Optimal Contracting with Altruistic Agents: Medicare Payments for Dialysis Drugs*

Martin Gaynor†
Carnegie Mellon University
and NBER

Nirav Mehta‡
University of Western Ontario

Seth Richards-Shubik§
Lehigh University
and NBER

September 12, 2022

Abstract

We study health care provider agency and optimal payment policy in the context of an expensive medication for dialysis patients. Using Medicare claims data we estimate a structural model of treatment decisions, in which providers differ in their altruism and marginal costs, and this heterogeneity is unobservable to the government. In a novel application of nonlinear pricing methods, we empirically characterize the optimal unrestricted contracts in this screening environment with multidimensional heterogeneity. The optimal contracts initially pay similar amounts as the one used by Medicare at the time, but the marginal payment rates decline precipitously at higher dosages. Adopting the optimal contracts would eliminate medically excessive dosages and substantially reduce expenditures, resulting in approximately $300 million in gains from better contracting. The approach we develop could be applied to a broad class of problems in health care payment policy.

Keywords: Medicare, payment policy, optimal contracts, incentives, asymmetric information, screening models, health care, dialysis

*We are grateful to David Dranove, Liran Einav, George-Levi Gayle, Josh Gottlieb, Kate Ho, Amanda Kowalski, Albert Ma, Bentley MacLeod, Ryan McDevitt, Bob Miller, Ariel Pakes, Mark Satterthwaite, Steven Stern, Lowell Taylor for helpful comments, and to audiences at presentations at Boston University, Carnegie Mellon, Georgia, Johns Hopkins, Michigan, NYU, North Carolina, Northwestern, Penn, Princeton, Stanford, University of British Columbia, Wisconsin, the 2018 International Industrial Organization Conference, the 2018 Annual Meeting of the Society of Labor Economists, the 2018 Cowles Structural Microeconomics Conference, the 2018 Southern Economic Association Annual Meeting, the 2019 Annual Health Econometrics Workshop, and the 2020 Health Economics Research Organization meeting. Dr. Christos Argyropolous provided us with important information on EPO dosing practices and the operation of dialysis facilities. We thank Ali Kamranzadeh, Martin Luccioni, and Cecilia Diaz Campo for excellent research assistance. The usual caveat applies.

†E-mail: mgaynor@cmu.edu
‡E-mail: nirav.mehta@uwo.ca
§E-mail: sethrs@lehigh.edu

1 Introduction

A central problem in health care is how to pay providers to treat patients. Asymmetric information is pervasive because providers often have substantial information that is not observed by payers, which are typically third parties. Therefore, payers have to decide how to contract with providers to deliver care, while recognizing that providers possess relevant information that they do not. Consequently, the effects of better or worse incentives can have profound impacts on treatments, expenditures, and health.

Economists have devoted a great deal of attention to understanding the impacts of payment incentives on health care spending and health outcomes, and to suggest better methods of payment (e.g., Cutler, 1995; Acemoglu and Finkelstein, 2008; McClellan, 2011; Clemens and Gottlieb, 2014; Ho and Pakes, 2014; Einav et al., 2018; Currie and MacLeod, 2020). However, while there has been extensive work examining the impacts of payment methods on provider behavior, to date the literature has not directly applied contract theory to find optimal payment arrangements for health care providers.

In this paper we use classic results for screening models (e.g., Myerson, 1981; Maskin and Riley, 1984; Goldman et al., 1984; Wilson, 1993) to empirically derive optimal payment contracts for an expensive and controversial medication used to treat anemia (a lack of red blood cells) in patients with end-stage renal disease (ESRD, also known as kidney failure). The medication, epoetin alfa (EPO), is administered by dialysis providers and is primarily paid for by the Medicare program, the dominant payer for the treatment of ESRD in the United States. The program spent more on EPO than on any other single medication for several years in the 2000s ($2 billion in 2010, U.S. Government Accountability Office, 2012), and there were strong financial incentives to administer EPO because provider margins were on the order of 30 percent (Whoriskey, 2012). There were also substantial consequences for health, with ongoing concerns about the risks of high dosages (Brookhart et al., 2010; Whoriskey, 2012), which include serious cardiovascular events and death.

Our results indicate that optimal payment contracts could generate gains on the order of $300 million per year and would eliminate medically excessive dosages—i.e., those which are harmful on the intensive margin of treatment. This approach could be relevant for provider-administered medications more broadly, and for other treatments where decisions are primarily about the quantity, as we discuss below.

Key features of the setting make a screening model the appropriate framework to study optimal contracting for EPO, and potentially for other provider-administered drugs and treatments as well. The medication is given intravenously to nearly all patients with ESRD, so the relevant choice is the quantity administered by the provider (the agent), as in the

classic theoretical models. Also as in those models, the quantity is observed, because dosages are reported on insurance claims submitted by the provider to the government (the principal), which runs the Medicare program. At the same time, there is likely to be hidden information about provider characteristics that affect treatment choices. Our model features two natural dimensions of such unobserved heterogeneity: altruism and marginal costs. By altruism we mean how much providers care about patient health versus their own compensation; however this could include other intrinsic motivations such as professionalism (e.g., Ash and MacLeod, 2015; Currie and MacLeod, 2020), or extrinsic motivations to keep a patient healthy, for example, so the patient can be treated in the future or to avoid malpractice liability. The marginal costs of providing the treatment pertain to purchasing and administering the drug. While the presence of asymmetric information generally results in a suboptimal outcome, altruism attenuates the distortion because providers put weight on patient health, which the principal values.

We use data from Medicare claims in 2008 and 2009, a period when the payment policy was stable and when there were no major informational shocks about EPO. As with most insurance claims, the treatments are observed, in this case the dosages administered to patients. Furthermore, quite uniquely, a key quantitative measure of the patient's condition is available in the claims data. Providers were required to report the patient's red blood cell level (i.e., the severity of their anemia) in order to be paid for the EPO, and these blood levels are recorded on the claims. Other Medicare data provide rich information on additional patient characteristics, such as the presence of relevant comorbidities. Because of these institutional features and rich data, we are able to use a relatively simple approach to estimate the structural parameters of our theoretical model. Our specification yields linear reduced forms of the structural model, which can be estimated by OLS, while having sufficient flexibility to fit the data, and the structural parameters are direct, closed-form functions of the reduced-form estimates.

Our estimates indicate that altruism is important in this context, and that there is substantial heterogeneity across dialysis providers, in both their degree of altruism and their marginal costs. Theory therefore implies that, in contrast to the observed linear payment contracts, the optimal contracts must be nonlinear, so that there are varying marginal incentives to help mitigate the distortions from asymmetric information. Furthermore, we show that the observed reimbursement rates were too high: they cannot be rationalized as optimal, even when restricting to linear contracts.

We derive the optimal contracts using the demand profile approach (Goldman et al., 1984; Wilson, 1993). This approach, which was developed for monopoly pricing problems and has not previously been applied to supply contracting, tractably accommodates multidimensional

heterogeneity, as opposed to standard methods for solving for optimal contracts, which only work with heterogeneity of one dimension. This approach enables us to characterize the unconstrained optimal contracts (which are conditional on patient characteristics), which would not only improve over the status quo, but would also in concept obtain the second-best allocations. We also show that the demand profile approach is broadly applicable to supply contracting problems like this one.

At low dosages, the optimal contracts are roughly similar to the observed contract (a traditional fee-for-service contract), with a fairly constant marginal payment rate that is close to, but below, the average reimbursement rate used at the time. However the optimal marginal payment rates drop rapidly around the median dosage, falling by 75% or more. Furthermore, there are important differences in where and how the optimal rates decline, depending on the patient's red blood cell level. The decline occurs at lower dosages for patients with higher levels, who benefit less from EPO, while it begins at higher dosages and proceeds more gradually for patients with lower levels. This reveals important qualitative features of the optimal contracts that could be useful for practical implementations of the policy, such as a set of tiered payment rates that depend on the red blood cell level.[1]

Our simulations of outcomes under the optimal contracts indicate that Medicare could substantially improve beneficiaries' health while reducing its expenditures. Seemingly unjustified variation in dosages, driven by the heterogeneity in provider altruism and marginal costs, is reduced by 27 percent, and the mean payment is reduced by 27 percent (the matching values are coincidental), for a patient with the median red blood cell level. This would improve the value of the government's objective, which depends on both patient health and total expenditures, by an amount equal to $1,500 per patient per year. Additionally, we can quantify the losses due to the asymmetric information about providers, perhaps for the first time in a health application. Those losses are substantial, equal to $2,200 per month for a patient with the median red blood cell level.[2]

The issues we address with this analysis are likely to be important for many provider-administered drugs, which cost Medicare (Part B) $39 billion in 2019 (Medicare Payment Advisory Commission, 2021) (over 12 percent of total Medicare spending) and is one of the most rapidly growing areas of Medicare spending, and which have been the object of ongoing concern and attempts at policy reform (Bach, 2009). This is also broadly relevant

[1] One might ask why Medicare would not simply impose a "forcing contract" that effectively required providers to administer an amount that would, for example, maximize a patient's health. We allow for such a contract, but it would not be optimal because it would be very costly to induce providers with high costs (or low altruism) to administer such an amount.

[2] This is in line with the results from studies of other contexts, which similarly find very large losses due to asymmetric information (e.g., Gayle and Miller, 2009; Abito, 2020, discussed below).

to longstanding concerns about financial incentives and excessive utilization of health care in general, as cited earlier. In fact, the approach we develop here is potentially applicable to a broad class of problems—the key features are that decisions relate to the quantity of treatment (as opposed to a choice among different types of treatment), and that the quantity of treatment is observable.

Our paper relates to the rich literature on health care provider agency (see e.g., McGuire, 2000; Chalkley and Malcomson, 2000, for overviews). The model of provider utility we employ is very similar to that in Ellis and McGuire (1986) and Gaynor et al. (2004), but allows for heterogeneity in altruism and costs. De Fraja (2000) and Jack (2005) theoretically study these forms of heterogeneity across physicians, although there are various distinctions between their models and ours.[3] Choné and Ma (2011) also consider how physician altruism may affect the design of optimal payment contracts, and Godager and Wiesen (2013) provide experimental evidence about heterogeneity in altruism among medical students. Like Clemens and Gottlieb (2014), we empirically examine the impact of Medicare payment incentives, although they look at payment incentives broadly, as opposed to our focus on a specific medical context. In the context of dialysis care, Eliason et al. (2019) examine the effects of corporate ownership on treatment decisions and patient outcomes, and find that drug dosages and other inputs change, and key outcomes worsen, after facilities are acquired by a chain.[4]

Some recent papers on financial incentives in health care examine the effects of counterfactual payment or insurance contracts on expenditures and patient outcomes. Einav et al. (2018) estimate a dynamic model to study how dynamic incentives in payments to long-term care hospitals affect the timing of discharges. Their model includes provider altruism, like ours, but asymmetric information is not a salient feature of their environment. Ho and Lee (2020) estimate a model of employee choice of health insurance plan and medical spending, and use their estimates to consider insurance plan offerings that raise average employee surplus at a single employer. Einav et al. (2021) examine provider selection into a voluntary bundled payments program, and simulate outcomes under alternative lump-sum payments for the bundle. All of these papers show that substantial improvements are possible by modifying the salient features of their observed contracting regimes (e.g., "short-stay" thresholds or coinsurance rates).

Our work also relates to the small number of existing papers that structurally estimate asymmetric information models. As noted by Chiappori and Salanié (2003), despite the

[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 and McDevitt (2017) similarly use the specific context of dialysis care to examine an issue of broad importance in health care, the tradeoff between quantity and quality.

rich theoretical literature on contracting in asymmetric information environments, there is little empirical work that leverages the power of contract theory to derive optimal payment contracts. Einav et al. (2010) discuss the small literature doing this for insurance contracts. Perhaps most closely related in terms of the modeling approach is the literature on optimal regulation, which considers screening models albeit in contexts that differ from ours in important ways (e.g., Wolak, 1994; Gagnepain and Ivaldi, 2002; Abito, 2020). As in the work by Gagnepain and Ivaldi and by Abito, our setting and data allow us to estimate structural parameters without imposing optimality of the observed contract, so we can test (and end up rejecting) the optimality of the observed contract. However in contrast to the models used in that literature, we allow for multidimensional heterogeneity, which requires a different approach to characterize the optimal contract. Furthermore, as we discuss in Section 3, the demand profile approach could be broadly applicable to supply contracting problems like ours. Other papers, on optimal compensation, consider hidden action environments. Paarsch and Shearer (2000) use optimal linear contracts to calculate the incentive effects of piece rates for tree planting, and Gayle and Miller (2009) quantify the welfare loss from moral hazard in executive compensation.[5]

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

2 Background on Dialysis Financing and Treatment

End-stage renal disease (ESRD), or kidney failure, is a chronic and life-threatening condition that affects over half a million individuals in the United States. Since 1973, the Medicare program has provided universal coverage for the treatment of ESRD, regardless of age. In 2009, at the end of our study period, Medicare spent $28 billion on health care for individuals with ESRD (over 7 percent of total Medicare spending), and of that amount, $1.74 billion

[5]Our environment also has similarities to those that studied in the literature on optimal taxation in hidden information environments, which was initiated by Mirrlees (1971). 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 empirical unconstrained optimal contract using our estimates of structural parameters.

was paid specifically for EPO.[6] The drug is used to treat anemia, a lack of red blood cells, which often accompanies chronic kidney disease.[7] 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. The level is commonly measured in terms of the *hematocrit*, which is the volume percentage of red blood cells in the blood.

An important biological fact is that the half-life of EPO is under 12 hours (Elliott et al., 2008), which motivates our use of a static framework to model this treatment decision. Additionally, patient hematocrit levels are highly variable over time. From one month to the next, more than half of the patients in our data experience a change of greater than one percentage point (see Appendix Table A6), which is a clinically relevant difference. Accordingly, providers regularly adjust the dosages for each patient to address these fluctuations.

For ESRD patients, EPO is typically administered intravenously during each dialysis session, which occur multiple times per week (typically three times per week for three to four hours each) at specialized facilities called dialysis centers. Because dialysis occurs so frequently, and because patients are often fairly debilitated by it, travel costs are quite high and patients regard facilities as highly differentiated with regard to location (Eliason, 2019), which limits selection.

The staff at dialysis centers consists of one medical director (a physician, usually a nephrologist), with additional physicians at larger facilities, and multiple nurses and medical technicians.[8] Physicians are independent practitioners who may endogenously match with dialysis facilities.[9] Physicians prescribe dosages of EPO for patients, and nurses or medical technicians administer the injection of the prescribed dosage. Payments are primarily made to the facilities, not the individual physicians or nurses, which is partly why we treat each dialysis center as a unitary provider. In what follows, we call the agent making dosage decisions for patients a "provider."[10]

[6]USRDS *2016 Annual Data Report*, volume 2, chapter 11; available at https://www.usrds.org/annual-data-report/previous-adrs/. 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 make copayments of up to 20%.

[7]EPO is a biological product, or "biologic," but we will typically refer to it as a drug. Another drug, injectable iron, is often used in conjunction with EPO to treat anemia in ESRD patients, but expenditures on iron were much smaller. In 2005, EPO accounted for 70% of expenditures on separately billable drugs for ESRD patients (GAO, 2006).

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

[9]Physicians may have a financial stake in a dialysis facility, e.g., by owning it themselves or through a joint venture. (Private communication from Christos Argyropoulos, M.D. and from an anonymous referee.)

[10]Treating a facility as a unitary provider is consistent with the health economics literature, which for the most part does not distinguish between health care organizations and physicians (e.g., Eliason et al., 2022, 2019; Einav et al., 2018), and with treatments of the firm in economics in general, which often presume (explicitly or implicitly) that firms provide incentives to workers to achieve firm objectives.

The main cost of providing EPO is acquiring the drug from the manufacturer (via a distributor), because its production involves an expensive biological process, and one manufacturer had a monopoly over this class of medications at the time. This motivates the assumption of constant marginal costs in our model, as the pricing in the purchasing contracts was largely per unit. Administering the drug to patients also involves non-trivial costs of staff time to prepare the dosages and monitor the injections (see Section 5.2), which is an additional source of cost heterogeneity.

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.[11] While dialysis itself was reimbursed with a prospective payment system known as the "composite rate," which paid a fixed amount of roughly $135 per session, EPO was a separately billable drug with its own per-unit reimbursement rate. Prior to 2005, that rate was held fixed at $10.00 per 1000 units. In 2006, Medicare adopted a new policy where the rate was based on average sales prices calculated from data reported by the manufacturer. This policy, which was in effect through 2010, set a limit on the reimbursement rate each quarter, equal to 106 percent of the national average sales price from roughly six months earlier (GAO, 2006).[12] This provides the variation we need to estimate the model parameters governing how providers respond to the marginal payment rate for EPO.

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

Alongside the concerns about overprovision, there had been substantial uncertainty about the benefits and risks of EPO (see, e.g., Foley, 2006). Many clinicians and medical researchers felt it was important to counteract severe anemia, to improve quality of life and address other specific risks associated with anemia. In the early 2000s, the National Kidney Foundation

[11] There were concerns both that dosages were supraoptimally high (i.e., marginal benefits less than marginal costs) and that dosages were high enough to harm patient health (i.e., negative marginal product). We will refer to the former as "overprovision" and the latter as "medically excessive".

[12] 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 (i.e., lump-sum) payment. See Eliason et al. (2022) for an analysis of the effects of this policy change.

considered whether to recommend higher targets for the hematocrit level (NKF-KDOQI, 2006). However, the risks associated with high dosages of EPO became clear by the mid 2000s. A major clinical trial found that patients who were given more EPO to achieve a higher target level of hematocrit suffered a greater risk of serious cardiovascular events and death (Singh et al., 2006).[13] This study was published in November 2006, and strong warnings ("black box warnings") were added to the drug's labels in 2007.[14]

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-KDOQI, 2007), and the FDA maintained its suggested range of hematocrit targets at 30–36%. Broadly, it seems that clinicians felt there were health benefits from providing EPO to patients with low red blood cell levels, as well as serious risks from administering high dosages of EPO. Consequently we assume the health production function in our model is first increasing in the dosage and is then decreasing after some point.

The dialysis industry was also undergoing rapid consolidation over the decade from 2000 to 2009, with the largest number of acquisitions occurring in 2006 (see Eliason et al., 2019, for an analysis of the impacts of this consolidation).[15] By 2009, two large chains treated a combined 60 percent of dialysis patients in the US.[16] However, there is scope for variation in dosing decisions across facilities within a chain; for example, federal regulations explicitly state that each facility should have "some authority to individualize corporate policies to address unique facility situations" (ESRD Program Interpretive Guidance 2008, p. 279). Costs also vary across facilities within a chain, including the cost of acquiring EPO. While both chains had purchasing agreements for EPO with its manufacturer (Amgen), there are a number of parts of the agreements that are consistent with there being variation in prices, rebates, and discounts across facilities owned by the same chain, as opposed to a single corporate rate.[17] Additionally, annual facility-level cost reports submitted to Medicare show this variation in per-unit prices for EPO within chains (see Section 4).

Accordingly, in the empirical analysis, we treat each dialysis center as an independent

[13] Eliason et al. (2022) confirms these risks and provides evidence on other health effects of EPO, outside the setting of a clinical trial, using a novel instrumental variable.

[14] During our study period of 2008 and 2009, there were no major informational shocks like this.

[15] During 2008 and 2009 there were a relatively small number of acquisitions, 26 and 104, respectively (private communication from Chris Ody).

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

[17] The purchasing agreements are on file with the Securities and Exchange Commission, and redacted versions are publicly available. The agreements covering our study period are located here https://www.sec.gov/Archives/edgar/data/927066/000119312508042304/dex1062.htm (for DaVita) and here https://www.sec.gov/Archives/edgar/data/1333141/000132693207000082/f01549exv4w18.htm (for Fresenius).

entity, with its own marginal cost and degree of altruism. This allows for heterogeneity both within and across chains, and fits naturally with our theoretical framework. However, we examine the robustness of our analysis to this assumption by also estimating a version of the reduced form that includes facility fixed effects, which allows for an arbitrary distribution of these effects within and across chains (e.g., an arbitrary correlation structure). The coefficient estimates are essentially unchanged, which suggests that this issue of independence is not a first-order concern for our analysis. Also, having one common distribution of unobserved provider characteristics fits with having one common contract for all providers.[18] Medicare does not write separate payment contracts for each provider, or for each health system or corporation. Indeed, having separate contracts could be very costly to administer, and could potentially allow for distortions from lobbying by individual organizations.

3 Model

Our model uses a static screening framework, describing an interaction between a principal and an agent. The government (the principal) pays a provider (the agent) to treat a patient. The government seeks to maximize the benefit for patient health minus the cost of a payment to the provider. 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 provider's utility also depends on patient health, weighted by the provider's degree of altruism, along with the cost of administering the treatment and the compensation received.

The patient arrives at the provider with a baseline health status, b, and other relevant characteristics, x. The provider 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.[19] In our application, b is the hematocrit level from the prior month, x represents other patient characteristics that may affect the health benefits and risks of EPO, and a is the total units of EPO administered over the current month; a, b, and x are all observed by the government (and the econometrician) because they are reported in the monthly claims.

Given the patient's health status and other characteristics, the treatment produces health according to the *health function*, $h(a; b, x)$. This function summarizes the overall health benefits and risks of EPO (as perceived by the provider) for a dialysis patient with anemia, such as relieving the effects of chronic anemia versus increasing the risk of cardiovascular

[18] We confirm that a single, unimodal, distribution fits the residuals from our model in Appendix K.3.

[19] Because the medication is administered intravenously while the patient is undergoing dialysis, there is no issue with patient compliance, as opposed to patient adherence to oral medications or diet and exercise.

events, as described in Section 2. Accordingly the function is initially increasing in a but is then decreasing in a after some point. We refer to dosages with negative marginal product ($h'(a;b,x) < 0$) as "medically excessive."[20] The marginal product also depends on the baseline health b and other characteristics x, because patients with lower hematocrit typically need more EPO, and certain characteristics modify the effectiveness and risks of EPO. Last, the function h is assumed to be twice differentiable and strictly concave in a, because patients with more severe anemia benefit more from EPO, while the serious health risks from the drug increase with larger dosages.

The degree of provider altruism, α, gives the provider's marginal rate of substitution between the patient's health and the provider's own income. The provider also has a constant marginal cost of treatment, z. These two attributes are unobserved by the government, and we refer to (α, z) as the provider's *type*. Heterogeneity in altruism captures differences between providers' preferences.[21] The treatment costs reflect the costs of acquiring and administering EPO, both of which can also be expected to be heterogeneous. However, while we allow for both altruism and cost heterogeneity, whether there is substantial heterogeneity along either dimension is an empirical question, which we address in our econometric analysis. The joint distribution of these attributes is $F(\alpha, z)$, with the associated density $f(\alpha, z)$ that is strictly positive and differentiable over a compact set $[\underline{\alpha}, \overline{\alpha}] \times [\underline{z}, \overline{z}] \subset \mathbb{R}^2_+$, where $\underline{\alpha}$ and $\underline{z}$ are strictly positive and $\overline{\alpha}$ and $\overline{z}$ are finite.

The government sets a *payment policy*, which specifies the payment to be made to the provider based on the treatment amount, the baseline hematocrit, and the other observed patient characteristics. The policy consists of a set of potentially nonlinear payment contracts for the treatment amount, $P(a;b,x)$, one for each possible value of (b,x). That a affects the payment amount means we are considering a general form of fee-for-service contracts, and the presence of (b,x) is analogous to risk adjustment in a broad sense. While both fee-for-service and risk adjustment are ubiquitous in health care payment systems, we permit unrestricted flexibility in payment contracts, in contrast to commonly analyzed (e.g., linear) contracts.

The timing is that of a typical screening model. The government sets the payment policy

[20] Note that "medically excessive" is a statement about the production technology h and is distinct from a normative economic concept. We use "overprovision" to refer to economically excessive amounts. However, these concepts are related because a medically excessive amount will always be economically excessive.

[21] As noted in the introduction, what we refer to as "altruism" could include other intrinsic or extrinsic motivations to care about patient health. Providers may also vary in their beliefs about the benefits and risks of EPO. Heterogeneity in beliefs could have similar implications for our analysis as heterogeneity in altruism, because both would be expected to remain invariant in the counterfactual payment contracts we consider. Furthermore, under some specifications, heterogeneity in beliefs could be observationally equivalent to heterogeneity in altruism.

$\{P(a;b,x)\}$, after which the provider's type (α, z) and the patient's baseline hematocrit level (b) and observed characteristics (x) are realized. The provider then decides whether to participate and, if the provider does participate, chooses a treatment amount (a). Finally, the outcomes occur and payoffs are received.

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

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

That is, the provider has quasilinear preferences, a standard assumption (Rochet and Stole, 2003). The provider's reservation utility is $\underline{u}$; this level of utility must be attainable in order for the provider to participate.[22]

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 provider.[23] The government's weight on patient health generically differs from the provider's weight if there is a nondegenerate distribution of α; furthermore, because the government represents the patient, α_g may be larger than the median of α, for example. The government's valuation of the outcome, where the patient has baseline hematocrit b and observed characteristics x, and receives treatment amount a, is as follows:[24]

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

Because the provider's type is not observed, the government considers the expectation of this valuation over the distribution of amounts that would be chosen by different types, given the patient's baseline hematocrit b and other characteristics x.

We use Bayesian Nash equilibrium to define behavior. The provider chooses a treatment amount to maximize utility function (1) given their type, the patient's baseline health,

[22]Note that $P(a;b,x) - za$ (which corresponds to profits if z is a purely 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, z includes non-pecuniary components; moreover, the dialysis centers provide many services, making it reasonable to allow for negative profits from the provision of EPO.

[23] As is standard in these models, the principal's objective does not include the agent's objective, meaning it does not represent social welfare. If the agent's objective were included, there would be no distortions from the efficient allocation. This is different from the optimal regulation literature, where distortions are introduced via asymmetric weights on consumer surplus and profits (Baron and Myerson, 1982) or a cost of funding the regulation program (Laffont and Tirole, 1986).

[24]This valuation does not include the costs of other "downstream" medical care, such as transfusions and hospitalizations, that may be affected by changes in dosages of EPO. We use a simple calculation to examine how those costs might change under the optimal contracts in Section 6, where the dosages decrease, and find that there would be a modest reduction in downstream costs as well.

and the payment policy (the incentive compatibility constraint). The provider also decides whether to participate, and does not participate if the maximum possible utility would be below the reservation utility (the voluntary participation constraint).[25] The government sets the payment contract for each (b, x), knowing how each provider type would respond. Thus, given (b, x), the government's problem is to maximize the expected value of (2), subject to the provider's incentive compatibility (IC) and voluntary participation (VP) constraints, which must hold for each type:

$$\max_{P \in \mathcal{P}} \int_{\alpha,z} \left[\alpha_g h(a^*(\alpha, z; b, x, P); b, x) - P\left(a^*(\alpha, z; b, x, P); b, x\right)\right] f(\alpha, z) d\alpha dz$$

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

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

where the set of possible payment contracts, $\mathcal{P}$, is the set of real functions.

The presence of the participation constraint means that a "forcing contract" that only reimbursed the provider for a specific treatment amount could not be optimal, even though it is in the set of possible payment contracts.[26] While requiring voluntary participation is standard in the literature, this assumption also speaks to the government's concern that providers be available to see patients. It is important to Medicare to have all dialysis providers accept Medicare patients, and a forcing contract that compensated the providers based on any type but the "worst" type, $(\underline{\alpha}, \overline{z})$, would lead to some providers choosing not to participate.[27]

Next, we turn to the solution of the model. First, we characterize the first-best allocation, which would occur under full information. We then solve the model under asymmetric information, starting with the provider's behavior, and then presenting our approach to derive the optimal contract, which results in the second-best allocation. This analysis is presented for a single value of the baseline hematocrit and patient characteristics, and so b and x are suppressed for the remainder of this section. Also, we focus on interior solutions here to clarify the exposition. When solving the model for the empirical analysis, we allow

[25] We make the natural assumption that the treatment amount is zero if the provider does not participate (this assumption only affects off-equilibrium behavior).

[26] For example, consider a contract that only compensated the provider for choosing the maximum full-information amount, which would be the treatment amount chosen by the "best" type, $(\overline{\alpha}, \underline{z})$, under full information (see page 15). While this contract could induce the efficient allocation for the best type (we show this also occurs under the optimal unrestricted contract), this type is only of measure zero. Meanwhile, all other types would have to be paid more than it was worth to the government to have them participate.

[27] Even without voluntary participation constraints, the government might still 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.

for corner solutions where some provider types administer zero units (see Appendix D). We follow the screening 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 provider type, denoted $a^{*\text{FI}}(\alpha, z)$. The interior optimality condition is

$$\alpha_g h'(a^{*\text{FI}}(\alpha, z)) = z - \alpha h'(a^{*\text{FI}}(\alpha, z)). \tag{3}$$

The efficient allocation equates the principal's marginal benefit (left side) with the agent's marginal cost (right side), as is standard, but in this case the relevant marginal cost is the effective, or "net," marginal cost, which includes the effect of altruism. Unlike typical asymmetric information models with non-altruistic agents, here the agent derives utility from the same outcome as the principal does, and so the agent's marginal benefit from that outcome appears in the condition because it reduces the total marginal cost experienced by the agent. The efficient allocation will never have medically excessive amounts (i.e., where $h' < 0$); therefore, the facts that the provider's altruism weight is positive and that h is strictly concave imply that treatment amounts in the efficient allocation are higher with altruism than without.

3.2 Provider Behavior

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

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

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

nonlinear contract, a^* is increasing in α and decreasing in z.

To see how the payment contract affects behavior by different types of providers, it helps to start with a linear contract. Let $P^L(a) \equiv p_0 + p_1 a$ denote an arbitrary linear contract, where p_0 is a lump-sum payment, and p_1 is a constant marginal payment (i.e., the per-unit payment rate). Then rearranging (4) to $\alpha h'(a^*) = z - p_1$, it is apparent that all provider types with marginal costs below p_1 would administer 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 provider administer a treatment amount closer to the health maximizing amount, due to the strict concavity of h.

3.3 Optimal Contract

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

Instead, we use an analog of the "demand profile" approach (Goldman et al., 1984; Wilson, 1993), which reformulates the principal's problem in terms of finding the marginal payments for each possible quantity. The power of this approach is that it projects a multidimensional distribution of agent types onto a one-dimensional distribution of quantities, and the solution for each quantity can be found separately when certain conditions are satisfied.

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

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

In essence, this integral is an infinite sum of the government's marginal valuation of each amount (i.e., the derivative of (2) with respect to a, which is inside the square brackets), where each amount is weighted by the probability of receiving that amount, $S(p, a)$. 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 provider is a type that will administer a treatment amount of at least a, given the payment contract. In that case, the government will receive its marginal valuation from amount a, which is $\alpha_g h'(a) - p(a)$.

The set of potential treatment amounts, A, is an interval spanning zero, which corresponds to the amounts from excluded types, to $\bar{a}^{*\text{FI}} \equiv a^{*\text{FI}}(\overline{\alpha}, \underline{z})$, the amount that would be provided by the "best" type (highest altruism, lowest cost) under full information. We show in Appendix C.3 that the standard "no distortion at the top" result is obtained (i.e., the highest amount is undistorted) and that all other types' treatment amounts are downwards-distorted in the second-best allocation. This means that $a^{*\text{FI}}(\overline{\alpha}, \underline{z})$ is the maximum equilibrium treatment amount under the optimal nonlinear contract.

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

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

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

Figure 1 suggests that this approach may be more broadly useful for solving screening problems with multidimensional heterogeneity. The demand profile approach has mainly been applied to monopoly pricing problems, but there the single-intersection condition can be more difficult to satisfy because both the consumer demand curves and the marginal price curve are typically downward sloping (see, e.g., Deneckere and Severinov, 2015, for discussion). By contrast, because marginal cost curves are typically upward sloping, the condition

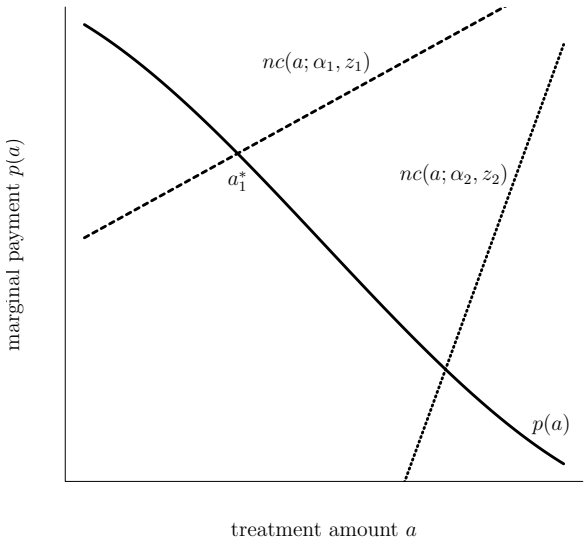


Figure 1: Example marginal payment contract and provider supply curves.

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., $z_1 = z_2$.

can be easier to satisfy in monopsony applications (i.e., purchasing goods or services).[28]

Next, using the distribution of treatment amounts generated by (6), the government's problem (5) is solved separately for each treatment amount. In addition to the single-intersection condition, this relies on the quasilinearity of the agent's preferences (i.e., no income effects), a standard assumption in screening models. Specifically, the provider's marginal utility at amount a does not depend on the marginal payment for any other amount, so the effect of $p(a)$ on the supply of amount a does not depend on the payments for other amounts.[29] The separate problems for each treatment amount are thus

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

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

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

[29]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.

Finally, the optimal contract is characterized by the first-order condition of (7) for each amount, treating $p(a)$ as a parameter:[30]

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

This equates the marginal benefit from increasing $p(a)$ (the change in the probability that at least amount a is provided, $\frac{\partial S(p^*(a),a)}{\partial p(a)}$, times the marginal valuation of that amount, $[\alpha_g h'(a) - p^*(a)]$) with the marginal cost (paying incrementally more if at least amount a is provided, which occurs with probability $S(p^*(a), a)$). The contract is constructed by first solving (8) for $p^*(a)$, for each $a \in A$, and then integrating the marginal payments to yield P^* (see Appendix C.2 for details). The level of P^* is fixed by setting the lowest equilibrium utility equal to the reservation utility, $\underline{u}$, making that type's participation constraint bind.

We present additional intuition about the optimal nonlinear contract and discuss normative aspects of the resulting allocation in Appendix C.3.

4 Data

We now turn to the empirical analysis. Our primary data come from Medicare outpatient claims from renal dialysis centers (freestanding or hospital-based) in 2008 and 2009, for the treatment of patients with ESRD. The raw sample (20% of patients) contains a total of 1.4 million ESRD claims, which are typically filed monthly. Almost 90% of the claims (1.25 million) bill for at least one injection of EPO or a related medication. All claims with an injection include a baseline hematocrit level from the previous month (or a comparable hemoglobin level), but claims without an injection do not report this. As a consequence, we exclude claims without any injections of EPO.[31] Also, in order to avoid extreme outliers, which often reflect data entry errors, we remove observations where the reported amount of EPO is above the 99th percentile. Finally, we restrict to observations where the baseline hematocrit is within a broadly recommended range for using EPO, which is between 30 and 39 percent.[32] This excludes 119,788 observations (10.6% of the remaining total) with

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

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

[32] The FDA-approved labeling for EPO stated a suggested target range for hematocrit of 30 to 36 percent (https://www.accessdata.fda.gov/scripts/cder/daf/index.cfm), and guidelines issued by the National Kidney Foundation recommended the use of hemoglobin targets from 11 to 12 g/dl, and not greater than 13 g/dl (NKF-KDOQI, 2007), which is comparable to hematocrit targets from 33 to 36 percent, and not greater

hematocrit levels below the recommend range, where treatment protocols may have differed, and 87,595 observations (7.8%) above the range, where certain restrictions on reimbursements may also have influenced dosages.[33] The final sample has 919,745 claims, for 74,260 unique patients, from 5,148 unique providers.

The unit of observation is the monthly claim, which reports the services given by provider i to patient j in period t. As discussed in Section 2, we use the dialysis centers as the providers, and the claims are submitted and the payments are received by them. The treatment amount, a_{ijt}, is total amount of EPO administered over the claim period, and the baseline hematocrit, b_{jt}, is the prior hematocrit level reported on the claim.[34] The payment rate, p_{1t}, is the national payment rate per 1,000 units of EPO for the quarter in which the claim was filed. These rates are listed in publicly available Medicare Part B Average Sales Price Drug Pricing Files.[35] Last, the observable patient characteristics, x_{jt}, which may affect the benefits and risks of EPO, are demographics and comorbidities, specifically age, sex, and the Charlson Comorbidity Index (CCI).[36]

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

Table 1 also shows the distribution of the annual average acquisition cost of EPO across dialysis centers, from publicly available Renal Dialysis Facilities Cost Report Data.[37] The

than 39 percent.

[33]Medicare reduced the reimbursement rate by half for EPO provided to patients whose hematocrit exceeded 39 percent for three consecutive months (https://www.cms.gov/medicare-coverage-database/details/medicare-coverage-document-details.aspx?MCDId=11).

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

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

[36]The CCI has been validated for dialysis patients (Beddhu et al., 2000). To construct the index, we apply the implementation from Quan et al. (2005) to Medicare inpatient claims (MEDPAR). Patient age and sex are taken from the Medicare Beneficiary Summary File.

[37]https://www.cms.gov/Research-Statistics-Data-and-Systems/Downloadable-Public-Use-Files/Cost-

Table 1: Summary Statistics

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

Notes: The EPO dosage, EPO payment rate, hematocrit level, and Charlson Comorbidity Index come from Medicare claims data. (†)The EPO acquisition costs are 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. The national payment limit comes from quarterly Medicare Part B ASP Drug Pricing Files for 2008 and 2009.

percentiles show potentially important heterogeneity in acquisition costs, even though the drug was produced by a single manufacturer.[38] As we discuss below in Section 5.2, there are also nontrivial costs of administering EPO (another component of the marginal cost), which are likely to vary across dialysis centers, but which are not well observed in the facility level cost report data.

5 Empirical Implementation

We now describe how we adapt the model from Section 3 to the empirical application, and how we recover the parameters of the empirical specification from the data. The model extends to an environment with many providers, each treating many patients, under the

Reports/Renal-Facility-265-1994-form. CMS requires dialysis centers to submit detailed annual cost reports, which include their total expenditures on EPO and the total number of units provided. From the total expenditures (less any rebates) and total units, we compute the average acquisition cost per 1,000 units of EPO for each center in the cost report data from 2008.

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

natural assumptions that the providers' utility functions and the government's objective function are additively separable across patients.[39] Therefore our earlier results can be used to characterize optimal contracts in this setting. Below, we first develop the empirical specification, then discuss identification and explain the approach used for estimation, and finally present our parameter estimates.

5.1 Empirical Specification

For the empirical analysis, we assume a quadratic specification of the health function, h. This captures the likely non-monotonicity of the effects of EPO, and yields simple, closed-form expressions for the treatment amounts. However, this specification is not crucial because h is nonparametrically identified up to location and scale, and our normative results are invariant to the choice of both (see Appendix E.1). Hence the sign of the marginal effect of treatment is identified (i.e., what dosages are health damaging on the margin).[40] The quadratic specification is as follows:

$$h(a; b, x) \equiv H - \frac{1}{2}[\delta a + b - \tau' x]^2. \tag{9}$$

Here δ is a linear technology that converts the amount of EPO provided, a, into an increase in hematocrit from the baseline level, b. The maximum health is achieved when $\delta a + b$ equals $\tau' x$. While the value of $\tau' x$ could be interpreted as a medical target level for patients with characteristics x, the estimated value should be interpreted with caution because the *level* of τ (i.e., its location) is identified by functional form—unlike the *marginal effects* of x, a, and b, and the shape of h. Finally, the health function includes a positive constant, $H \gg 0$, so that patient health enters positively into provider utility.[41,42]

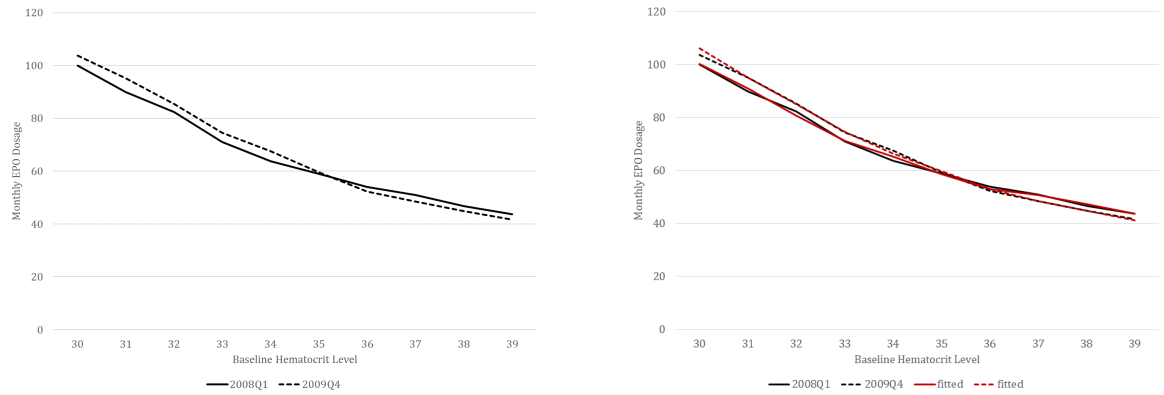
With this quadratic specification, and with a constant marginal payment rate (p_1) as in the linear contracts that were in place during our study period, the provider's first-order

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

[40] To be clear, a non-monotonic h is not necessary for our overall approach. Hence it would be equally relevant for applications where treatments never damage health.

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

[42] It is also worth noting that this specification is robust to certain alternatives, e.g., if providers were partially motivated to minimize their deviations from any particular treatment level.



(a) Observed dosages

(b) Observed and predicted

Figure 2: Mean monthly dosages of EPO in relation to baseline level of hematocrit, and predicted dosages from the estimated reduced form.

Notes: Means calculated by integer value of hematocrit, rounded down. Fitted values are predicted with the estimated versions of the reduced form reported in Appendix Table A3.

condition (4) yields a simple linear solution for the chosen treatment amounts:

$$a^*(\alpha, z; b, x, P^L) = \frac{\tau' x - b}{\delta} + \frac{p_1 - z}{\alpha \delta^2}. \tag{10}$$

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

Equation (10) implies a globally linear relationship between the patient's baseline hematocrit and the amount of EPO provided. To examine this, Figure 2(a) plots average dosages against the baseline hematocrit, separately for the first and last quarters in our data (when the national payment rates were respectively \$8.96 and \$9.58 per 1,000 units). Average dosages are monotonically decreasing in b, which is consistent with our model, but the relationship appears to be somewhat nonlinear, with a steeper slope at lower hematocrit levels. When the payment rate was higher (2009Q4), average dosages are larger for patients with low and medium hematocrit levels, which is also consistent with (10). 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 (10) would predict. While these aggregate plots do not provide ceteris paribus comparisons, they suggest that certain nonlinearities absent from (10) may be empirically relevant.

To capture those potential nonlinearities, our empirical specification adds flexibility in relation to the patient's baseline hematocrit. Specifically, we allow the model parameters to

take different values when b is in different intervals, denoted by k. As a consequence, each interval of baseline hematocrit can be treated separately in the estimation of the model. This approach maintains the linear, closed-form solution, while having sufficient flexibility to fit the nonlinearities quite well, as seen in panel (b) of Figure 2 (discussed further in Section 5.3). To provide some interpretation for this flexibility in the parameters, allowing different values of δ (i.e., δ_k) means that the productivity of EPO may depend on the baseline hematocrit,[43] and the flexibility in τ (i.e., τ_k) means that the benefits and risks of EPO related to patient characteristics may interact with the baseline hematocrit. There are also potentially different distributions of (α, z) (i.e., F_k) for different values of b, which allows there to be different altruism weights and marginal costs depending on the severity of a patient's anemia.

Finally, to allow for unexplained variation from the econometrician's perspective, we add an independent, mean-zero shock, η. Additionally, as we make clear below, it is useful to decompose the marginal cost as $z_{ik} = \mu_z + \zeta_{ik}$. With these extensions to (10), the observed dosage given by provider i to patient j in period t is

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

for a patient whose baseline hematocrit is in interval k. This is the empirical reduced form for the observed dosages, which we take to the data. It can be rearranged to yield reduced-form parameters and disturbances (structural parameters are in the body of the equation, reduced-form parameters are below the brackets):

$$a_{ijt} = \underbrace{\left[\frac{-1}{\delta_k}\right]}_{\beta_1^k} b_{jt} + \underbrace{\left[\frac{1}{\alpha_{ik}\delta_k^2}\right]}_{\beta_{2i}^k} \underbrace{[p_{1t} - \mu_z]}_{\tilde{p}_t} + \underbrace{\frac{\tau_k'}{\delta_k}}_{\beta_3^{k\prime}} x_{jt} + \underbrace{\left[\frac{-\zeta_{ik}}{\alpha_{ik}\delta_k^2}\right]}_{\nu_i^k} + \underbrace{\eta_{ijtk}}_{\epsilon_{ijt}^k}. \tag{11}$$

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 is a piecewise linear function, but it can be estimated separately within each interval.

5.2 Identification and Estimation

In this section, we explain the approach we take to identify and estimate the empirical model. The structural parameters to be recovered are the scalars δ_k, the vectors τ_k, and the joint distributions $F_k(\alpha, z)$, in each interval of baseline hematocrit, $k = 1 \ldots K$. One parameter of the joint distributions, μ_z, the mean of the marginal cost, is assumed to be the same across

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

intervals. As can be seen from the reduced form (11), μ_z is not separately identified from a constant term in τ_k.[44] To identify μ_z, we use external information on average per-unit costs of acquisition and administration of EPO, described later in this section. The other parameters are identified from the reduced-form estimates. The values of δ_k and τ_k follow immediately from the coefficients β_1^k and β_3^k, given a value of μ_z. The joint distribution of α and z in each interval, F_k, is identified from the joint distribution of the random coefficient and random effect, β_2^k and ν^k, as discussed next .

Multiple approaches to recover F_k are possible. For efficiency and computational tractability we use a parametric assumption, summarized as follows (details are in Appendix F). We specify $\ln \alpha$ and z to have a joint normal distribution, so that α has a lognormal distribution with strictly positive support. Then in each hematocrit interval k, there are four unknown parameters of the joint distribution, $\mu_{\alpha,k}$, $\sigma^2_{\alpha,k}$, $\sigma_{\alpha z,k}$, and $\sigma^2_{z,k}$ (while $\mu_{z,k} = \mu_z$ is treated as known from our external information on costs).[45] Using Stein's lemma (Stein, 1981) and properties of the lognormal distribution, these parameters are identified by, and can be recovered analytically from, the first and second moments of the random coefficient (β_2^k) and random effect (ν^k) in the reduced form (11). Those moments are semiparametrically estimated via an auxiliary regression of the residuals of (11), which is derived specifically for this purpose and takes advantage of the panel structure of the data (see Appendix F).

While this parametric approach is tractable and efficient, F_k is in fact nonparametrically identified under the assumption that the idiosyncratic shocks η_{ijtk} (equivalently, ϵ^k_{ijt}) are mean-independent of the observables b, p_1, and x. To provide some intuition, an alternative approach to recover the joint distribution of α and z would be to estimate (11) separately for each provider (within each interval), using the large number of observations per dialysis center. The resulting 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 E.3 for further discussion). We do not pursue this approach because it would be computationally intensive due to the large number of dialysis centers, and the resulting estimates would be much noisier. However, we are able to assess and confirm a key aspect of our parametric assumption about F_k, that a single, unimodal, distribution fits the heterogeneity across dialysis centers (see Appendix K.3). Regardless, this assumption plays no role in the OLS estimation of the mean reduced-form coefficients.

[44] Rearranging (11) and taking expectations, the intercept of the reduced form would be $\tau_{k,0} \cdot \delta_k^{-1} - \mu_z \cdot \delta_k^{-2} \mathrm{E}[\alpha_k^{-1}]$, where $\tau_{k,0}$ is the constant term in τ_k. Hence because $\tau_{k,0}$ and μ_z only appear in the intercept, only their weighted sum is identified.

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

Separately, as noted earlier, the quadratic specification that yields a linear reduced form is not crucial for the identification of our model or the derivation of the optimal contracts. As we show in Appendix E.1, the health function is nonparametrically identified given a single-index assumption (e.g., $\delta a + b - \tau' x$) and an exclusion restriction on costs. So, most importantly, the marginal effects of dosages on a provider's utility and on the government's objective are identified under these general assumptions. These marginal effects fully characterize provider behavior and the optimal contract (see Sections 3.2 and 3.3).

Next, given the empirical specification, the identification of the structural parameters naturally depends on the consistency of the reduced-form estimates. We use OLS to estimate the reduced form, so we are relying on the exogeneity of the observables, b, p_1, and x. The exogenous variation in the baseline hematocrit, b, comes from natural fluctuations within patients over time. There is substantial variation in hematocrit levels from month to month, and providers react to these fluctuations by adjusting dosages (see Section 2 and Appendix K.2). While this natural variation drives our estimates of β_1^k (and hence δ_k), one possible concern about the exogeneity of b and x would be selection of patients to providers, which could make these variables correlated with the provider-level unobservables (i.e., β_2^k and ν^k, which come from α_{ik} and z_{ik}). We assess this by comparing fixed effects estimates of (11) with our OLS results, and find that the coefficient estimates are quite similar (see Section 5.3 for those results and further discussion).

As for the payment rate, p_1, it was set nationally by Medicare each quarter, based on the average sales price of EPO from roughly six months earlier (see Section 2). An individual dialysis center could not affect the national average price, but if demand shocks were substantially correlated across centers and over time, there could be a correlation between p_{1t} and ϵ_{ijt}^k. 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 this absorbs the effects of systematic demand shocks from dialysis centers, the other potential sources of variation in lagged prices that could generate exogenous variation in p_1 would include supply shocks from the drug manufacturer, and demand shocks from other purchasers of EPO.[46]

Finally, as noted earlier, we use external information on costs to determine the value of the mean per-unit cost, μ_z. Given the high price of EPO, most of the cost is from acquisition (i.e., purchasing the drug from a distributor). The Renal Dialysis Facility Cost Report Data presented 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

[46]For example, EPO is also used extensively for chemotherapy patients and for surgery patients.

Table 2: Reduced-Form Coefficient Estimates

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

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

component of μ_z.[47] The cost of administering EPO is also non-trivial. Several time-and-motion studies have been published to assess the cost of administering EPO, and we use estimates from Schiller et al. (2008), which is the most thorough and relevant for our time period. The results from that study imply an average cost of staff time and non-drug supplies for administering EPO equal to \$1.05 per 1,000 units (see Appendix G.1 for details). Adding this to the acquisition cost, we set the value of μ_z equal to \$8.58 per 1,000 units.

The reduced form is estimated 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 separately in each interval, which yields estimates of the variances and covariance of β_2^k and ν^k (see Appendix F). The hematocrit intervals we use for estimation are three percentage points wide (e.g., $30 < b_{jt} \leq 33$), which provides a good balance between the flexibility of the specification and the precision of the estimates. The linear segments fit the global relationship well (Figure 2b), while the key parameter estimates are sufficiently precise (Tables 2 and 4).

5.3 Estimation Results

Our main estimates of the reduced-form coefficients on the baseline hematocrit (β_1^k) and the payment rate ($\bar{\beta}_2^k$) are shown in Table 2 (estimates of the coefficients on the other patient characteristics are shown in Appendix Table A3). To interpret these coefficients, for example

[47] 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.

Table 3: Robustness of Reduced-Form Estimates

Variable (Coefficient)	Provider Fixed Effects			No Patient Observables			Comorbidity Indicators		
	> 30-33	> 33-36	> 36-39	> 30-33	> 33-36	> 36-39	> 30-33	> 33-36	> 36-39
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Baseline hematocrit	-9.22	-6.51	-4.00	-9.61	-6.39	-3.46	-9.24	-6.32	-3.56
(β_1^k)	(0.19)	(0.13)	(0.12)	(0.24)	(0.15)	(0.13)	(0.24)	(0.15)	(0.13)
Reimbursement rate	9.42	5.99	4.67	9.81	6.13	4.26	9.40	6.09	4.08
($\bar{\beta}_2^k$)	(3.00)	(1.95)	(1.85)	(3.20)	(2.04)	(1.92)	(3.20)	(2.03)	(1.91)
Obs. in interval	231,702	405,019	283,024	231,702	405,019	283,024	231,702	405,019	283,024

Notes: Each column is a separate regression. Complete lists of variables and coefficient estimates for each regression appear in Appendix Tables A3 and A4. Asymptotic standard errors in parentheses, clustered on dialysis center.

in the middle interval, a patient with one unit higher baseline hematocrit (say 35 vs. 34) receives 6,320 less units of EPO per month on average. Also in that interval, a one dollar increase in the payment rate (per 1,000 units) would induce providers to increase dosages by 6,390 units per month on average. The linear segments for the three intervals fit the global relationship between the baseline hematocrit and the dosage very well, as shown earlier in Figure 2(b). The average predictions from the linear regressions in each interval are very close to the average observed doses, and there are no apparent discontinuities in the predictions from one interval to the next.

Next we examine the robustness of our reduced-form estimates, as well as the assumption of a common distribution for (α, z) across all dialysis centers. We start by estimating the reduced form with provider-level fixed effects. The results, shown in Table 3, columns 1-3, are quite similar to our main estimates. This suggests that any patient selection on time-invariant provider characteristics (i.e., α and z) does not affect the estimates substantially. As described in Section 2, we believe the high travel costs associated with dialysis care may limit this form of selection. Similarly, the robustness to provider fixed effects also suggests that any heterogeneity across providers in their hematocrit targets (or other treatment protocols) is not substantially affecting our estimates, because such heterogeneity would be absorbed by the fixed effects. In addition, beyond these endogeneity concerns, it is worth noting that the fixed effects allow for an arbitrary distribution of the provider-level unobservable, ν^k, including any correlation structure among these effects within chains. Hence, our key coefficient estimates do not appear to be sensitive to the assumed independence across facilities that comes with our parametric specification of $F_k(\alpha, z)$.

The rest of Table 3 shows the robustness of these coefficients under alternative specifications of the patient characteristics x (age, sex, and the CCI).[48] The regressions reported

[48] Appendix K.1 contains the full estimation results for these alternative specifications, as well as the

in columns 4-6 omit these characteristics entirely, while those in columns 7-9 use separate indicators for each comorbidity rather than for each value of the comorbidity index.[49] The similarity of the key coefficients across these specifications provides some reassurance against misspecification concerns.

We assess the assumption that one common distribution fits the provider-level heterogeneity across all dialysis centers in Appendix K.3. Appendix Figure A7 nonparametrically plots the distributions of the reduced-form residuals from each hematocrit interval. They appear to be unimodal; by contrast, if there was extremely strong dependence in α or z within chains we might instead expect to see more modes (e.g., one each for the large national chains, DaVita and Fresenius, and a third for the rest). We also formally test the unimodality of these distributions, and the assumption is not rejected. Hence we can be fairly confident that any dependence within chains is not so strong that it would invalidate our use of a single, unimodal distribution to describe the heterogeneity across providers.

Our estimates of the structural parameters are presented in Table 4. The estimated values of δ_k, the effect of EPO, can be compared with results from the medical literature, and they seem to be generally consistent with those results. For example, our estimate in the middle interval implies that 1,000 units of EPO raises hematocrit by 0.158 percentage points. This, and the estimates in the other intervals, are similar to estimates of the average productivity of EPO that can be derived from results from clinical trials.[50] 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 (see Figure 2a). The estimates of τ_k must be interpreted more cautiously because, as noted earlier, their identification is dependent on functional form. While the implied patient-level hematocrit targets ($\tau_k' x_{ijt}$) fall within the defined range for hematocrit (i.e., 0 to 100), the means reported in Table 4 are above what might be expected based on clinical guidelines.[51] However, as discussed earlier and detailed in Appendix E.1, our main

results for the main specification with asymptotic standard errors clustered on chains rather than facilities.

[49] We prefer our main specification with indicators for each value of the CCI, because it is a parsimonious way to include interactions among comorbidities (e.g., the coefficient on CCI=2 gives the effect of having two comorbidities). Moreover, the CCI has been validated for dialysis patients (Beddhu et al., 2000).

[50] For example, Tonelli et al. (2003) construct a dose-response curve based on results from five clinical trials, which indicates average productivities ranging from 0.135 to 0.241 depending on the resulting hematocrit level. Also, 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 (on hematocrit) for the two treatment groups in that study (our calculations). More recently, Eliason et al. (2022) have estimated a local average treatment effect of EPO on hematocrit, equal to to 0.64, using facility elevation as an instrument.

[51] 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-KDOQI, 2007), which is comparable to hematocrit targets from 33 to 36 percent, and not greater than 39 percent. These could be interpreted as possible values for the average target in our model ($\tau_k' \bar{x}_k$), assuming the guidelines ignored

Table 4: Structural Parameter Estimates

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

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

results depend on the marginal effects of EPO and the distribution of provider types, which are nonparametrically identified.

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

the cost of providing EPO.

would administer a medically excessive dosage such that $h'(a; b, x) = -1/18.4$. This would be 2,180 units (3.9%) more than the amount that maximizes patient health.[52] The variance of the log of altruism, $\sigma^2_{\alpha,k}$, is significantly greater than zero at conventional significance levels in all baseline hematocrit intervals, meaning that altruism itself varies significantly in each interval.[53] The marginal cost, z, is denominated in dollars, so the estimates of $\sigma^2_{z,k}$ imply standard deviations of marginal costs equal to \$0.69, \$0.93, and \$0.58, respectively, in the three intervals. For comparison, the interquartile range of acquisition costs reported in Table 1 is \$0.92.

It is also possible to make inferences about the values of α and z for individual providers, using the estimated distribution of types and the observed dosages and covariates. Specifically, we can compute a posterior distribution of α and z for each provider, based on the dosages administered to their patients. This is useful because we can then compare the posterior means across different groups of providers, such as the two large chains vs. others, or non-profits vs. for-profits, for example. The details are presented in Appendix H, but the overall results are consistent with widely held views about this industry. The posterior means of α are somewhat lower among DaVita and Fresenius facilities, on average, compared to other providers, as are the posterior means of z. Similarly, the posterior means of α and z are somewhat lower among for-profits compared to non-profits. These differences however are modest in comparison with the overall variation across providers.

Finally, to examine the importance of altruism versus marginal cost heterogeneity, we simulate the distributions of dosages that would occur if only one of these dimensions were to vary.[54] The results indicate that heterogeneity in altruism accounts for more of the variation in dosages. For example, in the middle interval, the standard deviation of dosages is 9.8 thousand units of EPO when both altruism and marginal cost are allowed to vary. When only altruism varies, the standard deviation falls to 6.3. When only marginal cost varies, the standard deviation is 1.6, which is smaller but not negligible. As we show in Section 6.3, the optimal nonlinear contract targets heterogeneity in altruism more than heterogeneity in marginal costs, but both are relevant.

[52] From (9), $h'(a; b, x) = -(\delta a + b - \tau' x)\delta$. Taking the difference between dosages that yield $h' = 0$ and $h' = -1/18.4$ gives $-(\delta \Delta_a)\delta = -1/18.4$, which solves to $\Delta_a = 18.4^{-1} \times 0.158^{-2} = 2.177$. The health-maximizing dosage for a patient with median baseline hematocrit and average characteristics is 56,300 units.

[53] Given that $\sigma^2_{\alpha,k}$ is a known, simple, transformation of $V(\beta_{2i})$ (the variance of the reduced-form coefficient on the reimbursement rate), this is fairly direct evidence of altruism heterogeneity among providers.

[54] Specifically, we simulate dosages allowing altruism to vary according to its marginal distribution, fixing the marginal cost at its mean value, and we simulate dosages allowing marginal cost to vary according to its marginal distribution, fixing altruism at its mean value. The simulations are done separately for each interval, using the median b and mean x in the interval.

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. The improvements we find indicate the potential value of adopting nonlinear payment contracts for certain health care services.

Two final steps are required to compute the optimal contracts.[55] First, we must truncate the estimated type distributions to render them compact, as in the model. We remove the bottom and top 0.5 percent from the symmetric marginal distributions of z_k and the bottom 0.5 percent from the asymmetric marginal distributions of α_k, and we remove an amount from the top of the latter distribution so that the means of $\delta_k^{-2}\alpha_k^{-1}$ still equal the estimated values of $\bar{\beta}_2^k$.[56] Second, we must fix a value for α_g, the weight placed by the government on health relative to money. We do not assume the observed contract is optimal—indeed, we prove that it is not possible to rationalize the observed payment rate with any value of α_g, given our parameter estimates (see Appendix B), so this parameter should not be recovered from the observed payment rates. Instead, we calibrate a value for α_g based on the value of a statistical life year, information on the relationship between hematocrit levels and mortality risk taken from clinical trials on EPO (see Appendix G.2). The resulting value of 52.6 is above the median value of α among the providers, meaning that the principal places more weight on patient health than do most agents.

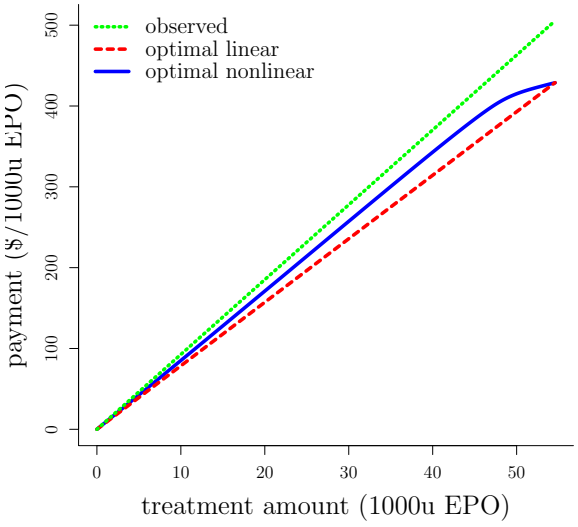
Below, we first present the optimal contracts and resulting dosages in detail (Section 6.1). We compare the optimal nonlinear contracts with the observed contract, and with optimal linear contracts[57] also computed using the estimated model parameters.[58] The optimal contracts are defined for each b and x—broadly analogous to risk adjustment—so we present the contracts for the median value of baseline hematocrit in each interval, using the mean patient characteristics from each interval. We then compare the outcomes under these contracts, to examine the gains from optimal contracting (Section 6.2). Finally, we show how the nonlinear contract screens among the different dimensions of physician types (Section 6.3).

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

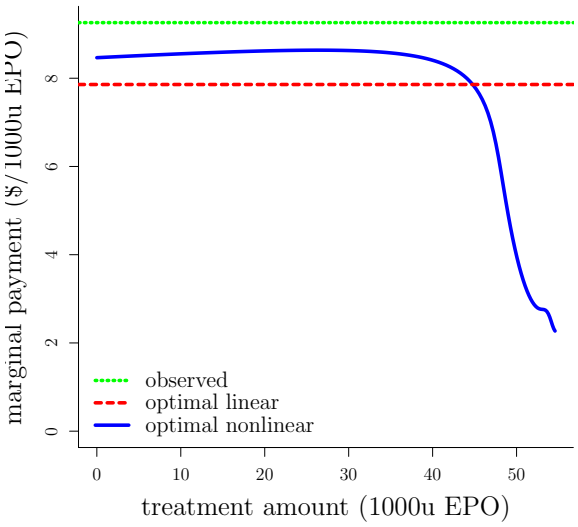
[56] Our results are not very sensitive to the choice of truncation points (see footnote 65).

[57] We show how to solve for optimal linear contracts in Appendix A.

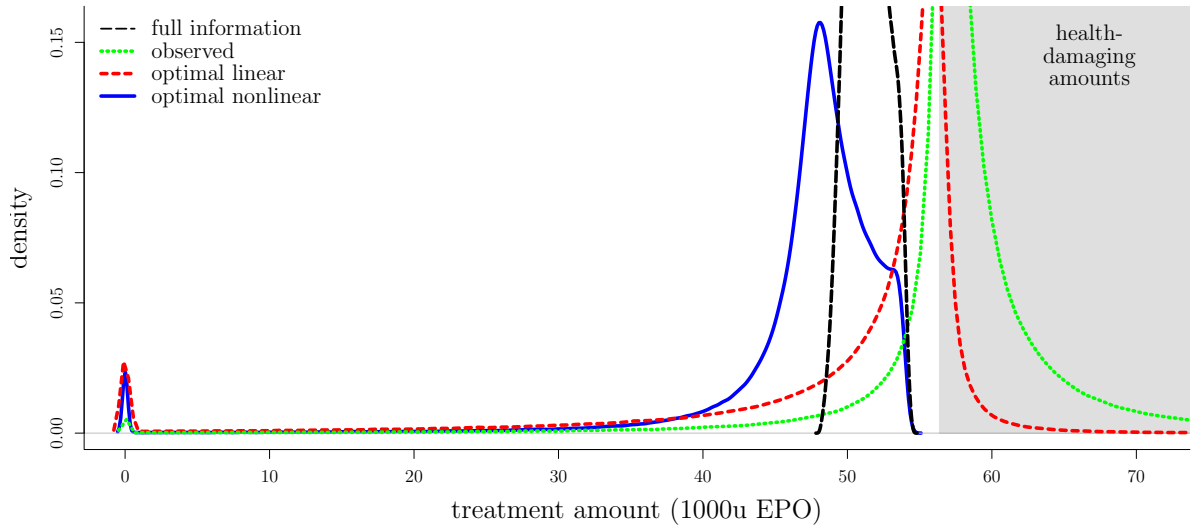
[58] We set the η shock equal to zero in all simulations.



(a) Payment as a function of treatment amount



(b) Marginal payment as function of treatment amount



(c) Distribution of treatment amounts

Figure 3: Treatment and payment amounts under observed and optimal contracts, 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 and mean characteristics in the middle hematocrit interval. 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.

6.1 Optimal Contracts and Distributions of Dosages

We start with the contract for a patient with the median hematocrit level and the mean characteristics in the middle interval. Figure 3 plots the total payments (panel a), the marginal payments (panel b), and the distributions of treatment amounts (panel c), with the optimal nonlinear contract in blue solid lines, the optimal linear contract in red dashed lines, and the observed contract in green dotted lines. For the observed contract, we use the mean of the payment rates in our sample, equal to \$9.26. All three contracts pay \$0 for zero provision. This 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.[59]

For positive treatment amounts, the total payments (panel a) from the optimal nonlinear contract are lower than from the observed contract, and may be higher or lower than the total payments from the optimal linear contract, depending on the treatment amount. The differences in these total payments can be non-trivial: for 45 thousand units, for example, the nonlinear contract would pay \$383.77, the linear contract would pay \$353.60, and the observed contract would pay \$416.70, per month. The marginal payment (panel b) in the nonlinear contract is roughly constant below 40 thousand units, where it lies between the fixed marginal rates of the observed and linear contracts. However, most dosages induced by the nonlinear contract are between 40 and 55 thousand units, where the marginal payment changes substantially, falling from above \$8 to about \$2 per 1,000 units.

The gray shaded area in panel c indicates medically excessive dosages, i.e., treatment amounts with a negative marginal product. This plot also includes the distribution of treatment amounts in the full-information solution for comparison (black, dashed line), which naturally lies strictly below the health-damaging treatment amounts. It is readily apparent that the treatment amounts under the observed contract are typically too high, exceeding the point where the marginal product becomes negative. This accords with concerns that were raised about high payment rates encouraging medically excessive (not just economically excessive) provision of EPO. The optimal linear contract offers a lower payment rate, so the treatment amounts under this contract are less than those under the observed contract. However, it does not eliminate health-damaging amounts, which still occur with 19 percent of providers (see Table 5). That is because, as noted in Section 3.2, any providers with marginal costs below the payment rate will be induced to provide dosages that yield

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

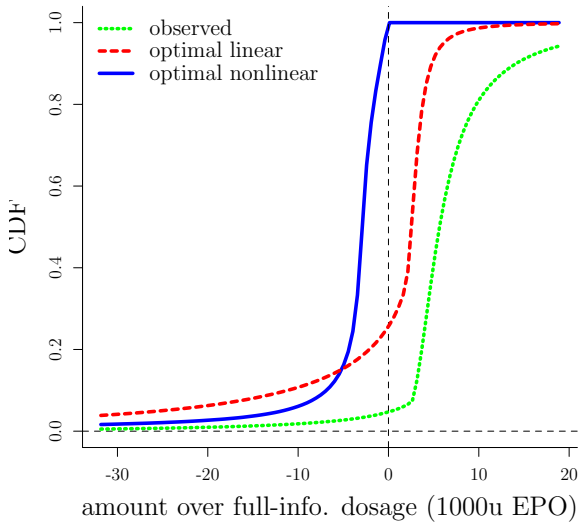


Figure 4: CDFs of deviations from full-information treatment amounts, for patients with median severity of anemia.

Notes: Figure plots the CDFs of the deviations from full-information treatment amounts under the optimal nonlinear contract (blue, solid line), optimal linear contract (red, dashed line), and baseline contract (green, dotted line), for patients with median baseline hematocrit and mean characteristics in the middle hematocrit interval.

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 providers with high marginal costs.

Next, to directly examine over- and underprovision, Figure 4 plots the distributions (across provider types) of the deviations of the treatment amounts provided under each contract from their full information amounts.[60] Overprovision is nearly universal with the observed contract (95.3% of provider types), and it remains very common with the optimal linear contract (74.3% of provider types). In other words, under the optimal linear contract, most providers still administer dosages where the marginal benefit to the principal is below the net marginal cost for the agent. By contrast, there is no economic overprovision with the optimal nonlinear contract: the highest treatment amount equals the maximum in the full-information allocation, and all other treatment amounts are distorted downward (this is a standard result; see Appendix C.3). This further indicates the value of having flexible marginal incentives, because any overprovision is strictly dominated by underprovision in an

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

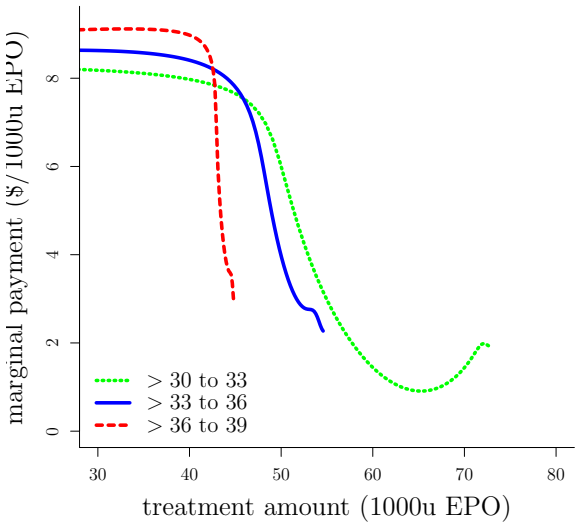


Figure 5: Marginal payments in optimal nonlinear contracts for patients with median severity of anemia in each estimation interval.

Notes: Figure plots marginal payments under the optimal nonlinear contracts for patients with median baseline hematocrit and mean characteristics in the lowest (green, dotted line), middle (blue, solid line), and highest (red, dashed line) baseline hematocrit intervals. Each plot extends up to the maximum dosage provided in equilibrium, which naturally differs with these characteristics.

equal amount, which yields the same health at lower cost.

We now turn to the examples from the low and high hematocrit intervals, to examine how the optimal nonlinear contracts change with the patient's need for treatment. Figure 5 plots the marginal payments in the optimal nonlinear contracts for a baseline hematocrit of 32.0 (green, dotted line) and 37.4 (red, dashed line), along with the contract for the median level of 34.8 (blue, solid line) discussed above. In all cases the payment rate is fairly constant until at least 40 thousand units, and is somewhat close to the observed reimbursement rate. It then drops rapidly, from above $8 per 1,000 units to below $3. However this drop in the marginal payment rate occurs at lower dosages for patients with higher baseline hematocrit (red, dashed line), who need less EPO. Also, notably, the reduction is more gradual for patients with lower baseline hematocrit (green, dotted line), which would induce greater variation in dosages. The fact that these optimal nonlinear contracts differ across the intervals provides a useful insight for policy, which cannot be obtained without computing the unconstrained optimal contracts. For example, the optimal contracts could be approximated with a set of tiered payment rates, where the number of tiers and their levels depended on patient characteristics.

Table 5: Summary of Outcomes under Alternative Contracts in Each Hematocrit Interval

Contract	Mean Payment	Mean Dosage	Std. Dev. Dosage	Share Med. Excessive	Share Overprov.	Gain in Govt. Obj.
Baseline hematocrit >30 to 33						
Observed	740	79.9	12.9	82%	98%	
Optimal Linear	409	60.5	20.6	0%	64%	$184
Optimal Nonlinear	387	54.6	12.9	0%	0%	$219
Baseline hematocrit >33 to 36						
Observed	542	58.6	9.8	75%	95%	
Optimal Linear	396	50.4	11.8	19%	74%	$98
Optimal Nonlinear	393	47.1	7.2	0%	0%	$125
Baseline hematocrit >36 to 39						
Observed	437	47.2	5.3	86%	97%	
Optimal Linear	383	44.6	5.1	46%	87%	$60
Optimal Nonlinear	384	42.9	2.5	0%	0%	$88

Note: Table shows summary statistics of outcomes occurring under the observed, optimal linear, and optimal nonlinear contracts for patients with median baseline hematocrit and mean characteristics in each baseline hematocrit interval. Mean and SD of dosage are in 1,000 units/month. Medically excessive dosages are those that damage health, on the margin, while overprovision refers to economically excessive amounts. The gain in the government objective is computed relative to the observed payment contract.

6.2 Outcomes under Optimal Contracts

Next we consider the outcomes that occur under these contracts, summarized in Table 5. The mean dosages and payments are lower under the optimal contracts than under the observed contract. In all cases, the dosages are lowest under the optimal nonlinear contract, as are the payments in two of the three cases. For the median hematocrit, for example, the mean monthly dosage is 11.5 thousand units lower and the mean monthly payment is $149 lower under the optimal nonlinear contract compared to the observed contract.[61]

Because the medical need is held constant in each example (i.e., b and x are fixed), the variation in dosages indicates the extent to which these contracts address the unobserved heterogeneity across providers. Compared to the observed contract, the optimal nonlinear contract reduces the standard deviation of dosages by 27% and 53% at the medium and high

[61]This reduction in expenditures does not include possible changes in "downstream" medical care, such as transfusions and hospitalizations, which could be affected by changes in dosages of EPO. Making a rough calculation with estimates from other sources, we find that these changes would be predicted to yield an additional net savings of $27 per patient per month (see Appendix K.4 for details). This suggests that the direct savings on EPO may be a somewhat conservative lower bound for the total savings.

baseline hematocrit levels, respectively.[62] By contrast, the optimal linear contract typically does not reduce the variation in dosages, because it provides a constant marginal incentive, just like the observed contract.[63] For comparison, with full information the standard deviations would be much smaller (3.2, 1.3, and 0.4 thousand units for the low, middle, and upper intervals, respectively). The variation that remains with full information reflects the (in this case observable) heterogeneity in altruism and costs, which still affects the optimal amounts, but without any distortions due to informational frictions.

The reduction in the mean and variation of dosages is clearly beneficial to patients. Under the observed contract, around 80 percent of providers would give medically excessive dosages to patients with these baseline hematocrit levels. The optimal linear contract does not eliminate this obvious inefficiency: in the middle and upper intervals, respectively, 19 and 46 percent of providers would give medically excessive dosages to patients under this contract. This inefficiency does not occur with the optimal nonlinear contract because treatment amounts are below their full-information values, all of which are strictly below what would be medically excessive (due to positive marginal costs of treatment and positive, finite, altruism).

The last column of Table 5 shows the government's gains from better contracting, by calculating the increases in the government's objective relative to its values 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 discussed above.[64] There are substantial gains from using the optimal nonlinear contract, ranging from \$88 to \$219 (or roughly \$1,050 to \$2,600 per patient per year) in these examples.[65] Compared to the mean monthly payments under the observed contracts of \$437 to \$740, these gains would represent clear improvements for the government and the patients it represents. The optimal linear contracts achieve 70 to 85 percent of the gains from optimal nonlinear contracts. The mean payments are similar, but the mean dosages are higher under the linear contracts, and the excessive dosages that

[62]The optimal nonlinear contract does not reduce the standard deviation of dosages for the low baseline hematocrit interval, because it excludes a nontrivial share of types, which places a point mass at zero.

[63]The optimal linear contract excludes more types in the bottom and middle intervals than it does in the upper interval, which increases the variation relative to the observed contract.

[64]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. The constant H drops out from the differences shown here.

[65] The gains to the government from using the optimal nonlinear contract instead of the observed contract are very similar even when doubling or halving the truncation tail probabilities for the lower tail of α and both tails of z (we continue to truncate the upper tail of α asymmetrically to maintain the estimated values of $\bar{\beta}_2^k$). In the example from the middle hematocrit interval, the government gains from using the optimal nonlinear contract would be \$121 per patient per month when the doubling truncation tail probabilities and \$130 per patient per month when halving them.

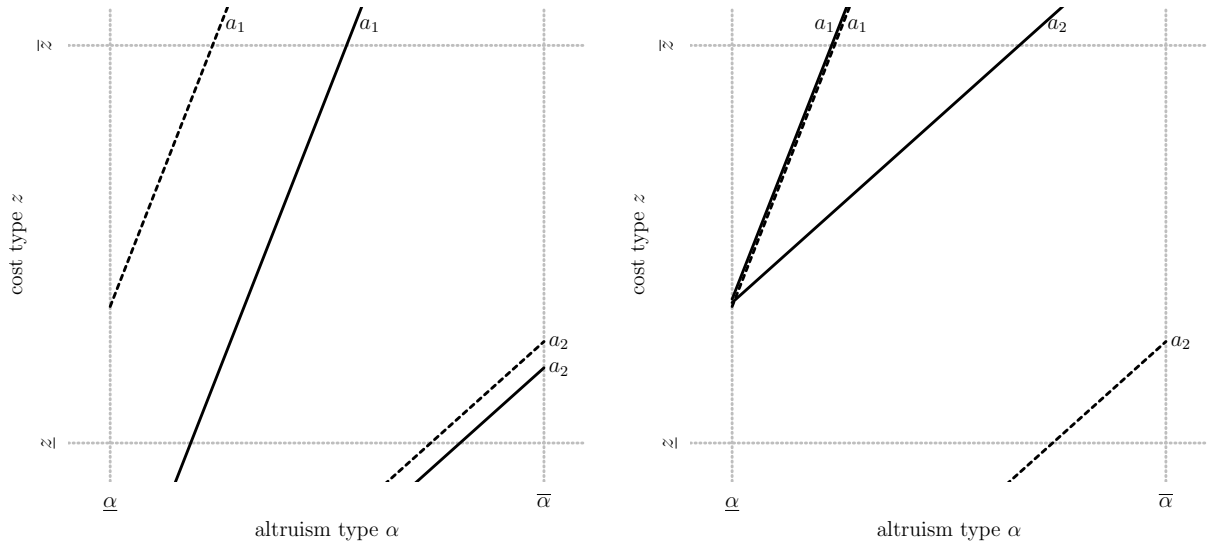
still occur reduce the average gains to health. We can also calculate the gains in the full information scenario. Comparing them to the gains under the optimal contracts provides a measure of the losses due to asymmetric information, which are substantial. The differences between the gains in the full information scenario and the best feasible gains under the optimal nonlinear contract range from \$1,750 to \$3,740 per patient per month.

Given the reduction in variation in dosages (and the elimination of medically excessive treatment amounts) achieved by better contracting, one might be curious about the performance of a forcing contract. To examine this, we have computed the forcing contract implementing the maximum dosage under the full-information allocation, and associated gains to the government over the observed contract for the middle hematocrit interval.[66] To satisfy the voluntary participation constraint for all types, the payment under the forcing contract is larger than even under the observed payment contract, leaving massive information rents to "better" (i.e., higher-altruism and/or lower-cost) types. Accordingly, the gain in the government objective over the observed contract is \$24 per patient per month, a fifth of that under the optimal nonlinear contract. This may not be surprising, as this (and any other) forcing contract was in the set of contracts considered by the principal when solving for the optimal unrestricted contract. The presence of asymmetric information is quite important even when considering only dosages that are not medically excessive.

6.3 Multidimensional Screening in the Nonlinear Contract

Finally, we show how the flexible marginal incentives in a nonlinear contract allow the government to better screen among the different dimensions of unobserved heterogeneity. First note that with either the linear or nonlinear contract, the set of types that will provide some treatment amount a (i.e., an isoquant) is a line in the support of (α, z), because the provider's first-order condition (4) rearranges to $z = p(a) + h'(a)\alpha$. Isoquants under the linear contract rotate around an intercept defined by the constant marginal incentive, while those under the nonlinear contract may have different intercepts, associated with variable marginal incentives (we discuss this in more detail in Appendix C.1). Using the case of the median baseline hematocrit, Figure 6a plots isoquants under the full-information (dashed lines) and second-best allocations (solid lines) for the 75th and 99.99th percentile treatment amounts under the second best, which we respectively denote a_1 and a_2. The higher amount (a_2) is very close to the full-information maximum ($\bar{a}^{*\text{FI}}$) because there is no distortion at the top (see Appendix C.3). The provider types that choose at least a_1 or a_2 lie below (i.e., lower z and higher α) the corresponding isoquants. The isoquants under the optimal

[66] We focus on this treatment amount because it is the highest dosage the government would ever wish to implement. See Appendix L for details about how we computed these results.



(a) Optimal nonlinear contract

(b) Optimal linear contract

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

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

nonlinear contract are below the corresponding isoquants under full information because of the downward distortions induced by the optimal contract, which are larger at the lower amount (a_1). This is in contrast to the optimal linear contract (panel b), which can result in overprovision. In particular, under the linear contract the isoquant for a_2 lies far above the full-information isoquant, indicating that a considerable share of types provide more than the full-information maximum ($\bar{a}^{*\text{FI}}$). (The virtual coincidence of the isoquants for a_1 under full information and the optimal linear contract is itself coincidental.)

One way to see how the nonlinear contract better screens among types is to project the set of types choosing treatment amounts of at least some particular amount onto each axis. First, under full information, there exist combinations of (α, z) such that all altruism types and all cost types would provide at least a_1, while only strict subsets of both dimensions would provide at least a_2. The optimal nonlinear contract discriminates more among altruism types than cost types for a_1, because all cost types would provide at least this amount while only a strict subset of altruism types would. Also the optimal nonlinear contract gets quite close to the full-information allocation for a_2. The optimal linear contract discriminates far less among types along either dimension because the isoquants rotate around a single intercept.

Hence the sets of altruism types and cost types that would provide at least a_1 or a_2 equal the full ranges in each dimension. Thus the flexible incentives provided by the nonlinear contract allow the government to better address the multidimensional heterogeneity, and we learn that the optimal contract discriminates more on provider altruism.[67] This points to the value of using more flexible contracts for health care providers, who likely differ in multiple ways that are unobservable to a payer.

7 Summary and Conclusions

In this paper we examine contracting in health care, a large sector of the economy where asymmetric information is pervasive and where providers' responses to incentives can have important impacts on both health and costs. We specifically examine the provision to dialysis patients of an important and expensive drug used to treat anemia.

By empirically applying results from the literature on screening models, we are able to characterize optimal payment contracts, which in concept induce provision of the best feasible dosages of the drug. Health care providers are likely heterogeneous in multiple ways (as are agents in many other applications), and our use of the demand profile approach naturally accommodates this. Our results indicate there is significant asymmetric information, and hence substantial potential for Medicare (and in principle other payers) to generate considerable savings and improve patient outcomes via better contracting with providers.

We find that moving from the observed contract used by Medicare to the optimal contract completely eliminates medically excessive dosages (given to the overwhelming majority of patients under the observed contract) and reduces spending by 12%, 27%, and 48%, respectively, for the lower, middle, and upper baseline hematocrit intervals, leading to substantial gains from better contracting. Multiplying the gains in our examples by the total number of patient-months in each interval to make a rough approximation, we find that the total gains could be on the order of $300 million per year.[68] To put this in context, Medicare spent

[67] Given the importance of altruism it would be natural to ask whether the heterogeneity in cost types matters. To assess this, we substantially reduced the variance of z from its baseline value, and recomputed the optimal nonlinear contract in this counterfactual environment (see Appendix M). The government's gain from moving from the observed contract to the optimal nonlinear contract would be 10% higher when using the baseline optimal nonlinear contract than it would be when instead the government moved to the contract derived under the counterfactually low variance of z. That is, the government would have a 10% higher objective when designing the optimal payment policy to take into account both dimensions of unobserved heterogeneity.

[68] To arrive at this number, we multiply the gains in each of our examples by the number of patient-months in each interval, divide by two to get an annual average (because there are two years of data), and then multiply by five (as we have a 20% sample of beneficiaries). These calculations use the median b and mean x in each interval, and can thus be interpreted as the gains for a representative patient.

almost $2 billion per year on EPO for ESRD patients during our study period. We also find that there are substantial costs borne due to asymmetric information, ranging from $1,750 to $3,740 per patient per month.

This approach to contracting could prove particularly valuable in improving how Medicare pays for provider-administered drugs (through the Part B program), which is widely acknowledged to be problematic with regard to both dosing and spending. While the general nonlinear contracts we derive may seem complex, these results can provide guidance for simple approximations of the optimal contracts, such as a set of tiered payment rates. Moreover, as Clemens et al. (2017) show, private insurers commonly benchmark their payment contracts to Medicare for many services, so if Medicare adopted these new forms of contracts private insurers might very well follow suit. This approach can also extend more broadly to other forms of treatment. The key requirements are that medical decisions primarily relate to the quantity of treatment, not the type of treatment, and that the quantity of treatment is observable; both are likely satisfied in a wide variety of important applications. Combined with the results in this paper, this suggests that further exploration by economists of optimal contracting in health care, and other areas, could prove valuable to real world policymakers.

References

Abito, J. M., "Measuring the Welfare Gains from Optimal Incentive Regulation," *Review of Economic Studies*, 87(5):2019–2048, 2020.

Acemoglu, D. and A. Finkelstein, "Input and Technology Choices in Regulated Industries: Evidence from the Health Care Sector," *Journal of Political Economy*, 116(5):837–880, 2008.

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

Ash, E. and B. MacLeod, "Intrinsic Motivation in Public Service: Theory and Evidence from State Supreme Courts," *Journal of Law and Economics*, 58(4):863–913, 2015.

Bach, P. B., "Limits on Medicare's Ability to Control Rising Spending on Cancer Drugs," *New England Journal of Medicine*, 360(6):626–633, 2009.

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

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

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

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.

Brookhart, M., S. Schneeweiss, J. Avorn, B. Bradbury, J. Liu and W. Winkelmayer, "Comparative Mortality Risk of Anemia Management Practices in Incident Hemodialysis Patients," *Journal of the American Medical Association*, 303(9):857–864, 2010.

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

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

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

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

Clemens, J., J. D. Gottlieb and T. L. Molnár, "Do Health Insurers Innovate? Evidence from the Anatomy of Physician Payments," *Journal of Health Economics*, 55:153–167, 2017.

Currie, J. and W. B. MacLeod, "Understanding Doctor Decision Making: The Case of Depression Treatment," *Econometrica*, 88(3):847–878, 2020.

Cutler, D., "The Incidence of Adverse Medical Outcomes under Prospective Payment," *Econometrica*, 63(1):29–50, 1995.

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

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

Einav, L., A. Finkelstein, Y. Ji and N. Mahoney, "Voluntary Regulation: Evidence from Medicare Payment Reform*," *The Quarterly Journal of Economics*, 137(1):565–618, 2021, ISSN 0033-5533, doi: 10.1093/qje/qjab035.

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

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

Eliason, P., "Market Power and Quality: Congestion and Spatial Competition in the Dialysis Industry," 2019, unpublished manuscript.

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

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

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

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

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

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

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

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

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

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

Ho, K. and R. S. Lee, "Health Insurance Menu Design for Large Employers," 2020, unpublished manuscript.

Ho, K. and A. Pakes, "Physician Payment Reform and Hospital Referrals," *American Economic Review*, 104(5):200–205, 2014.

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

Laffont, J.-J. and J. Tirole, "Using Cost Observation to Regulate Firms," *Journal of Political Economy*, 94(3, Part 1):614–641, 1986.

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.

McClellan, M., "Reforming Payments to Healthcare Providers: The Key to Slowing Healthcare Cost Growth While Improving Quality?" *Journal of Economic Perspectives*, 25(2):69–92, 2011.

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

Medicare Payment Advisory Commission, "July 2021 Data Book: Health Care Spending and the Medicare Program," 2021.

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

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

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

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

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.

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.

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

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

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.

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

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.

U.S. Government Accountability Office, "Medicare Part B Drug Spending," 2012, publication No. GAO-13-46R.

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

Whoriskey, P., "Anemia drug made billions, but at what cost?" *Washington Post*, 2012.

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

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

Optimal Contracting with Altruistic Agents: Medicare Payments for Dialysis Drugs

SUPPLEMENTAL APPENDIX

Martin Gaynor
Carnegie Mellon University
and NBER

Nirav Mehta
University of Western Ontario

Seth Richards-Shubik
Lehigh University
and NBER

September 12, 2022

Contents

A Optimal Linear Contract

A.1 Optimal Linear Contract when there is No Exclusion

In this section we solve for the optimal linear contract for the case where no physician types are excluded in equilibrium, i.e., all physicians would choose strictly positive treatment amounts. Although we allow for corner solutions for treatment amounts in our quantitative results, in Section 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 (10), the government's problem can be written as

$$\max_{\{(p_0,p_1)\in\mathbb{R}^2\}} \int_{\underline{\alpha}}^{\overline{\alpha}} \int_{\underline{z}}^{\overline{z}} \left[\alpha_g h(a) - p_0 - p_1 a^*(\alpha, z; p_1)\right] f(\alpha, z) dz d\alpha \tag{A1}$$

s.t.

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

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

We can eliminate the participation constraints for all types but

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

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

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

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

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

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

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

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

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

$$\Rightarrow p_1^* = \frac{\alpha_g \, \mathrm{E}\left[\frac{z}{\alpha^2}\right] + \mathrm{E}\left[\frac{z}{\alpha}\right] - \frac{\ddot{z}}{\ddot{\alpha}}}{\alpha_g \, \mathrm{E}\left[\frac{1}{\alpha^2}\right] + 2 \, \mathrm{E}\left[\frac{1}{\alpha}\right] - \frac{1}{\ddot{\alpha}}}. \tag{A2}$$

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

A.2 Optimal Linear Contract when there is Exclusion

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

$$\max_{\{(p_0, p_1) \in \mathbb{R}^2\}} \mathrm{E}\left[u_g(a(\alpha, z; p_1); p_0, p_1)\right] = \int_{\underline{\alpha}}^{\overline{\alpha}} \int_{\underline{z}}^{\tilde{z}^0(\alpha, p_1)} \left[\alpha_g h(a^*(\alpha, z; p_1)) - p_0 - p_1 a^*(\alpha, z; p_1)\right] f(\alpha, z) dz d\alpha$$
$$+ \int_{\underline{\alpha}}^{\overline{\alpha}} \int_{\tilde{z}^0(\alpha, p_1)}^{\overline{z}} \left[\alpha_g h(0) - p_0\right] f(\alpha, z) dz d\alpha \tag{A3}$$

s.t.

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

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

[2] Note that $\tilde{z}^0 \equiv \tilde{z}(\alpha; p_1, a = 0)$, where $\tilde{z}$ is defined in equation (A6), in Appendix C.2.

(Note that, while we use the more general h notation for the health production function when it simplifies expressions, results here were obtained using the quadratic-loss parameterization of h, specified in Section 5.)

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

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

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

$$\int_{\underline{\alpha}}^{\overline{\alpha}} \int_{\underline{z}}^{\tilde{z}^0(\alpha,p_1^*)} \left[\frac{z[\alpha_g + \alpha]}{\alpha^2}\right] f(\alpha, z) dz d\alpha - \delta[\tau - b] \int_{\underline{\alpha}}^{\overline{\alpha}} \int_{\underline{z}}^{\tilde{z}^0(\alpha,p_1^*)} f(\alpha, z) dz d\alpha = p_1^* \int_{\underline{\alpha}}^{\overline{\alpha}} \int_{\underline{z}}^{\tilde{z}^0(\alpha,p_1^*)} \left[\frac{\alpha_g + 2\alpha}{\alpha^2}\right] f(\alpha, z) dz d\alpha. \tag{A4}$$

B Rationalizability of Observed Payment Rate

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

Unlike the case where there is no equilibrium exclusion under the optimal linear contract

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

(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 $\tilde{z}^0(\alpha; p_1) \equiv \alpha\delta[\tau - b] + p_1$ denote the cost type indifferent between providing treatment and not, given altruism type α and payment rate p_1.[4] Let $p_1^*(\alpha_g; \tilde{z}^0(\cdot, p_1^*))$ denote the solution to (A4), 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 (A2) as $\alpha_g \to \infty$, we have $p_1^*(\infty; \overline{z}) = \mathrm{E}\left[\frac{z}{\alpha^2}\right] / \mathrm{E}\left[\frac{1}{\alpha^2}\right]$, which is a very simple explicit expression that can be evaluated using only model primitives.

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

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

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

[4] Note that $\tilde{z}^0(\alpha; p_1) \equiv \tilde{z}(\alpha; p_1, a = 0)$, where $\tilde{z}$ is defined in equation (A6), in Appendix C.2. This is the same definition as in Appendix A.2, and is reproduced here for convenience.

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

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

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

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

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

Proposition 2 $(p_1^*(\infty;\tilde{z}^0(\cdot,p_1^*)) < p_1^*(\infty;\overline{z}))$**.** *Taking the limit of* (A4) *as $\alpha_g \to \infty$, and after some manipulation and dropping the vanishing terms, we have*

$$\int_{\underline{\alpha}}^{\overline{\alpha}} \int_{\underline{z}}^{\tilde{z}^0(\alpha,p_1^*)} \frac{z}{\alpha^2} f(\alpha,z)dzd\alpha = p_1^* \int_{\underline{\alpha}}^{\overline{\alpha}} \int_{\underline{z}}^{\tilde{z}^0(\alpha,p_1^*)} \frac{1}{\alpha^2} f(\alpha,z)dzd\alpha. \tag{A5}$$

Treating $\tilde{z}^0$ as a parameter, consider how an increase in $\tilde{z}^0$ (towards $\overline{z}$) would affect p_1^* defined in (A5)*. The derivative of the left side with respect to $\tilde{z}^0$ is $\int_{\underline{\alpha}}^{\overline{\alpha}} \frac{\tilde{z}^0(\alpha,p_1^*)}{\alpha^2} f(\alpha,\tilde{z}^0(\alpha,p_1^*))d\alpha$. The derivative of the double-integral expression on the right side with respect to $\tilde{z}^0$ is $\int_{\underline{\alpha}}^{\overline{\alpha}} \frac{1}{\alpha^2} f(\alpha,\tilde{z}^0(\alpha,p_1^*))d\alpha$. Because we have $\tilde{z}^0(\cdot,\cdot) \geq \underline{z} > 1$,[5] the left side will increase more than the double integral on the right side, meaning $\frac{\partial p_1^*}{\partial \tilde{z}^0} > 0$ and, therefore, $p_1^*(\infty;\tilde{z}^0(\cdot,p_1^*)) < p_1^*(\infty;\overline{z})$.* □

Table A1 shows that the upper bound derived above for the optimal linear payment rate is lower than the observed payment rate, 9.26, for the median baseline HCT level in each of the three baseline HCT intervals. Combining this with Propositions 1-2, there cannot exist a value of α_g that rationalizes the observed payment rate for any of these baseline HCT levels. That is, $p_1^*(\alpha_g;\tilde{z}^0(\cdot,p_1^*)) \leq p_1^*(\alpha_g=\infty;\tilde{z}^0(\cdot,p_1^*)) \leq p_1^*(\alpha_g=\infty;\tilde{z}^0(\cdot,p_1^*)=\overline{z}) = \mathrm{E}\left[\frac{z}{\alpha^2}\right] / \mathrm{E}\left[\frac{1}{\alpha^2}\right] < 9.26$.

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

Table A1: Upper bound for optimal linear payment rate

	Baseline HCT interval		
	30-33	33-36	36-39
$p_1^*(\infty;\overline{z})$	8.96	9.10	8.95

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

C Model Details

C.1 Restrictiveness of Linear Contracts

Figure A1 illustrates how the two-dimensional physician types map into treatment amounts, under an arbitrary linear contract and an arbitrary nonlinear contract. With either contract, the set of types that will provide the treatment amount a is a line in the support of (α, z): see that (4) rearranges to $z = p(a) + h'(a)\alpha$. The figure plots two such isoquants for amounts a_1 and a_2, where a_2 is medically excessive.[6] The immediately apparent difference between the linear and nonlinear contracts is that with a linear contract (panel a), the intercept of the isoquants is fixed at p_1, while it can change with the nonlinear contract (panel b) because the marginal payment can vary (e.g., $p(a_1) > p(a_2)$).[7] This suggests the difficulty of designing a linear contract that induces appropriate treatment amounts. For example, a linear contract would have difficulty avoiding medically excessive amounts because the payment rate (p_1) would have to be below the marginal cost of the lowest-cost type ($\underline{z}$) to avoid downward slopes, which would likely exclude a nontrivial share of higher-cost types. Nonlinear contracts can avoid this particular tension because, as illustrated by the isoquant for a_2 in the right panel, the marginal payments for medically excessive amounts (e.g., $p(a_2)$) can be set below the marginal cost of the lowest-cost type ($\underline{z}$), which places such isoquants entirely outside the support of (α, z).

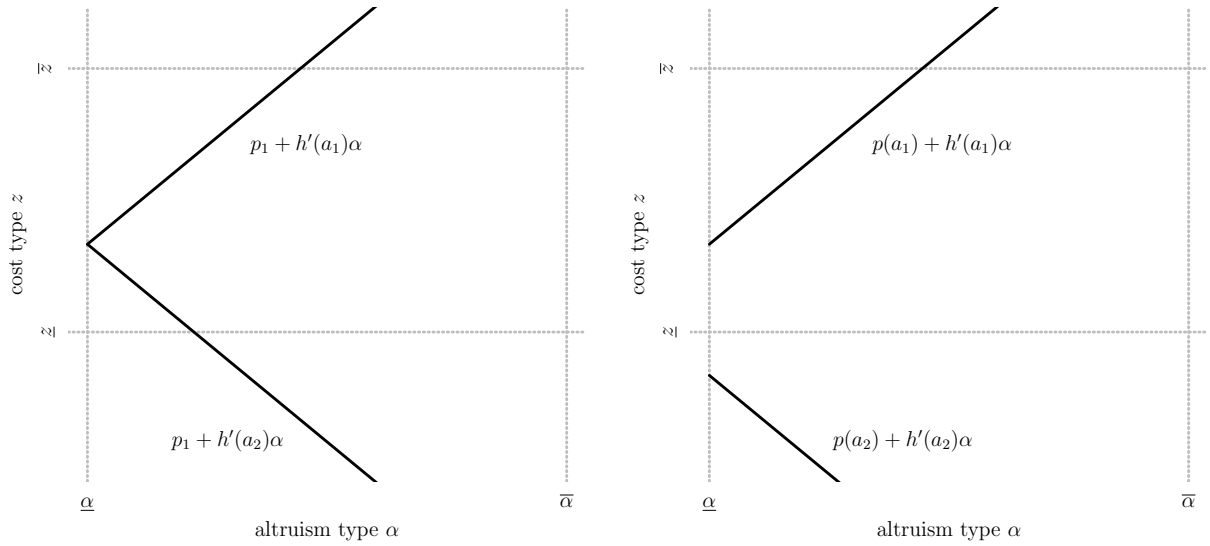
C.2 Details for Solution of Optimal Nonlinear Contract

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

$$\tilde{z}(\alpha; p, a) \equiv p + \alpha h'(a). \tag{A6}$$

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

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



(a) Linear contract (b) Nonlinear contract

Figure A1: Isoquants for example contracts.

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

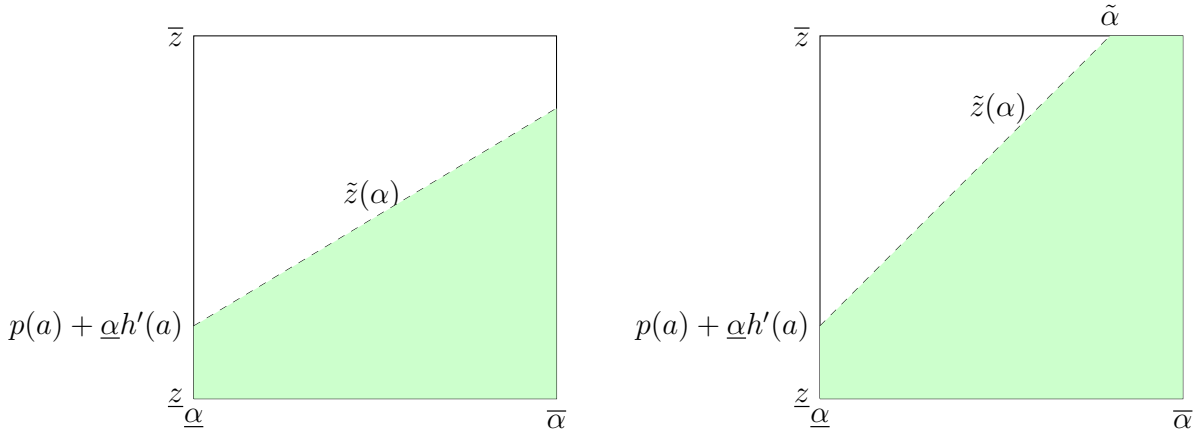
Note that $\tilde{z}$ has intercept p and slope of $h'(a)$, both of which must be non-negative at an optimal solution $p^*(a)$.[8] We also define $\tilde{\alpha}(p, a) = \frac{\overline{z} - p(a)}{h'(a)}$ as the altruism type satisfying $\tilde{z}(\tilde{\alpha}) = \overline{z}$. Suppose that $\tilde{z}(\underline{\alpha}) \geq \underline{z}$. As Figure A2 shows, there are two cases, corresponding to $\tilde{\alpha}$. If $\tilde{\alpha} \geq \overline{\alpha}$, as depicted on the left, then

$$S(p, a) = \Pr\{\underbrace{\underline{\alpha} h'(a) + p}_{\tilde{z}(\underline{\alpha}; p, a)} \geq z\} = \int_{\underline{\alpha}}^{\overline{\alpha}} \int_{\underline{z}}^{\tilde{z}(\alpha; p, a)} f(\alpha, z) dz d\alpha, \tag{A7}$$

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

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

Figure A2: $\tilde{\alpha}$ cases



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

$$S(p,a) = \int_{\underline{\alpha}}^{\tilde{\alpha}(p,a)} \int_{\underline{z}}^{\tilde{z}(\alpha;p,a)} f(\alpha,z)dzd\alpha + [1 - F_\alpha(\tilde{\alpha})], \tag{A8}$$

where F_α denotes the marginal CDF of α.

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

$$\frac{\partial S(p,a)}{\partial p} = \int_{\underline{\alpha}}^{\overline{\alpha}} f(\alpha, \tilde{z}(\alpha;p,a)) \underbrace{\frac{\partial \tilde{z}(\alpha;p,a)}{\partial p}}_{1} d\alpha. \tag{A9}$$

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

$$\frac{\partial S(p,a)}{\partial p} = \int_{\underline{\alpha}}^{\tilde{\alpha}} f(\alpha, \tilde{z}(\alpha;p,a))d\alpha. \tag{A10}$$

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

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

[10] Although not depicted in Figure A2, when $\tilde{\alpha}(p,a) \geq \underline{\alpha}$, it is possible that $\tilde{z}(\underline{\alpha}) < \underline{z}$. Here, the integration limits for α must be adapted to account for $\tilde{z}(\alpha)$ crossing the α axis from below. Let $\check{\alpha}(p,a) \equiv \frac{\underline{z}-p}{h'(a)}$ denote the altruism type satisfying $\tilde{z}(\check{\alpha}) = \underline{z}$. (Note that the condition $\tilde{z}(\underline{\alpha}) < \underline{z}$ is equivalent to $\check{\alpha}(p,a) > \underline{\alpha}$.) There are two subcases. First, if $\check{\alpha}(p,a) > \overline{\alpha}$, then even the most altruistic physician type would not provide the level of treatment under consideration at marginal transfer p, meaning $S(p,a) = 0$ and $\frac{\partial S(p,a)}{\partial p} = 0$. Second,

C.3 Intuition and Normative Aspects of the Optimal Contract

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

$$\frac{\alpha_g h'(a) - p^*(a)}{p^*(a)} = \frac{1}{\eta(a)}, \tag{A13}$$

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 (A13) to the Lerner Index for monopoly pricing, i.e., $\frac{p-c'}{p} = \frac{1}{\eta}$, where p and c' are, respectively, the marginal price and marginal cost and η is the elasticity of demand. Our expression differs from that because the government is a monopsonist and, instead of a marginal cost of production c', the government has a marginal valuation of treatment, $\alpha_g h'$. Intuitively, the principal's objective is lower (i.e., it extracts less surplus) where supply is more responsive to price changes (i.e., the elasticity of supply is larger).

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

$$\underbrace{\alpha_g h'(a)}_{\text{Principal's MB}} = \underbrace{z_i - \alpha_i h'(a)}_{\text{Agent's net MC}} + \underbrace{\frac{S(p^*(a),a)}{\frac{\partial S(p^*(a),a)}{\partial p}}}_{\text{distortion}}, \tag{A14}$$

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 (A14) to show that the allocation under the optimal nonlinear contract will be downward-distorted from the first-best for all but the highest-amount type, $(\overline{\alpha}, \underline{z})$.[11] Equivalently, for any amount $a < \overline{a}^{*\text{FI}}$, fewer types choose a in the second-best because they

---

if $\check{\alpha}(p,a) \in (\underline{\alpha}, \overline{\alpha}]$ then, if $\tilde{\alpha} \geq \overline{\alpha}$ then (A7) becomes

$$S(p,a) = \int_{\check{\alpha}}^{\overline{\alpha}} \int_{\underline{z}}^{\tilde{z}(\alpha;p,a)} f(\alpha,z) dz d\alpha, \tag{A11}$$

and if, instead, $\tilde{\alpha} \in [\underline{\alpha}, \overline{\alpha})$, then (A8) becomes

$$S(p,a) = \int_{\check{\alpha}}^{\tilde{\alpha}(p,a)} \int_{\underline{z}}^{\tilde{z}(\alpha;p,a)} f(\alpha,z) dz d\alpha + [1 - F_\alpha(\tilde{\alpha})]. \tag{A12}$$

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

are being distorted downwards. To see this, first recall that $S(p(a), a)$ is the probability the physician would choose at least a. Hence, the numerator of the distortion, $S(p^*(a), a)$, is strictly positive for all but the maximum treatment amount, which is only provided by the highest-amount type (which has a measure of zero). Also the denominator of the distortion, $\frac{\partial S(p^*(a),a)}{\partial p(a)}$, is positive because the probability in (6) increases with $p(a)$. Hence the right side of (A14) is larger than the right side of (3) for all but the highest-amount type. Because h is strictly concave, the second-best treatment amount is therefore below the first-best amount for all but the maximum treatment amount. $S(p^*(a), a)$ increases as we consider lower dosages, and the distortion typically increases, as well.

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

D Computational Details

D.1 Computation of Optimal Linear Contract

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

$$\max_{\{(p_0,p_1)\in\mathbb{R}^2\}} \mathrm{E}\left[u_g(a(\alpha,z;p_1);p_0,p_1)\right] = \int_{\underline{\alpha}}^{\overline{\alpha}} \int_{\underline{z}}^{\overline{z}} \left[\alpha_g h(a^*(\alpha,z;p_1)) - p_0 - p_1 a^*(\alpha,z;p_1)\right] f(\alpha,z) dz d\alpha \tag{A15}$$

s.t.

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

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

(Note that, while we use the more general h notation when it simplifies expressions, these results were obtained using the quadratic-loss parameterization of h, in Section 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.2, using the BBoptim subroutine contained in the BB package in R (Varadhan and Gilbert, 2009). We solve (8) for a grid of 100 amounts. The lowest value of the grid is zero because we allow for optimal exclusion via the nonlinear contract. The maximum value of the grid is 0.01 below the full-information amount for the highest-treatment-choice type; we use this as the maximum point due to the numerical issues incumbent in evaluating derivatives at the upper corner of the treatment amount space (which is the same as the upper bound of the full-information treatment amount space, due to the downwards-distortion of equilibrium amounts under the optimal nonlinear contract). Finally, we fit a spline to the grid of treatment amounts, which is what we use for our quantitative results.

E Identification

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

We start by studying identification of more general specifications for the health function and the provider type distribution than we use in our empirical implementation (specified in Section 5). We maintain the assumption of quasilinear utility for providers. We also allow for provider-level heterogeneity in the intercept of marginal cost functions, though here we also allow for (homogeneous) convexity in marginal cost functions, which allows for cost functions with heterogeneous convexity. We show that the marginal product of treatment on health, $h'(a; b, x)$, is identified to scale under a single-index specification for the arguments of h',[13] and, therefore, that the sign of the marginal effect of treatment on health is identified. The

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

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

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

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

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

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

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

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

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

The first order condition to maximize utility is

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

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

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

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

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

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

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

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

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

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

With the normalizations, we now have

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

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

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

$$p_1 = \underbrace{\left[c_{0i} + c_{1i} \cdot a_i^* + c_{2i} \cdot (a_i^*)^2 + \cdots\right]}_{\approx \frac{\partial c_i(a_i^*)}{\partial a}} - \underbrace{\alpha_i \left[h_0 + h_{1a} \cdot a_i^* + h_{1b} \cdot b + h_{2a} \cdot (a_i^*)^2 + h_{2b} \cdot b^2 + h_{2ab} \cdot a_i^* \cdot b + \cdots\right]}_{\approx \alpha_i \frac{\partial h(a_i^*; b)}{\partial a}}. \tag{A16}$$

For concreteness, consider the case in which both polynomials were of degree two. Higher-degree polynomials would also be identified (as would be lower-degree ones, like those we use in our empirical implementation). We assume $c_i'(a) = c_{0i} + c_1 a + c_2 a^2$, i.e., we allow for heterogeneity in the intercept of marginal costs and also allow for non-constant marginal

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

costs within provider.[19]

With an infinite number of patients per provider and the mean-independence of η, the equilibrium treatment choices $a_i^*(p_1, b)$ are identified for each observed value of (p_1, b). With the observed variation in (p_1, b), this directly identifies the reduced-form parameters (θ^i, γ^i) listed in braces under the rearranged version of (A16) below:

$$p_1 = \underbrace{(c_{0i} - \alpha_i h_0)}_{\theta_0^i} + \underbrace{(c_1 - \alpha_i h_{1a})}_{\theta_1^i} a_i^* + \underbrace{(c_2 - \alpha_i h_{2a})}_{\theta_2^i}(a_i^*)^2 + \underbrace{(-\alpha_i h_{1b})}_{\gamma_1^i} b + \underbrace{(-\alpha_i h_{2b})}_{\gamma_2^i} b^2 + \underbrace{(-\alpha_i h_{2ab})}_{\gamma_{2a}^i} a_i^* b \tag{A17}$$

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

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

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

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

$$\begin{aligned}
\overline{\theta}_0 &= \mu_{c_0} - \mu_\alpha h_0 \\
\overline{\theta}_1 &= c_1 - \mu_\alpha h_{1a} \\
\overline{\theta}_2 &= c_2 - \mu_\alpha h_{2a} \\
\overline{\gamma}_1 &= -\mu_\alpha h_{1b} \\
\overline{\gamma}_2 &= -\mu_\alpha h_{2b} \\
\overline{\gamma}_{2a} &= -\mu_\alpha h_{2ab}.
\end{aligned} \tag{A18}$$

Identification then proceeds as follows:

1. Test any of the restrictions $\overline{\gamma}_1 = 0$, $\overline{\gamma}_2 = 0$, or $\overline{\gamma}_{2a} = 0$. If we reject then we can set the

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

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

scale of α by choosing a positive value for μ_α.[21] We have identified (up to the scale of μ_α) the parameters

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

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

   The single-index assumption implies that

$$\frac{\partial^2 h(a; b)}{\partial a^2} = \kappa_{ab} \frac{\partial^2 h(a; b)}{\partial a \partial b}. \tag{A19}$$

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

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

   so, with the index assumption we have at an optimum

$$[h_{1a} + 2h_{2a} \cdot a_i^* + h_{2ab} \cdot b] = \kappa_{ab} [h_{1b} + 2h_{2b} \cdot b + h_{2ab} \cdot a_i^*]. \tag{A20}$$

   Step 1 identified the parameters on the right of (A20), other than κ_{ab}. We then need to observe at least three vectors of $(b, p_1, a_i^*(b, p_1))$ to exactly identify $h_{1a}, h_{2a}, \kappa_{ab}$; more than three would yield overidentification and, thus, better estimates. (The same argument holds with higher-degree polynomials, but the number of points required increases.)

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

[21] Recall that α_i is non-negative.

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

3. Using the second and third lines of (A18), we can identify c_1 and c_2:

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

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

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

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

   Then, using the first line of (A18), we have

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

   i.e., h_0 is identified to scale.

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

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

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

Identification of the sign of h' Here we note that the identification of the sign of h' (i.e., whether treatments are health improving or health damaging on the margin) does not rely on the scale normalization. Rearranging (A16), we have

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

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

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

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

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

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

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

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

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

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

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

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

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

We then calibrate α_g from the expression

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

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

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

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

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

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

Recall our assumption that η is mean-independent of (b, x, p_1): $\mathrm{E}(\eta|b, x, p_1) = 0$.[23] Then OLS estimation of the reduced form (11), separately for each provider, yields consistent estimates of β_1, β_{2i}, β_3, and ν_i for arbitrary provider i. The structural parameters and provider types are continuous functions of reduced-form parameters and variables, as follows:

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

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

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

As noted in Section 5.2, we recover $F_k(\alpha, z)$ under a distributional assumption, where $\ln \alpha$ and z have a joint normal distribution. Here we show how we estimate the parameters of that distribution, which are recovered from the first and second moments of the random coefficient (β_2^k) and random effect (ν^k) in the reduced form (11). First we present an auxiliary regression of the residuals of (11) that yields the second moments of β_2^k and ν^k (while the mean of β_2^k comes directly from (11), and the mean of ν^k is zero). Then we derive closed-form expressions for the parameters of $F_k(\alpha, z)$ as functions of these moments.

To develop the auxiliary regression, let $\bar{\beta}_2^k$ denote the mean of β_2^k, and decompose the

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

random coefficient as $\beta_2^k = \bar{\beta}_2^k + \tilde{\beta}^k$. Then (11) can be rearranged as

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

(for b_{jt} in interval k). The OLS coefficient on $\tilde{p}_t$ is a consistent estimate of the mean of the random coefficient, $\mathrm{E}(\beta_2^k)$, under the assumptions discussed in Section 5.2. The auxiliary regression then uses the composite residual, r_{ijt}^k, times the provider-level mean residual, $\bar{r}_i^k$ (taken within interval k), as its dependent variable. This yields consistent estimates of the second moments, $\mathrm{V}(\beta_2^k)$, $\mathrm{V}(\nu^k)$, and $\mathrm{Cov}(\beta_2^k, \nu^k)$, as we show next.[24]

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

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

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

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

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 $\tilde{p}_t \overline{\tilde{p}}_i$, $\tilde{p}_t + \overline{\tilde{p}}$, a constant, and $\frac{1}{n_i}$.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Therefore,

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

results together, we have

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

Therefore

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

□

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

G Calibrations

G.1 Calibration of μ_z

As described in the paper, we use external information on the costs of acquiring and administering EPO to calibrate the value of the mean per-unit cost, μ_z. For the acquisition cost, we use the median across facilities of the per-unit cost of purchasing the drug from a distributor (net of discounts and rebates), computed from Renal Dialysis Facility Cost Report Data, which is equal to \$7.53 per 1,000 units (Table 1). For the administration cost, we compute an average per-unit cost of staff time and non-drug supplies based on results from Schiller et al. (2008), as follows. Schiller et al. (2008) reports an average cost for EPO administration of \$3.63 per dialysis session, and an average of 13.0 sessions per month, for a total cost of \$47.19 per month. From our claims data, the median dosage per month is 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.) Dividing \$47.19 by 45,000, we arrive at an average administration cost of \$1.05 per 1,000 units. Adding this to the acquisition cost, we obtain a value of μ_z equal to \$8.58 per 1,000 units.

G.2 Calibration of α_g

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

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

If we assume how the targets used in the trial relate to the true point where health is maximized (i.e., where $h'(a; b, x) = 0$), we can compute the difference in health from the two targets, as defined by our quadratic specification. We assume that the lower target used in the trial is equal to τ, where health is maximized, implying that the difference in health from the two targets is equal to 7.6, as follows:

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

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

If we further assume that the government's value of this difference in health comes entirely from the difference in the risk of death or a debilitating health event, we can find the monetary value of this difference in health by multiplying a VSLY estimate by the difference in these risks from the two target levels. Aldy and Viscusi (2008) provides VSLY estimates

of approximately \$300,000 (p. 580), so the annual value of the difference in risks would be $0.016 \times \$300,000 = \$4,800$. Finally, because the time periods in our model are months, this would equal the government's value of the above difference in health over twelve periods. Therefore, we have

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

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

H Posterior Means of α and z

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

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

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

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

$$\hat{\eta}_{ijt}^s = a_{ijt} - \left[\frac{-1}{\delta_k}\right] b_{jt} + \left[\frac{1}{\hat{\alpha}_{ik}^s \delta_k^2}\right] p_{1t} + \left[\frac{\tau_k'}{\delta_k}\right] x_{jt} + \left[\frac{-\hat{z}_{ik}^s}{\hat{\alpha}_{ik}^s \delta_k^2}\right] \tag{A21}$$

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

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

and similarly for the posterior mean of z. To complete these computations, the estimated

parameters are used in (A21), and the variance $\sigma^2_{\eta,k}$ is set equal to the variance of the reduced-form residuals in that interval.

Table A2 presents summary statistics on these provider-level posterior means, by ownership type and by chain affiliation. Among for-profit dialysis centers, for example, the median of the center-specific posterior means of α is 31.9 and the mean is 36.3, in the bottom interval of baseline hematocrit. By comparison, among non-profit and governmental centers, the median is 32.8 and the mean is 41.5 in that interval, indicating somewhat greater weight placed on patient health, on average. The posterior means of z, the marginal cost, are noticeably lower among for-profit centers, with medians and means below \$8.60 in all intervals, while the medians and means among non-profit and governmental centers are roughly between \$8.70 and \$8.90. However the distributions of α and z also overlap substantially between these two groups of providers. In all intervals, the standard deviations of the provider-level posterior means within each group are much larger than the differences between the medians or means of the two groups.

We see similar patterns comparing providers in the two large chains against all other providers. In almost all cases, the posterior means of α are somewhat lower in DaVita and Fresenius centers, compared to all other centers. The marginal costs are also consistently lower for centers in the two large chains, compared to other centers. The variation in marginal costs is lower within the large chains as well, typically by one quarter to one third. This is broadly consistent with the variation in acquisition costs observed in Medicare cost report data (see footnote 38 in the paper).

I Check of Regularity Condition

Figure A3 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.[27]

J Full Estimation Results and Counterfactuals

This section presents the complete results on the optimal contracts and outcomes under those contracts for the median baseline hematocrit and mean patient characteristics in each of the three hematocrit intervals (30–33, 33–36, and 36–39), using the government's valuation of health, α_g, calibrated as described above.[28] In addition, Table A3 provides the full estimation

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

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

Table A2: Distribution of Provider-Level Posterior Means

	Altruism (α)			Marginal Cost (z)		
	Interval of Baseline Hematocrit			Interval of Baseline Hematocrit		
	>30 to 33,	>33 to 36,	>36 to 39	>30 to 33,	>33 to 36,	>36 to 39
I) Ownership Type						
a) Non-profit and governmental						
Median	32.8	14.8	20.9	8.68	8.72	8.61
Mean	41.5	16.7	30.0	8.88	8.95	8.75
Std. Dev.	88.1	24.2	71.3	0.59	0.81	0.42
b) For-profit						
Median	31.9	15.5	20.5	8.55	8.50	8.58
Mean	36.3	17.3	27.4	8.55	8.55	8.59
Std. Dev.	59.1	25.4	64.0	0.38	0.64	0.28
II) Chain Affiliation						
c) DaVita						
Median	28.9	15.9	22.2	8.49	8.44	8.58
Mean	33.2	17.5	26.3	8.44	8.43	8.58
Std. Dev.	56.4	24.8	41.8	0.29	0.56	0.22
d) Fresenius						
Median	31.1	14.4	18.8	8.55	8.47	8.56
Mean	35.8	16.7	26.4	8.56	8.51	8.53
Std. Dev.	55.4	24.9	74.6	0.34	0.59	0.23
e) Other/Indep.						
Median	35.4	16.7	20.4	8.61	8.65	8.59
Mean	40.0	17.6	28.3	8.73	8.80	8.68
Std. Dev.	66.4	26.3	70.5	0.51	0.74	0.36

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

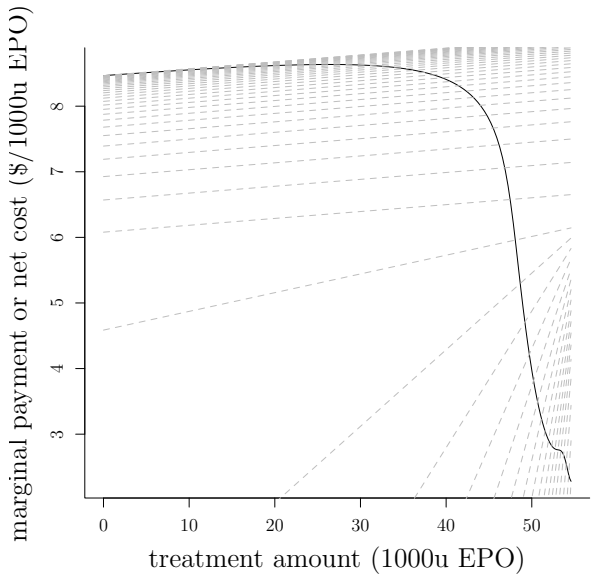


Figure A3: Regularity condition check, for patients with median severity of anemia.

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

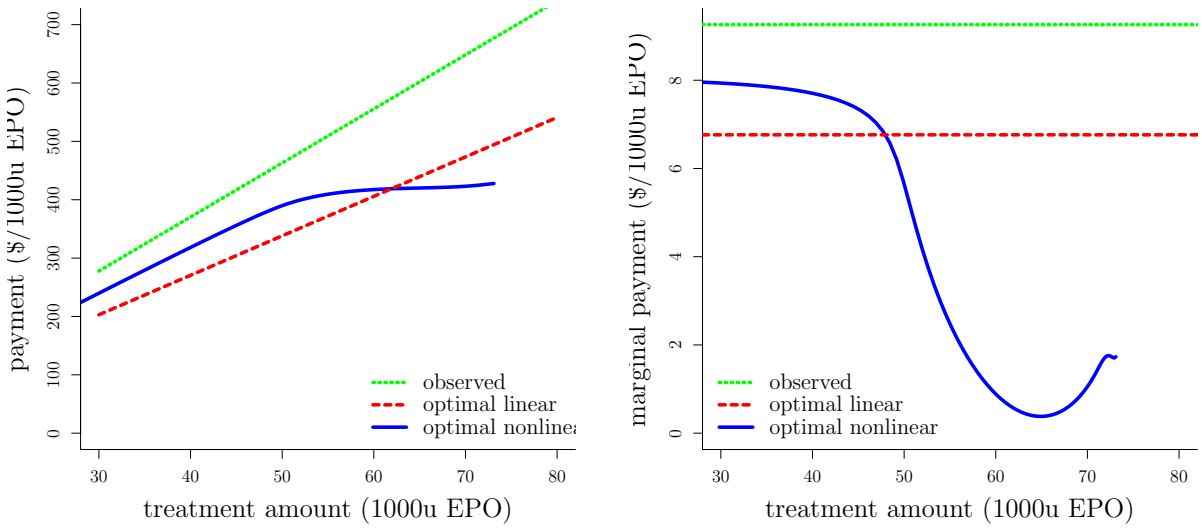
results for our reduced form.

Figures A4 to A6 show the contracts (i.e., the total payments as a function of the treatment amounts), the marginal payments, and distributions of treatment amounts, separately for each interval. They have similar patterns, as discussed in the main text, with the optimal nonlinear contract below the observed contract and intersecting the optimal linear contract. All contracts start at zero dollars for zero units. The reduction in the marginal payment is more gradual in the contract for the low baseline hematocrit, and it occurs at a higher dosage. On the other hand, in the optimal linear contract, the payment rate is smaller for the low baseline hematocrit, 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.

K Sensitivity Analyses and Other Assessments

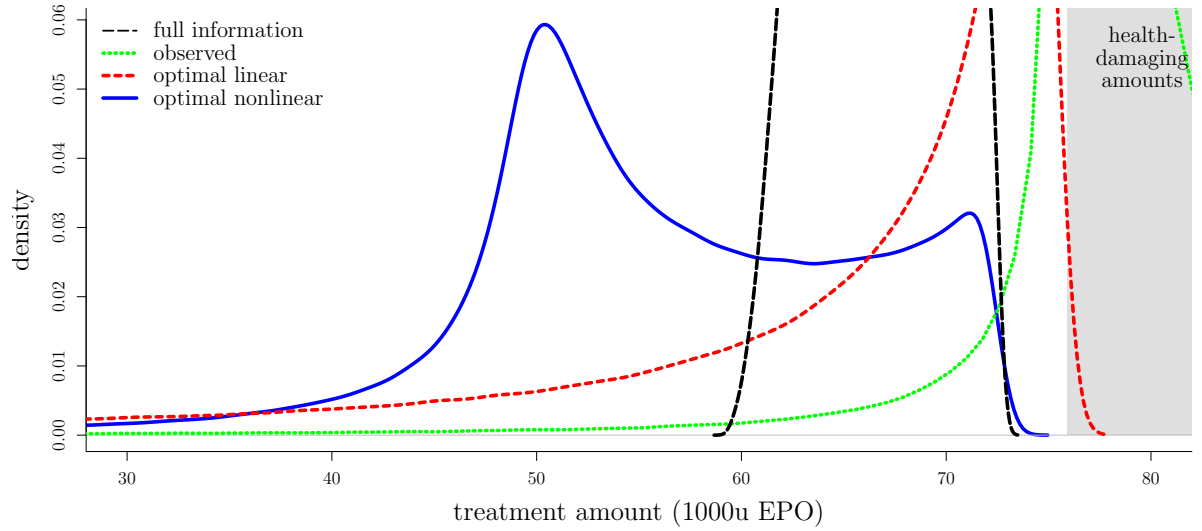
K.1 Robustness of the Reduced Form

Table A4 provides the full estimation results for the alternative specifications of the reduced form reported in Table 3, columns 4 to 9. Table A5 provides the results for the main specification with asymptotic standard errors clustered on chains rather than facilities.



(a) Payment as a function of the treatment amount

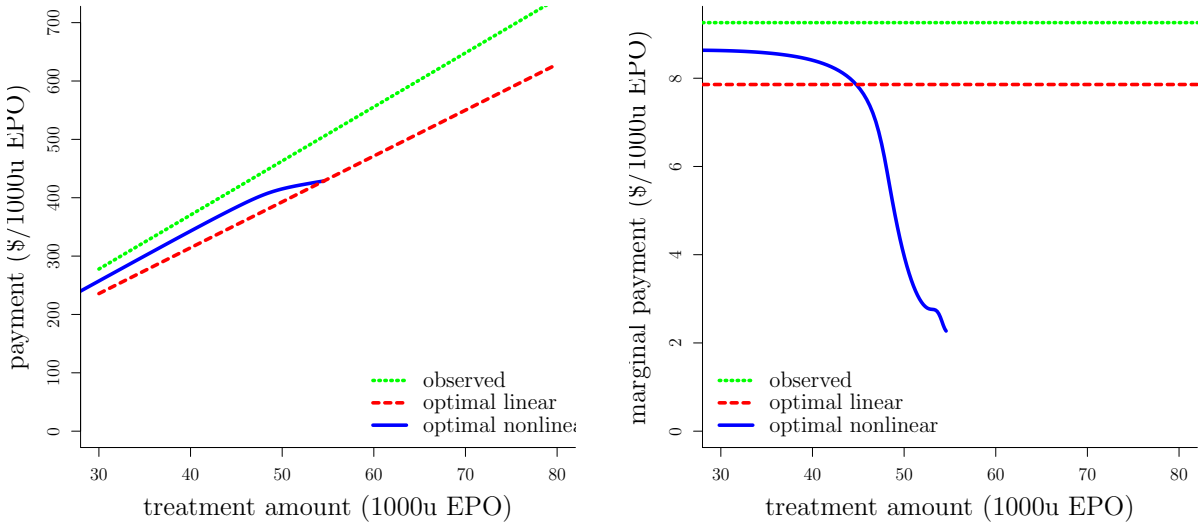
(b) Marginal payment as function of treatment amount



(c) Distribution of treatment amounts

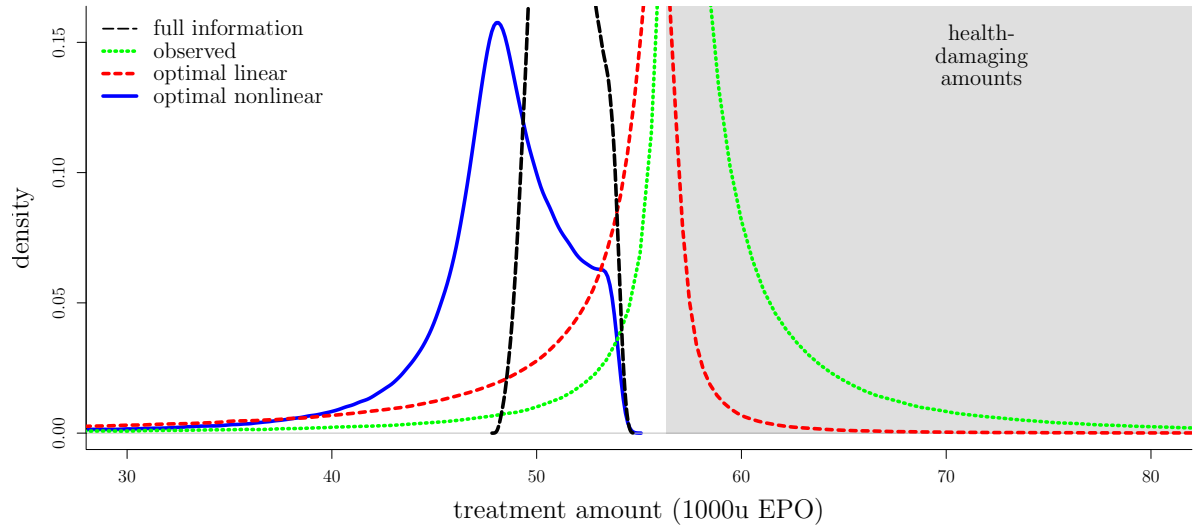
Figure A4: Optimal nonlinear contract treatment amounts and payments, baseline hematocrit 30-33

Notes: Figure plots treatment and payment amounts under the optimal nonlinear contract (blue, solid lines) for patients with median baseline hematocrit and mean characteristics in the lower hematocrit interval. 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.



(a) Payment as a function of the treatment amount

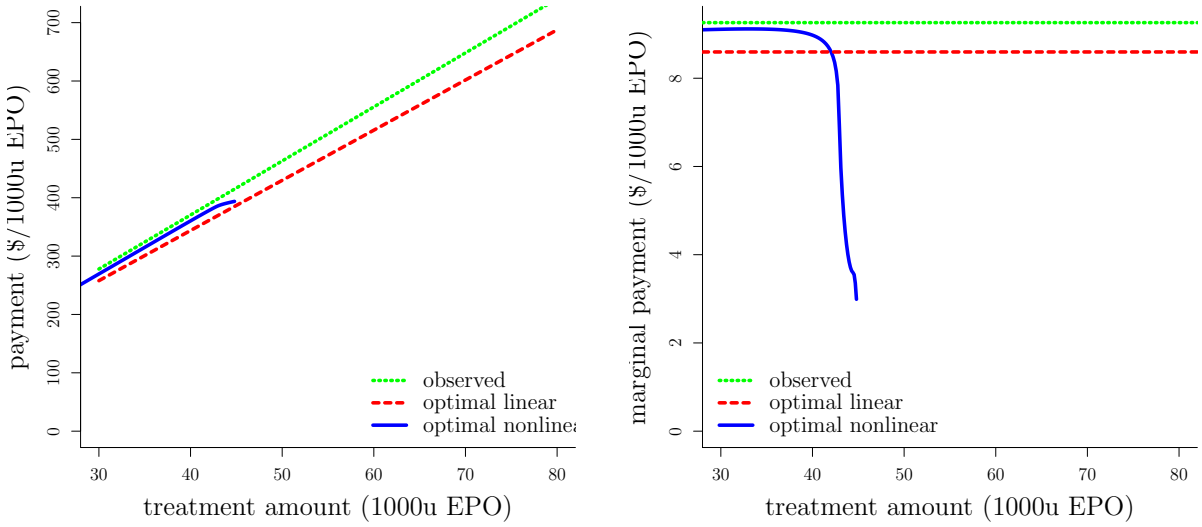
(b) Marginal payment as function of treatment amount



(c) Distribution of treatment amounts

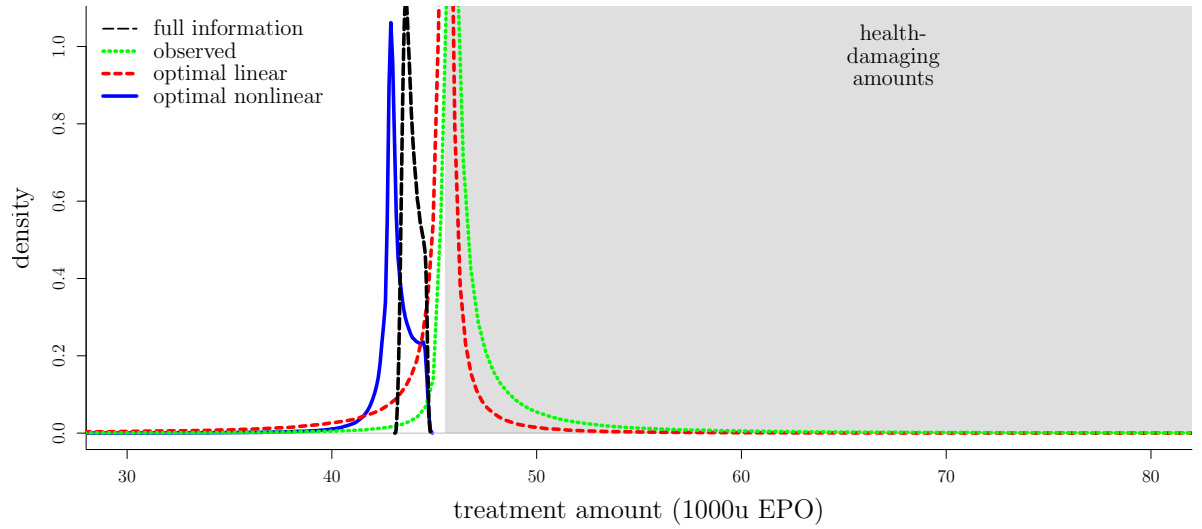
Figure A5: Optimal nonlinear contract treatment amounts and payments, baseline hematocrit 33-36

Notes: Figure plots treatment and payment amounts under the optimal nonlinear contract (blue, solid lines) for patients with median baseline hematocrit and mean characteristics in the middle hematocrit interval. 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.



(a) Payment as a function of the treatment amount

(b) Marginal payment as function of treatment amount



(c) Distribution of treatment amounts

Figure A6: Optimal nonlinear contract treatment amounts and payments, baseline hematocrit 36-39

Notes: Figure plots treatment and payment amounts under the optimal nonlinear contract (blue, solid lines) for patients with median baseline hematocrit and mean characteristics in the upper hematocrit interval. 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.

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

	OLS			Fixed Effects		
Interval:	> 30 to 33,	> 33 to 36,	> 36 to 39	> 30 to 33,	> 33 to 36,	> 36 to 39
Variable	(1)	(2)	(3)	(4)	(5)	(6)
Hematocrit	-9.29	-6.32	-3.56	-9.22	-6.51	-4.00
	(0.24)	(0.15)	(0.13)	(0.19)	(0.13)	(0.12)
Reimb. rate	9.53	6.39	3.92	9.42	5.99	4.67
	(3.19)	(2.03)	(1.91)	(3.00)	(1.95)	(1.85)
Age in years	-0.41	-0.37	-0.26	-0.37	-0.33	-0.24
	(0.02)	(0.02)	(0.01)	(0.02)	(0.01)	(0.01)
Female sex	-0.89	1.54	2.89	-1.53	1.21	2.38
	(0.55)	(0.40)	(0.34)	(0.49)	(0.38)	(0.33)
Charlson=1	9.06	8.05	7.37	7.97	7.08	6.50
	(0.96)	(0.69)	(0.60)	(0.86)	(0.65)	(0.59)
Charlson=2	10.76	10.25	8.21	10.30	9.72	7.93
	(0.90)	(0.67)	(0.59)	(0.81)	(0.63)	(0.57)
Charlson=3	13.87	11.87	8.58	12.60	11.09	8.73
	(0.94)	(0.72)	(0.60)	(0.88)	(0.70)	(0.58)
Charlson=4	15.55	13.93	10.83	15.06	13.77	10.64
	(1.22)	(0.86)	(0.73)	(1.05)	(0.82)	(0.70)
Charlson=5	16.56	15.03	11.89	16.20	14.53	11.27
	(1.40)	(1.08)	(0.93)	(1.26)	(1.01)	(0.89)
Charlson=6	18.63	18.52	13.84	17.82	18.05	13.44
	(1.87)	(1.48)	(1.21)	(1.61)	(1.35)	(1.14)
Charlson=7	26.23	26.02	20.39	23.46	24.37	19.95
	(3.02)	(2.48)	(2.19)	(2.61)	(2.30)	(2.12)
Charlson=8	23.96	24.27	14.52	23.02	22.00	15.70
	(3.94)	(3.06)	(2.51)	(3.56)	(3.09)	(2.50)
Charlson=9	32.00	32.43	22.86	31.54	32.96	23.44
	(4.98)	(4.17)	(3.81)	(4.97)	(4.08)	(3.98)
Charlson=10	23.91	28.48	32.24	22.57	27.65	29.77
	(7.02)	(6.71)	(6.96)	(6.16)	(6.46)	(6.76)
Charlson=11	39.13	43.64	39.81	40.92	40.83	39.65
	(11.01)	(8.79)	(7.31)	(8.45)	(8.04)	(7.07)
Charlson=12	38.42	33.52	25.67	27.82	27.22	16.10
	(12.51)	(8.06)	(9.82)	(10.18)	(7.17)	(10.17)
Constant	392.18	294.37	192.16	388.18	299.35	207.52
	(7.93)	(5.29)	(4.98)	(6.18)	(4.60)	(4.58)
Observations	231,702	405,019	283,024	231,702	405,019	283,024
R-squared	0.029	0.028	0.021	0.029	0.027	0.021
RMSE	71.43	58.46	49.01	65.78	55.05	46.29

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

Table A4: Alternative Specifications of the Reduced Form

	No Patient Observables			Comorbidity Indicators		
Interval:	> 30 to 33,	> 33 to 36,	> 36 to 39	> 30 to 33,	> 33 to 36,	> 36 to 39
Variable	(1)	(2)	(3)	(4)	(5)	(6)
Hematocrit	-9.61	-6.39	-3.46	-9.24	-6.32	-3.56
	(0.24)	(0.15)	(0.13)	(0.24)	(0.15)	(0.13)
Reimb. rate	9.81	6.13	4.26	9.40	6.09	4.08
	(3.20)	(2.04)	(1.92)	(3.20)	(2.03)	(1.91)
Age in years				-0.39	-0.36	-0.26
				(0.02)	(0.02)	(0.01)
Female sex				-0.73	1.61	2.95
				(0.55)	(0.40)	(0.34)
Myocardial inf.				-0.62	0.31	-0.74
				(1.09)	(0.88)	(0.74)
Cong. hrt. failure				9.38	9.16	7.06
				(0.80)	(0.59)	(0.50)
Periph. vasc. dis.				4.16	3.63	3.12
				(1.01)	(0.78)	(0.66)
Cerebro vasc. dis.				-2.35	-0.22	-0.43
				(1.19)	(0.98)	(0.74)
Dementia				-2.93	0.06	0.18
				(2.73)	(1.96)	(1.58)
Chron. pulm. dis.				3.63	3.11	1.99
				(0.88)	(0.65)	(0.58)
Rheumatic dis.				6.74	8.67	5.36
				(2.18)	(1.81)	(1.50)
Peptic ulcer dis.				9.62	7.32	6.35
				(2.15)	(1.71)	(1.41)
Mild liver dis.				6.78	4.18	3.33
				(2.24)	(1.62)	(1.37)
Diabetes w/out comp.				4.86	4.56	3.65
				(0.72)	(0.56)	(0.48)
Diabetes w/chron. comp.				1.64	0.93	0.74
				(0.80)	(0.59)	(0.51)
Hemi/para-plegia				3.58	3.03	0.96
				(3.26)	(2.39)	(2.03)
Any malignancy				12.70	10.77	8.30
				(1.95)	(1.57)	(1.38)
Mod/severe liver dis.				18.14	21.84	17.08
				(5.18)	(3.77)	(3.47)
Metastatic tumor				14.63	10.88	11.07
				(4.55)	(3.60)	(3.45)
AIDS/HIV				20.96	22.05	18.22
				(4.00)	(3.22)	(2.96)
Constant	383.73	280.61	178.15	390.51	294.56	192.74
	(7.88)	(5.24)	(4.96)	(7.89)	(5.26)	(4.98)
Observations	231,702	405,019	283,024	231,702	405,019	283,024
R-squared	0.014	0.009	0.005	0.030	0.030	0.022
RMSE	71.98	59.01	49.40	71.38	58.40	48.98

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

Table A5: Alternative Clusters for the Standard Errors

	Clustered on Dialysis Centers			Clustered on Chains		
Interval:	> 30 to 33,	> 33 to 36,	> 36 to 39	> 30 to 33,	> 33 to 36,	> 36 to 39
Variable	(1)	(2)	(3)	(4)	(5)	(6)
Hematocrit	-9.29	-6.32	-3.56	-9.29	-6.32	-3.56
	(0.24)	(0.15)	(0.13)	(0.46)	(0.97)	(0.40)
Reimb. rate	9.53	6.39	3.92	9.53	6.39	3.92
	(3.19)	(2.03)	(1.91)	(7.83)	(6.50)	(4.25)
Age in years	-0.41	-0.37	-0.26	-0.41	-0.37	-0.26
	(0.02)	(0.02)	(0.01)	(0.03)	(0.02)	(0.02)
Female sex	-0.89	1.54	2.89	-0.89	1.54	2.89
	(0.55)	(0.40)	(0.34)	(1.13)	(0.55)	(0.54)
Charlson=1	9.06	8.05	7.37	9.06	8.05	7.37
	(0.96)	(0.69)	(0.60)	(1.27)	(1.05)	(0.67)
Charlson=2	10.76	10.25	8.21	10.76	10.25	8.21
	(0.90)	(0.67)	(0.59)	(1.51)	(0.99)	(0.62)
Charlson=3	13.87	11.87	8.58	13.87	11.87	8.58
	(0.94)	(0.72)	(0.60)	(1.75)	(1.04)	(0.64)
Charlson=4	15.55	13.93	10.83	15.55	13.93	10.83
	(1.22)	(0.86)	(0.73)	(2.48)	(1.68)	(0.95)
Charlson=5	16.56	15.03	11.89	16.56	15.03	11.89
	(1.40)	(1.08)	(0.93)	(2.97)	(1.96)	(1.31)
Charlson=6	18.63	18.52	13.84	18.63	18.52	13.84
	(1.87)	(1.48)	(1.21)	(2.84)	(3.22)	(1.50)
Charlson=7	26.23	26.02	20.39	26.23	26.02	20.39
	(3.02)	(2.48)	(2.19)	(4.02)	(4.03)	(3.89)
Charlson=8	23.96	24.27	14.52	23.96	24.27	14.52
	(3.94)	(3.06)	(2.51)	(3.51)	(2.93)	(2.54)
Charlson=9	32.00	32.43	22.86	32.00	32.43	22.86
	(4.98)	(4.17)	(3.81)	(5.79)	(5.85)	(2.72)
Charlson=10	23.91	28.48	32.24	23.91	28.48	32.24
	(7.02)	(6.71)	(6.96)	(5.32)	(7.77)	(5.02)
Charlson=11	39.13	43.64	39.81	39.13	43.64	39.81
	(11.01)	(8.79)	(7.31)	(8.76)	(7.06)	(6.64)
Charlson=12	38.42	33.52	25.67	38.42	33.52	25.67
	(12.51)	(8.06)	(9.82)	(12.21)	(6.21)	(9.20)
Constant	392.18	294.37	192.16	392.18	294.37	192.16
	(7.93)	(5.29)	(4.98)	(16.00)	(33.99)	(12.94)
Observations	231,702	405,019	283,024	231,702	405,019	283,024
R-squared	0.029	0.028	0.021	0.029	0.028	0.021
RMSE	71.43	58.46	49.01	71.43	58.46	49.01

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

Table A6: Distribution of Hematocrit on Current and Prior Month Claims

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

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

K.2 Variability of Hematocrit within Patients over Time

Table A6 describes the variability of hematocrit levels within patients over time, by showing the distribution of hematocrit values reported on patients' prior monthly claims given the values on their current monthly claims. Each column shows this distribution for a one-percentage-point interval in the current hematocrit. For example, among patients with current hematocrit greater than 34 and less than or equal to 35 (the column labeled ">34 - 35"), 16.4% had hematocrit in that same interval reported on their prior monthly claim, while 11.2% and 5.5% had hematocrit levels of >33 - 34 and >31 - 32, respectively (the corresponding rows in that column).

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

K.3 Distributions of Facility Residuals and Test of Unimodality

Figure A7 shows the distributions of the facility-level mean residuals ($\bar{r}_i^k$, defined in Appendix F) in each hematocrit interval. We formally test the null hypothesis of unimodality for these three distributions using a "dip test" (Hartigan and Hartigan, 1985), implemented with the user-written command `diptest` in Stata (Cox, 2009). The test statistics (p-values) in each interval are as follows: 0.0033 (0.9930), 0.0035 (0.9930), and 0.0029 (1.0000). Thus the null hypothesis of unimodality is not rejected in any interval, and indeed test statistics are quite small with p-values quite close to one.

K.4 Downstream Medical Costs

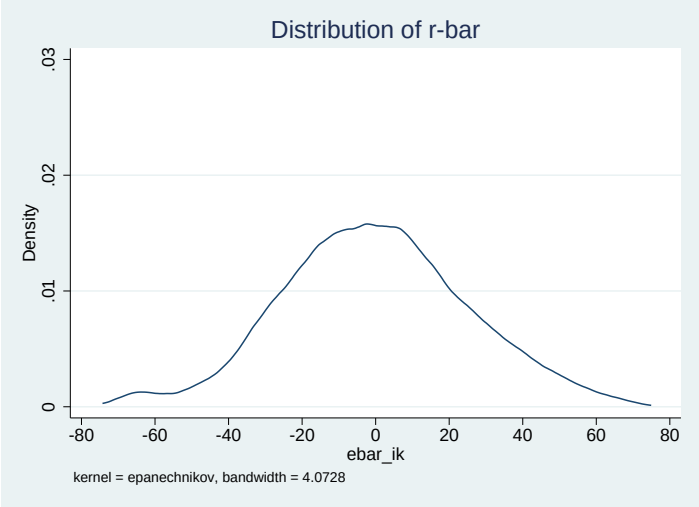
We combine estimates of the effects of EPO on transfusions and hospitalizations from Eliason et al. (2022) with estimates of the average costs of transfusions and hospitalizations from other sources noted below, to calculate a rough estimate of the the change in downstream medical costs under the optimal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

The exact sources and values are as follows:

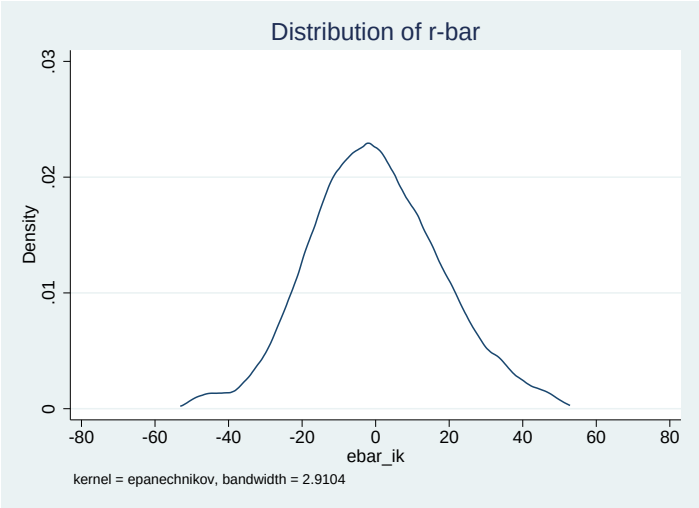
- Effect of 1,000u of EPO on monthly transfusion rate: -0.000586 (Eliason et al. (2022), Table 7, column 4 – IV estimate of the effect on transfusions)
- Effect of 1,000u of EPO on monthly hospitalization rate: 0.000205 (Eliason et al. (2022), Table 8, column 2 – IV estimate of the effect on hospitalization for any cause)
- Mean expenditure per outpatient transfusion episode among a sample of chronic dialysis patients: $854 (Gitlin et al., 2012, Table 2)
- Mean per-person per-year Medicare inpatient expenditures for ESRD patients in 2009: $25,244 (United States Renal Data System, 2020, Figure 9.6)
- Mean per-person per-year number of hospitalizations for ESRD patients in 2009: 1.82 (United States Renal Data System, 2020, Figure 4.1)
- Mean Medicare inpatient expenditures per hospitalization: $25,244 / 1.82 = $13,870 (derived from above)

With these values, we calculate the change in downstream costs that would result from the change in the mean monthly dosage of EPO under the optimal nonlinear contract, equal to -11.5 thousand units, as follows:

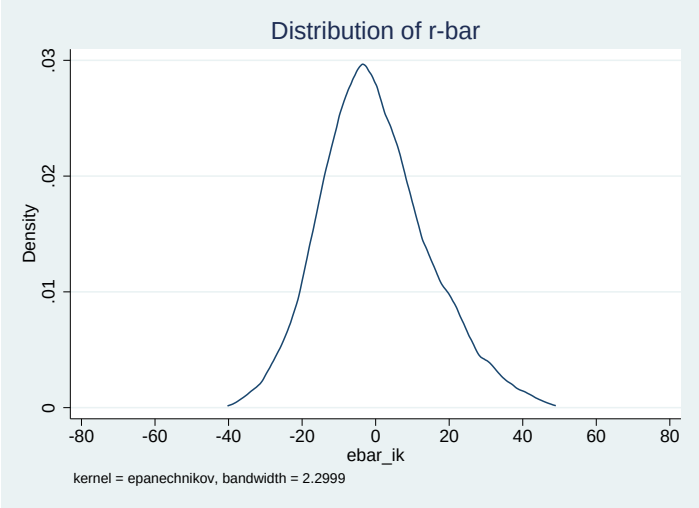
$$-11.5 \times \left[(-0.000586 \times \$854) + (0.000205 \times \$13,870)\right] = -\$26.95 \text{ per patient per month.}$$



(a) Lower hematocrit interval (30-33)



(b) Middle hematocrit interval (33-36)



(c) Upper hematocrit interval (36-39)

Figure A7: Distribution of facility-level mean residuals ($\bar{r}_i^k$)

L Forcing Contract

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

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

$$P^{\text{force}}(a) = \begin{cases} \overline{P} & \text{for } a = \overline{a}^{*\text{FI}} \\ -\infty & \text{else} \end{cases} . \tag{A22}$$

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

The results for the middle baseline hematocrit interval are presented in the bottom row of Table A7. While there are no medically excessive treatments under this forcing contract, the payment is larger than even under the observed payment contract, leaving massive information rents to better types. Indeed, the gain in the government objective over the observed contract (presented in the last column) is a fifth of that under the optimal nonlinear contract. This makes sense, as this (and any other) forcing contract was in the set of contracts considered by the principal when solving for the optimal unrestricted contract. Intuitively, while this forcing contract does implement a desired treatment amount for one particular agent type (highest altruism, lowest cost), the cost of getting the vast majority of agents to implement this amount is larger than the principal's valuation of any associated health benefit.

Table A7: Summary of Outcomes under Forcing and other Contracts for Patients with Median Severity of Anemia

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

M Importance of Both Dimensions of Heterogeneity

One of the strengths of our framework is that we are not beholden to an assumption that there is only one dimension of heterogeneity (or, for that matter, that there exists multidimensional heterogeneity). Rather, the model can recover the variation in different dimensions and we can quantify the importance of different types of unobserved heterogeneity. Given that we found altruism heterogeneity to be more substantial than heterogeneity in marginal costs, a natural question is whether the latter type of heterogeneity matters, from a normative perspective. Accordingly, we have examined the importance of heterogeneity in z by reducing the variance of z from its estimated value of 0.858 (which is different from zero at standard significance levels) to 0.10, and then solving for the optimal nonlinear contract in this counterfactual environment.[29]

The contracts are shown in Figure A8, for dosages of 40,000 units of EPO and greater (this corresponds to over 90% of treatment amounts). Figure A8 plots the marginal payment rates of the optimal nonlinear contract under our baseline parameterization (solid, blue, line) and when the variance of z is reduced (dashed, red, line). The main difference is that the marginal payment is higher for dosages up to about 48,000 units. This reflects an increase in the marginal costs of formerly low-cost providers, when z is shrunk toward the mean. (Note that very high-cost providers are not pictured here because they provide treatment amounts lower than 40,000 units.) The dosages above 48,000 units come from types with sufficiently high altruism that their behavior is not substantially affected by changes in marginal costs.

We have also computed how the optimal nonlinear contract based on the counterfactual parameterization featuring less heterogeneity in z would affect the gains to the government from better contracting. We compute that the government would on average gain \$125 per patient/month from moving to the optimal nonlinear contract from the observed contract.[30] Using instead the optimal nonlinear contract resulting from misspecifying the model with less heterogeneity in z, the government would gain \$113 per patient-month. Some of the reduction in the gain comes directly from the higher payments under the misspecified nonlinear contract, which do not outweigh the government's valuation of the resulting increases in patient health. Thus, taking into account the full extent of the variation in z would improve the government's gain by just over 10%.

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

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

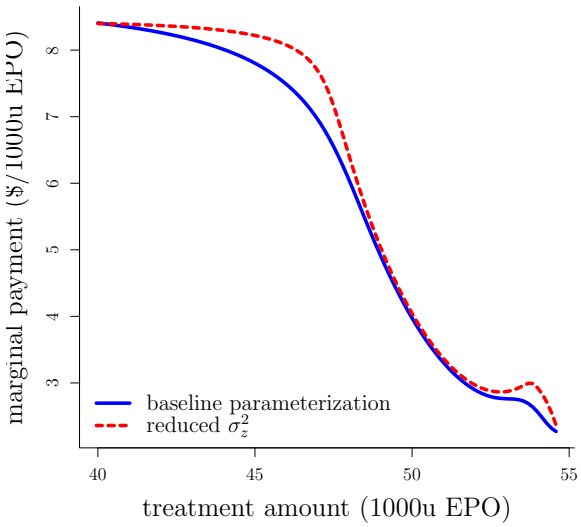


Figure A8: Comparison of marginal payments for nonlinear contracts under baseline parameterization and under parameterization with reduced σ_z^2.

References

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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