composite residual, r_{ijt}^k, times the provider-level mean residual, $\bar{r}_i^k$ (within interval k), then yields consistent estimates of the other moments: $\mathrm{V}(\beta_2^k)$, $\mathrm{V}(\nu^k)$, and $\mathrm{Cov}(\beta_2^k, \nu^k)$.[60]

To derive the auxiliary regression, first expand the product of the composite residual and the provider-level mean residual as follows:

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

[60] $\mathrm{E}(\nu^k) = 0$ by construction.

(The variables of the form $\bar{x}_i^k$ 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 reimbursement rates and the number of observations is as follows:

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

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 $\frac{1}{n_i}$.

Next, we show that the structural parameters $\mu_\alpha(k)$, $\sigma_\alpha^2(k)$, $\sigma_{\alpha z}(k)$, and $\sigma_z^2(k)$ are functions of the 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 structural parameters are uniquely identified by these moments, and it provides analytic expressions to compute their values. For simplicity, the derivations below omit the index for the hematocrit interval (k).

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

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

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

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

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

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

b) Next we obtain $\sigma_{\alpha z}$ from $\mathrm{Cov}(\beta_2, \nu)$, along with $\mathrm{E}(\beta_2)$ and $\mathrm{V}(\beta_2)$. For this we first use the definitions of $\beta_2 \equiv \delta^{-2}\alpha^{-1}$ 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 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 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.[61] 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

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

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,[62]

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

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

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

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

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

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

To do this we apply Stein's lemma to $-\ln\alpha$; however to simplify the derivation, we use X in place of $-\ln\alpha$. Here the lemma is stated as follows: if X is normally distributed, g is differentiable, and the relevant expectations exist, then $\mathrm{E}[(X-\mu_X)g(X)] = \mathrm{V}(X)\cdot\mathrm{E}[g'(X)]$.

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

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

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

□

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_{ijt}. We then show the semiparametric identification of h, specifically the curvature of the function, if its arguments enter via a known index specification.

F.1 F

Let $n_i \to \infty$, and further suppose that the number observations within each interval of baseline hematocrit goes to infinity for each provider: $\sum_{j=1}^{n_i} 1\{\bar{b}_{k-1} < b_{0jt} \leq \bar{b}_k\} \to \infty$. Assume that η_{ijt} is mean-independent of (b_{0jt}, x_{jt}, p_t): $\mathrm{E}(\eta_{ijt}|b_{0jt}, x_{jt}, p_t) = 0$. 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. 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 h

We consider a particular baseline hematocrit interval k, and omit the k subscript for the remainder of this section. We start by noting that only the scale of $\alpha_i h$ is identified, meaning we could fix the scale of α_i (e.g., set μ_α to a constant) or h. To be consistent with our empirical specification, we fix the scale of h; see Appendix F.1 for an argument for nonparametric recovery of the joint distribution of (α_i, z_i), given an h with a fixed scale.

We make use of an index assumption to show semiparametric identification of h. With some abuse of notation, let $h(a; b_0, x) = h(\delta a + b_0 - \tau' x)$, where the function h and the values of δ and τ are unknown. The 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 for an individual 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. Given the strict concavity of h, the index inside h must, therefore, take the same value for all patients receiving treatment. This fixes the index up to scale and location, neither of which are identified because they cannot be separated from h. Therefore, we set the coefficient on b_0 to one, which gives the index a natural interpretation: red blood cell count. Because the intercept of τ is arbitrary, it may be set to zero.[63]

Then, given δ and τ, the curvature 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 index, which now has a known value. Taking the difference in the first-order conditions for provider i in periods t and s, we have

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

Hence the derivative of h is known (given that the scale of h has been fixed and that α_i was estimated, e.g., by a modified version of the argument in Appendix F.1), for T points of support (where T is the number of periods with different payment rates). Furthermore, the values of the index y are different for different providers, because they have different values of (α, z)—which imply different optimal choices of a_{ijt} and, hence, different values of y_{ijt}, given the same (b_{0jt}, x_{jt})—so the shape of h can be traced out at many points of support.

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) between health, specified as a squared loss in hematocrit levels, and dollars in the government's objective. The clinical trial gives estimates of the mortality risk associated with different hematocrit levels, so under certain assumptions (noted below), we can find a value of α_g that multiplies the VSLY by the

[63] This is in contrast to our ability to estimate the intercept for τ when using our parametric specification for h.

mortality risk. ALT:converts the difference in h implied by the different hematocrit levels in the clinical trial to the difference in the (monetary) value of statistical life years implied by the clinical trial.

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. (This also assumes the utility loss from a serious cardiovascular event is equal to the expected loss from 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 $\alpha_g = 52.6$.

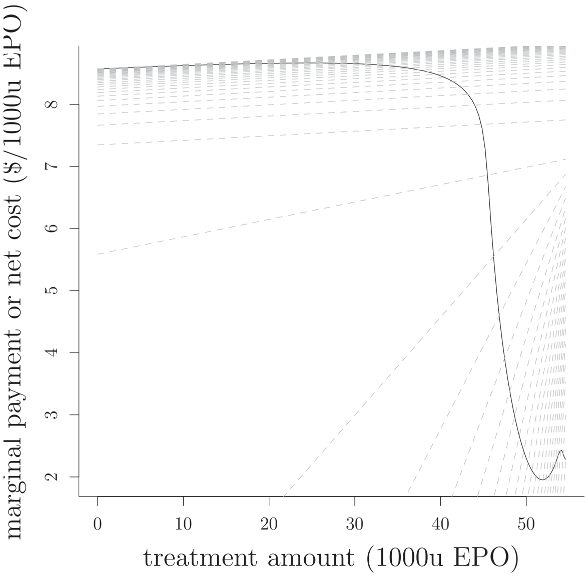


Figure 8: Regularity condition check, for patients with median severity of anemia.
Notes: Figure plots marginal payment curve (solid, black line) and physician supply curves (dashed, grey lines) for patients with median baseline hematocrit ($b_0 = 34.8$) and mean target hematocrit ($\tau_k' \bar{x}_k = 43.4$).

H Check of Regularity Condition

Figure 8 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.[64]

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.[65] Figure 9 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

[64] We have also verified that this regularity condition is satisfied in the other baseline hematocrit intervals.
[65] The values are 32, 34.8, and 37.4 for the lower, middle, and upper intervals, respectively.

Table 6: Summary of Outcomes under Optimal Contracts

Baseline HCT 30-33

	Mean Pmt	Mean Dosage	SD Dosage	Share above τ (%)
Observed	750	80.9	12.5	87
Optimal Linear	427	60.7	20.9	0
Optimal Nonlinear	404	54.6	12.3	0

Note: Table shows summary statistics of outcomes corresponding to contracts plotted in Figure 9a. Mean and SD of dosage are in 1,000 units/month. Treatment choice to get to hematocrit target: 76.3

Baseline HCT 33-36

	Mean Pmt	Mean Dosage	SD Dosage	Share above τ (%)
Observed	548	59.2	7.8	88
Optimal Linear	420	51.7	10.2	15
Optimal Nonlinear	389	46.8	6.1	0

Note: Table shows summary statistics of outcomes corresponding to contracts plotted in Figure 9b. Mean and SD of dosage are in 1,000 units/month. Treatment choice to get to hematocrit target: 56.4

Baseline HCT 36-39

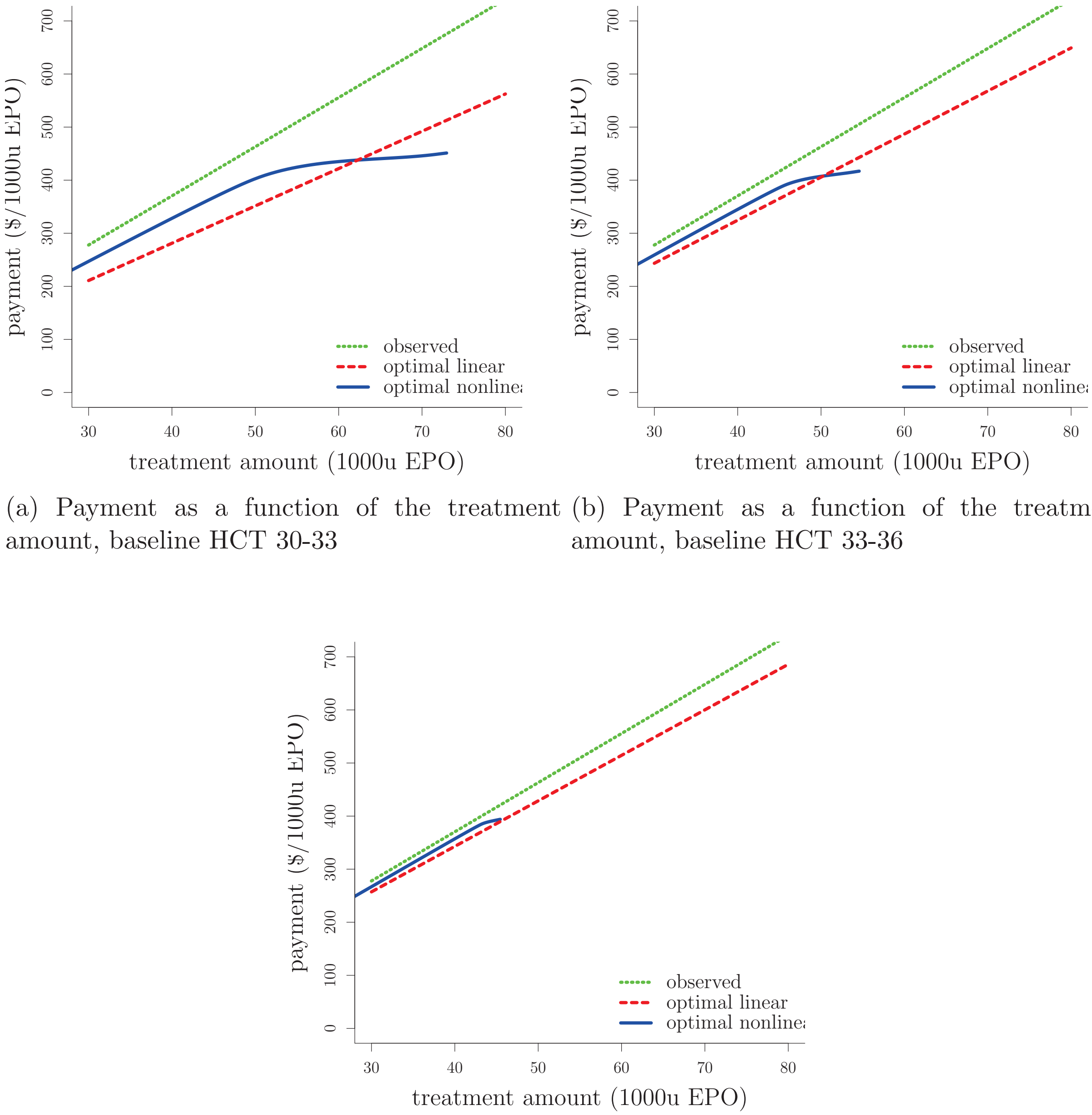
	Mean Pmt	Mean Dosage	SD Dosage	Share above τ (%)
Observed	442	47.7	4.8	88
Optimal Linear	388	45.2	4.8	43
Optimal Nonlinear	384	43.3	2.6	0

Note: Table shows summary statistics of outcomes corresponding to contracts plotted in Figure 9c. Mean and SD of dosage are in 1,000 units/month. Treatment choice to get to hematocrit target: 46.2

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 6 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 almost 90 percent of providers would give medically excessive dosages (i.e., negative marginal product) to patients with these baseline hematocrit levels, according to our estimates. The optimal linear contract does not eliminate this obvious inefficiency: to patients with the median hematocrit in the middle and upper intervals, respectively 15 and 43 percent of providers would give medically excessive dosages under it. This inefficiency does not occur with the optimal nonlinear contract because, as

Figure 9: Optimal Nonlinear Contracts for median of the three hematocrit intervals



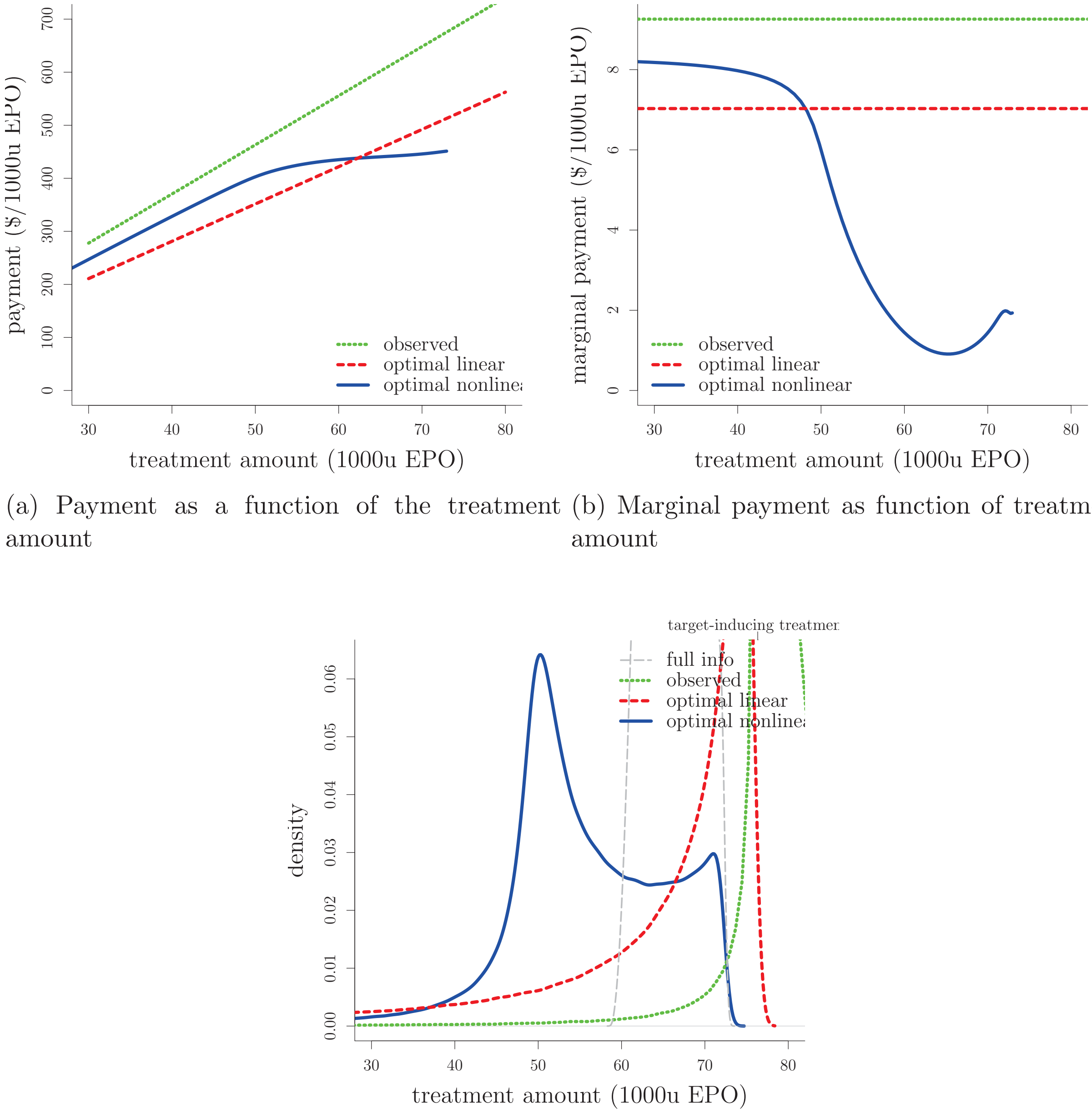
(a) Payment as a function of the treatment amount, baseline HCT 30-33

(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 seen in Figure 11c, 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 in each example/interval). The optimal nonlinear contract reduces the standard deviation of dosages, compared to the observed contract, by 2%, 22%, and 46% at the low, medium, and high baseline hematocrit levels. By contrast the optimal linear contract increases the variation in dosages, however, because it provides a constant marginal incentive, just like the observed contract; in fact, there is *higher* variation than under the observed contract because some types are (optimally) excluded, which puts a non-negligible mass at zero. In contrast, under the full information scenario the standard deviations are substantially smaller, but some variation remains, which reflects the variation in altruism and marginal costs.

Figure 10: Optimal Nonlinear Contract Treatment Amounts and Payments, baseline HCT 30-33

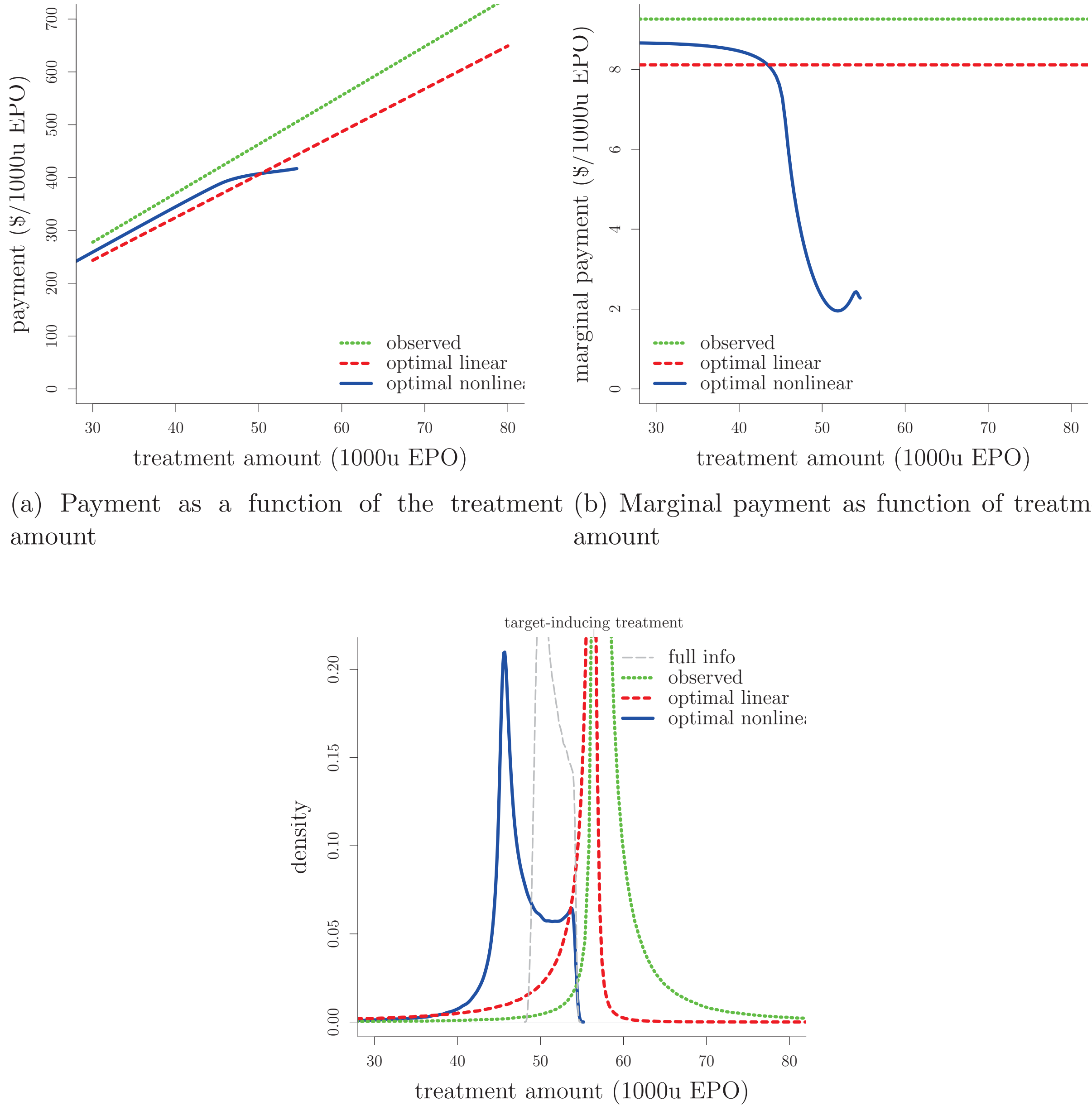


(a) Payment as a function of the treatment amount

(b) Marginal payment as function of treatment amount

(c) Distribution of treatment amounts

Figure 11: Optimal Nonlinear Contract Treatment Amounts and Payments, baseline HCT 33-36

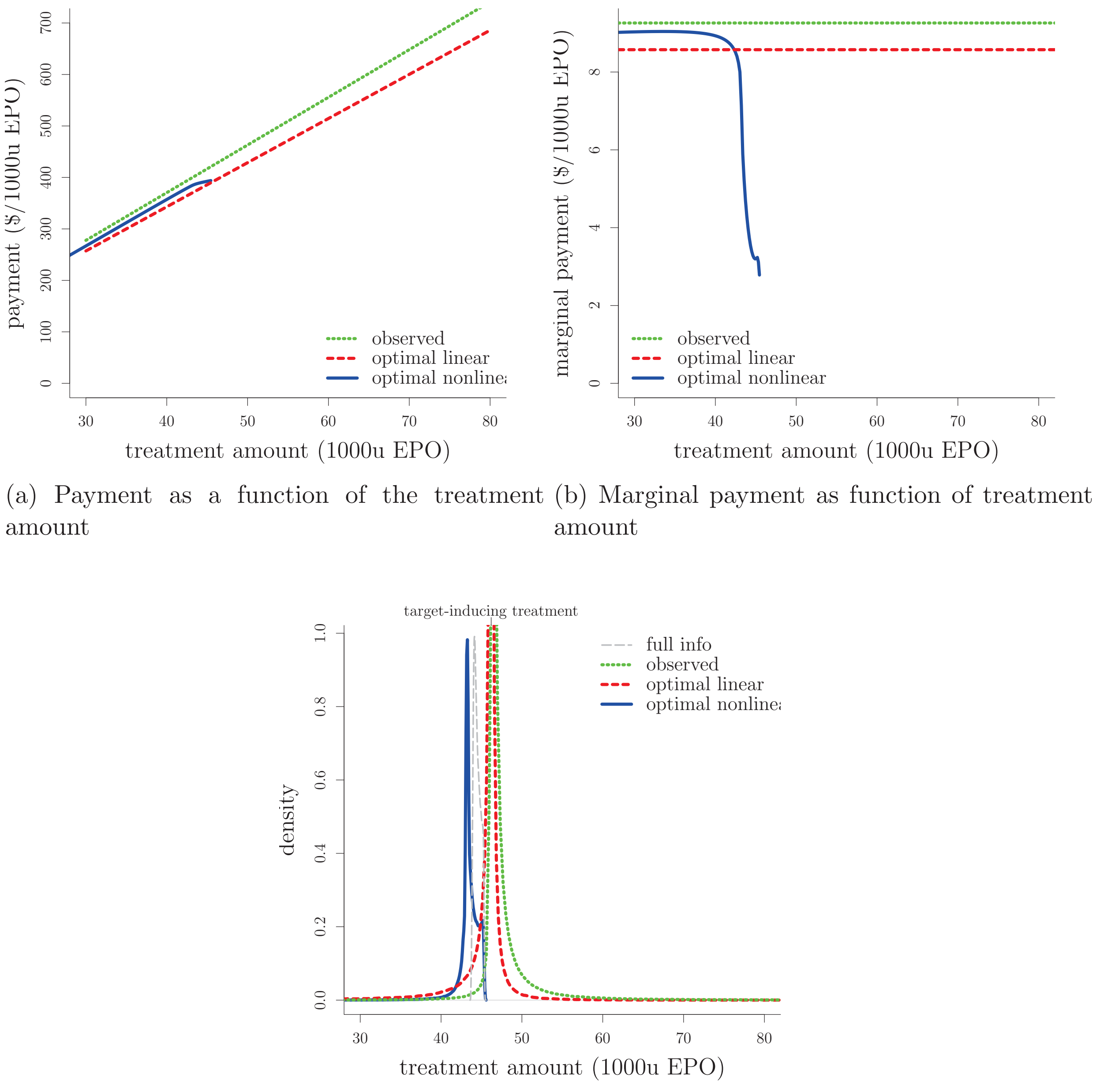


(a) Payment as a function of the treatment amount

(b) Marginal payment as function of treatment amount

(c) Distribution of treatment amounts

Figure 12: 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 A1. OLS ESTIMATES

	Hematocrit Interval					
	(1)	(2)	(3)	(4)	(5)	(6)
VARIABLES	≤ 27	> 27 to 30	> 30 to 33	> 33 to 36	> 36 to 39	> 39
hematocrit	-7.55	-7.68	-9.60	-6.54	-3.67	-1.76
	(0.11)	(0.44)	(0.24)	(0.15)	(0.13)	(0.15)
reimb_rate	-1.99	12.80	11.16	6.99	3.72	8.48
	(5.49)	(6.20)	(3.30)	(2.11)	(1.98)	(2.76)
age	-0.39	-0.34	-0.40	-0.36	-0.26	-0.18
	(0.04)	(0.03)	(0.02)	(0.02)	(0.01)	(0.01)
female	0.04	-2.85	-1.09	1.44	2.84	2.84
	(1.08)	(0.86)	(0.55)	(0.41)	(0.35)	(0.39)
1.Charlson	6.67	5.90	9.20	8.22	7.67	5.01
	(1.85)	(1.49)	(0.99)	(0.71)	(0.61)	(0.66)
2.Charlson	6.73	8.34	11.10	10.63	8.31	5.88
	(1.77)	(1.39)	(0.91)	(0.68)	(0.61)	(0.65)
3.Charlson	9.36	12.50	14.41	12.12	8.67	5.52
	(1.70)	(1.45)	(0.97)	(0.74)	(0.62)	(0.67)
4.Charlson	8.11	10.70	16.41	14.23	11.07	7.45
	(2.06)	(1.72)	(1.23)	(0.90)	(0.75)	(0.78)
5.Charlson	13.05	15.30	16.95	15.11	11.85	8.97
	(2.33)	(2.10)	(1.45)	(1.12)	(0.96)	(1.05)
6.Charlson	10.65	16.00	19.35	19.07	13.96	9.13
	(3.07)	(2.73)	(1.89)	(1.52)	(1.23)	(1.27)
7.Charlson	14.75	20.80	27.57	26.73	20.92	16.50
	(4.67)	(3.82)	(3.06)	(2.59)	(2.28)	(2.57)
8.Charlson	10.44	17.47	26.03	25.39	15.79	13.05
	(5.09)	(5.65)	(4.05)	(3.13)	(2.62)	(2.60)
9.Charlson	24.77	25.64	32.07	31.37	22.73	19.96
	(6.99)	(6.71)	(5.21)	(4.31)	(3.89)	(4.30)
10.Charlson	21.67	19.32	23.41	27.44	31.98	24.74
	(17.83)	(9.33)	(7.15)	(6.73)	(7.01)	(8.45)
11.Charlson	43.08	53.85	39.13	41.47	38.45	15.62
	(20.15)	(15.37)	(11.56)	(9.08)	(7.44)	(9.42)
12.Charlson	8.45	40.56	47.15	35.69	31.33	-3.10
	(22.65)	(15.39)	(10.97)	(8.18)	(9.82)	(5.78)
Constant	337.49	342.34	401.56	302.33	196.40	115.98
	(4.64)	(13.20)	(7.90)	(5.34)	(5.13)	(6.40)
Observations	97,983	82,639	216,390	379,527	267,245	83,344
R-squared	0.231	0.016	0.030	0.028	0.021	0.016
RMSE	69.20	84.50	71.82	58.88	49.32	41.71

Robust standard errors in parentheses

TABLE A2. FIXED EFFECTS ESTIMATES

	Hematocrit Interval					
	(1)	(2)	(3)	(4)	(5)	(6)
VARIABLES	≤ 27	> 27 to 30	> 30 to 33	> 33 to 36	> 36 to 39	> 39
hematocrit	-5.61	-7.29	-9.37	-6.66	-4.03	-2.38
	(0.17)	(0.36)	(0.20)	(0.13)	(0.12)	(0.15)
reimb_rate	-2.47	15.40	11.53	7.11	4.69	9.27
	(5.14)	(5.68)	(3.11)	(2.03)	(1.91)	(2.71)
age	-0.32	-0.36	-0.38	-0.33	-0.24	-0.18
	(0.03)	(0.03)	(0.02)	(0.02)	(0.01)	(0.01)
female	-1.35	-3.96	-1.48	1.21	2.39	2.27
	(0.92)	(0.77)	(0.51)	(0.39)	(0.34)	(0.39)
1.Charlson	6.08	5.36	7.98	7.12	6.60	4.15
	(1.58)	(1.37)	(0.90)	(0.68)	(0.61)	(0.66)
2.Charlson	6.80	8.01	10.41	9.93	8.05	4.82
	(1.49)	(1.23)	(0.84)	(0.66)	(0.58)	(0.65)
3.Charlson	8.81	10.32	12.89	11.27	8.72	5.06
	(1.63)	(1.37)	(0.91)	(0.73)	(0.60)	(0.66)
4.Charlson	8.95	9.48	15.30	13.93	10.81	7.65
	(1.68)	(1.60)	(1.10)	(0.86)	(0.73)	(0.77)
5.Charlson	10.24	12.61	16.36	14.62	11.18	7.37
	(2.36)	(2.01)	(1.32)	(1.05)	(0.92)	(1.00)
6.Charlson	9.86	15.58	18.04	18.32	13.47	7.89
	(2.49)	(2.44)	(1.67)	(1.41)	(1.18)	(1.27)
7.Charlson	15.53	17.61	23.77	24.61	20.65	15.53
	(4.02)	(3.15)	(2.76)	(2.42)	(2.21)	(2.55)
8.Charlson	2.90	13.54	24.81	23.37	17.03	12.44
	(4.12)	(5.04)	(3.74)	(3.22)	(2.62)	(2.53)
9.Charlson	22.26	23.72	31.70	32.16	23.54	19.33
	(6.75)	(6.82)	(5.24)	(4.22)	(4.09)	(4.00)
10.Charlson	32.65	14.35	22.70	27.22	30.10	22.96
	(9.12)	(9.60)	(6.30)	(6.53)	(6.80)	(7.37)
11.Charlson	35.23	31.41	41.02	38.59	37.81	18.81
	(17.46)	(11.76)	(8.77)	(8.29)	(7.17)	(8.35)
12.Charlson	21.60	32.34	31.65	27.59	17.16	-7.72
	(5.99)	(14.03)	(10.89)	(7.52)	(11.57)	(5.51)
Constant	273.28	331.26	393.28	305.02	209.04	141.78
	(6.14)	(10.85)	(6.43)	(4.77)	(4.76)	(6.28)
Observations	97,983	82,639	216,390	379,527	267,245	83,344
R-squared	0.080	0.017	0.030	0.027	0.021	0.016
N. providers	4,507	4,638	4,785	4,814	4,769	4,523
RMSE	60.66	75.35	66.31	55.54	46.65	38.68

Robust standard errors in parentheses