Optimal Contracting with Altruistic Agents: Medicare Payments for Dialysis Drugs

By Martin Gaynor and Nirav Mehta and Seth Richards-Shubik*

We study health care provider agency and optimal payments, considering 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 contracts in this environment. The optimal contracts eliminate medically excessive dosages and reduce expenditures, resulting in approximately $300 million in annual gains from better contracting. This approach could be applied to a broad class of problems in health care payment policy.

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 third party payers. Therefore, payers have to decide how to contract with providers to deliver care, while recognizing that providers possess relevant information that they do not. Consequently, the effects of better or worse incentives can have profound impacts on 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, Finkelstein and Mahoney, 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, Leland and Sibley, 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

* Gaynor: Heinz College, Carnegie Mellon University and NBER, 5000 Forbes Ave. Pittsburgh PA 15213, mgaynor@cmu.edu. Mehta: Department of Economics, University of Western Ontario, Social Science Centre 4063 London, ON, N6A 5C2 Canada, nirav.mehta@uwo.ca. Richards-Shubik: Department of Economics, Lehigh University and NBER, 621 Taylor St. Bethlehem PA 18015, sethrs@lehigh.edu. Acknowledgements: We are grateful to Liran Einav and to four anonymous referees for comments and suggestions that greatly improved the paper, and to David Dranove, 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 BU, Cambridge, Carnegie Mellon, Georgia, Johns Hopkins, Michigan, NYU, North Carolina, Northwestern, Penn, Princeton, Stanford, UBC, Wisconsin, York, the 2018 International Industrial Organization Conference, the 2018 Annual Meeting of the Society of Labor Economists, the 2018 Healthcare Markets Conference, 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, Tian Liu, Martin Luccioni, and Cecilia Diaz Campo for excellent research assistance. The usual caveat applies.

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 margin. This approach could be relevant for provider-administered medications more broadly, and for other treatments where decisions are primarily about the quantity given to each patient, 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. 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, Leland and Sibley, 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, comprising over 12 percent of total Medicare spending (Medicare Payment Advisory Commission, 2021). This is one of the most rapidly growing areas of Medicare spending, and has been the object of ongoing concern and attempts at policy reform (Bach, 2009). The issues are also broadly relevant 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, Rebitzer and Taylor (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 unobserved 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 treatment. In the context of dialysis care, Eliason et al. (2019) examine the effects of corporate ownership on treatment decisions and patient

[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).

[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).

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, Finkelstein and Mahoney (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 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, Finkelstein and Levin (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 II, 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 I). In Section II, we introduce the model and then derive the optimal payment contract. Section III presents the data we use for our empirical analysis, and Section IV 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 V.

I. 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

[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.

[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.

of total Medicare spending), and of that amount, $1.74 billion 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, Pham and Macdougall, 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 Online Appendix Table O6), 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 (see Shinkman, 2016, for a useful overview of how dialysis centers are run). Physicians are independent practitioners who may endogenously match with dialysis facilities.[8] 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. Accordingly, we call the agent making dosage decisions for patients a "provider."

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 IV.B), 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.[9] 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

[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, while injectable iron accounted for 15% (U.S. Government Accountability Office, 2006).

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

[9] 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".

sales price from roughly six months earlier (U.S. Government Accountability Office, 2006).[10] 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 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). This study was published in November 2006, and strong warnings ("black box warnings") were added to the drug's labels in 2007. There were no such major informational shocks in 2008 or 2009, our study period.

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).[11] By 2009, two large chains treated a combined 60 percent of dialysis patients in the US (United States Renal Data System, 2011). 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."[12] In our data, chain effects account for only a small proportion of the variation in mean dosages across facilities. 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.[13] Additionally, annual facility-level cost reports submitted to

[10] In 2011, Medicare adopted a comprehensive "bundled" prospective payment system for dialysis that included EPO, so the payment policy for the drug switched from fee-for-service (per-unit) to prospective (lump-sum) payment. See Eliason et al. (2022) for an analysis of the effects of this policy change.

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

[12] ESRD Program Interpretive Guidance 2008, p. 279, https://www.cms.gov/Medicare/Provider-Enrollment-and-Certification/GuidanceforLawsAndRegulations/downloads/esrdpgmguidance.pdf

[13] The purchasing agreements are on file with the Securities and Exchange Commission. For the (redacted) agreements

Medicare show this variation in per-unit prices for EPO within chains (see Section III).

Accordingly, in the empirical analysis, we treat each dialysis center as an independent entity, with its own marginal cost and degree of altruism. This allows for heterogeneity both within and 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 (see Section IV.C). Also, having one common distribution of unobserved provider characteristics fits with having one common contract for all providers. 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.

II. 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.[14] 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 events, as described in Section I. 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."[15] 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

covering our study period, see https://www.sec.gov/Archives/edgar/data/927066/000119312508042304/dex1062.htm (for DaVita) and https://www.sec.gov/Archives/edgar/data/1333141/000132693207000082/f01549exv4w18.htm (for Fresenius).

[14] Because the medication is administered intravenously while the patient is undergoing dialysis, patient compliance is less of a concern here than it is for, say, oral medications taken at home.

[15] 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.

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.[16] 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 $\{P(a; b, x)\}$, 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.[17]

The government's objective is also a function of the patient's resulting health, weighted by a parameter, α_g, minus the payment to the provider.[18] 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

[16] 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 (similarly, we assume the distribution of cost types would be invariant under counterfactual payment contracts). Furthermore, under some specifications, heterogeneity in beliefs could be observationally equivalent to heterogeneity in altruism.

[17] 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.

[18] 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).

and observed characteristics x, and receives treatment amount a, is as follows:

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

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, 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).[19] 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:

$$\begin{aligned}
\max_{P \in \mathcal{P}} \int_{\alpha,z} & \left[\alpha_g h(a^*(\alpha, z; b, x, P); b, x) - P\left(a^*(\alpha, z; b, x, P); b, x\right)\right] f(\alpha, z) d\alpha dz \\
\text{s.t.} \quad & a^*(\alpha, z; b, x, P) = \arg\max_{a \geq 0} \; u(a; \alpha, z, b, x, P), \quad \forall \alpha, z && \text{IC} \\
& u(a^*(\alpha, z; b, x, P); \alpha, z, b, x, P) \geq \underline{u}, \quad \forall \alpha, z && \text{VP},
\end{aligned}$$

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.[20] 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.[21]

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 for corner solutions where some provider types administer zero units (see Online Appendix D). We follow the screening literature in referring to this as *exclusion* (see, e.g., Armstrong, 1996), which is distinct from non-participation.

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

[20] 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 11). 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.

[21] 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.

A. *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 $a^{*\mathrm{FI}}(\alpha, z)$. The interior optimality condition is

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

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.

B. *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 II.C). 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.

C. *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, Leland and Sibley, 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 $\bar{a}^{*\mathrm{FI}} \equiv a^{*\mathrm{FI}}(\overline{\alpha}, \underline{z})$, the amount that would be provided by the "best" type (highest altruism, lowest cost) under full information. We show in Online 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 $a^{*\mathrm{FI}}(\overline{\alpha}, \underline{z})$ 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^*

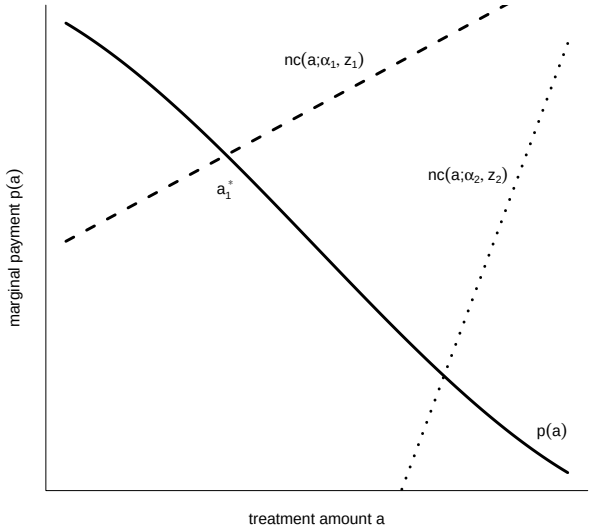


Figure 1. : Example marginal payment contract and provider supply curves.

Note: 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., $z_1 = z_2$.

(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 can be easier to satisfy in monopsony applications (i.e., purchasing goods or services).[22]

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.[23] The separate problems for each treatment amount are thus

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

[22] 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.

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

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

Finally, the optimal contract is characterized by the first-order condition of (7) for each amount, treating $p(a)$ as a parameter:[24]

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

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 Online 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 Online Appendix C.3.

III. 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 (Centers for Medicare and Medicaid Services, 2008, 2009*e*). 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.[25] 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.[26] This excludes 119,788 observations (10.6% of the remaining total) with hematocrit levels below the recommended 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.[27] 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 I, 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

[24] 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.

[25] 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.

[26] 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 than 39 percent.

[27] 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).

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)†	*	*	7.14	7.32	7.53	8.15	9.05
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)	9.44 (Q2)	9.62 (Q3)	9.58 (Q4)

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

total amount of EPO administered over the claim period, and the baseline hematocrit, b_{jt}, is the prior hematocrit level reported on the claim.[28] 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 (Centers for Medicare and Medicaid Services, 2008, 2009*b*).[29] 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).[30]

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 across all the claims in our sample 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 (Centers for Medicare and Medicaid Services, 2008, 2009*a*).[31] The percentiles show potentially important heterogeneity

[28] 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).

[29] 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.

[30] 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, Centers for Medicare and Medicaid Services (2007, 2008, 2009*c*)). Patient age and sex are taken from the Medicare Beneficiary Summary File (Centers for Medicare and Medicaid Services, 2007, 2008, 2009*d*).

[31] 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

in acquisition costs, even though the drug was produced by a single manufacturer.[32] As we discuss below in Section IV.B, 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.

IV. Empirical Implementation

We now describe how we adapt the model from Section II 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 natural assumptions that the providers' utility functions and the government's objective function are additively separable across patients.[33] 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.

A. 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 Online Appendices E.1-E.2). Hence the sign of the marginal effect of treatment is identified (i.e., what dosages are health damaging on the margin).[34]

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.[35]

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 condition (4) yields

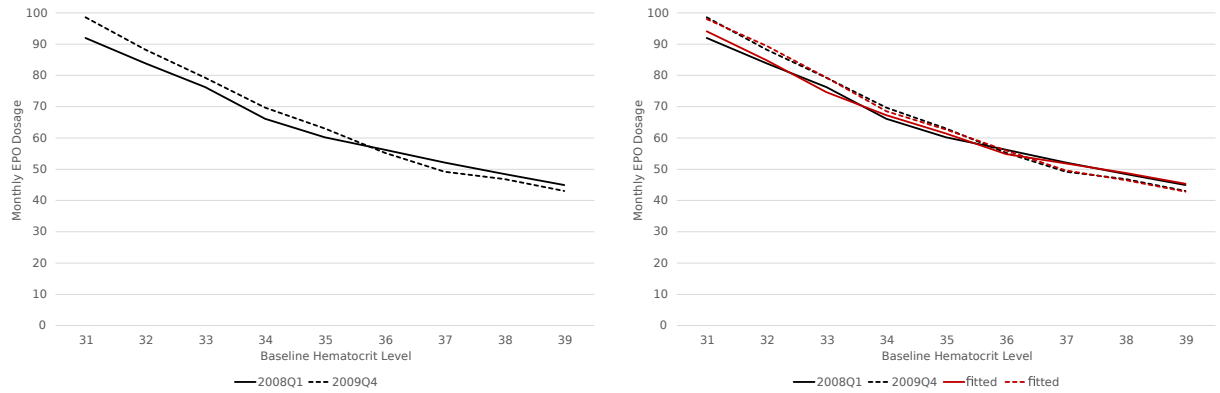
acquisition cost per 1,000 units of EPO for each center in the cost report data from 2008.

[32] 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 \$0.83 ($= 8.15 - 7.32$) across all centers shown in Table 1.

[33] The static framework can be applied to multiple time periods if there are no dynamic effects of EPO (as noted in Section I), 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 is restricted to observations in this range.

[34] 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.

[35] 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.



(a) Observed dosages (b) Observed and predicted

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

Note: Means calculated by integer value of hematocrit, rounded up. Fitted values are predicted with the estimated versions of the reduced form reported in Appendix Table A1.

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 III, 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 V.

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 IV.C). 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,[36] and the flexibility in τ (i.e., τ_k) means that

[36]Because patients with lower baseline hematocrit are given higher dosages on average, this could approximate diminishing

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.

B. *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 intervals. As can be seen from the reduced form (11), μ_z is not separately identified from a constant term in τ_k.[37] 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, because it is nonparametrically identified under the assumption that the idiosyncratic shocks η_{ijtk} (equivalently, ϵ_{ijt}^k) are mean-independent of the observables b, p_1, and x (see Online Appendix E.3). For efficiency and computational tractability, we use a parametric assumption. We specify $\ln \alpha$ and z to have a joint normal distribution, so that α has a lognormal distribution with strictly positive support. Then in each hematocrit interval k, there are four unknown parameters of the joint distribution, $\mu_{\alpha,k}$, $\sigma_{\alpha,k}^2$, $\sigma_{\alpha z,k}$, and $\sigma_{z,k}^2$ (while $\mu_{z,k} = \mu_z$ is treated as known from our external information on costs).[38] Using Stein's lemma

returns, for example.

[37] 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.

[38] 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 $\ln \alpha$.

(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 Online Appendix F for the full description of this approach.)

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 Online 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 II.B and II.C).

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 I and Online Appendix J.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 is the possible 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 IV.C).

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 I). 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.[39]

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 III allows us to compute per-unit acquisition costs by facility and year, and we use the median reported in Table 1, equal to \$7.53 per 1,000 units, as the acquisition component of μ_z.[40] 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 Online 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

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

[40] 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 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 (0.23)	-6.32 (0.15)	-3.56 (0.13)
Reimbursement rate ($\bar{\beta}_2^k$)	9.53 (3.11)	6.39 (2.12)	3.92 (1.89)
Obs. in interval	231,702	405,019	283,024

Note: 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.

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 Online 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).

C. 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 A1). To interpret these coefficients, for example in the middle interval, a patient with one unit higher baseline hematocrit (say 35 vs. 34) receives 6,320 less units of EPO per month on average. Also in that interval, a one dollar increase in the payment rate (per 1,000 units) would induce providers to increase dosages by 6,390 units per month on average. 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 (see also Online Appendix Table O3 for results using either physician-level or patient-level fixed effects). 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 I, 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. 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

Table 3—: Robustness of Reduced-Form Estimates

Variable (Coefficient)	Provider Fixed Effects		
	> 30-33	> 33-36	> 36-39
	(1)	(2)	(3)
Baseline hematocrit (β_1^k)	-9.22	-6.51	-4.00
	(0.19)	(0.13)	(0.12)
Reimbursement rate ($\bar{\beta}_2^k$)	9.42	5.99	4.67
	(3.00)	(1.95)	(1.85)
Obs. in interval	231,702	405,019	283,024
Variable (Coefficient)	No Patient Observables		
	> 30-33	> 33-36	> 36-39
	(4)	(5)	(6)
Baseline hematocrit (β_1^k)	-9.61	-6.39	-3.46
	(0.24)	(0.15)	(0.13)
Reimbursement rate ($\bar{\beta}_2^k$)	9.81	6.13	4.26
	(3.20)	(2.04)	(1.92)
Obs. in interval	231,702	405,019	283,024
Variable (Coefficient)	Comorbidity Indicators		
	> 30-33	> 33-36	> 36-39
	(7)	(8)	(9)
Baseline hematocrit (β_1^k)	-9.25	-6.32	-3.56
	(0.24)	(0.15)	(0.13)
Reimbursement rate ($\bar{\beta}_2^k$)	9.39	6.07	4.07
	(3.20)	(2.03)	(1.91)
Obs. in interval	231,702	405,019	283,024

Note: Each column is a separate regression. Complete lists of variables and coefficient estimates for each regression appear in Online Appendix Tables O3 and O4. Asymptotic standard errors in parentheses, clustered on dialysis center.

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).[41] The regressions reported 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.[42] 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 Online Appendix J.3. Online Appendix Figure O4 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

[41] Online Appendix J.1 contains the full estimation results for these alternative specifications, as well as the results for the main specification with asymptotic standard errors clustered on chains, rather than on facilities.

[42] 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).

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)	(0.010)
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
	(0.87)	(0.93)	(1.34)
$\sigma^2_{\alpha,k}$	2.68	2.14	3.64
	(0.90)	(1.00)	(1.34)
$\sigma_{\alpha z,k}$	-0.343	-0.437	-0.371
	(0.012)	(0.036)	(0.011)
$\sigma^2_{z,k}$	0.473	0.861	0.332
	(0.145)	(0.339)	(0.068)
Obs.	231,702	405,019	283,024

Note: 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.

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.[43] Also, the larger values of δ_k in intervals with higher baseline hematocrit are consistent with diminishing marginal productivity of the drug, because patients

[43]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.

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.[44] However, as discussed earlier and detailed in Online Appendix E.1, our main 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 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.[45] 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. 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 \$0.83.

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 (DaVita and Fresenius) vs. others, or non-profits vs. for-profits, for example. The details are presented in Online 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.[46] 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. As we show in Section V.C, the optimal nonlinear contract targets heterogeneity in altruism more than heterogeneity in marginal costs, but both are relevant.

[44] 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 the cost of providing EPO.

[45] 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.

[46] 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.

V. 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.[47] 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 $\delta_k^{-2}\alpha_k^{-1}$ $(= \beta_2^k)$ still equal the estimated values of $\bar{\beta}_2^k$. 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 Online 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, and information on the relationship between hematocrit levels and mortality risk taken from clinical trials on EPO (see Online 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 V.A). We compare the optimal nonlinear contracts with the observed contract, and with optimal linear contracts also computed using the estimated model parameters.[48] 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 V.B). Finally, we show how the nonlinear contract screens among the different dimensions of physician types (Section V.C).

A. Optimal Contracts and Distributions of Dosages

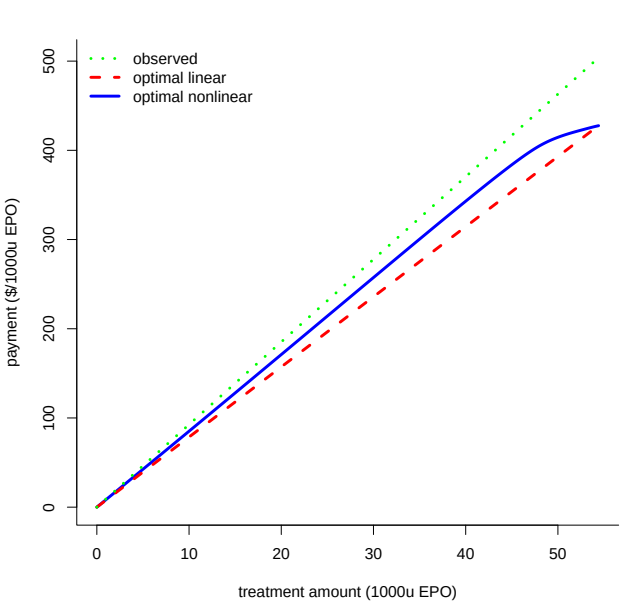
We start with the contract for a patient with the median hematocrit level and the mean characteristics in the middle interval. Figure 3 plots the total payments (panel a), the marginal payments (panel b), and the distributions of treatment amounts (panel c), with the optimal nonlinear contract in blue 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.[49]

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 \$383.79, the linear contract would pay \$353.73, and the observed contract would pay \$416.70, per month. The marginal payment (panel b) in the nonlinear contract is roughly constant below 40

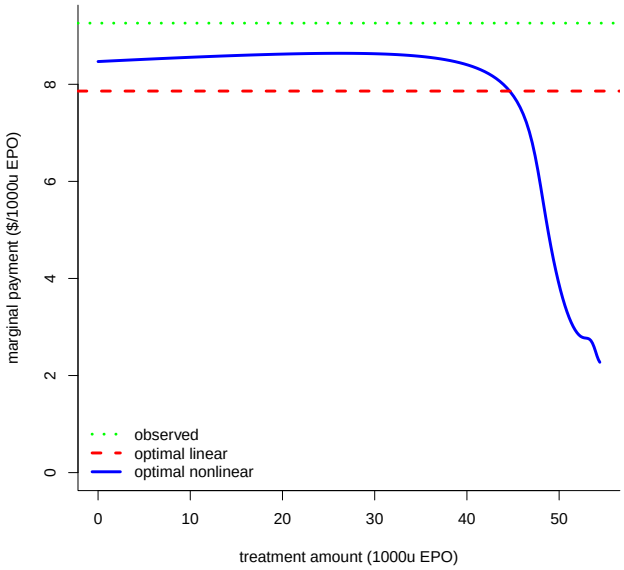
[47] See Online Appendix D for computational details. Also, we assess the regularity condition, that no provider types' supply curves intersect the marginal payment curve under the optimal nonlinear contract more than once, and find that it is not violated (Online Appendix I).

[48] The shock η is set equal to zero in all simulations. We show how to solve for optimal linear contracts in Online Appendix A.

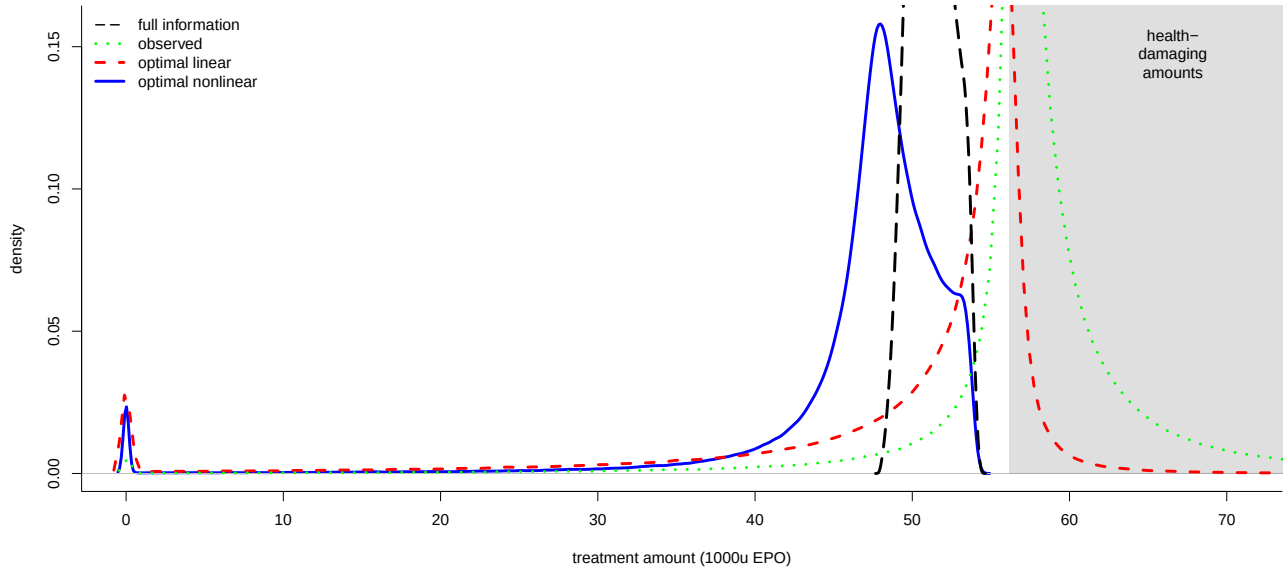
[49] The reservation utility $\underline{u}$ is set equal to the lowest utility obtained under the observed contract. A very small share of physicians (0.2%) are excluded in the simulation of the observed contract, which fixes $\underline{u}$ at the utility of a treatment amount of zero and zero payment,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



(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.

Note: 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.

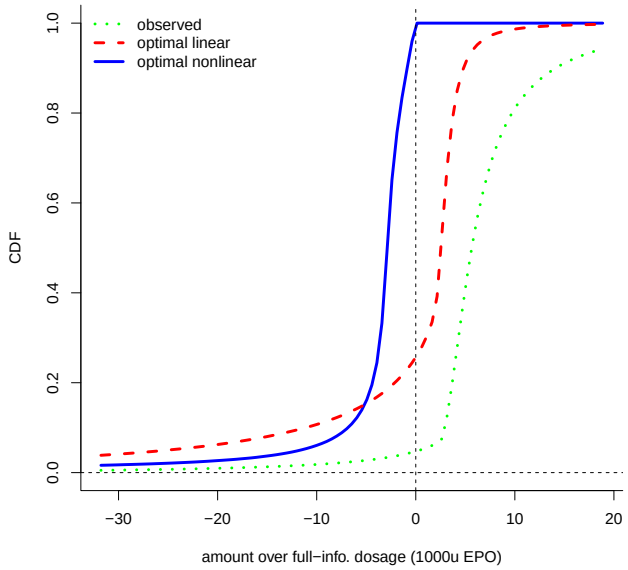


Figure 4. : CDFs of deviations from full-information treatment amounts, for patients with median severity of anemia.

Note: 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.

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 II.B, any providers with marginal costs below the payment rate will be induced to provide dosages that yield a negative marginal product of health, regardless of their degrees of altruism. Because the linear contract has only a single marginal payment, the government accepts these excessive dosages in order to avoid further underprovision by other 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.[50] Overprovision is nearly universal with the observed contract

[50]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

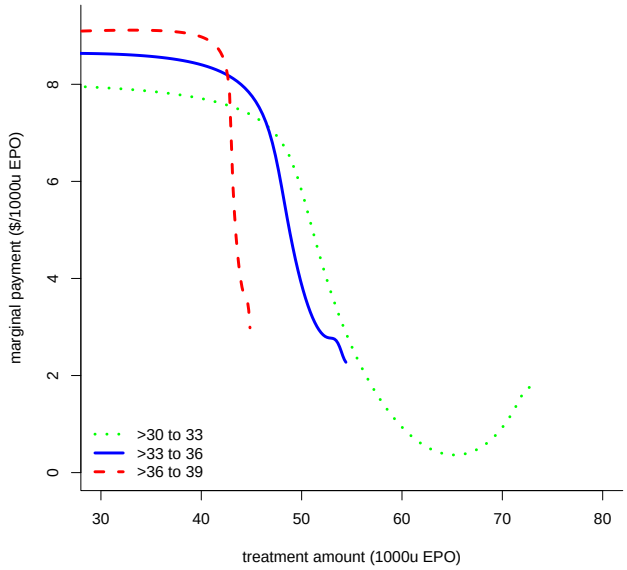


Figure 5. : Marginal payments in optimal nonlinear contracts for patients with median severity of anemia in each estimation interval.

Note: 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.

(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 Online Appendix C.3). This further indicates the value of having flexible marginal incentives, because any overprovision is strictly dominated by underprovision in an 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

equilibrium treatment amount provided by type (α, z) under the second-best and $a^{*\mathrm{FI}}(\alpha, z)$ is defined in (3).

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	744	80.4	12.9	82%	98%	
Optimal Linear	411	60.9	20.7	0%	64%	$185
Optimal Nonlinear	388	54.9	13	0%	0%	$220
Baseline hematocrit >33 to 36						
Observed	541	58.4	9.7	75%	95%	
Optimal Linear	395	50.2	11.7	19%	74%	$98
Optimal Nonlinear	392	47	7.2	0%	0%	$124
Baseline hematocrit >36 to 39						
Observed	437	47.2	5.2	86%	97%	
Optimal Linear	384	44.7	5.1	45%	87%	$60
Optimal Nonlinear	384	42.9	2.5	0%	0%	$87

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.

be approximated with a set of tiered payment rates, where the number of tiers and their levels depended on patient characteristics.

B. 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 11.4 thousand units lower and the mean monthly payment is $148 lower under the optimal nonlinear contract compared to the observed contract.[51]

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 26% and 52% at the medium and high baseline hematocrit levels, respectively.[52] 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.[53] For

[51] 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 Online Appendix J.4 for details). This suggests that the direct savings on EPO may be a somewhat conservative lower bound for the total savings.

[52] 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.

[53] 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.

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 45 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).

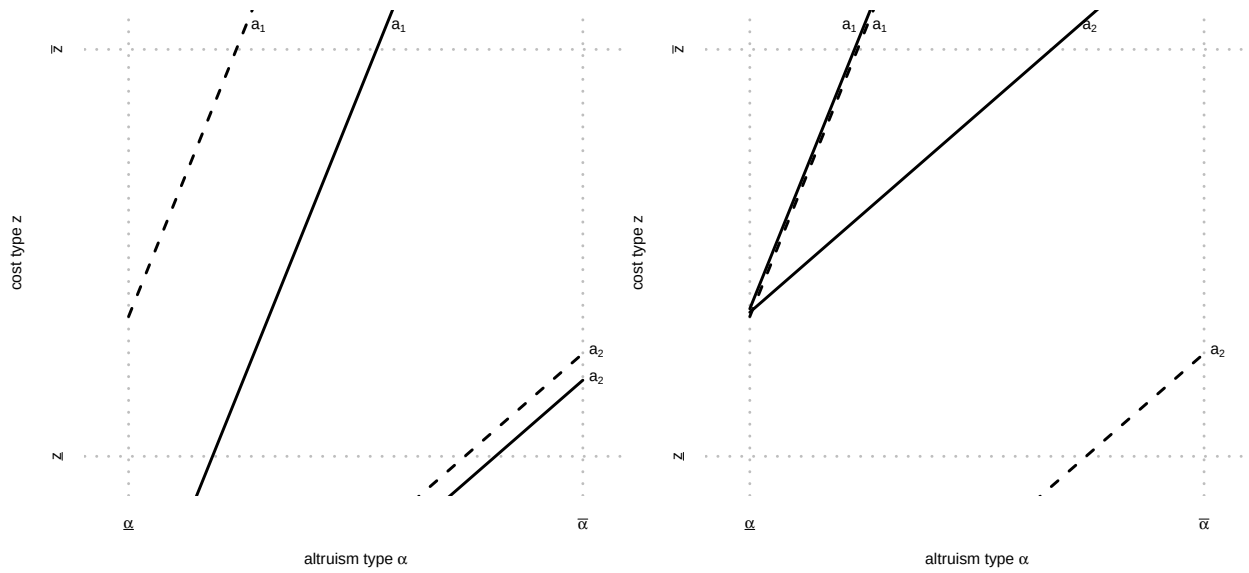
The last column of Table 5 shows the government's gains from better contracting, by calculating the increases in the government's objective relative to its values under the observed contract. This provides a summary measure, in dollars per patient per month, of the potential benefit to the government (and by extension, the patients represented by the government) from the changes in outcomes discussed above.[54] There are substantial gains from using the optimal nonlinear contract, ranging from \$87 to \$220 (or roughly \$1,050 to \$2,600 per patient per year) in these examples.[55] Compared to the mean monthly payments under the observed contracts of \$437 to \$744, 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 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 \$1,739 to \$3,752 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.[56] 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.

[54] 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.

[55] 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.

[56] We focus on this treatment amount because it is the highest dosage the government would ever wish to implement. See Online Appendix K 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.

Note: 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.

C. 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 Online 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 Online 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 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.[57] 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.

VI. 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 48%, 27%, and 12%, 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.[58] To put this in context, Medicare spent 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 $1,739 to $3,752 per patient per month.

This approach to contracting could prove particularly valuable in improving how Medicare pays

[57] 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 Online Appendix L). 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.

[58] 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.

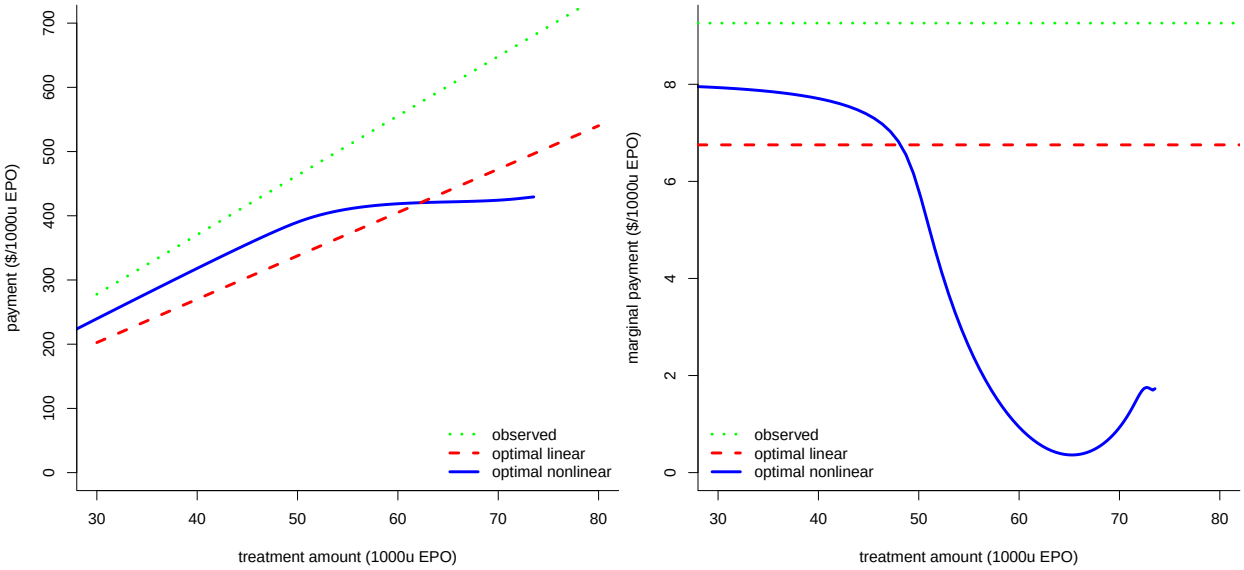
for provider-administered drugs (through the Part B program), which is widely acknowledged to be problematic with regard to both dosing and spending. While the general nonlinear contracts we derive may seem complex, these results can provide guidance for simple approximations of the optimal contracts, such as a set of tiered payment rates. Moreover, as Clemens, Gottlieb and Molnár (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.

Appendix: Full Estimation Results and Counterfactuals

This appendix 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.[59] In addition, Table A1 provides the full estimation results for our reduced form.

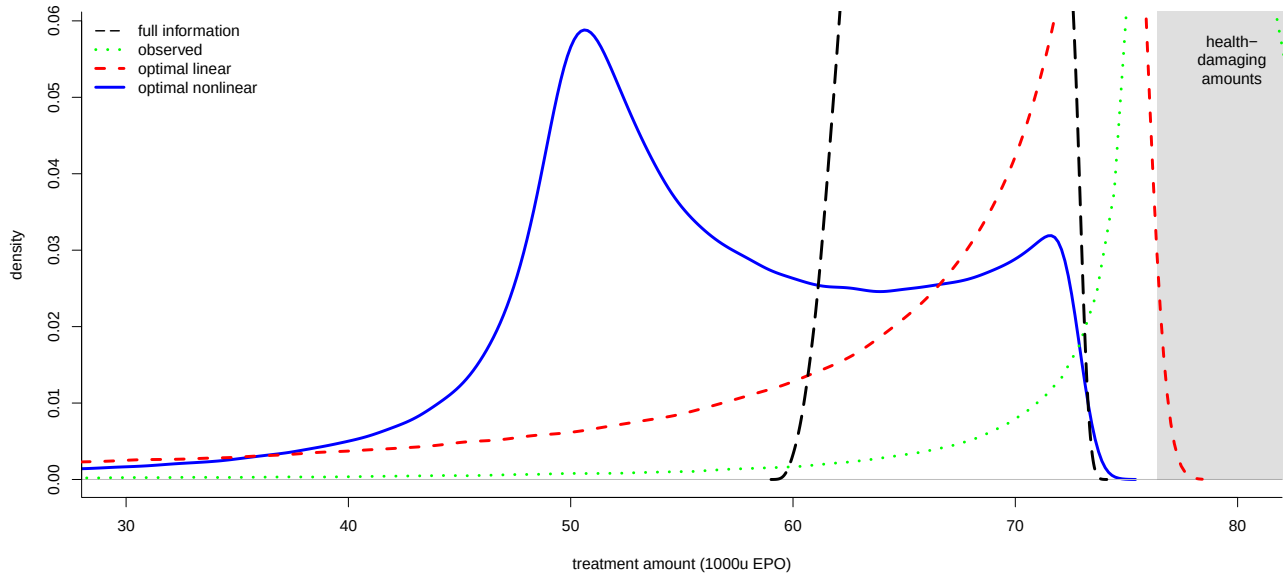
Figures A1 to A3 show the contracts (i.e., the total payments as a function of the treatment amounts), the marginal payments, and distributions of treatment amounts, separately for each interval. They have similar patterns, as discussed in the main text, with the optimal nonlinear contract below the 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.

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



(a) Payment as a function of the treatment amount

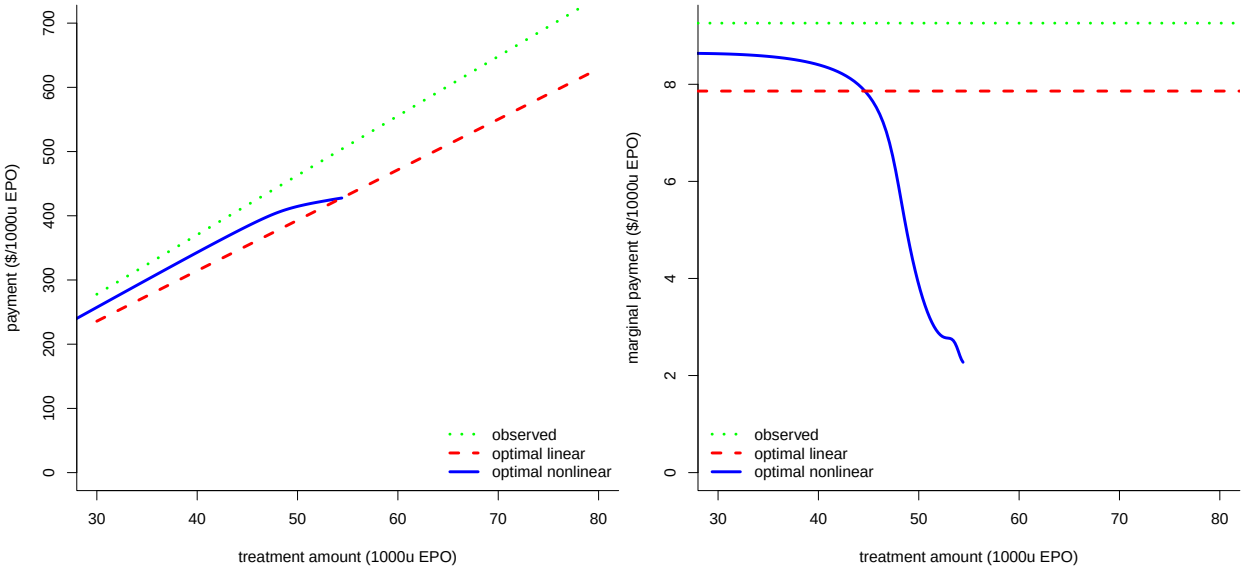
(b) Marginal payment as function of treatment amount



(c) Distribution of treatment amounts

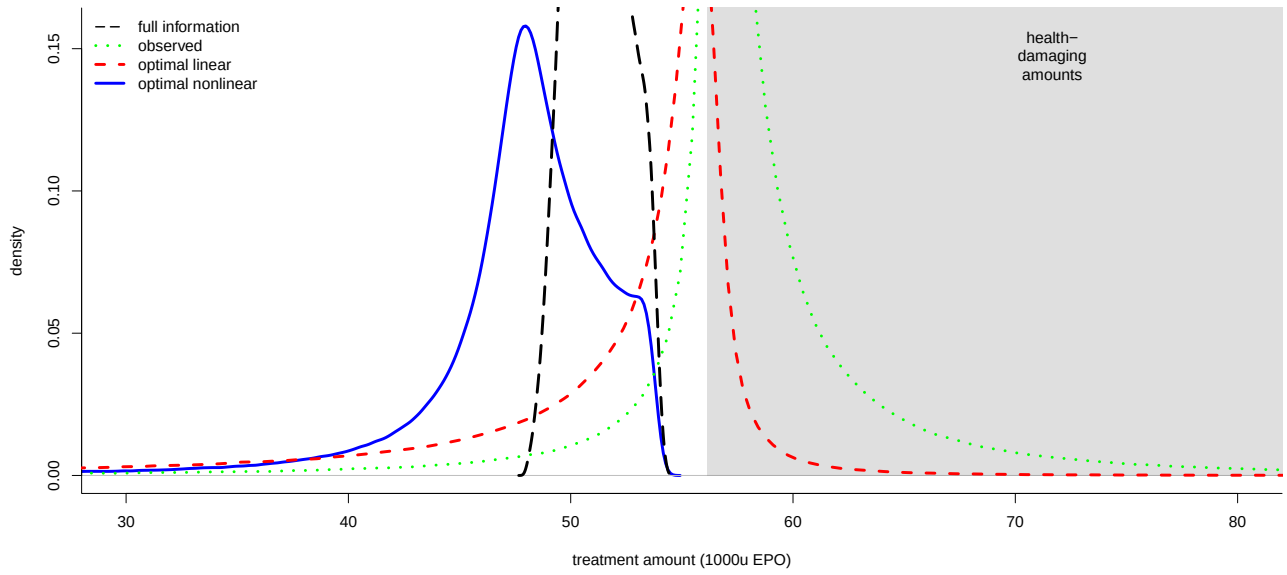
Figure A1. : Optimal nonlinear contract treatment amounts and payments, baseline hematocrit 30-33

Note: 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

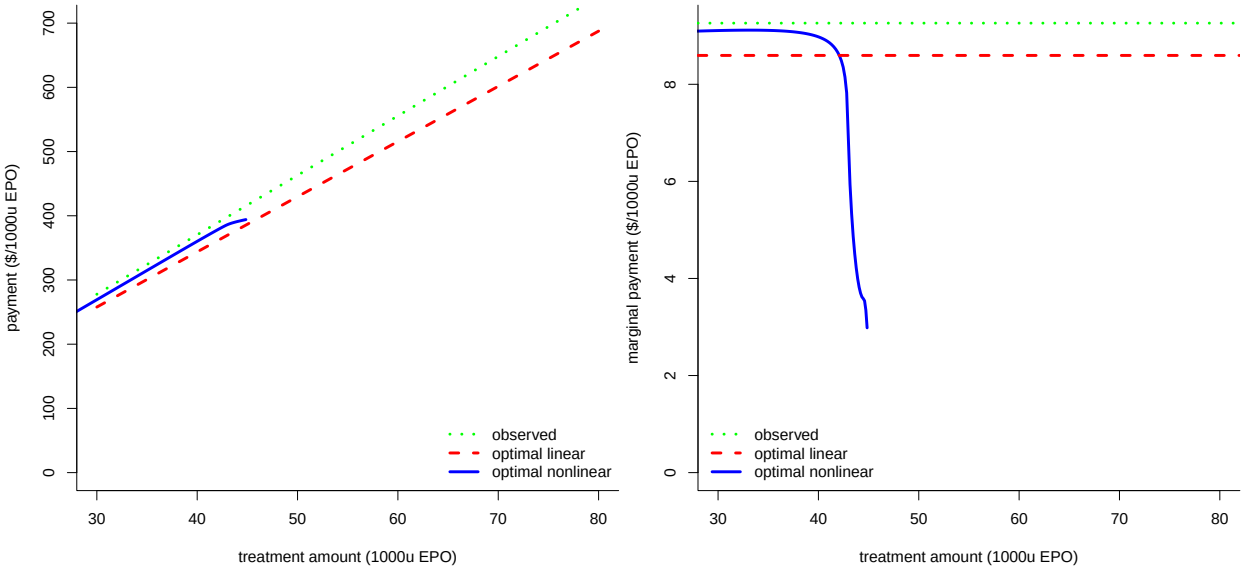
(b) Marginal payment as function of treatment amount



(c) Distribution of treatment amounts

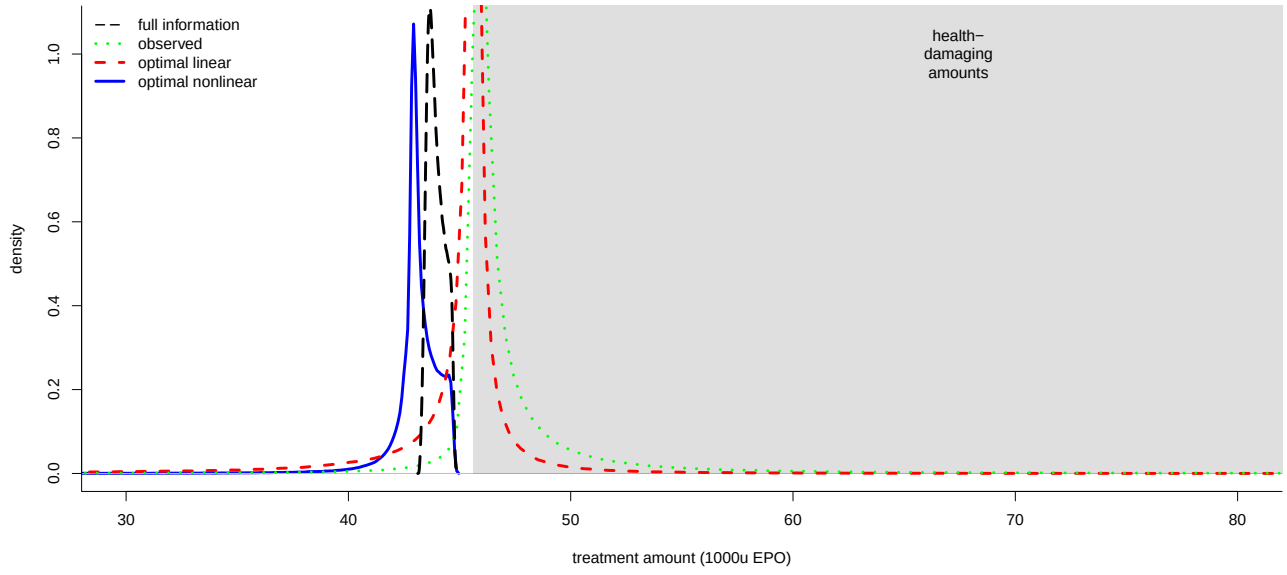
Figure A2. : Optimal nonlinear contract treatment amounts and payments, baseline hematocrit 33-36

Note: 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 A3. : Optimal nonlinear contract treatment amounts and payments, baseline hematocrit 36-39

Note: 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 A1—: Main Estimates of the Reduced Form

Variable	Interval of Baseline Hematocrit > 30 to 33,	> 33 to 36,	> 36 to 39
Baseline hematocrit	-9.29	-6.32	-3.56
	(0.24)	(0.15)	(0.13)
Reimbursement rate	9.53	6.39	3.92
	(3.19)	(2.03)	(1.91)
Age in years	-0.41	-0.37	-0.26
	(0.02)	(0.02)	(0.01)
Female sex	-0.88	1.55	2.89
	(0.55)	(0.40)	(0.34)
CCI=1	9.03	8.03	7.36
	(0.96)	(0.69)	(0.60)
CCI=2	10.73	10.23	8.20
	(0.90)	(0.67)	(0.59)
CCI=3	13.84	11.85	8.57
	(0.94)	(0.72)	(0.60)
CCI=4	15.52	13.91	10.82
	(1.22)	(0.86)	(0.73)
CCI=5	16.53	15.00	11.88
	(1.40)	(1.08)	(0.93)
CCI=6	18.61	18.50	13.83
	(1.87)	(1.48)	(1.21)
CCI=7	26.20	26.00	20.38
	(3.02)	(2.48)	(2.19)
CCI=8	23.92	24.24	14.50
	(3.94)	(3.06)	(2.51)
CCI=9	31.98	32.42	22.85
	(4.98)	(4.17)	(3.81)
CCI=10	23.88	28.45	32.23
	(7.02)	(6.71)	(6.96)
CCI=11	39.10	43.62	39.80
	(11.01)	(8.79)	(7.31)
CCI=12	38.39	33.50	25.66
	(12.51)	(8.06)	(9.82)
Obs. in interval	231,702	405,019	283,024

Note: Each column is a separate regression, estimated via OLS. Regressions also include month and year dummies. Asymptotic standard errors in parentheses, clustered on dialysis center.

*

REFERENCES

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

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

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

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

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

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

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

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

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

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

Centers for Medicare and Medicaid Services. 2007, 2008, 2009*c*. "Medicare Inpatient Claims Data." Research Data Assistance Center. Accessed via the National Bureau of Economic Research.

Centers for Medicare and Medicaid Services. 2007, 2008, 2009*d*. "Medicare Master Beneficiary Summary File base segment." Research Data Assistance Center. Accessed via the National Bureau of Economic Research.

Centers for Medicare and Medicaid Services. 2008. "Medicare 2008 ASP Drug Pricing Files." https://www.cms.gov/Medicare/Medicare-Fee-for-Service-Part-B-Drugs/McrPartBDrugAvgSalesPrice/01a_2008aspfiles, downloaded 12/06/2022.

Centers for Medicare and Medicaid Services. 2008, 2009*a*. "Healthcare Cost Report Information System (HCRIS) Dataset – Renal." https://www.cms.gov/Research-Statistics-Data-and-Systems/Downloadable-Public-Use-Files/Cost-Reports/Renal-Facility-265-1994-form, downloaded 12/05/2022.

Centers for Medicare and Medicaid Services. 2008, 2009*e*. "Medicare Outpatient Claims Data." Research Data Assistance Center. Accessed via the National Bureau of Economic Research.

Centers for Medicare and Medicaid Services. 2009*b*. "Medicare 2009 ASP Drug Pricing Files." https://www.cms.gov/Medicare/Medicare-Fee-for-Service-Part-B-Drugs/McrPartBDrugAvgSalesPrice/01a1_2009aspfiles, downloaded 12/06/2022.

Chalkley, Martin, and James M. Malcomson. 2000. "Government Purchasing of Health Services." In *Handbook of Health Economics*. Vol. 1, Part A of *Handbook of Health Economics*, , ed. Anthony J. Culyer and Joseph P. Newhouse, 847–890. Elsevier.

Chiappori, Pierre-Andre, and Bernard Salanié. 2003. "Testing Contract Theory: A Survey of Some Recent Work." In *Advances in Economics and Econometrics: Eighth World Congress*. Vol. 1, , ed. M. Dewatripont, L. P. Hansen and S. Turnovky, 115–149. Cambridge University Press.

Choné, Philippe, and Ching-to Albert Ma. 2011. "Optimal Health Care Contract Under Physician Agency." *Annals of Economics and Statistics/Annales d'Économie et de Statistique*, 229–256.

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

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

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

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

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

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

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

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

Einav, Liran, Amy Finkelstein, Yunan Ji, and Neale Mahoney. 2021. "Voluntary Regulation: Evidence from Medicare Payment Reform*." *The Quarterly Journal of Economics*, 137(1): 565–618.

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

Eliason, Paul J, Benjamin Heebsh, Riley J League, Ryan C McDevitt, and James W Roberts. 2022. "The Effect of Bundled Payments on Provider Behavior and Patient Outcomes: Evidence from the Dialysis Industry." unpublished manuscript.

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

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

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

Foley, Robert N. 2006. "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.

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

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

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

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

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

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

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

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

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

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

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

Maskin, Eric, J. J. Laffont, J.C. Rochet, T. Groves, R. Radner, and S. Reiter. 1987. "Optimal Nonlinear Pricing with Two-Dimensional Characteristics." *Information, Incentives and Economic Mechanisms (essays in honor of Leonid Hurwicz)*, 256–266. Minneapolis:University of Minnesota Press.

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

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

McGuire, Thomas G. 2000. "Physician Agency." In *Handbook of Health Economics*. Vol. 1, Part A of *Handbook of Health Economics*, , ed. Anthony J. Culyer and Joseph P. Newhouse, 461–536. Elsevier.

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

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

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

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

NKF-KDOQI. 2007. "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.

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

Quan, Hude, Vijaya Sundararajan, Patricia Halfon, Andrew Fong, Bernard Burnand, Jean-Christophe Luthi, L. Duncan Saunders, Cynthia A. Beck, Thomas E. Feasby, and William A. Ghali. 2005. "Coding Algorithms for Defining Comorbidities in ICD-9-CM and ICD-10 Administrative Data." *Medical Care*, 43(11): 1130–1139.

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

Rochet, Jean-Charles, and Lars A. Stole. 2003. "The Economics of Multidimensional Screening." In *Advances in Economics and Econometrics: Theory and Applications, Eighth World Congress*. Vol. 1 of *Econometric Society Monographs*, , ed. Mathias Dewatripont, Lars Peter Hansen and Stephen J. Turnovsky, 150–197. Cambridge University Press.

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

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

Shinkman, Ron. 2016. "The big business of dialysis care." *NEJM Catalyst*, 2(3). https://catalyst.nejm.org/the-big-business-of-dialysis-care/.

Singh, Ajay K., Lynda Szczech, Kezhen L. Tang, Huiman Barnhart, Shelly Sapp, Marsha Wolfson, and Donal Reddan. 2006. "Correction of Anemia with Epoetin Alfa in Chronic Kidney Disease." *New England Journal of Medicine*, 355(20): 2085–2098.

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

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

United States Renal Data System. 2011. *USRDS Annual Data Report: Atlas of chronic kidney disease and end-stage renal disease in the United States.* Bethesda, MD:National Institutes of Health, National Institute of Diabetes and Digestive and Kidney Diseases.

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

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

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

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

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

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