Insurance with Multiple Actors

Jason Abaluck
Oren Sarig
Jintaek Song*

March 26, 2026

Abstract

In most insurance contracts, multiple parties jointly determine expenditures: patients and doctors in health insurance, drivers and mechanics in auto insurance, or homeowners and contractors in property insurance, among others. In such settings, we analyze how an insurer or social planner divides cost-sharing between risk-averse consumers (e.g. patients) and firms (e.g. doctors). With no contracting frictions, only "visit copays" are necessary on the consumer-side; firm cost-sharing otherwise achieves first-best expenditure levels. Intensive margin consumer cost-sharing is necessary only as a solution to contracting frictions, such as asymmetric information about firm type. A key sufficient statistic is the marginal rate of substitution between consumer and firm cost-sharing; this determines the optimal level of firm cost-sharing, as well as optimal visit copays. For Medicare outpatient spending, we calibrate that a physician reimbursement schedule which offered lower profits when spending was higher would reduce risk for both patients and providers. For prescription drugs, we estimate that a dollar of physician paperwork costs is 33–120 times more impactful than a dollar of consumer cost-sharing, implying that insurers should use small physician penalties as the primary means of cost-control.

*Abaluck: Yale University and NBER, jason.abaluck@yale.edu, Sarig: Assured Allies sarig.oren@gmail.com, Song: University of Florida jintaek.song@ufl.edu. For helpful comments and suggestions, we thank Zarek Brot, Tatyana Deryugina, Amy Finkelstein, Jon Gruber, Yunan Ji, Timothy Layton, Barry Nalebuff, Julian Reif, Molly Schnell, and Mark Shepard. The authors have no conflicts of interest to disclose.

1 Introduction

In most settings where consumers purchase insurance, multiple parties determine costs in the case of a negative shock. For example, in health insurance, patients choose whether to see a doctor (and which doctor to see), but doctors (mostly) decide which medical procedures are warranted. Following a car accident, drivers choose car mechanics but car mechanics (mostly) decide what repairs are necessary. When their roof is damaged in a fire, homeowners choose roofers who are responsible for assessing and repairing the relevant damage (subject to homeowner approval).

Insurers who want to mitigate moral hazard can therefore incentivize consumers or firms. In health insurance, consumers face coinsurances but physicians are incentivized by risk-sharing contracts (Delbanco, 2014) or by "prior authorization" requirements that effectively tax them for more expensive care (Huckfeldt et al., 2024). In auto insurance, drivers face deductibles, and car mechanics receive performance-based incentives from insurers to keep repair costs low (Huetter, 2017). In homeowners insurance, property owners have deductibles, and property insurers use fixed-price payments to incentivize contractors (Morton, 2022).

This paper investigates how insurers optimize cost-sharing for consumers and firms in settings where both jointly determine expenditures. We theoretically characterize when consumer cost-sharing is warranted, and then apply our model to two empirical settings. For Medicare outpatient spending, we calibrate where optimal physician reimbursement lies between "fee-for-service" reimbursement at cost and up-front "prospective payment." For prescription drugs, we compute optimal cost-sharing given new estimates of the impacts of both consumer cost-sharing and firm side penalties (prior authorization) on expenditures.

Our theory develops the problem of a social planner insurer who seeks to maximize consumer surplus for fixed firm profits. In the model, consumers receive a negative shock that can be addressed by spending money (e.g. obtaining medical treatment or getting their car repaired). They choose to either stay home or visit one of several symmetric firms who compete on services to attract consumers. Consumers pay premiums to insurers, who in turn reimburse firms if they provide services to a consumer. The insurer can combat moral hazard either by taxing consumers for each dollar of services or taxing firms for each dollar of services. A "tax" for firms refers to the totality of the incentives created by their contract: a firm is being taxed on net if their per-consumer profit is lower the more they spend. The desired "first-best" service level is the service level consumers would choose for themselves were they paying out of pocket.

Our first result is that, if utility is transferable – meaning that firms can charge a price or make side-payments to consumers – there is "tax neutrality". This means that service levels and net transfers are the same regardless of which side bears cost-sharing. This result is analogous to conventional neutrality results in the public economics literature (Kotlikoff and Summers, 1987). If you tax consumers for larger service levels, demand for services falls and firms offer consumers a better deal (i.e. a lower price or a larger side payment). The side payments required for this neutrality result are often absent in multi-party insurance settings. In healthcare, anti-kickback laws prevent doctors from paying consumers to seek services. More generally, insurers prefer (if they can) to ban side payments so that utility is nontransferable. This is because side payments undermine the ability of insurers to efficiently combat moral hazard. This result motivates consideration of the nontransferable case.

With nontransferability, the situation is radically different. When spending is within the purview of firms, firm contracts can achieve first-best service levels with zero consumer cost-sharing for intensive margin services and only "visit costs" for extensive margin services. The case for consumer coinsurances thus relies on contracting frictions. The basic intuition is that firm-side incentives transfer risk from consumers to firms. This follows immediately if firms are risk neutral, but similar arguments hold for risk averse firms when either i) contracts can be frontloaded, ii) firms are sufficiently uncompetitive, or iii) firms see a large number of consumers with uncorrelated risks. The challenge for insurers is structuring firm incentives in such a way that firms endogenously choose the right service levels for all hidden types concurrently (e.g. doctors spend the right amount on both sick and healthy patients). Our result shows that this is possible, and optimal firm contracts become weaker and more linear when firms have more market power. When optimal contracts are linear, firms can achieve first-best service levels simply by reducing patient cost-sharing and increasing firm cost-sharing in a way that holds service levels constant.

There are three types of contracting frictions that restore some role for consumer cost-sharing for intensive-margin services: if contracts must be linear (and firms are competitive), if there is asymmetric information about firm type, or if firms are risk-averse and lump sum payments are not possible. In each case, we discuss how the optimal mix of firm and consumer cost-sharing changes.

In our model, a few sufficient statistics determine the optimal mix of consumer and firm cost-sharing. Intensive margin moral hazard – the amount spent conditional on seeing a given firm – is controlled by firm incentives in our baseline case. The shape of the optimal firm contract to deal with intensive margin moral hazard depends on the relative impact of firm and consumer cost-sharing on service levels for different types (i.e. the marginal rate of substitution), which is a function of the

underlying competitiveness of firms. Optimal visit copays depend on the marginal effect of consumer cost-sharing on extensive marginal moral hazard; i.e. the probability that a consumer visits a firm at all.

We apply our model to analyze optimal cost-sharing design for outpatient care and prescription drugs. For Medicare outpatient care, we calibrate our model using existing estimates of utilization effects from Cabral et al. (2018) as well as the impact of physician reimbursement on expenditures from Clemens and Gottlieb (2014). In the status quo, physicians earn higher profits per patient when they spend *more*, so schedules which hold average profits the same but reduce profits when spending is higher can reduce *physician* risk as well as patient risk. We find that physicians should be paid more up front per visit, but with profits declining the more they spend, while consumer cost-sharing outside of visit copays should be reduced from 20% to near 0. The optimal physician schedule offers higher profits up front with profits asymptoting towards zero for the highest cost patients – the schedule resembles a "physician deductible", with marginal penalties much higher for low cost than high cost patients. For prescription drugs, we estimate the impact of both prior authorization (penalizing doctors) and consumer cost-sharing on expenditures. We find that a dollar of physician cost-sharing (in the form of paperwork costs) is 33–120 times more impactful than a dollar of consumer cost-sharing on drug expenditures. As a result, our model implies that small physician penalties should be the principal means of controlling moral hazard.

Our theoretical results revise the conventional view of both consumer and firm cost-sharing. With perfect competition, we recover the traditional Baily-Chetty result where consumer cost-sharing alone is used to control moral hazard (e.g., Baily, 1978; Chetty, 2006). But with even an ϵ amount of imperfect competition, insurers switch to using firm cost-sharing as the principal means of cost control, with consumer cost-sharing serving only a second-best role for intensive margin services. Our results suggest that firm cost-sharing should likewise be applied cautiously. A common incentive scheme in practice are up-front payments, or "prospective payment" in healthcare. When firms have substantial market power, our results raise the concern that prospective payment will lead firms to curtail spending too much because they are not disciplined by consumer choice. It is also commonly argued that strong incentives for firms are more acceptable when competition disciplines quality.[1] Our

[1]Ellis and McGuire (1986) note, "It is frequently asserted that 'competitive pressures' under prospective payment will keep hospitals from undersupplying services... [but] one has to recognize that not all patients are attractive to hospitals under prospective payment – only patients whose costs are less than the prospective payment amount. How are hospitals to attract low-cost patients and discourage high-cost ones?" Our model formalizes that, while prospective payment impacts all types proportionately to firm cost-sharing when firms are monopolistic, Ellis and McGuire (1986) is exactly right that prospective payment disproportionately distorts care for high cost enrollees when firms are competitive.

model shows that this is not correct either: if competitive firms earn low profits, the incentives from prospective payment are too strong and would lead low-margin firms to turn away high cost patients. Prospective payment is generally only optimal at intermediate levels of competition.

Our model relates to several literatures beyond the aforementioned Baily-Chetty results on optimal social insurance. A central question in health economics has been whether it is necessary to control costs with consumer cost-sharing, or whether provider incentives suffice; our model clarifies and unifies an earlier literature on this topic. The most closely related paper is Ellis and McGuire (1990), which analyzes a model where doctors altruistically weight consumer utility with homogeneous consumers, and so consumer and firm cost-sharing are perfect substitutes. The baseline case of their model exhibits the same behavior as the monopolistic limit of our model: firms transfer risk to consumers. In their model, consumer cost-sharing may be justified by a cap on the allowable firm cost-sharing (they assume firms cannot be taxed more than 100% of expenditures). In our model, consumer cost-sharing is needed either because of the extensive margin or contracting failures. Pauly and Ramsey (1999) model a managed care insurance plan that can use either cost-sharing or a uniform cap on expenditures to control costs. Unexpectedly, this model is equivalent to the special case of our model where firms have an ϵ amount of market power, and firms are constrained to use linear contracts: in this case firm cost-sharing pushes all high-expenditure consumers down to a uniform level of spending. This literature is reviewed in McGuire (2011).

There is also a recent literature studying specifically the efficiency of prior authorization and consumer cost-sharing. These studies call into question the efficiency of consumer cost-sharing, finding that it leads consumers to cut back on potentially valuable care (Brot-Goldberg et al., 2017) with negative health impacts (Chandra et al., 2024). Brot-Goldberg et al. (2023) and Sarig (2024) explicitly find that the cost savings from prior authorization exceed the downsides. This paper complements these approaches, providing a general theory for when firm cost-sharing (such as prior authorization) is preferred to consumer cost-sharing. Our substantive conclusion in the case of Medicare Part D is consistent with earlier studies: the theoretical and empirical rationale for consumer cost-sharing in this setting is weak.

Section 2 lays out our model, Section 3 characterizes the solution with transferability and discusses when insurers might prefer non-transferability, Section 4 characterizes the solution with non-transferability, Section 5 calibrates the optimal physician contract for Medicare outpatient services, Section 6 estimates optimal cost-sharing for prescription drugs in Medicare Part D, and Section 7 concludes.

2 Model Set-up

In our model, a risk-averse consumer experiences a shock whose consequences can be mitigated with costly services. The consumer chooses among several symmetric firms based on the level of services they offer. Consumers pay premiums to an insurer who in turn reimburses firms for services provided to consumers. The insurer can combat moral hazard either by implementing cost-sharing on the consumer side or firm side; that is, they can tax consumers for each dollar that is spent or they can tax firms for each dollar that is spent.

We can think of this model as subsuming many different types of insurance with "ex post moral hazard": consumers choosing a doctor to treat an illness, drivers choosing an auto mechanic following an accident, or homeowners choosing a contractor after property damage.

2.1 Consumers

Our specification of consumer utility follows Einav et al. (2013). A consumer experiences a shock λ that reduces her utility, but this reduction can be mitigated through services costing m. For example, an ill patient seeks medical treatment costing m, or a driver gets in a car accident requiring m dollars worth of repairs.

In our model, it will be useful to distinguish three stages. In "Stage 0", consumers do not yet know their realization of λ; social planners or competitive insurers (as proxies for consumers) thus care about ex ante expected utility in this stage (we call this "Stage 0" because consumers make no active choice at this stage). In "Stage 1", consumers make an extensive margin decision of whether to patronize a firm to receive treatment. For example, a patient decides whether to visit a doctor, or a policy-holder decides whether to see a car mechanic. Among consumers who decide to visit, firms compete for consumers by offering a menu consisting of services costing m, and in some cases, kickbacks or side payments. In Stage 2, consumers who chose to visit receive i.i.d. taste shocks and decide which firm to patronize.[2]

We will now describe each of these stages in reverse order.

Let i index consumers and j firms. Consumer i's ex post utility in Stage 2 should i choose firm j

[2]Two timing subtleties simplify the problem dramatically. First, consumers decide to visit before firms offer their menus (e.g. a consumer decides if they are sick enough they need to see a doctor); firms thus compete to attract visiting consumers, and the visit decision depends only on *expected* rather than realized service levels. Second, firms cannot observe taste shocks when they choose service levels for each type. Service levels are thus functions of λ, but not idiosyncratic preferences.

is given by:

$$u_{ij} = W_i - \alpha_j(\lambda_i) - \tau_1(m_j(\lambda_i)) - p + h(m_j(\lambda_i) - \lambda_i) + e_{ij} \tag{1}$$

where W_i is wealth, $\alpha_j(\lambda_i)$ is the side payment required of firm j for consumers of type λ_i (this can be positive or negative), τ_1 is consumer cost-sharing (a potentially nonlinear function of service costs), p are premiums (explained below as part of the firm problem), and h is an increasing but concave function with $h'(0) = 1$, so that λ represents the first best level of spending that would be chosen if consumers paid out of pocket. By a "side payment", we mean a (sometimes illegal) payment not intermediated by the insurer at all; for example, a doctor gives patients who visit \$100, knowing she will receive \$200 in reimbursement from insurers. The e_{ij} are i.i.d. extreme value shocks representing the consumer's idiosyncratic preference for firms (e.g. her convenience and taste for a given doctor). Below, we suppress the i subscripts in most cases to ease notation.

In Stage 1, consumers make an extensive margin decision of whether to visit any firm at all. Consumers make this decision after observing $(m(\lambda), \alpha(\lambda))$ for each firm, but prior to observing firm-specific taste shocks. If consumers stay home, they receive utility $U^0(\lambda) = W + h(0 - \lambda) - p$. If they visit a firm, they receive expected utility given by:

$$U^{visit}(\lambda) = E_e(\max_j W - \alpha_j - \tau_1(m_j) + h(m_j - \lambda) - p + e_j) \tag{2}$$

$$= \sigma \log \sum_{j=1}^{J} \exp\left(\frac{1}{\sigma}(W - \alpha_j - \tau_1(m_j) + h(m_j - \lambda) - p)\right) \tag{3}$$

If they decide to patronize a firm, the e_j shocks are realized and consumers choose the firm that maximizes equation 1. The parameter σ corresponds to the variance of the e_j shocks (given the normalization of a constant marginal utility of wealth). This variance determines the competitiveness of firms.

We define the indirect utility of a type λ consumer in stage 1 as:

$$U(\lambda) = \max\{U^{visit}(\lambda), U^0(\lambda)\} \tag{4}$$

Finally, in Stage 0 (where no active decision is made by consumers), risk averse consumers have

ex ante utility given by:

$$E_\lambda\left(v(U(\lambda))\right) \tag{5}$$

for some strictly concave v and where the expectation is taken over $F(\lambda)$, the distribution of the λ types. This concave wrapper is necessary so that consumers value insurance.

2.2 Firms

There are J symmetric firms, each of which offers a menu $\left(m_j(\lambda), \alpha_j(\lambda)\right)$ to consumers with realized shock λ who have decided to visit. $m_j(\lambda)$ is the amount of service provided to consumers, and $\alpha_j(\lambda)$ is the side payment that firms require (or make). Both are chosen prior to the realization of the idiosyncratic e_{ij} shocks.

We normalize the total market size to 1. Knowing consumer demand, firm j chooses $\left(m_j(\lambda), \alpha_j(\lambda)\right)$ that maximizes profits for each type, $\pi(\lambda)$, given by:

$$\begin{aligned} m_j(\lambda), \alpha_j(\lambda) &= \text{Argmax}_{m,\alpha}\, \pi(\lambda) = \text{Argmax}_{m,\alpha}\, s_j(m,\alpha,\lambda)\cdot(R-\tau_2(m)+\alpha) \\ \text{s.t.} \quad & m \geq k \end{aligned} \tag{6}$$

where R is the fixed component of firm reimbursement,[3] τ_2 are firm cost-sharing rates and $s_j(m,\alpha,\lambda)$ is firm j's market share among type λ consumers who visit a firm:

$$s_j(m,\alpha,\lambda) = Prob(i : j = \text{Argmax}_j u_j(m,\alpha,\lambda))$$

The equilibrium menu $(m_j^*(\lambda), \alpha_j^*(\lambda))$ offerred by each firm for each λ is a symmetric Nash equilibrium in which firms maximize profits given the maximizing choices of rival firms. The constraint $m \geq k$ means that once the consumer chooses a firm other than the outside option, the firm must spend at least k dollars. This captures extensive margin spending such as fees paid per doctor visit that are incurred whenever a patient sees a doctor rather than choosing to stay home, or the cost of a checkup by a car mechanic (we also consider the case where there are no such fees). In the non-transferable case, this spending will provide an important potential rationale for consumer cost-sharing.

In our baseline model, we assume that firms maximize profits. An alternative specification, espe-

[3] More formally, R is the per-consumer profits if the firm were to provide no services in a setting where $k = 0$.

cially common in healthcare, allows providers to be altruistic and choose service level to maximize a weighted average of profits (given by equation 6) and beneficiary ex post utility (given by equation 1). While our baseline model does not include altruism to ease the notational burden, we consider an extension below which allows for this possibility.

2.3 Insurers

In our model, competitive insurers choose consumer cost-sharing $\tau_1(m_j)$ and firm cost-sharing $\tau_2(m_j)$ schedules to maximize consumer ex ante welfare given by:

$$E_\lambda \left[v(U(\lambda, \{\alpha_j^*(\lambda, \tau_1, \tau_2), m_j^*(\lambda, \tau_1, \tau_2)\}, p(\tau_1, \tau_2), \tau_1, \tau_2)) \right] \tag{7}$$

where α_j^* and m_j^* are the optimal side payments and service levels which solve equation 6, p are premiums (defined below), and consumers solve equation 4 to decide whether to visit firms.

The level of R is pinned down by the additional constraint $E(\pi(\lambda)) = \bar{\pi}$, a negotiated average per consumer profit. The idea of the contract is that firms will on average earn $\bar{\pi}$, but are also incentivized by firm cost-sharing, so they will be paid more per consumer the less that they spend. This formulation allows us to separate the results of firm/insurer bargaining (which ultimately determines $\bar{\pi}$) from the incentive contract that insurers use to control service levels via firm cost-sharing. We will treat $\bar{\pi}$ as a constant in most of our analysis.[4] Since firm profits are exogenously fixed, insurers that maximize consumer surplus are also maximizing total surplus.

We further define side-payments, service levels, and fixed fees for consumer i inclusive of the extensive margin choice as (respectively): $\alpha_i^* = 1(U_i^{visit} \geq U_i^0)\alpha_{j(i)}^*$, $m_i^* = 1(U_i^{visit} \geq U_i^0)m_{j(i)}^*$, and $R_i^* = 1(U_i^{visit} \geq U_i^0)R$ where $j(i)$ is the plan chosen by i in equilibrium.

In a symmetric equilibrium, premiums are given by:

$$p(\tau_1, \tau_2) = [m_i^* - E(\tau_1(m_i^*)) - E(\tau_2(m_i^*)) + E(R_i^*) + JS] \tag{8}$$

$$= E(m_i^*) - E(\tau_1(m_i^*)) + J\bar{\pi} - E(\alpha_i^*) \tag{9}$$

where J is the total number of firms. In other words, premiums finance services not paid out of pocket, and firm profits less any side payments (since these side payments are received directly by firms).

[4] This assumption is not impactful for our empirical exercises below, but it would be important to relax when studying the empirical relationship between competition and contracts, since competition would impact contracts both via $\bar{p}i$ and via the modeled impact of σ.

In equilibrium, the average financial burden of all consumers (paid either via premiums, side payments, or out of pocket costs) is the sum of p^*, $E(\alpha_i^*)$, and $E(\tau_1(m_i^*))$ which is given by:

$$E(m_i^*) + J\bar{\pi}$$

That is, consumers pay for services and per capita firm profits.

It may be helpful to recap the model with an example: consider the case of a car mechanic in a world with no side payments ($\alpha = 0$). A type λ_h consumer requires engine repairs which (in equilibrium) cost $m = \$2,000$. The insurer fully reimburses this amount (in this example, there is no consumer cost-sharing). Suppose the average consumer in this market requires $\$1,500$ worth of repairs in equilibrium (so the λ_h consumer is more expensive than average). The insurer negotiates with the mechanic that they will reimburse at a rate where the mechanic averages $200 in profits per consumer (this is $\bar{\pi}$). Insurers also say that all mechanics will be incentivized to spend less by paying 2% (τ_2) of any repair costs (m) back to the insurer. This cost-sharing tax averages $30 per consumer ($\$1,500 \cdot 0.02$), so the mechanic is paid at a rate $R = \$230$ per consumer. For this consumer (with $m = \$2,000$), the mechanic nets $190 (230-0.02 $\cdot$ 2000). Due to the incentive contract, higher type consumers are less profitable on average, counteracting moral hazard.

2.4 Equivalence with a Bargaining Model

Ellis and McGuire (1990) consider a model in which patients and a single altruistic doctor bargain over service levels and discuss how the optimal contract depends on the relative bargaining power of patients and physicians.

We show in Appendix A that, in a special case of our model, the equilibrium service-level condition is equivalent to the outcome of Nash bargaining in which consumers are randomly assigned a firm and bargain given the outside option of switching to an alternative (symmetric) firm. Specifically, the firm's Nash bargaining weight equals $\sigma/(1+\sigma)$, so the market power parameter σ microfounds the relative bargaining position of consumers and firms.

Our model also differs from Ellis and McGuire (1990) in modeling the moral hazard problem as arising from heterogeneous, unobservable consumer types λ. In a model with a single type of consumer, insurers could solve the moral hazard problem by agreeing to reimburse only the first-best level of services. In our model, firms know each consumers' type but insurers do not, and so

insurers try to set cost-sharing in such a way that firms will select efficient service levels for all types simultaneously given the distribution of types in the population.

Finally, Ellis and McGuire (1990) assume that firms cannot be given a steeper incentive than prospective payment – in other words, firms cannot be taxed more than \$1 for each dollar they spend. We do not make this assumption in our baseline model, but we discuss below circumstances when it might be binding.

3 Theory with Transferability

With transferability, firms can make side payments to consumers ($\alpha_j \neq 0$). These can be positive or negative; firms might subsidize consumers to visit, or alternatively, they might charge consumers an extra fee not reimbursed by the insurer.

3.1 Transferability Implies Neutrality

The main result in this section is that, with transferable utility, there is tax neutrality: the level of services chosen depends only on the sum of consumer and firm cost-sharing. Additionally, the side payment adjusts so that the incidence of the tax is the same regardless of which side is taxed.

The formal result is as follows:

Proposition 1: Assume that $(m_j^*(\lambda), \alpha_j^*(\lambda))$ is a symmetric equilibrium where consumers make visit decisions according to Equation 4, choose among firms according to Equation 1, and firms choose service levels and side payments according to Equation 6. Assume further that the combined cost-sharing schedule $T(m) \equiv \tau_1(m) + \tau_2(m)$ is strictly increasing. Then, m_i^*, the consumer's net out-of-pocket payment $\tau_1(m_i^*) + \alpha_i^*$, and the firm's net profit per beneficiary $R - \tau_2(m_i^*) + \alpha_i^*$ all depend only on $T(m) = \tau_1(m) + \tau_2(m)$, not on $\tau_1(m)$ and $\tau_2(m)$ separately.

We prove this result in Appendix B. This proposition states that there is "tax neutrality" in the transferable model. It does not matter whether the insurer imposes cost-sharing on consumers or firms. Equilibrium service levels, consumer utility and firm profits will be the same regardless.

The intuition is straightforward: consumers at the end of the day care about their net out of pocket costs, $\tau_1(m_j(\lambda)) + \alpha_j(\lambda)$ and service levels $m_j(\lambda)$ while firms care about their profits. If the cost-sharing rate is shifted from the firm to the customer at any given level of m_j while keeping the

total tax rate constant, the firm can reduce the side payment charged to offer the same m_j, while keeping consumer net out-of-pocket costs and their own profits unchanged.

3.2 Reasons for Non-transferability

There are a variety of reasons that transferability is generally not present in insurance markets, meaning that the neutrality result above no longer holds.

A general theoretical reason is implicit in the above argument: insurers would like, if possible, to use firm-side cost-sharing in lieu of consumer cost-sharing in order to transfer risk from risk-averse consumers to risk-neutral firms. In the presence of side payments, the neutrality result above makes this risk transfer impossible. Preventing side payments enables insurers to use cost-sharing for firms as a separate instrument without incurring the usual costs of cost-sharing for consumers. We prove a formal result along these lines in a special case of our model in Appendix C.[5]

In many markets, there are also more practical reasons for non-transferability. For example, anti-kickback laws (perhaps motivated by the theory above) prevent side payments between physicians and consumers.

4 Theory with Non-transferability

In light of the results in the previous section, we next consider how an insurer chooses consumer and firm cost-sharing in settings where type-specific transfers are prohibited (so $\alpha = 0$).

We first show that when contracting is unrestricted, consumer cost-sharing is used only to control the decision of whether or not to visit, and only via a copay (i.e. a visit cost that is not proportional to expenditures). Firm cost-sharing achieves a suitably defined constrained first-best among all beneficiaries who visit. We discuss how, in this context, the shape of the optimal contract varies with the competitiveness of firms. Next, we discuss valid rationales for consumer cost-sharing: when contracts must be linear, when there is asymmetric information about firm type, or if firms are risk-averse and lump sum payments infeasible. Finally, we consider extensions of the model that do not in general justify positive consumer cost-sharing: altruism, behavioral hazard, and imperfect adherence.

[5]This result does importantly rely on the exogeneity of $\bar{\pi}$. In other words, we are not modeling the determination of firm profits in equilibrium, and with endogenous $\bar{\pi}$, transferability could impact the division of surplus between consumers and firms. In that more general case, our partial equilibrium result shows that insurers have one reason to prefer non-transferability which would need to be weighed against other considerations.

4.1 First Best with Unrestricted Contracting

The next proposition provides an organizing principle for the remainder of our analysis. It shows that, in the absence of contracting frictions, the optimal contract is such that consumer cost sharing is needed only to curb extensive-margin moral hazard. Further, full consumption smoothing is achieved among visitors, as their out of pocket costs do not vary with their service costs.

In this setting, optimal firm cost-sharing among visitors leads to the same expenditures a social planner would have chosen with full information on visitor types given their constraints. We refer to this allocation as a "constrained first best." The constraint is that the insurer cannot enforce visit decisions, and so introduces a visit copay that sacrifices some consumption smoothing between visitors and non-visitors in order to deter unnecessary visits. Conditional on this visit copay, the allocation is equivalent to an optimal Arrow-Debreu state contingent contract among visitors.[6] The constrained first-best gives the expenditure levels achievable by a social planner who uses an Arrow-Debreu state-dependent contract among visitors, plus consumer coinsurances to control the visit schedule. We formally characterize the optimal contract and and its equivalence to a "constrained first best" in Proposition 2.

A direct corollary is that, in a hypothetical Arrow–Debreu first-best world where it is optimal for everyone to visit (for example, if $\lambda > k$ for all individuals), the optimal τ_v, τ_1, and τ_2 also implement the *unconstrained* first-best allocation. In this case there is no extensive-margin moral hazard to deter, so the constrained first best coincides with the full Arrow–Debreu first best.

Proposition 2: Define the constrained first-best service levels $m^*_{cfb}(visit)$ as the service levels the social planner would choose if they could directly choose service levels and a visit copay (τ_v),[7] but visit decisions are made endogenously according to equation 4. Let τ^*_v denote the visit copay they would choose (a constant copay paid if you visit, which does not otherwise depend on service levels m). Assume $\alpha_j = 0$. Let $g(m) = R - \tau_2(m)$ denote firm profits per consumer (see equation 6). When $\bar{\pi} > 0$, then $m^*_{cfb}(visit)$ is achievable for a given level of profits with coinsurances given by: (τ_v, τ_1, τ_2)

[6] Assuming no contract leaves visitors with more money than they are left with after paying their per visit fee.

[7] These differ from the true first-best service levels, $m = \lambda$, because state-dependent utility means that consumers who have to pay a visit copay will prefer slightly different service levels than consumers with no such copay.

$= (\tau_v^*, 0, \tau_2^*(m))$. τ_v^* and $\tau_2^*(m)$ are characterized by the following equations:

$$[\tau_v]: \; P(\text{visit}) \cdot k \cdot \frac{E[v'(U \mid \text{visit})] - E[v'(U \mid \text{all})]}{E[v'(U \mid \text{all})]} = -E[m_{min} - \tau_v k] \frac{\partial P(\text{visit})}{\partial \tau_v},$$

$$[\tau_2(m)]: \; g(m) = A \cdot \exp\left(-\frac{\left(1 - \frac{1}{J}\right)\phi}{\sigma} m \right),$$

where

$$\phi = \begin{cases} 1, & m_{\min} > k, \\ \dfrac{E[v'(U) \mid \text{all}]}{E[v'(U) \mid m > k]}, & m_{\min} = k, \end{cases}$$

and $m_{\min}$ is the minimum equilibrium service level among all visiting types under the optimal policy. $U(\lambda)$ is defined as in equation 4 at the policy regime $(\tau_v^*, 0, \tau_2^*)$.

Intuitively, the τ_v equation is a conventional Chetty-Bailey equation for the visit copay. The left-hand side gives the risk protection harm from increasing τ_v (and redistributing away from sicker visitors), and the right-hand side gives the moral hazard benefits, which are proportional to $m_{min} - \tau_v k$ for each visitor, the portion of service costs not paid out of pocket.

$g(m)$ characterizes profits for each type as a function of service levels for that type under the optimal contract. Under this contract, firms make smaller profits when consumers spend more, with the marginal penalty larger for lower-cost consumers to offset the fact that firms want to increase service levels especially for these consumers because they are more profitable. The specific functional form for $g(m)$ results from the choice of logit errors. The ϕ term characterizes the marginal utility of health in monetary terms (a term which arises due to state-dependence because $h(\cdot)$ enters the marginal utility for visiting beneficiaries).

Note that if there is no constraint on the use of front-loading in firm contracting, the optimal contract can in principle expose firms to arbitrarily little risk. To reduce risk while holding fixed firm incentives, we can scale $g(m)$ down by a constant, while paying firms more up front. In Section 4.2, we discuss various reasons this might not be feasible in practice.

The contract $g(m)$ looks very different depending on the degree of competition σ. Figure 1 graphs this contract for high and low σ. When σ is small and firms highly competitive, incentives are in general very strong and the optimal contract is convex: larger marginal penalties are needed at lower service levels. For low cost consumers, firms earn especially large margins, so they have a strong incentive to increase spending to attract more such consumers, and large marginal penalties

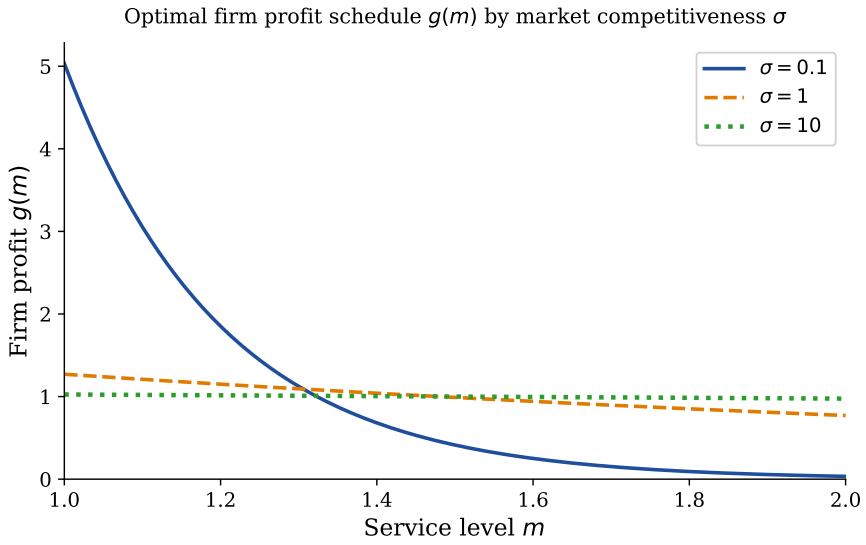


Figure 1: Optimal firm profit schedule $g(m)$ under the optimal contract (Proposition 2), for three values of the demand parameter σ governing firm market power. Each curve is normalized so that average profits equal 1. Higher σ (more market power) yields a flatter, more linear schedule; lower σ (more competition) yields a steeper, convex schedule requiring stronger incentives at low service levels.

are necessary to offset this. When σ is large and firms have a lot of market power, weaker incentives suffice to reduce firm spending. When $g(m)$ is flat, this means that margins are similar regardless of consumer type (since firm cost-sharing is a small fraction of profits). As a result, the same marginal incentives work to constrain spending at each spending level and $g(m)$ looks more linear. We return to this observation below as it suggests both that linear contracts will be optimal when firms are uncompetitive, and also that firms can achieve the constrained first best by targeting only expected service costs.

None of this yet implies that the first-best contract is practical. One might worry about a contract which hinges on a specific parametric model of firm competition, or that requires insurers to know effort costs for all firms.[8] We next discuss these considerations.

4.2 Reasons for Consumer Cost-Sharing

Several extensions to our model restore some role for intensive margin consumer cost-sharing. We summarize these in Table 1, along with tempting rationales that our model suggests do not rationalize

[8]The effort cost assumption is not completely outlandish – programs like Medicare go to great effort to estimate region-specific provider marginal costs for each procedure so that they can reimburse providers at a mark-up over those costs (Centers for Medicare & Medicaid Services, 2023).

consumer cost-sharing when firm cost-sharing can also be used.

Table 1: Rationales for intensive-margin consumer cost-sharing

Setting	**Key mechanism**	**Role for consumer cost-sharing**
Valid rationales (contracting frictions)		
Doctor heterogeneity	Consumers choose among heterogeneous firms; without cost-sharing they ignore firm cost-effectiveness	Firm-specific copays for choice of firm; firm contracts still handle intensive margin
Linear contracts	A single linear rate cannot simultaneously achieve first-best service levels for all types	Prevents moral hazard for low-cost types when firms are competitive
Asymmetric information	Some firms disproportionately respond to cost-sharing, harming consumers	Smooths treatment across firm types
Firm risk aversion	Optimal risk transfer to firms may be limited	Positive τ_1 when upfront payments infeasible, market power low, and consumer pooling insufficient
Insufficient rationales		
Altruism	Altruistic firms need steeper incentives but still respond to firm cost-sharing	Not a rationale; steeper τ_2 suffices
Behavioral hazard	Consumers over- or under-use services; differential firm cost-sharing can correct this	Not a rationale; firm contracts can target over- or under-used services
Non-adherence	Consumers fail to follow prescribed treatment	Only can justify *negative* consumer cost-sharing

Doctor Heterogeneity In the baseline version of our model, a visit copay addresses extensive margin heterogeneity while firm contracts address intensive margin heterogeneity.

More generally, when firms are heterogeneous in their productivity, there is a second kind of extensive margin moral hazard which must be addressed by consumer copays. Specifically, if some firms can provide services more cost-effectively but consumers are fully insured, then consumers will not take this into account in their choice of firms. We show in Appendix E that heterogeneity creates a need for firm-specific consumer copays to deal with moral hazard in the choice of firms. Firm contracts still address remaining intensive margin moral hazard, although the steepness of the optimal contract is firm-specific as well.

Linear Contracts Nonlinear contracts are infrequently seen in practice due both to practical complexities and to concerns about robustness to more general information environments (Holmstrom and Milgrom, 1987; Carroll, 2015). We do frequently see nonlinear cost-sharing for consumers in

the form of deductibles and coverage limits, so similar nonlinear contracts for firms may not be completely far-fetcched. Nonetheless, in this section, we consider optimal cost-sharing when insurers are restricted to linear contracts.

Suppose that firm profits per consumer should they choose service level m are given by $g(m) = R - \tau_2 m$. As with the nonlinear contract, the resulting incentive is invariant up to a scaling factor: if we double R and τ_2, firms would choose the same m. The strength of the incentive thus depends not on τ_2 alone but on τ_2/R, cost-sharing per dollar of service costs as a fraction of expected earnings per patient. The first-order condition of the firm problem with respect to m in a symmetric equilibrium simplifies to:

$$h'(m-\lambda) - \tau_1 = \frac{\sigma\tau_2}{(1-1/J)}\frac{1}{R-\tau_2 m} \tag{10}$$

In general, linear contracts cannot achieve constrained first-best service levels without some consumer cost-sharing. Recall that (with no extensive margin) the constrained first-best is achieved at $m = \lambda$, which would require the RHS of equation 10 to be equal for all types, which cannot hold because $R - \tau_2\lambda$ varies by type. A given amount of firm cost-sharing will disproportionately impact service levels for high λ (high service-level) types. Faced with firm cost-sharing, firms cut back service levels for these consumers more dramatically, since they earn low margins and would prefer expanding margins to attracting more consumers. For low λ types, firm cost-sharing is less impactful since firms want to increase their market share of these high-margin consumers. Hence, with nonlinear contracts, the optimal contract was convex, more steeply penalizing service levels for low service-level types than high service-level types. As noted above, a linear contract does achieve the constrained first-best in the limit as market-power becomes arbitrarily large (we show this formally in Appendix F).

The behavior of the model in the perfectly competitive limit is also notable. In the case of exactly perfect competition $\sigma = 0$, firm cost-sharing is completely ineffectual and we are in the classic Baily-Chetty model where consumer cost-sharing is balanced against risk protection. However, the limit as $\sigma \to 0$ does not generally converge to the case of perfect competition. When $\sigma = \epsilon > 0$ firm cost-sharing is effective, but disproportionately so for the highest expenditure types.[9] As a result, the solution as $\sigma \to 0$ is exactly the same as if insurers could impose a uniform cap on expenditures

[9]This follows because, in the competitive limit, firm cost-sharing can only change behavior when firm margins become arbitrarily small. Otherwise, if $R - \tau_2 m(\lambda) > 0$ (rather than weakly greater) for some λ type, firms never want to scare consumers away by reducing m because they will get a finite gain, but scare away all but an arbitrarily small number of consumers. Given that m is (weakly) monotonic in λ, firm cost-sharing will therefore only be impactful for high-cost types whose expenditures are identical are equilibrium (i.e. if $R - \tau_2 m(\lambda) = 0$ and $m(\lambda') < m(\lambda^*)$, then markups will be strictly positive for λ').

(which they achieve with firm cost-sharing), and then use consumer cost-sharing to control moral hazard for types not at the cap. We prove in appendix G:

1. There is some λ^* such that, for all $\lambda \geq \lambda^*$, $\lim_{\sigma \to 0} m^*(\lambda, \tau_1^*, \tau_2^*) = \frac{R}{\tau_2}$
2. For the same λ^*, for all $\lambda < \lambda^*$, τ_1^* converges to the consumer cost-sharing insurers would choose if the only instruments available to them were consumer cost-sharing and a uniform cap on service levels for types $\lambda \geq \lambda^*$

In other words, in this limit, firm cost-sharing is used to control costs for high-cost enrollees, and consumer cost-sharing is used to control costs for low-cost enrollees. In fact, a uniform cap plus coinsurances is precisely the framework that Pauly and Ramsey (1999) use to model how managed care plans control costs. Our model microfounds this combination of tools as optimal in a setting where firms compete intensively for consumers.

Prospective payment as a special case Prospective payment—the practice of paying firms a fixed amount per case regardless of services rendered—corresponds to the corner case $\tau_2 = 1$ in the linear contract $g(m) = R - \tau_2 m$. Under prospective payment, the firm bears the full marginal cost of effort and receives no additional reimbursement beyond the upfront payment. This contract is widely used in practice: Medicare's DRG system pays hospitals a fixed amount per inpatient admission, leaving hospitals to absorb any costs above the DRG rate.

Whether prospective payment is optimal depends on the degree of firm market power. In the monopolistic limit ($\sigma \to \infty$), a linear contract achieves first-best service levels at a much smaller τ_2, making prospective payment unnecessarily aggressive. Conversely, in the competitive limit ($\sigma \to 0$), the optimality of prospective payment depends on target profits. If target profits also converge to zero (as is the case in many competitive models), then prospective payment also cannot be optimal among competitive firms. Only at intermediate levels of competition can prospective payment be justified.

This logic connects our model to the institutional divide between inpatient and outpatient reimbursement. Hospital inpatient care markets are highly competitive, and DRG-based prospective payment is standard. Our analysis suggests that here, weaker cost-sharing may be welfare-improving. Outpatient physician markets exhibit more market power: patients maintain ongoing relationships with their physicians and reimbursement is predominantly fee-for-service ($\tau_2 \approx 0$). Our calibration in Section 5 finds that the optimal physician contract calls for $\hat{\tau}_2 = 0.016$, confirming that strong cost-sharing is far from optimal but also implying that pure fee-for-service goes too far: even a

small penalty for excessive spending substantially improves welfare. The analysis thus suggests that both extremes of observed reimbursement—prospective payment for inpatient care and pure fee-for-service for outpatient care—are likely not the theoretical optimum.

Asymmetric Information A conventional reason why contracting fails is asymmetric information – in our case, insurers may be imperfectly informed about firms. One tractable way to model asymmetric information about firm heterogeneity is to assume that doctors vary in their degree of altruism, but the insurer cannot distinguish between them. In such a case, there is a role for consumer cost sharing because firm cost sharing disproportionately affects patients who are assigned to less altruistic doctors. As a result, patients assigned to less altruistic doctors receive relatively less care than patients assigned to more altruistic doctors. By contrast, consumer cost sharing affects patients more equally across doctors. Replacing consumer cost sharing with firm cost sharing will make patients who are already receiving relatively less treatment because of their doctor receive even less treatment, while patients who are receiving more treatment will receive even more. This creates inefficiency in the allocation of treatment across doctor types. In Appendix H, we show that when firms are altruistic in this way, the optimal coinsurance rate is positive, and a first-order approximation depends on the variance of the marginal rate of substitution, which captures how heterogeneous firms are in their relative responses to firm and consumer cost-sharing.

Firm Risk Aversion Hospitals are sufficiently large that it may make sense to consider them as risk-neutral with respect to the amount that they are reimbursed for any individual patient. The same is generally not the case for auto mechanics or contractors or doctors.

We show in Appendix I that even if firms are risk averse, it is generally optimal to transfer risk to firms provided either: a) up-front payments (i.e. salaries) which do not depend on the number of patients are feasible, b) firms have sufficient market power that optimal incentives are weak, c) the law of large numbers allows firms to mitigate risk by pooling over many consumers. Additionally, if firms are currently being paid more the *more* they do ($\tau_2 < 0$), penalizing firms more by increasing τ_2 (making it less negative) would reduce firm risk exposure while controlling costs.

In our empirical application to physician reimbursement, we consider how b)-d) impact physician's risk exposure, assuming that shifting compensation to salaries independent of the number of patients who visit is not feasible.

4.3 Bad Reasons for Consumer Cost-Sharing

In this section, we show that altruism and behavioral hazard are not in general good reasons for positive cost-sharing.

Altruism A common intuition (in our experience) is that if firms are altruistic towards consumers, then consumer cost-sharing may become necessary to control costs. In Appendix H, we show that this intuition is not correct. Ellis and McGuire (1990) considers the case of an altruistic monopolist. In this case, linear firm cost-sharing achieves first best outcomes, with stronger incentives (more negative τ_2) necessary as firms become more altruistic. More generally, if we embed altruism in the competitive model of firms, then steeper incentives are necessary to achieve first-best outcomes with altruistic firms, but altruism does not provide a rationale for consumer cost-sharing.

Behavioral Hazard The framework of "value-based insurance design" (Chernew et al., 2007; Baicker et al., 2015) recommends lower consumer copays for high-value treatments, on the grounds that patients systematically overvalue inferior alternatives. In Appendix J, we extend our model to a setting with two treatments—a high-value and a low-value option—where patients overvalue the inferior one. We show that differential firm reimbursements (paying providers more for the high-value treatment) implement the same treatment allocation as differential consumer copays. The case for consumer-side instruments therefore rests on contracting frictions—such as an inability to verify which treatment was delivered—rather than on behavioral hazard per se. In healthcare, there is also substantial concern with patient non-adherence to prescribed drugs, which would typically be a rationale for *negative* consumer cost-sharing, as positive cost-sharing would exacerbate non-adherence.

5 Application #1: Medicare Outpatient Services

This section evaluates the feasibility of replacing Medicare's current patient coinsurance with the optimal physician fee schedule characterized in Section 4. The exercise compares two regimes: the status quo, characterized by the estimated physician payment schedule and patient coinsurance rates which are low for individuals with supplemental insurance and more substantial otherwise; and the optimum, in which physicians face a declining profit schedule $g(m)$ and patient coinsurance is eliminated ($\tau_1^* = 0$), aside from a per visit copay.

Conducting this comparison requires four objects: the distribution of patient types $F(\lambda)$, the

health-benefit curvature (which we parametrize with ω), the physician market-power parameter σ, and the current implicit penalty (or reward) for physician spending τ_2. We observe the episode revenue distribution directly from claims data and use this to back out the distribution of patient types by inverting the physician first-order condition. We calibrate σ externally from quasi-experimental estimates in the literature. We treat the fixed per-patient payment R as bounded above by an observable benchmark from the fee schedule. We calibrate (ω, τ_2) by matching two moments from the model's first-order conditions to external quasi-experimental elasticity estimates: the response of physician revenue to the elimination of patient coinsurance (Cabral et al., 2018) identifies ω, and the elasticity of physician revenue with respect to reimbursement rates (Clemens and Gottlieb, 2014) identifies τ_2. We describe the data, calibration, and identification in turn.

5.1 Data and Sample Construction

The primary data source for the episode expenditure distribution is the IBM MarketScan Medicare Supplemental and Coordination of Benefits database, which covers traditional Medicare fee-for-service beneficiaries who carry employer-sponsored supplemental insurance.[10] We use data for 2014–2018 and restrict to Level-1 outpatient encounters: physician-billed evaluation and management (E/M) office visits. Because we observe separately Medicare payments, out of pocket costs, and supplemental coverage, we can also simulate how welfare would change for the same individuals were they to lack supplemental coverage.

We take as the key observable the total allowed amount per episode Rev_i—the physician's gross reimbursement from all payers, or equivalently, the total outlays necessary from insurers and consumers to finance the episode of care.[11] All moment conditions themselves are written directly in terms of revenue. Physician effort m is a latent variable, which can be inverted from the function that maps revenue and m (which we will discuss in the next subsection).

Episodes are defined as contiguous spells of physician-billed outpatient care for a given patient, constructed from MarketScan claims. For each beneficiary-episode we record the total allowed amount Rev_i (the data primitive) and the episode duration.

Table 2 reports the observed per episode revenue distribution across deciles used in the structural estimation.

[10] We restrict the sample to DATATYP = 3 (fee-for-service claims) and exclude Medicare Advantage enrollees (MEDADV = 0).

[11] Concretely, Rev_i equals the sum of COB (Medicare FFS payment), NETPAY (employer supplemental), DEDUCT, COPAY, and COINS in MarketScan. Using only subsets (e.g. NETPAY alone) yields an incorrect measure.

Table 2: Observed revenue distribution from MarketScan episode data (10 deciles)

Decile	Mean revenue Rev ($)
1	76.59
2	104.94
3	120.86
4	152.82
5	202.22
6	272.26
7	396.70
8	601.53
9	1,079.82
10	4,404.80
Kink revenue Rev(k)	73.30
Episodes at kink	15.1%

Notes: Revenue is the total allowed amount per episode (Medicare FFS payment plus supplemental payer contributions), observed directly from MarketScan 2014–2018. The kink point corresponds to the minimum episode revenue in the data; 15.1% of episodes fall at or below this threshold. Physician effort m is a latent variable recovered via $m = (\text{Rev} - R)/(1-\tau_2)$ after estimation.

5.2 Structural Model and Identifying Equations

Under the logit demand model of Section 3, a physician in symmetric equilibrium chooses episode spending $m \geq k$ to maximize market share times profit (equation 6). This symmetric-equilibrium problem can equivalently be represented as a behavioral objective in which the physician maximizes a weighted sum of patient health benefit and log profit:[12]

$$\max_{m \geq k} \quad h(m-\lambda) - \tau_1 m + \sigma \ln\big[A \cdot \text{Rev}(m) - m\big], \tag{11}$$

where $h(x) = x - x^2/(2\omega)$ is the quadratic health benefit function, τ_1 is the patient coinsurance rate, $A \cdot \text{Rev}(m) - m$ is per-patient profit, and $\sigma > 0$ is the physician market-power parameter (which absorbs the factor $1/(1-1/J)$ from the symmetric logit equilibrium, so that σ here equals $\sigma_0/(1-1/J)$ where σ_0 is the logit scale parameter in equation 2). The scalar $A \geq 0$ is a proportional reimbursement shifter with baseline $A = 1$; we introduce it to formalize the variation exploited by Clemens and Gottlieb (2014), who use Medicare's fee schedule updates as quasi-experimental shifts in the overall level of reimbursement. In our model, a change in A scales the entire revenue function proportionally, leaving its slope unchanged. All expressions below are evaluated at $A = 1$ unless otherwise noted; A enters only when computing the Clemens–Gottlieb elasticity for Moment 2 (Section 5.4).

[12] The equivalence follows from the logit market-share structure. In symmetric equilibrium, maximizing $s_j \cdot g(m_j)$ yields the FOC $(1-1/J)/\sigma_0 \cdot [h'(m-\lambda) - \tau_1]\, g(m) + g'(m) = 0$, where σ_0 is the logit scale parameter from equation (2). Defining $\sigma \equiv \sigma_0/(1-1/J)$ and rearranging gives the behavioral representation below, with the same first-order condition.

Linear revenue schedule and the link from revenue to effort. To connect the physician's choice problem (11) to the observed revenue data, we assume that the physician revenue schedule is linear in effort:

$$\text{Rev}(m) = R + (1 - \tau_2)\, m, \tag{12}$$

where R is a fixed per-patient payment and τ_2 is the implicit marginal penalty (or subsidy, if $\tau_2 < 0$) on physician spending, so that per-patient profit is $\pi(m) = R - \tau_2 m$. This linear specification corresponds to the linear contract analyzed in Section 4.2 and approximates Medicare's resource-based fee schedule, in which each additional RVU of services is reimbursed at approximately the same marginal rate. Given structural parameters (R, τ_2), latent effort is recovered from observed revenue as

$$m_i = \frac{\text{Rev}_i - R}{1 - \tau_2}. \tag{13}$$

First-order condition in terms of revenue. Under the linear revenue schedule (12), the physician's FOC for an interior optimum ($m > k$) at baseline $A = 1$ can be written entirely in terms of observable revenue (see Section 4.2, equation 10):[13]

$$1 - \frac{1}{\omega}\left(\frac{\text{Rev} - R}{1 - \tau_2} - \lambda\right) - \tau_1 - \frac{\sigma\,\tau_2(1 - \tau_2)}{R - \tau_2\,\text{Rev}} = 0. \tag{14}$$

Inverting (14), for each observed revenue $\text{Rev}_i > \text{Rev}_k$ we recover the patient's type:

$$\lambda_i = \frac{\text{Rev}_i - R}{1 - \tau_2} - \omega\left[1 - \tau_1 - \frac{\sigma\,\tau_2(1 - \tau_2)}{R - \tau_2\,\text{Rev}_i}\right]. \tag{15}$$

Minimum effort, bunching, and the types at the kink. The constraint $m \geq k$ in the firm problem (equation 6) implies a minimum effort level k, representing the cost of providing a basic visit. Two type cutoffs determine who is treated at this minimum. Patients with $\lambda \in [\lambda_2, \lambda_1]$ visit but receive exactly $m = k$—they are "bunched" at the minimum. In the data, this bunching mass accounts for 24.2% of episodes.

The upper cutoff λ_1 is the type at which the physician's interior choice equals k—types above λ_1

[13] This is the linear-contract FOC from Section 4.2, with m replaced by $(\text{Rev} - R)/(1 - \tau_2)$ and the $(1 - 1/J)$ factor absorbed into σ. The factor $(1 - \tau_2)$ in the last term arises from converting the profit denominator: $R - \tau_2 m = (R - \tau_2\,\text{Rev})/(1 - \tau_2)$.

receive $m > k$, while types below are constrained to $m = k$:

$$\lambda_1 = \frac{\text{Rev}_k - R}{1-\tau_2} - \omega\left[1 - \tau_1 - \frac{\sigma\,\tau_2(1-\tau_2)}{R - \tau_2\,\text{Rev}_k}\right], \tag{16}$$

where $\text{Rev}_k \equiv \text{Rev}(k)$ is the revenue at minimum effort. The lower cutoff λ_2 is the marginal type who is just willing to visit at $m = k$. Applying the extensive-margin visit decision (equation 4) at $m = k$, and allowing for a nonfinancial visit cost $c_{\text{visit}} \geq 0$ (e.g. travel time, hassle of scheduling), the marginal visiting type is:

$$\lambda_2 = \frac{k}{2} - \omega(1-\tau_1) + \frac{\omega\, c_{\text{visit}}}{k}. \tag{17}$$

5.3 External Calibration

In the outpatient application, σ is an episode-level logit scale parameter governing physician choice at the start of an outpatient episode. We therefore calibrate σ from evidence on the market share of *new* physician episodes, rather than from visit-level copay differentials or from switching away from incumbent physicians. Sinaiko and Rosenthal (2014) study tiered physician networks in the Massachusetts GIC and find that average-tier specialists had a new-patient market share of 0.63% versus 0.56% for bottom-tier specialists. Patients choosing a new physician saw that physician three times on average, implying approximately $30 of episode-level copay savings from choosing an average-tier over a bottom-tier physician. Mapping these moments through the logit formula $\Delta \ln s = -\Delta c/\sigma$ implies

$$\sigma \approx \frac{\Delta c_{\text{episode}}}{|\Delta \ln s|} = \frac{30}{\ln(0.63) - \ln(0.56)} \approx \$255 \text{ per episode.}$$

Using all physicians rather than specialists gives $\sigma \approx \$229$ (market shares 0.57% vs. 0.50%). We use $\sigma = \$250$ per episode as our baseline. This parameter should be interpreted as an effective episode-start competition parameter; incumbent physician attachment operates on a conceptually separate margin and would be captured by an explicit switching-cost term in a richer model (Sinaiko and Rosenthal, 2014).

Rather than calibrating the intercept R (physician profit at $m = 0$, an unobserved counterfactual), we anchor the revenue schedule at the observable minimum spending level. Under the Medicare Physician Fee Schedule, the revenue for a basic visit at minimum effort –analogous to our minimum effort– corresponds to a standard evaluation and management office visit (CPT 99213/99214), reimbursed at approximately $100 at the CY2014–2018 conversion factor.[14] We set revenue level

[14] Sensitivity to $\text{Rev}_k \in \{\$50, \$100, \$200\}$ is reported in Table 3.

at k, $\text{Rev}_k = \$100$ as our baseline. The minimum effort level k itself is not directly observed; we calibrate it as a fraction of Rev_k: our baseline assumes $k = 0.35 \cdot \text{Rev}_k = \$35$, reflecting physician practice-expense overhead (roughly 65% of visit revenue goes to non-physician direct costs such as staff, supplies, and facilities).

5.4 Identification of (ω, τ_2)

Given the externally calibrated (σ, R', k), the remaining structural parameters (ω, τ_2) are identified by matching two empirical moments derived from the physician's FOC(equation 14).

Moment 1: Revenue change from Medigap (identifies ω). Cabral et al. (2018) estimate that Medigap supplemental insurance (which sets $\tau_1 \approx 0$, eliminating the 20% patient coinsurance) increases total annual Medicare Part B revenue by approximately 17.2%. In the model, eliminating τ_1 causes three responses: (i) doctors above the kink increase m; (ii) doctors at the kink become unconstrained and increase to $m > k$; (iii) new patients whose participation threshold $\lambda_2^{\text{cf}} = k/2 - \omega$ falls below $\lambda_2 = k/2 - \omega(1 - \tau_1)$ are brought in on the extensive margin. With a 20% extensive-margin response estimated by Chandra et al. (2010), the moment condition is:

$$\frac{\text{Rev}^{\text{cf}} - \text{Rev}^{\text{data}}}{\text{Rev}^{\text{data}}} = 0.172, \tag{18}$$

where $\text{Rev}^{\text{data}} = 0.758\,\overline{\text{Rev}}(\mathbf{m}_{\text{data}}) + 0.242\,\text{Rev}(k)$ and the counterfactual averages are computed by re-solving the physician FOC at $\tau_1 = 0$ for each patient type recovered from equation (15).

Moment 2: Revenue elasticity with respect to reimbursement (identifies τ_2). Clemens and Gottlieb (2014) exploit geographic variation in Medicare payment rates generated by fee-schedule consolidation and estimate a long-run (2001–2005) elasticity of total physician revenue with respect to the conversion factor of $e = 1.45$. In the model, this is matched by the revenue elasticity with respect to the reimbursement scalar A:

$$\left.\frac{d \log \text{Rev}(A)}{d \log A}\right|_{A=1} = 1.45, \tag{19}$$

where $\text{Rev}(A)$ is mean revenue when physicians re-optimize at the scaled reimbursement $A \cdot \text{Rev}(m)$. The revenue elasticity is smaller than the spending elasticity because of the fixed component $R > 0$: $\epsilon_{\text{rev}} = \epsilon_m \cdot \frac{(1-\tau_2)\bar{m}}{R+(1-\tau_2)\bar{m}} < \epsilon_m$.

We solve the two-equation system (18)–(19) via a nonlinear equation solver with multiple start-

ing values. Table 3 reports structural estimates across specifications. At the baseline specification ($\text{Rev}(k) = \$114$, $k/\text{Rev}(k) = 0.35$), $\hat{\tau}_2 = -0.143$ and $\hat{\omega} \approx \$1{,}440$; across alternative specifications, $\hat{\tau}_2$ ranges from -0.165 to -0.055 while $\hat{\omega}$ ranges from \$1,395 to \$3,667. The implied spending distribution and kink point depend on $\hat{\tau}_2$ through equation (13).

Interpretation of $\hat{\tau}_2 < 0$. The negative estimate $\hat{\tau}_2 = -0.143$ means that physicians currently face a marginal *subsidy* on episode spending: each additional dollar of episode expenditure raises per-patient profit by approximately 14.3 cents. Equivalently, the status-quo fee schedule rewards physicians for delivering more intensive episodes.

Medicare's physician fee schedule is explicitly resource-based: each CPT code's allowed amount is constructed from work, practice-expense, and malpractice RVUs calibrated to reimburse the resources a *typical* service requires. This design superficially targets $\tau_2 = 0$, attempting to compensate physicians at the marginal cost of effort. A marginal subsidy ($\tau_2 < 0$) can nonetheless arise through two related channels. First, billing intensity: physicians who bill additional procedure codes or select higher-complexity evaluation-and-management codes will effectively increase the revenue they receive for a given amount of effort, leading to a marginal subsidy. This interpretation accords with a large empirical literature documenting physician responses to fee-for-service incentives and upcoding (Geruso and Layton, 2020). Second, the practice-expense RVU component of the fee schedule is designed in part to cover physicians' fixed overhead costs. Practice-expense RVUs typically scale proportionately with other RVUs which cover solely marginal expenses. This means that physicians are actually being compensated at a rate higher than their marginal costs, precisely as the $\tau_2 < 0$ finding implies.

The status-quo schedule itself exposes physicians to substantial income risk from the composition of their patient panel. Under $\hat{\tau}_2 < 0$, a physician who draws a panel of relatively healthy, low-spending patients earns less per episode than one who draws high-cost patients—the mirror image of the risk under $\tau_2 > 0$, where high-cost patients reduce profit. Either sign of τ_2 generates physician income risk from spending variation; only the direction differs. The optimal contract also exposes physicians to income variation across patient types, but limits the downside: profits under $g(m) = C\exp(-m/\sigma)$ are bounded below at zero, so physicians treating very high-cost patients earn nothing on net but are always fully compensated for their time.

Table 3: Sensitivity of structural estimates to $\mathrm{Rev}(k)$ and $k/\mathrm{Rev}(k)$

$\mathrm{Rev}(k)$ (\$)	$k/\mathrm{Rev}(k)$	k (\$)	R (\$)	$\hat{\tau}_2$	$\hat{\omega}$ (\$)
Panel A: $\mathrm{Rev}(k) = \$100$					
100	0.30	30.0	65.0	−0.165	1,415
100	0.40	40.0	53.7	−0.157	1,528
100	0.50	50.0	42.7	−0.146	1,667
100	0.60	60.0	32.2	−0.131	1,849
100	0.70	70.0	22.2	−0.112	2,113
100	0.80	80.0	13.0	−0.088	2,470
100	0.90	90.0	5.0	−0.056	3,020
Panel B: $\mathrm{Rev}(k) = \$114$					
114	0.30	34.2	74.8	−0.145	1,395
114	0.40	45.6	62.0	−0.141	1,484
114	0.50	57.0	49.4	−0.134	1,615
114	0.60	68.4	37.2	−0.123	1,800
114	0.70	79.8	25.7	−0.107	2,044
114	0.80	91.2	15.0	−0.085	2,383
114	0.90	102.6	5.8	−0.055	2,967
Panel C: $\mathrm{Rev}(k) = \$150$					
150	0.30	45.0	99.4	−0.125	1,550
150	0.40	60.0	82.4	−0.127	1,618
150	0.50	75.0	65.7	−0.125	1,713
150	0.60	90.0	49.4	−0.118	1,891
150	0.70	105.0	33.9	−0.105	2,138
150	0.80	120.0	19.7	−0.085	2,490
150	0.90	135.0	7.5	−0.055	3,104
Panel D: $\mathrm{Rev}(k) = \$200$					
200	0.30	60.0	133.9	−0.101	1,981
200	0.40	80.0	111.2	−0.110	2,024
200	0.50	100.0	88.6	−0.114	2,121
200	0.60	120.0	66.5	−0.112	2,311
200	0.70	140.0	45.5	−0.103	2,585
200	0.80	160.0	26.3	−0.085	2,989
200	0.90	180.0	9.9	−0.056	3,667

Notes: Each row solves the two-moment system (Cabral Medigap response and Clemens–Gottlieb elasticity) under the τ_1-flip identification (status quo $\tau_1 \approx 0$ for Medigap enrollees; counterfactual $\tau_1 = 0.114$). R is derived as $R = \mathrm{Rev}(k) - (1 - \hat{\tau}_2)\,k$; it is not a free parameter. The extensive margin uses 20% from Chandra–Gruber–McKnight (2010). $\sigma = 250$ throughout. All specifications converge (sum of squared residuals $< 10^{-17}$). $\hat{\omega}$ is remarkably stable across $k/\mathrm{Rev}(k)$ within each panel, confirming that Moment 1 tightly identifies ω. $\hat{\tau}_2$ decreases as $k/\mathrm{Rev}(k)$ rises because higher k lowers R, requiring a larger marginal subsidy to match the observed elasticity.

5.5 Extensive Margin and Distribution of Patient Types

Given the calibrated $(\hat{\omega}, \hat{\tau}_2)$, we recover the full distribution of patient types λ using two sources of variation.

For episodes with $m_i > k$, equation (15) invertibly maps observed spending to patient type λ_i, identifying the truncated distribution $F(\lambda \mid \text{visit} = 1)$ among visitors. For patients who do not visit ($m_i \leq k$), no spending is observed, so type inversion cannot be applied. A single extensive-margin moment identifies only the density $f(\lambda^*)$ at the visit/no-visit cutoff; recovering the full distribution of non-visitor types requires an additional assumption on the shape of $f(\cdot)$ below λ^*. We assume that illness severity is *uniformly distributed* among non-visitors:

Assumption 1 (Uniform non-visitor types). *The density $f(\lambda)$ is constant on $[\underline{\lambda}, \lambda^*]$, where $\underline{\lambda}$ is the lowest illness severity in the population.*

Under Assumption 1, $f(\lambda^*)$ pins down the entire non-visitor distribution. We recover $f(\lambda^*)$ from the extensive-margin response estimated by Chandra et al. (2010): Medigap supplemental insurance, which effectively lowers the visit copay from roughly 20% to zero, raises visit probability by approximately 20%, Differentiating the physician's participation condition of our model,

$$\frac{\partial \Pr(\text{visit})}{\partial \tau_{1,k}} = -\frac{k\, f(\lambda^*)}{h'(k - \lambda^*)}, \tag{20}$$

where λ^* is the marginal type and $f(\lambda^*)$ is the (common) density, we can identify $f(\lambda^*)$ from the left-hand side estimate of Chandra et al. (2010) (as $h'(k - \lambda^*)$ is known from $(\hat{\omega}, \hat{\tau}_2)$).

5.6 Results: Optimal Physician Contract

Given the calibrated structural parameters $(\hat{\omega}, \hat{\tau}_2, R, \sigma)$ and the distribution of patient types, the optimal physician profit schedule from Proposition 2 of the main text takes the exponential-logit form:

$$g(m) = C \cdot \exp\left(-\frac{(N-1)}{\sigma N}\, m\right), \tag{21}$$

where N is the number of competing physicians and the constant C is chosen so that the expected per-patient payment to physicians equals the per-patient rent R:

$$C = \frac{R}{\mathbb{E}\left[e^{-m/\sigma_{\text{eff}}} \mid m \geq k\right]}, \qquad \sigma_{\text{eff}} = \frac{\sigma N}{N-1}. \tag{22}$$

The resulting profit schedule $g(m) = C\exp(-m/\sigma)$ declines steeply with episode expenditure at low levels of spending but is flatter at high levels of spending—structurally analogous to a physician-side deductible. Table 4 illustrates this. Under the status quo, per-episode profit ranges from \$97 (at m = \$200) to \$355 (at m = \$2,000), rising steadily because $\hat{\tau}_2 < 0$. Under the optimal contract, the same episodes yield profits of \$558 and \$0, respectively; doctors earn about 0.25 cents less for each dollar they earn, but profits are never negative. The optimal schedule is back-loaded toward low-spending episodes in precisely the way a consumer deductible is front-loaded toward high-cost states: just as a deductible makes patients the residual claimants on low-cost care, the optimal physician contract makes physicians the residual claimants on low-spending episodes.

The economic logic, however, differs fundamentally from the consumer case. Consumer deductibles trade off insurance value against moral hazard; physicians in our baseline calibration are risk-neutral firms for whom insurance has no direct value. Instead, the steep profit decline is necessary because low-cost episodes are profitable to attract under the optimal schedule. A physician who keeps a \$200 episode at \$200 earns \$558; one who allows it to reach \$600 earns only \$113. The \$445 gap is the incentive. Conversely, for high-cost episodes profit is near zero regardless, so further penalization adds little incentive at the margin; hence the contract concentrates its steepness at the low end of the spending distribution where it matters most.

The finding that $\tau_2 < 0$ implies that an alternative contract at $\tau_2 = 0$ could reduce risk for both patients and providers. The optimal contract exposes providers to more risk, but more effectively combats moral hazard. The total risk exposure is limited because profits never become negative: the contract thus still involves on average considerably weaker incentives than prospective payment and considerably less risk exposure.

6 Application #2: Controlling Costs for Prescription Drugs

In this section, we apply our model to determine optimal cost-sharing for prescription drugs in Medicare Part D. We produce new estimates of the impact of consumer coinsurance (τ_1) and physician prior authorization requirements (τ_2) on total drug expenditures, and use these to characterize the optimal combination of both instruments.

Medicare Part D is the federal prescription drug benefit for US residents aged 65 and older. Physicians have substantial market power over patient drug choices: patients typically cannot self-prescribe and rely on physician recommendations, so insurers use both consumer coinsurance and prior autho-

Table 4: Status-Quo vs. Optimal Physician Profit Schedule

	Status Quo ($\hat{\tau}_2 = -0.143$)		Optimal contract $g(m)$	
Episode spending (m)	Revenue ($)	Profit ($)	Revenue ($)	Profit ($)
$200	297	97	758	558
$400	526	126	651	251
$600	754	154	713	113
$1,000	1,212	212	1,023	23
$2,000	2,355	355	2,000	0
	Average: $189		Average: $189	

Notes: Status-quo columns assume $R = \$68$ and $\hat{\tau}_2 = -0.143$ (calibrated with $\sigma = 250$, $\text{Rev}(k) = \$114$, $k/\text{Rev}(k) = 0.35$): Revenue $= R + (1 - \hat{\tau}_2)m$; Profit $= R - \hat{\tau}_2 m$. Optimal columns use the nonlinear contract $g(m) = C\exp(-m/\sigma)$ with $C = 1{,}241$ and $\sigma = 250$, calibrated so that average profit equals the status-quo average (Revenue $= m + g(m)$; Profit $= g(m)$). The status-quo profit *increases* with spending (since $\hat{\tau}_2 < 0$, physicians receive a marginal subsidy), while the optimal profit schedule declines steeply, concentrating incentives at low service levels and achieving first-best expenditure control without consumer cost-sharing.

rization to control costs. Our theory predicts that physician-side cost-sharing will be disproportionately more impactful per dollar than consumer cost-sharing in settings with high physician market power.

Prior authorization (PA) is not formally identical to firm cost-sharing in our baseline model. A linear incentive contract taxes physicians the more they spend (i.e. $\tau_2 > 0$ means profit $= R - \tau_2 m$), whereas prior authorization imposes paperwork and time costs on physicians who prescribe more expensive drugs rather than transferring money to insurers. Prior authorization therefore "burns" money rather than redistributing it. We show that this distinction has minimal quantitative consequences. The dollar costs of prior authorization are small relative to their impact on expenditures, consistent with high physician market power, so the welfare analysis is nearly identical whether the physician's compliance cost is transferred to the insurer or dissipated.[15]

6.1 Data

We use data from the Center for Medicare and Medicaid Services (CMS) for 2011–2019.[16] The dataset includes the universe of Medicare Part D beneficiaries, the complete set of plans and their formularies, beneficiary enrollment by plan, and prescription drug claims for a 20% random subsample.

Medicare Part D in 2015 insured approximately 39 million beneficiaries in 1,001 distinct prescription drug insurance plans (Jacobson et al., 2019), covering a total of $90 billion in expenditures

[15]The key sufficient condition is that $\tau_2 \cdot \bar{m} \ll R$, i.e. the PA compliance cost per patient is small relative to physician rent, which holds empirically (see Section 6.3).

[16]The data are accessible through the CMS Research Data Assistance Center (ResDAC) at the University of Minnesota under DUA 57931.

(Medicare Board of Trustees, 2016).

Drugs covered by Medicare Part D include most prescription drugs sold at retail pharmacies, but not over-the-counter drugs or drugs primarily administered at physician offices or hospitals (e.g., chemotherapy). A "drug" in this paper refers to a unique brand-generic-class combination: we aggregate over varieties of the same brand (different package sizes, strengths, and dosage forms). Brand and generic names are defined using First DataBank MedKnowledge; therapeutic classes use the US Pharmacopeia Medicare Model Guidelines (MMG), which is Medicare Part D's reference classification.

Medicare Part D plans are organized around four benefit phases: the deductible phase, the pre-ICL (Initial Coverage Limit) phase, the "donut hole", and catastrophic coverage. Beneficiaries begin the year in the deductible phase (paying full drug costs), transition to the pre-ICL phase where plan cost-sharing rules apply, then enter the donut hole, and finally reach catastrophic coverage with low coinsurance. While out-of-pocket cost-sharing varies by benefit phase, utilization-management tools—including prior authorization, step therapy, and quantity limits—apply uniformly across all phases.

For each beneficiary i in year t, we define:

- M_{it}: total drug expenditures (out-of-pocket plus plan payment), the outcome variable.
- $\tau_{1,it}$: the *realized* coinsurance rate, defined as out-of-pocket costs divided by total drug expenditures.
- $\tau^*_{2,it}$: the expenditure-weighted share of i's drugs subject to prior authorization in year t, evaluated using the $t-1$ utilization bundle at i's base-year plan.

Because prior authorization is not expressed in dollar terms per unit of expenditure, we relate the raw prior authorization rate τ_2^* to our theoretical τ_2 using external estimates of the dollar cost per drug episode subject to PA. Existing literature estimates this cost at approximately \$14–\$50 per drug per authorization request.[17] This translates the expenditure-weighted PA share τ_2^* into a dollar cost per unit of expenditure:

$$\tau_2 = \tau_2^* \times \frac{\text{PA cost per drug}}{\bar{P}}, \tag{23}$$

where $\bar{P}$ is the mean drug expenditure per beneficiary.

Table 5 reports summary statistics for the sample.

[17] The \$14 estimate reflects physician time costs for a typical PA form completion; the \$50 estimate incorporates additional administrative burden. We use \$50 as our baseline and report sensitivity to \$14 and to alternative values.

Table 5: Summary Statistics: Medicare Part D (2011–2019)

Variable	Mean	Std. Dev.
Utilization characteristics		
Avg OOP costs	505	787
Coinsurance rate (τ_1)	0.26	0.18
% of Drugs Subject to Prior Auth	0.22	0.38
Beneficiary characteristics		
Male	0.43	0.49
Age	73.33	8.75
White	0.88	0.32
% of Inertial	0.82	0.41
Number of plans per year	2,300	
Number of beneficiary-year pairs	34,408,529	

Notes: Unit of observation is plan-year for plan characteristics and beneficiary-year for individual characteristics. Data pooled 2011–2019. CMS Research Data Assistance Center (ResDAC), DUA 57931.

6.2 Identification Strategy

Both coinsurance and prior authorization are mechanically endogenous. Consumer coinsurance varies with benefit-phase position (the deductible, donut-hole, and catastrophic-coverage phases have different cost-sharing rules), and individuals who consume more will have different realized coinsurance rates. Prior authorization may be applied more aggressively by plans that expect high-cost members, and plan formulary changes may be correlated with anticipatory spending patterns.

To address these concerns, we exploit *plan-level changes* in cost-sharing and prior authorization rates across years, following Abaluck et al. (2018). The key observation is that most Medicare Part D beneficiaries remain in the same plan from year to year due to inertia (Abaluck and Gruber, 2016). For inertial enrollees, changes in their plan's cost-sharing design represent a plausibly exogenous shock to the cost-sharing they face.

The identifying assumption is that plans altering their cost-sharing or prior authorization requirements would not otherwise have experienced different expenditure trends from other plans. This assumption is plausible: plan formulary changes are typically driven by drug manufacturer negotiations and PBM contracting decisions rather than by expectations about individual-level spending dynamics. In our results below, we find some evidence that plans may increase cost-sharing following large increases in expenditure. If anything, this would bias us towards overstating the impacts of cost-sharing, which we find are quite small relative to the impacts of prior authorization.

We estimate the following first-difference specification:

$$\Delta M_{it} = \beta_{\tau_1} \Delta \tau_{1,it} + \beta_{\tau_2^*} \Delta \tau^*_{2,it} + X'_{it}\,\alpha + e_{it}, \tag{24}$$

where X_{it} includes plan and year fixed effects and controls for deciles of prior-year expenditures $M_{i,t-1}$, prior-year coinsurance $\tau_{1,i,t-1}$, prior-year PA rate $\tau^*_{2,i,t-1}$, and the beneficiary's prior-year plan characteristics.

We instrument the two endogenous variables with the changes in plan-level coinsurance and prior authorization that the beneficiary's *base-year plan* underwent in year t, evaluated at a fixed national utilization bundle from $t-1$:

$$\mathrm{IV}_{1,it} = \tau_{1,\,\mathrm{plan}(i,t-1),\,t} - \tau_{1,\,\mathrm{plan}(i,t-1),\,t-1},$$
$$\mathrm{IV}_{2,it} = \tau^*_{2,\,\mathrm{plan}(i,t-1),\,t} - \tau^*_{2,\,\mathrm{plan}(i,t-1),\,t-1}.$$

These instruments capture only variation arising from changes in plan coverage design, not from endogenous shifts in the beneficiary's own utilization. Standard errors are clustered at the plan level.

6.3 Reduced-Form Results

Table 6 reports IV estimates of equation (24), progressively adding controls across columns.

Table 6: Effect of Coinsurance and Prior Authorization on Drug Expenditures

	(1) ΔM_{it}	(2) ΔM_{it}
$\Delta\tau_{1,it}$ (coinsurance)	−987	−987
	(137)	(137)
$\Delta\tau^*_{2,it}$ (PA share)	−3,058	−3,058
	(508)	(508)
Plan FE	No	Yes
Year FE	Yes	Yes
$M_{i,t-1}$ deciles	Yes	Yes
Plan chars$_{t-1}$	No	No
Individual controls	No	No
N	34,408,310	34,408,310
# clusters (plans)	2,906	2,906

Notes: Standard errors clustered at the plan level in parentheses. All specifications instrument $\Delta\tau_{1,it}$ and $\Delta\tau^*_{2,it}$ with the change in the plan-level counterparts for the beneficiary's base-year plan. The dependent variable is the first difference of total (plan plus out-of-pocket) drug expenditures.

A 10 percentage-point increase in the coinsurance rate reduces total drug expenditures by ap-

proximately \$99 per beneficiary per year, or 4.3% of mean annual drug spending of \$2,300. Because higher coinsurance shifts costs to consumers, this \$99 reduction comes at a consumer out-of-pocket cost of roughly \$230 per year, implying a spending reduction of about \$0.43 per dollar of additional consumer cost.

A 10 percentage-point increase in the share of drugs subject to prior authorization reduces expenditures by approximately \$306 per beneficiary per year, or 13.3% of mean annual spending. Physician time costs for completing PA forms are estimated at \$14–\$50 per request, implying a total physician compliance cost of roughly \$6–\$21 per beneficiary per year for a 10 pp increase in the PA share (0.43 new drugs subject to PA per patient), and a spending reduction of \$14–\$51 per dollar of compliance cost.

Table 7 summarizes the comparison. Prior authorization is roughly 33–120 times more impactful per dollar than consumer coinsurance. This ratio corresponds directly to the structural ratio σ/R in the model's monopolistic limit, confirming that physicians have substantial market power in prescription drug markets.

Table 7: Expenditure Impact of Coinsurance and Prior Authorization

	Spending reduction (10 pp ↑)		Cost imposed	Reduction
	\$/yr	% of mean	(10 pp ↑, \$/yr)	per dollar
Consumer coinsurance	\$99	4.3%	\$230 [a]	\$0.43
Prior authorization	\$306	13.3%	\$6–\$21 [b]	\$14–\$51
PA relative to coinsurance	3.1×	3.1×	—	33–120×

Notes: [a] Additional consumer out-of-pocket expenditure at mean annual drug spending of \$2,300. [b] Physician time cost per authorization request estimated at \$14–\$50; see text. Range reflects low-end (\$14) and high-end (\$50) per-request cost assumptions. Reduction per dollar for coinsurance is computed as \$99/\$230; for prior authorization as \$306/\$6–\$21.

The finding is robust to the assumed PA cost. Even at the high-end estimate of \$50 per authorization, prior authorization is approximately 14 times more impactful per dollar than coinsurance; the ratio would need to fall to 1:1—requiring PA costs above \$1,500 per authorization—to equalize the two instruments. If classical measurement error attenuates both IV estimates proportionally, the spending-reduction ratio $\hat{\beta}_{\tau_2^*}/\hat{\beta}_{\tau_1} \approx 3.1$ is unaffected, and attenuation would need to favor coinsurance by a factor exceeding 50:1 to reverse the per-dollar ranking.

6.4 Structural Identification and Parameter Estimates

Under the quadratic approximation $h(x) = x - x^2/(2\omega)$, the firm FOC implies that at the observed (τ_1, τ_2):

$$\frac{\partial E(m)}{\partial \tau_1} \approx -\omega, \tag{25}$$

$$\frac{\partial E(m)}{\partial \tau_2} \approx -\frac{\sigma}{\omega R}, \tag{26}$$

where R is the per-patient physician rent.[18] The linearized first-difference specification (24) estimates these derivatives consistently under the assumption that $\tau_2 \cdot \bar{m} \ll R$, which holds in our sample.

From the regression estimates $\hat{\beta}_{\tau_1} = -987$ and $\hat{\beta}_{\tau_2^*} = -3{,}058$:

$$\hat{\omega} = -\hat{\beta}_{\tau_1}^{-1} \approx \$987 \text{ per year} \quad \text{(inverse curvature)}, \tag{27}$$

$$\hat{\sigma} = \hat{\omega} \cdot R \cdot \frac{|\hat{\beta}_{\tau_2^*}|}{|\hat{\beta}_{\tau_1}|} \cdot \kappa, \tag{28}$$

where κ is the conversion factor from PA share to τ_2 in model units (equation (23)).

In the prescription drug context, R represents the physician's per-patient profit from attracting a patient for whom no drug is ultimately prescribed—the baseline profit from the physician-patient relationship before any drug spending occurs. We interpret this as the profit from the associated office visit (Part B revenue), which is the same quantity calibrated in Section 5.

These two estimates are independent of R and of each other. $\hat{\omega} = \$987$ per year is identified directly from $\hat{\beta}_{\tau_1}$ and does not depend on the PA cost assumption. $\hat{\sigma}/R = |\hat{\beta}_{\tau_2^*}| \times \text{PA cost}/\bar{P} = 3{,}058 \times$ PA cost/\$2,300, yielding $\hat{\sigma}/R = 66.5$ under the baseline PA cost of \$50 and $\hat{\sigma}/R = 18.6$ under the alternative of \$14.

The ratio $\hat{\sigma}/R = 66.5$ (baseline) is substantially greater than one, indicating that physician market power is large relative to per-patient rent. This is consistent with the reduced-form finding in Table 7: when $\sigma/R \gg 1$, even a small reduction in physician profit per patient has a large effect on prescribing, making physician-side cost-sharing far more impactful per dollar than consumer cost-sharing. As we show in Section 6.5, a large $\hat{\sigma}/R$ also has strong implications for optimal policy.

[18] Equations (25)–(26) hold at $\tau_2 = 0$ (the $\tau_2 = 0$ limit from Proposition 1 of the main text). With the observed PA rate $\tau_2^* \approx 0.03$, the correction term for $\tau_2 \neq 0$ is proportional to $(\tau_2/R)^2$ and is numerically negligible. Importantly, because τ_2 in this section is *calibrated* from external estimates of PA compliance costs (equation (23)) rather than recovered from the data, we can verify a priori that $\tau_2 \cdot \bar{m} \ll R$: the calibrated $\tau_2 \approx 0.03$ implies a PA cost-sharing burden of at most a few dollars per beneficiary per year, which is small relative to any plausible value of R. This contrasts with Section 5, where $\hat{\tau}_2$ is recovered from data and the condition $\tau_2 \cdot \bar{m} \ll R$ must be checked ex post.

6.5 Optimal Policy

Given the identified structural parameters $(\hat{\omega}, \hat{\sigma})$, the optimal linear (τ_1^*, τ_2^*) are characterized by the insurer's first-order conditions for the linear-contract case (Section 4). We solve these numerically for each value of the CARA risk-aversion parameter γ from Handel et al. (2015), subject to the constraint that total drug expenditures remain approximately equal to their status-quo level.

Table 8: Counterfactual Welfare Simulations

Panel A: Optimal coinsurance rate and PA			
	Coinsurance rate (τ_1)	Exp. share subject to PA	PA costs as % of exp.
Status quo	22.1%	29.0%	3.0%
1×10^{-4} (\$909)	0.6%	35.9%	3.7%
2×10^{-4} (\$832)	0.3%	35.0%	3.6%
Panel B: Welfare gain			
		Welfare gain (\$)	
1×10^{-4} (\$909)		12	
2×10^{-4} (\$832)		1,250	

Notes: Status-quo values from 2015 Medicare Part D data. "Exp. share subject to PA" is the share of drug expenditures for which prior authorization is required. "PA costs as % of exp." is total physician PA compliance cost (including denied prescriptions) as a fraction of drug expenditure. Certainty equivalence (CE) is the financial loss \$X that makes a person with the corresponding CARA risk preference indifferent between accepting and declining a 50/50 gamble of winning \$1,000 and losing \$X. Welfare gain is the certainty-equivalent consumption gain from moving from the status-quo coinsurance to the optimal physician contract.

Table 8 reports the results. Since the model sets physician risk aversion to zero, the theory already establishes that physician cost-sharing dominates consumer cost-sharing: the optimal consumer coinsurance rate is near zero across all risk-aversion levels. What the empirical estimates add is a quantification of how much PA is needed. The optimal PA rate—measured as prior authorization costs as a share of drug expenditure—is close to the current status-quo level of 3%, and even this modest rate delivers substantial welfare gains. The primary inefficiency in current policy is therefore not insufficient PA, but the excessive reliance on consumer cost-sharing, which should be reduced substantially—from 26% to near zero.

The intuition follows directly from the 50:1 effectiveness ratio. A given reduction in drug expenditures can be achieved at far lower welfare cost through physician-side incentives than through patient cost-sharing, because PA imposes small dollar burdens on physicians relative to the spending it deters, while coinsurance imposes large financial risk on patients. The status-quo PA rate of 3% is therefore already bearing approximately the right cost-control burden; what is welfare-reducing is the additional 26% coinsurance burden imposed on patients when the same (or better) cost control is

already achievable through the existing PA requirement.

Moving from the status quo ($\tau_1 = 0.26$, $\tau_2^* = 0.03$) to the optimum ($\tau_1^* \approx 0$, $\tau_2^* \approx \tau_{2,\text{SQ}}^*$) while holding total expenditures approximately constant delivers welfare gains whose magnitude is extremely sensitive to the assumed level of consumer risk aversion: Panel B of Table 8 shows gains ranging from \$12 at low risk aversion ($\gamma = 1 \times 10^{-4}$) to \$1,250 at higher risk aversion ($\gamma = 2 \times 10^{-4}$). The primary source of the gain is the *reduction in financial risk* borne by beneficiaries: lowering consumer coinsurance from 26% to near zero substantially reduces out-of-pocket expenditure volatility while maintaining cost control through the existing PA channel.

These results are consistent with the asymptotic predictions of the model. When $\hat{\sigma}/R \gg 1$, the optimal intensive-margin coinsurance approaches zero in the limit: physician market power is so large that a modest PA requirement controls expenditures effectively, leaving no role for consumer cost-sharing at the intensive margin. The nonlinear contract result (Proposition 2) makes this exact—with nonlinear physician profit schedules, intensive-margin consumer coinsurance is zero, and the insurer addresses any remaining extensive-margin moral hazard through a visit copay $\tau_{1,k}$.

The finding that $\tau_1^* \approx 0$ is robust to wide variation in assumptions. Even at \$50 per authorization (the high-end PA cost estimate), prior authorization is $\approx 14\times$ more impactful per dollar than coinsurance; the ratio would need to fall to 1:1—implying PA costs above \$1,500 per authorization—to reverse the policy ranking. The conclusion that optimal coinsurance is near zero is also qualitatively stable across the full distribution of risk-aversion parameters from Handel et al. (2015); even at very low risk aversion, the dominant instrument is PA. A further implication of the small optimal PA rate is that physician risk aversion is not a concern: because the optimal PA rate is close to the status quo, the implied increase in physician cost exposure is modest, and the welfare ranking of physician over consumer cost-sharing is insensitive to reasonable assumptions about physician risk preferences.

7 Conclusion

This paper develops a framework for optimal cost-sharing in settings where consumers and firms jointly determine expenditures. Firm contracts can achieve constrained first-best service levels with zero intensive-margin consumer coinsurance; only visit copays are needed to address extensive-margin moral hazard. With risk neutral firms, consumer coinsurance for intensive-margin services is justified only by contracting frictions such as asymmetric information or linear contracts. Firm risk aversion can be mitigated when contracts can be front-loaded, when firms have sufficient market

power that optimal incentives are weak, or when firms pool risk across many consumers. In settings where firm contracts currently offer higher per-consumer profits at higher service levels, contracts which instead penalize firms for doing more can even reduce risk exposure.

We apply the framework to two settings. For Medicare outpatient care, our model suggests that a "physician deductible" type schedule is optimal, where physician profits per patient fall the more they spend and asymptote towards zero at high expenditures. Such a schedule could replace patient coinsurances while holding physician profits fixed via higher up front payments. For Medicare Part D prescription drugs, prior authorization is 33–120 times more impactful per dollar than consumer coinsurances, implying that the current 26% consumer coinsurance rate should be reduced to near zero while the current PA rate is approximately optimal.

Our results have both positive and normative implications. Positively, the model predicts that firm cost-sharing is especially simple when firms have market power: in that case, small linear firm penalties suffice to control costs. This at least superficially matches the recent trend in physician markets, where provider consolidation has accompanied the rise of prior authorization. Normatively, our model provides guidance for government agencies like Medicare or Medicaid, as well as private insurers about when consumer cost-sharing should be used: for extensive margin choices, or when contracting frictions are paramount.

Our model studies social planner insurers who seek to maximize consumer welfare. In practice, private insurers will not necessarily use cost-sharing in a way that benefits consumers, due to cream-skimming or because consumers cannot perceive differences in plan quality when they choose insurance plans (Abaluck and Gruber, 2011; Abaluck et al., 2021). Understanding the equilibrium behavior of competing private insurers with endogenous coverage decisions is an important, albeit challenging problem due to the difficulties of modeling equilibrium in insurance markets with selection (Handel et al., 2015).

References

Abaluck, Jason and Jonathan Gruber, "Choice inconsistencies among the elderly: evidence from plan choice in the Medicare Part D program," *American Economic Review*, 2011, *101* (4), 1180–1210.

_ and _, "Evolving choice inconsistencies in choice of prescription drug insurance," *American Economic Review*, 2016, *106* (8), 2145–84.

_ , _ , and Ashley Swanson, "Prescription drug use under Medicare Part D: A linear model of nonlinear budget sets," *Journal of public economics*, 2018, *164*, 106–138.

_ , Mauricio Caceres Bravo, Peter Hull, and Amanda Starc, "Mortality effects and choice across private health insurance plans," *The quarterly journal of economics*, 2021, *136* (3), 1557–1610.

Baicker, Katherine, Sendhil Mullainathan, and Joshua Schwartzstein, "Behavioral hazard in health insurance," *The Quarterly Journal of Economics*, 2015, *130* (4), 1623–1667.

Baily, Martin Neil, "Some aspects of optimal unemployment insurance," *Journal of public Economics*, 1978, *10* (3), 379–402.

Brot-Goldberg, Zarek C., Amitabh Chandra, Benjamin R. Handel, and Jonathan T. Kolstad, "What does a Deductible Do? The Impact of Cost-Sharing on Health Care Prices, Quantities, and Spending Dynamics*," *The Quarterly Journal of Economics*, 04 2017, *132* (3), 1261–1318.

Brot-Goldberg, Zarek C, Samantha Burn, Timothy Layton, and Boris Vabson, "Rationing medicine through bureaucracy: authorization restrictions in medicare," Technical Report, National Bureau of Economic Research 2023.

Cabral, Marika, Michael Geruso, and Neale Mahoney, "Do larger health insurance subsidies benefit patients or producers? Evidence from Medicare Advantage," *American Economic Review*, 2018, *108* (8), 2048–2087.

Carroll, Gabriel, "Robustness and linear contracts," *American Economic Review*, 2015, *105* (2), 536–563.

Centers for Medicare & Medicaid Services, "Medicare Payment Systems," `https://www.cms.gov/Outreach-and-Education/Medicare-Learning-Network-MLN/MLNProducts/html/medicare-payment-systems.html` 2023. Accessed: 2025-03-30.

Chandra, Amitabh, Evan Flack, and Ziad Obermeyer, "The health costs of cost-sharing," *The Quarterly Journal of Economics*, 2024, p. qjae015.

_ , **Jonathan Gruber, and Robin McKnight**, "Patient cost-sharing and hospitalization offsets in the elderly," *American Economic Review*, 2010, *100* (1), 193–213.

Chernew, Michael E, Allison B Rosen, and A Mark Fendrick, "Value-Based Insurance Design," *Health Affairs*, 2007, *26* (2), w195–w203.

Chetty, Raj, "A general formula for the optimal level of social insurance," *Journal of Public Economics*, 2006, *90* (10-11), 1879–1901.

Clemens, Jeffrey and Joshua D Gottlieb, "Do physicians' financial incentives affect medical treatment and patient health?," *American Economic Review*, 2014, *104* (4), 1320–1349.

Delbanco, Suzanne, "The Payment Reform Landscape: An Overview," *Health Affairs Forefront*, 2014. Accessed: 2024-11-03.

Einav, Liran, Amy Finkelstein, Stephen P Ryan, Paul Schrimpf, and Mark R Cullen, "Selection on moral hazard in health insurance," *American Economic Review*, 2013, *103* (1), 178–219.

Ellis, Randall P and Thomas G McGuire, "Provider behavior under prospective reimbursement: Cost sharing and supply," *Journal of health economics*, 1986, *5* (2), 129–151.

_ **and** _ , "Optimal payment systems for health services," *Journal of health economics*, 1990, *9* (4), 375–396.

Geruso, Michael and Timothy J. Layton, "Upcoding: Evidence from Medicare on Squishy Risk Adjustment," *Journal of Political Economy*, 2020, *128* (3), 984–1026.

Handel, Ben, Igal Hendel, and Michael D Whinston, "Equilibria in health exchanges: Adverse selection versus reclassification risk," *Econometrica*, 2015, *83* (4), 1261–1313.

Holmstrom, Bengt and Paul Milgrom, "Aggregation and linearity in the provision of intertemporal incentives," *Econometrica: Journal of the Econometric Society*, 1987, pp. 303–328.

Huckfeldt, Peter J., Victoria Shier, José J. Escarce, Brendan Rabideau, Tyler Boese, Helen M. Parsons, and Neeraj Sood, "Medicare Advantage Has Become Notorious for Prior AuthorizationâCMS and Lawmakers Are Taking Action," *JAMA*, 2024, *332* (13), 1345–1347.

Huetter, John, "CARSTAR: Insurers shifting to performance-based agreements with auto body shops," *Repairer Driven News*, January 2017. Accessed: 2024-11-03.

Jacobson, G., M. Freed, A. Damico, and T. Neuman, "A Dozen Facts About Medicare Advantage in 2019," Technical Report, Kaiser Family Foundation 2019.

Kotlikoff, Laurence J and Lawrence H Summers, "Tax incidence," in "Handbook of public economics," Vol. 2, Elsevier, 1987, pp. 1043–1092.

McGuire, Thomas G, "Demand for health insurance," *Handbook of health economics*, 2011, *2*, 317–396.

Morton, Sherrie, "The Case for Using Cost-Plus Contracts on Property Insurance Claims," 2022. Fast Forward.

Pauly, Mark V and Scott D Ramsey, "Would you like suspenders to go with that belt? An analysis of optimal combinations of cost sharing and managed care," *Journal of Health Economics*, 1999, *18* (4), 443–458.

Sarig, Oren, "Pharmaceutical demand response to utilization management," *Journal of Health Economics*, 2024, *93*, 102830.

Sinaiko, Anna D. and Meredith B. Rosenthal, "The Impact of Tiered Physician Networks on Patient Choices," *Health Services Research*, 2014, *49* (4), 1348–1363.

A Equivalence to Bargaining

This appendix establishes that the equilibrium service-level condition of the main model is equivalent to the outcome of Nash bargaining between consumers and firms. The result is derived under four simplifying assumptions relative to the full model:

1. *No extensive margin.* All consumers visit a firm; there is no outside option of staying home. The consumer's outside option in the bargaining problem is switching to an alternative (symmetric) firm, not forgoing a visit entirely.

2. *No side payments.* Transfers between firms and consumers are prohibited ($\alpha = 0$), so the equivalence applies to the non-transferable case.

3. *Linear cost-sharing.* Consumer and firm cost-sharing rates τ_1 and τ_2 are constants rather than functions of m.

4. *No minimum service floor.* There is no constraint $m \geq k$, so equilibrium service choices are always interior.

Under these conditions, we show that the firm's equilibrium first-order condition for service levels in the logit competition model coincides with the Nash bargaining solution when the firm's bargaining weight is $\gamma = \sigma/(1+\sigma)$. Thus the logit demand parameter σ—which governs firm market power—microfounds the relative bargaining position of consumers and firms.

Suppose there are $N \geq 2$ ex ante identical firms. A consumer of type λ is randomly matched with one firm, and the matched pair bargains over the service level m.

If the pair agrees on service level m, the consumer's utility and the firm's payoff are:

$$u_C(m, \lambda) = W - \tau_1 m + h(m - \lambda), \qquad u_F(m) = R - \tau_2 m. \tag{29}$$

If bargaining breaks down, the firm's outside option is zero. The consumer can instead switch to an alternative firm. Let $m^o(\lambda)$ denote the service level the consumer would receive from an alternative firm, and suppose that switching entails search cost:

$$c_N = \frac{1}{N-1}. \tag{30}$$

Hence the consumer's outside option is:

$$U_0(\lambda) = W - \tau_1 m^o(\lambda) + h(m^o(\lambda) - \lambda) - \frac{1}{N-1}. \tag{31}$$

Taking $U_0(\lambda)$ as given in the bilateral bargaining problem, the matched consumer-firm pair solves the Nash program:

$$m^*(\lambda) \in \arg\max_m \left[u_C(m,\lambda) - U_0(\lambda)\right]^{1-\gamma} \left[u_F(m)\right]^{\gamma}, \tag{32}$$

where $\gamma \in (0,1)$ is the firm's Nash bargaining weight. Equivalently, the pair solves:

$$m^*(\lambda) \in \arg\max_m (1-\gamma)\ln(u_C(m,\lambda) - U_0(\lambda)) + \gamma \ln(u_F(m)). \tag{33}$$

Assuming an interior solution, the first-order condition is:

$$(1-\gamma)\frac{h'(m^*(\lambda) - \lambda) - \tau_1}{u_C(m^*(\lambda),\lambda) - U_0(\lambda)} - \gamma \frac{\tau_2}{R - \tau_2 m^*(\lambda)} = 0. \tag{34}$$

Rearranging gives:

$$h'(m^*(\lambda) - \lambda) - \tau_1 = \frac{\gamma}{1-\gamma}\left(u_C(m^*(\lambda),\lambda) - U_0(\lambda)\right)\frac{\tau_2}{R - \tau_2 m^*(\lambda)}. \tag{35}$$

A symmetric equilibrium is a service rule $m^*(\cdot)$ such that, for every λ, $m^*(\lambda)$ solves the bargaining problem above when the outside option is formed using $m^o(\lambda) = m^*(\lambda)$. In a symmetric equilibrium:

$$\begin{aligned} u_C(m^*(\lambda),\lambda) - U_0(\lambda) &= \left[W - \tau_1 m^*(\lambda) + h(m^*(\lambda) - \lambda)\right] \\ &\quad - \left[W - \tau_1 m^*(\lambda) + h(m^*(\lambda) - \lambda) - \frac{1}{N-1}\right] \\ &= \frac{1}{N-1}. \end{aligned} \tag{36}$$

Substituting this into the first-order condition yields:

$$h'(m^*(\lambda) - \lambda) - \tau_1 = \frac{\gamma}{(1-\gamma)(N-1)}\frac{\tau_2}{R - \tau_2 m^*(\lambda)}. \tag{37}$$

Suppressing the superscript *, the equilibrium condition can be written as:

$$h'(m - \lambda) - \tau_1 = \frac{\gamma}{(1-\gamma)(N-1)}\frac{\tau_2}{R - \tau_2 m}. \tag{38}$$

This matches the equilibrium condition of our main model,

$$h'(m-\lambda)-\tau_1=\frac{\sigma}{N-1}\frac{\tau_2}{R-\tau_2 m}, \tag{39}$$

when the firm's bargaining weight satisfies $\gamma/(1-\gamma)=\sigma$, i.e. $\gamma=\sigma/(1+\sigma)$. Thus the logit demand parameter σ in the main model microfounds the relative bargaining position of consumers and firms.

B Proof of Neutrality with Transferable Utility

We prove Proposition 1 for the baseline model: firms are profit-maximizing (no altruism), cost-sharing schedules $\tau_1(\cdot)$ and $\tau_2(\cdot)$ may be nonlinear, and consumers make both an intensive-margin choice (which firm to visit) and an extensive-margin choice (whether to visit at all).

Step 1: Reparameterizing the firm's intensive-margin problem.

For a given consumer type λ who visits, firm j solves (Equation 6):

$$\max_{m_j,\,\alpha_j} \; s_j(m_j,\alpha_j,\lambda)\cdot\big(R-\tau_2(m_j)+\alpha_j\big) \quad \text{s.t.} \quad m_j\geq k,$$

where the logit market share is (Equation 1):

$$s_j(\lambda)=\frac{\exp\big(\frac{1}{\sigma}(W-\tau_1(m_j)-\alpha_j+h(m_j-\lambda))\big)}{\sum_{j'}\exp\big(\frac{1}{\sigma}(W-\tau_1(m_{j'})-\alpha_{j'}+h(m_{j'}-\lambda))\big)}.$$

(The premium p is common across all firms and therefore cancels from the share formula.)

Define the consumer's net out-of-pocket cost and the firm's net profit per consumer as:

$$\begin{aligned}\mathrm{OOP}_j(\lambda) &\equiv \tau_1(m_j(\lambda))+\alpha_j(\lambda),\\ \Pi_j(\lambda) &\equiv R-\tau_2(m_j(\lambda))+\alpha_j(\lambda).\end{aligned}$$

Subtracting yields:

$$\mathrm{OOP}_j-\Pi_j=\tau_1(m_j)+\tau_2(m_j)-R.$$

Let $T(m)\equiv\tau_1(m)+\tau_2(m)$ denote the combined cost-sharing schedule. Provided T is strictly increas-

ing (so its inverse T^{-1} exists), we can recover m_j uniquely:

$$m_j = T^{-1}\big(R + \text{OOP}_j - \Pi_j\big).$$

This is a one-to-one mapping between (m_j, α_j) and (OOP_j, Π_j) for any given schedule $T(\cdot)$.

Substituting into the firm's problem, the market share becomes:

$$s_j(\lambda) = \frac{\exp\big(\frac{1}{\sigma}(W - \text{OOP}_j + h(T^{-1}(R + \text{OOP}_j - \Pi_j) - \lambda))\big)}{\sum_{j'} \exp\big(\frac{1}{\sigma}(W - \text{OOP}_{j'} + h(T^{-1}(R + \text{OOP}_{j'} - \Pi_{j'}) - \lambda))\big)},$$

and the constraint $m_j \geq k$ becomes $R + \text{OOP}_j - \Pi_j \geq T(k)$. The firm's problem is therefore:

$$\max_{\text{OOP}_j, \Pi_j} \; s_j(\text{OOP}_j, \Pi_j, \lambda) \cdot \Pi_j \quad \text{s.t.} \quad R + \text{OOP}_j - \Pi_j \geq T(k).$$

This problem depends on $\tau_1(\cdot)$ and $\tau_2(\cdot)$ only through the combined schedule $T(\cdot) = \tau_1(\cdot) + \tau_2(\cdot)$. Therefore the equilibrium values $\text{OOP}^*(\lambda)$, $\Pi^*(\lambda)$, and

$$m^*(\lambda) = T^{-1}\big(R + \text{OOP}^*(\lambda) - \Pi^*(\lambda)\big)$$

all depend only on $T(\cdot)$.

Step 2: Premiums depend only on $T(\cdot)$.

In a symmetric equilibrium, premiums satisfy (from the premium formula in the text):

$$p(\tau_1, \tau_2) = E[m_i^*] - E[\tau_1(m_i^*)] + J\bar{\pi} - E[\alpha_i^*] = E[m_i^*] - E[\text{OOP}_i^*] + J\bar{\pi}.$$

The second equality uses $E[\tau_1(m^*)] + E[\alpha^*] = E[\text{OOP}^*]$. Since $E[m^*]$ and $E[\text{OOP}^*]$ both depend only on $T(\cdot)$, so does p.

Step 3: The extensive margin is also neutral.

A consumer of type λ visits if and only if $U^{\text{visit}}(\lambda) \geq U^0(\lambda)$ (Equation 4), where (Equation 2):

$$U^{\text{visit}}(\lambda) = \sigma \log \sum_j \exp\Big(\tfrac{1}{\sigma}\big(W - \text{OOP}_j^* + h(m_j^* - \lambda) - p\big)\Big),$$

and $U^0(\lambda) = W - p + h(0 - \lambda)$.

Since m^*, OOP^*, and p all depend only on $T(\cdot)$, both U^{visit} and U^0 depend only on $T(\cdot)$. The visit

decision — and hence the equilibrium set of types who visit — therefore also depends only on the combined schedule $T(\cdot) = \tau_1(\cdot) + \tau_2(\cdot)$, not on $\tau_1(\cdot)$ and $\tau_2(\cdot)$ separately. □

C Proof that Insurers Prefer Non-transferability

This appendix proves that an insurer with non-transferable utility achieves strictly higher consumer welfare than one with transferable utility. The result is established under the following simplifying assumptions relative to the full model:

1. *Linear, constant consumer cost-sharing.* τ_1 is a constant (not a function of m).
2. There exist consumer types λ for whom the interior solution $m^* > k$ holds, so that consumer out-of-pocket costs $\tau_1 m^*(\lambda)$ vary with λ.

We use "NT" to denote the non-transferable world ($\alpha_j = 0$ for all j) and "T" the transferable world (side payments α_j unrestricted). Our proof proceeds in three steps:

1. We show that at $\tau_2 = 0$, the NT insurer achieves weakly higher consumer welfare than the T insurer. Even without firm cost-sharing, side payments directly harm consumers: they create out-of-pocket risk (by making α increasing in λ) and induce wasteful marginal visits.
2. By Proposition 1 (neutrality), the T insurer's global optimum equals its optimum at $\tau_2 = 0$ — increasing τ_2 while holding $\tau_1 + \tau_2$ fixed leaves all T-world equilibrium outcomes unchanged. Therefore the T insurer's global optimum is weakly dominated by the NT insurer's optimum, by Step 1.
3. By Proposition 2, the optimal NT contract sets $\tau_1 = 0$ and $\tau_2 > 0$, achieving full consumption smoothing among visitors (zero OOP variation). Any $\tau_2 = 0$ policy with $\tau_1 > 0$ instead exposes visitors to OOP variation $\tau_1 m^*(\lambda)$ that varies with λ. With strictly concave $v(\cdot)$, the NT optimum therefore strictly dominates any $\tau_2 = 0$ outcome, making the final inequality strict.

Step 1: At $\tau_2 = 0$, non-transferability achieves weakly higher welfare.

NT equilibrium ($\alpha = 0$, $\tau_2 = 0$).

With $\alpha_j = 0$ and $\tau_2 = 0$, the firm earns R per consumer regardless of m. The firm therefore chooses m to maximize its market share, which is equivalent to maximizing consumer utility (Equation 1):

$$m^*_{nt}(\lambda) = \arg\max_{m \geq k} \left[h(m - \lambda) - \tau_1 m \right].$$

The interior first-order condition $h'(m-\lambda)=\tau_1$ gives $m = h'^{-1}(\tau_1)+\lambda$. Therefore:

$$m^*_{nt}(\lambda) = \begin{cases} h'^{-1}(\tau_1)+\lambda & \text{if } h'^{-1}(\tau_1)+\lambda \geq k, \\ k & \text{otherwise.} \end{cases}$$

Let λ^*_{nt} denote the marginal type who is indifferent between visiting and staying home:

$$W - \tau_1 m^*_{nt}(\lambda^*_{nt}) + h\big(m^*_{nt}(\lambda^*_{nt}) - \lambda^*_{nt}\big) + \sigma \log J - p = W + h(0-\lambda^*_{nt}) - p.$$

All consumers with $\lambda \geq \lambda^*_{nt}$ visit and receive $m^*_{nt}(\lambda)$.

T equilibrium ($\tau_2 = 0$).

With $\tau_2 = 0$, the firm's per-consumer profit is $R+\alpha_j$, and its problem for m_j is identical to the NT case (market-share maximization). Therefore:

$$m^*_t(\lambda) = m^*_{nt}(\lambda).$$

The firm additionally chooses $\alpha_j(\lambda)$ to maximize profits $\pi_j(\lambda) = (R+\alpha_j)\, s_j(\lambda)$. We show that the equilibrium has two cutoffs $\lambda^*_t < \lambda'_t$:

Interior regime ($\lambda > \lambda'_t$). For types who visit regardless of α_j, the unconstrained first-order condition for α_j (from Equation 6 with $\beta = 0$) in a symmetric equilibrium with $s_j = 1/J$ gives:

$$\alpha^* = \frac{J\sigma}{J-1} - R.$$

Note $\alpha^* > -R$ since $J\sigma/(J-1) > 0$ for $J \geq 2$.

Indifference regime ($\lambda^*_t \leq \lambda \leq \lambda'_t$). For types who would not visit at the interior α^* but can be induced to visit at a lower charge, the firm sets $\alpha_j(\lambda)$ to make the consumer just indifferent:

$$\alpha^*_t(\lambda) = h\big(m^*_t(\lambda) - \lambda\big) - h(0-\lambda) - \tau_1 m^*_t(\lambda) + \sigma \log J.$$

This is feasible as long as $\alpha^*_t(\lambda) \geq -R$.

Cutoffs λ^*_t and λ'_t. λ^*_t is the lowest type for whom the firm can break even while inducing a visit (i.e., the type indifferent when the firm offers its maximum subsidy $\alpha = -R$); λ'_t is the type indifferent at the interior charge α^*.

To confirm $\lambda_t' > \lambda_t^*$: define $\Delta(\lambda) \equiv U^{\text{visit}}(\lambda;\alpha^*) - U^0(\lambda)$. Since $\alpha^* > -R$, we have $\Delta(\lambda) < \Delta(\lambda;\alpha = -R)$ for all λ. By definition of λ_t^*, $\Delta(\lambda_t^*;\alpha = -R) = 0$, so $\Delta(\lambda_t^*) < 0$. Moreover $\partial\Delta/\partial\lambda > 0$: for interior types, $\partial\Delta/\partial\lambda = -\tau_1 + h'(-\lambda) > 0$ since $h'(-\lambda) > h'(0) = 1 > \tau_1$ by concavity of h; for corner types ($m^* = k$), $\partial\Delta/\partial\lambda = h'(-\lambda) - h'(k-\lambda) > 0$ by concavity. Hence there exists $\lambda_t' > \lambda_t^*$ with $\Delta(\lambda_t') = 0$.

Welfare comparison at $\tau_2 = 0$.

Because competitive insurers pass all surplus to consumers via the premium, expected consumer welfare is determined by the distribution of ex-post utility across types. We compare the NT and T cases along two dimensions.

Efficiency. $\lambda_t^* < \lambda_{nt}^*$: more consumers visit in the T world. For $\lambda \in [\lambda_t^*, \lambda_{nt}^*)$, the firm sets $\alpha_t^*(\lambda)$ to make these consumers exactly indifferent, so their ex-post utility equals their outside option $U^0(\lambda)$ —identical to what they would receive without visiting. Yet their visit incurs service costs $m_t^*(\lambda) > 0$ that are ultimately borne by all consumers via the premium. This is a pure waste: resources are consumed with no net benefit to the marginal visitors, raising premiums and reducing welfare for all types.

Risk protection. For $\lambda \geq \lambda_{nt}^*$ (types who visit in both worlds), since $m_{nt}^* = m_t^*$, the difference in consumer out-of-pocket costs is simply $\alpha_t^*(\lambda)$. As shown above, $\alpha_t^*(\lambda)$ is weakly increasing in λ: sicker consumers face higher OOP in the T world. This positive covariance between OOP and illness severity increases the variance of consumer consumption. With strictly concave $v(\cdot)$, this reduction in risk protection strictly reduces ex-ante welfare relative to the NT world, where $\alpha = 0$ and OOP varies with λ only through $\tau_1 m^*(\lambda)$.

Combining these two comparisons, the NT world achieves weakly higher consumer welfare than the T world when $\tau_2 = 0$.

Step 2: Extending to all τ_2.

By Proposition 1 (neutrality), for any policy (τ_1, τ_2) available to the T insurer, the equilibrium outcome is identical to the outcome under $(\tau_1 + \tau_2, 0)$ with transferable utility. Therefore:

$$\max_{\tau_1,\,\tau_2} \text{Welfare}^T(\tau_1,\tau_2) = \max_{\tau_1} \text{Welfare}^T(\tau_1,0) \leq \max_{\tau_1} \text{Welfare}^{NT}(\tau_1,0) \leq \max_{\tau_1,\,\tau_2} \text{Welfare}^{NT}(\tau_1,\tau_2),$$

where the middle inequality follows from Step 1.

Step 3: Non-transferability is strictly preferred.

By Proposition 2, the optimal NT contract sets $\tau_1 = 0$ and $\tau_2 = \tau_2^*(m) > 0$. Under this contract,

consumer out-of-pocket costs among visitors are $\tau_1 \cdot m^*(\lambda) = 0$ for all λ —full consumption smoothing across visitor types. In contrast, any $\tau_2 = 0$ policy with $\tau_1 > 0$ exposes visitors to OOP costs $\tau_1 m^*(\lambda)$, which vary strictly with λ under the intensive margin assumption. Since $v(\cdot)$ is strictly concave, the reduction in OOP variance achieved by the optimal NT contract yields strictly higher ex-ante welfare than any $\tau_2 = 0$ policy. Therefore the final inequality in Step 2 is strict, and the optimal NT insurer achieves strictly higher consumer welfare than any T insurer. □

D Definition of Constrained First Best

The constrained first best is the solution to the insurer's problem of maximizing consumers' expected utility subject to budget balance and the following constraints on the constrained Arrow-Debreu state contingent contract:

- The visit decision is not enforceable by the insurer. Each agent can always choose the minimum treatment level k as long as she is willing to pay the visit copay $c_v = \tau_v k$ set by the insurer.
- Arrow–Debreu contract among visitors for the intensive margin only. Once an individual has paid the visit copay and decides to visit, her type λ is revealed to the insurer, who can then offer an Arrow–Debreu state-contingent contract among visitors. In particular, the insurer can freely choose $m(\lambda) \geq k$ and a state-dependent reimbursement schedule for the intensive margin $m(\lambda) - k$, up to full reimbursement of the intensive-margin treatment. Equivalently, for each realized visitor define patient out-of-pocket spending on the intensive margin by

$$x(\lambda) \in [0, m(\lambda) - k].$$

- Budget balance. The premium is chosen to satisfy the insurer's (or planner's) budget constraint.

Thus a realized visitor of type λ receives utility

$$U_v(\lambda) = W - p - c_v - x(\lambda) + h(m(\lambda) - \lambda) + \sigma \log J,$$

while a non-visitor receives

$$U_0(\lambda) = W - p + h(0 - \lambda).$$

Visit decisions remain endogenous according to equation (4). The planner chooses the common visit

copay c_v, the expenditure rule $m(\lambda)$ for realized visitors, and the intensive-margin out-of-pocket schedule $x(\lambda)$ for realized visitors, subject to budget balance.

Lemma 1. *In any optimal allocation problem that maximize expected utility of consumer, the planner sets*

$$x^*(\lambda) = 0 \qquad \textit{for every realized visitor } \lambda.$$

Equivalently, the planner fully reimburses the intensive margin for every realized visitor.

Proof. Fix a realized visitor λ with $m(\lambda) > k$. Let

$$\mu \equiv E[v'(U(\lambda))]$$

denote the social value of one additional dollar of premium revenue, and let $\eta(\lambda) \geq 0$ be the multiplier on the participation constraint for type λ.

For an interior choice of $m(\lambda)$, the planner's first-order condition is

$$\big(v'(U(\lambda)) + \eta(\lambda)\big)h'(m(\lambda) - \lambda) = \mu. \tag{40}$$

Now consider a feasible perturbation that increases both $m(\lambda)$ and $x(\lambda)$ by the same small amount $d > 0$. This leaves insurer cost, and hence the premium, unchanged. Utility in state λ changes by

$$dU_v(\lambda) = \big(h'(m(\lambda) - \lambda) - 1\big)d.$$

Optimality therefore implies

$$h'(m(\lambda) - \lambda) \leq 1. \tag{41}$$

Combining (A1) and (A2) yields

$$v'(U(\lambda)) + \eta(\lambda) \geq \mu. \tag{42}$$

Now increase $x(\lambda)$ holding $m(\lambda)$ fixed. This lowers utility in state λ by one dollar, but lowers insurer cost by one dollar and therefore lowers the common premium by one dollar for everyone. Hence

$$\frac{\delta \mathcal{L}}{\delta x(\lambda)} = \mu - \big(v'(U(\lambda)) + \eta(\lambda)\big) \leq 0$$

by (A3). So welfare is weakly decreasing in $x(\lambda)$, and the planner weakly prefers the lowest feasible

value, namely

$$x(\lambda) = 0.$$

If instead $m(\lambda) = k$, then there is no intensive margin, so feasibility already implies

$$x(\lambda) \in [0, m(\lambda) - k] = \{0\}.$$

Therefore $x^*(\lambda) = 0$ for every realized visitor. □

Proposition 2. *Then the constrained-first-best allocation is characterized as follows.*

1. *For any fixed visit copay τ_v, there exists a constant $e(\tau_v)$ such that the planner's optimal expenditure rule is*

$$m^{cfb}(\lambda; \tau_v) = \begin{cases} 0, & \lambda < \lambda^*(\tau_v), \\ k, & \lambda^*(\tau_v) \le \lambda \le \lambda^{**}(\tau_v), \\ \lambda + e(\tau_v), & \lambda > \lambda^{**}(\tau_v), \end{cases}$$

   where

$$\lambda^{**}(\tau_v) = k - e(\tau_v),$$

   and $e(\tau_v)$ solves

$$h'(e(\tau_v)) = \frac{E[v'(U)]}{E[v'(U) \mid m^{cfb} > k]}.$$

   Equivalently, all interior visitors satisfy the same marginal condition

$$h'(m^{cfb}(\lambda) - \lambda) = \phi \qquad \text{for all } \lambda \text{ such that } m^{cfb}(\lambda) > k,$$

   where

$$\phi \equiv \frac{E[v'(U)]}{E[v'(U) \mid m^{cfb} > k]}.$$

2. *The visit cutoff $\lambda^*(\tau_v)$ is determined by indifference at the corner $m = k$:*

$$W - p - \tau_v k + h(k - \lambda^*) + \sigma \log J = W - p + h(0 - \lambda^*).$$

3. *The optimal visit copay τ_v^{cfb} satisfies*

$$P(visit)\,k\,\frac{E[v'(U)\mid visit]-E[v'(U)\mid all]}{E[v'(U)\mid all]}=-E[m_{\min}-\tau_v k]\,\frac{\partial P(visit)}{\partial \tau_v}.$$

4. *This allocation is implemented by the contract*

$$(\tau_v^*,\tau_1^*,\tau_2^*(\cdot))=(\tau_v^{cfb},0,\tau_2^*(\cdot)),$$

where $g(m)\equiv R-\tau_2(m)$ satisfies

$$\frac{g'(m)}{g(m)}=-\frac{1-1/J}{\sigma}\phi.$$

Hence

$$g(m)=A\exp\left(-\frac{1-1/J}{\sigma}\phi\, m\right)$$

for some constant $A>0$ chosen to match the target profit level.

Proof. By the zero-OOP lemma, once a consumer visits, the planner fully reimburses the intensive margin. Therefore the planner chooses only a common visit copay τ_v and an expenditure rule $m(\lambda)\geq k$ for visiting types.

The planner's objective is

$$E_\lambda[v(U(\lambda))],$$

where

$$U(\lambda)=\max\{U_{visit}(\lambda),U_0(\lambda)\},$$

$$U_0(\lambda)=W-p+h(0-\lambda),$$

and

$$U_{visit}(\lambda)=W-p-\tau_v k+h(m(\lambda)-\lambda)+\sigma\log J.$$

Suppressing the fixed profit term, the premium is

$$p=\int 1\{\lambda \text{ visits}\}\big(m(\lambda)-\tau_v k\big)\,dF(\lambda).$$

Fix τ_v, and define

$$\mu\equiv E[v'(U(\lambda))].$$

Consider any type λ with an interior allocation $m(\lambda) > k$. A marginal increase in $m(\lambda)$ raises utility in state λ by $h'(m(\lambda) - \lambda)$ and raises the premium by one dollar for everyone. Hence the first-order condition is

$$v'(U_{visit}(\lambda))\, h'(m(\lambda) - \lambda) = \mu. \tag{43}$$

Now, for an interior visitor,

$$U_{visit}(\lambda) = W - p - \tau_v k + h(m(\lambda) - \lambda) + \sigma \log J,$$

so the left-hand side of (1) depends on λ only through the gap

$$e(\lambda) \equiv m(\lambda) - \lambda.$$

Because h is concave and v is concave, the map

$$e \mapsto v'(W - p - \tau_v k + h(e) + \sigma \log J)\, h'(e)$$

is strictly decreasing. Therefore (1) implies that all interior visitors must have the same gap $e(\tau_v)$:

$$m(\lambda) = \lambda + e(\tau_v) \qquad \text{for all } \lambda \text{ such that } m(\lambda) > k. \tag{44}$$

Let

$$\lambda^{**}(\tau_v) \equiv k - e(\tau_v).$$

Then the lower bound $m(\lambda) \geq k$ implies

$$m^{cfb}(\lambda; \tau_v) = \begin{cases} 0, & \lambda < \lambda^{*}(\tau_v), \\ k, & \lambda^{*}(\tau_v) \leq \lambda \leq \lambda^{**}(\tau_v), \\ \lambda + e(\tau_v), & \lambda > \lambda^{**}(\tau_v). \end{cases} \tag{45}$$

Because all interior visitors have the same gap $e(\tau_v)$, they also have the same utility level. Therefore, on the interior set,

$$v'(U_{visit}(\lambda)) = E[v'(U) \mid m^{cfb} > k].$$

Substituting into (1) yields

$$h'(e(\tau_v)) = \frac{E[v'(U)]}{E[v'(U) \mid m^{cfb} > k]}. \tag{46}$$

The visit cutoff $\lambda^*(\tau_v)$ is pinned down by indifference at the corner $m = k$:

$$W - p - \tau_v k + h(k - \lambda^*) + \sigma \log J = W - p + h(0 - \lambda^*). \tag{47}$$

Now optimize with respect to τ_v. By envelope, we may hold the expenditure rule fixed when differentiating the planner's value function. Let $P(visit)$ denote the visit probability. Then

$$\frac{dp}{d\tau_v} = -k\, P(visit) + E[m_{\min} - \tau_v k]\, \frac{\partial P(visit)}{\partial \tau_v}, \tag{48}$$

where $m_{\min}$ denotes the expenditure a consumer receives if she visits at the visit margin. Since marginal types are indifferent between visiting and not visiting, there is no direct first-order welfare effect from switching types. Hence

$$0 = \frac{d}{d\tau_v} E[v(U)] = -k\, P(visit)\, E[v'(U) \mid visit] - \frac{dp}{d\tau_v} E[v'(U) \mid all].$$

Substituting (6) and rearranging gives

$$P(visit)\, k\, \frac{E[v'(U) \mid visit] - E[v'(U) \mid all]}{E[v'(U) \mid all]} = -E[m_{\min} - \tau_v k]\, \frac{\partial P(visit)}{\partial \tau_v}. \tag{49}$$

This characterizes the constrained-first-best visit copay.

We now show implementation. Set

$$\tau_1 = 0,$$

and define firm profits per consumer by

$$g(m) \equiv R - \tau_2(m).$$

In the competitive-firm problem, firm j chooses $m_j(\lambda)$ to solve

$$\max_{m_j \geq k} s_j(\lambda)\, g(m_j),$$

where $s_j(\lambda)$ is the logit share among visitors. The first-order condition is

$$\frac{\partial s_j(\lambda)}{\partial m_j}\, g(m_j) + s_j(\lambda)\, g'(m_j) = 0.$$

Using the logit derivative,

$$\frac{\partial s_j(\lambda)}{\partial m_j} = \frac{1}{\sigma}\big(h'(m_j(\lambda) - \lambda) - \tau_1\big) s_j(\lambda)(1 - s_j(\lambda)).$$

Hence, in symmetric equilibrium where $s_j = 1/J$,

$$\frac{1}{\sigma}\big(h'(m(\lambda) - \lambda) - \tau_1\big)\left(1 - \frac{1}{J}\right) g(m) + g'(m) = 0.$$

Since $\tau_1 = 0$, this becomes

$$h'(m(\lambda) - \lambda) = -\frac{\sigma}{1 - 1/J}\frac{g'(m)}{g(m)}. \tag{50}$$

Now choose g so that

$$\frac{g'(m)}{g(m)} = -\frac{1 - 1/J}{\sigma}\phi, \qquad \phi \equiv h'(e(\tau_v^{cfb})). \tag{51}$$

Then (8) implies

$$h'(m(\lambda) - \lambda) = \phi,$$

so every interior firm choice satisfies

$$m(\lambda) = \lambda + h'^{-1}(\phi) = \lambda + e(\tau_v^{cfb}).$$

If this is below k, the constraint $m \geq k$ binds, so the firm chooses $m(\lambda) = k$. Therefore the firm's expenditure rule coincides exactly with (3).

Integrating (9) gives

$$g(m) = A \exp\left(-\frac{1 - 1/J}{\sigma}\phi\, m\right)$$

for some $A > 0$. The constant A determines the level of profits but does not affect incentives.

Finally, because $\tau_1 = 0$, consumers pay only the visit copay on the consumer side. The visit cutoff is therefore determined by the same indifference equation (5). Thus choosing

$$\tau_v = \tau_v^{cfb}$$

reproduces the same visit threshold λ^* as in the constrained first best. Since the expenditure rule and the out-of-pocket schedule coincide, the premium also coincides by budget balance. Therefore the contract

$$(\tau_v^{cfb}, 0, \tau_2^*(\cdot))$$

implements the constrained-first-best allocation. □

E Doctor Heterogeneity

We characterize the first-best allocation when the social planner can choose treatment for each patient, the allocation of patients to doctors, and out-of-pocket payments for each patient. We then show that the same outcome can be decentralized with doctor-specific copays and an optimal τ_2 schedule: within each doctor, out-of-pocket payments do not vary with illness severity λ, and cross-doctor differences in copays equal their productivity differences.

Suppose doctors j are heterogeneous in productivity θ_j, so that a patient of type λ treated by doctor j with effort m achieves health $h(m+\theta_j-\lambda)$. There is no extensive margin: every patient visits exactly one doctor. A patient is characterized by illness severity λ and idiosyncratic doctor tastes $e = (e_1, \ldots, e_J)$. If assigned to doctor j, who exerts effort m, the patient pays premium p and out-of-pocket cost $c_j(\lambda)$ and obtains utility:

$$h(m+\theta_j-\lambda) - c_j(\lambda) - p + e_j.$$

A planner who observes the patient's type (λ, e) and each doctor's productivity θ_j chooses:

- an assignment rule $j(\lambda, e) \in \{1, \ldots, J\}$,
- effort $m_j(\lambda, e)$,
- a patient out-of-pocket schedule $c_j(\lambda)$,
- and a common premium p,

to maximize expected patient welfare:

$$W = \mathbb{E}_\lambda \left[v \left(\mathbb{E}_e \left[h\big(m_{j(\lambda,e)}(\lambda,e) + \theta_{j(\lambda,e)} - \lambda\big) + e_{j(\lambda,e)} - c_{j(\lambda,e)}(\lambda) \,\middle|\, \lambda \right] - p \right) \right],$$

subject to budget balance:

$$p + \mathbb{E}_{\lambda,e}\left[c_{j(\lambda,e)}(\lambda)\right] = \mathbb{E}_{\lambda,e}\left[m_{j(\lambda,e)}(\lambda,e)\right].$$

Assume h is concave, the solution is interior, and $h'(x) = 1$ has the unique solution $x = 0$. Then:

1. the first-best effort is:

$$m_j^*(\lambda,e) = \lambda - \theta_j,$$

   so it depends on patient type and doctor productivity but not on e;

2. all doctors deliver the same first-best health level,

$$h\Big(m_j^*(\lambda,e) + \theta_j - \lambda\Big) = h(0);$$

3. the efficient doctor assignment is:

$$j^*(\lambda,e) \in \arg\max_j \{e_j + \theta_j\};$$

4. if $\{e_j\}$ are i.i.d. Type-I extreme value with scale parameter σ, the efficient market share of doctor j is:

$$s_j^* = \frac{\exp\big((h(0)+\theta_j)/\sigma\big)}{\sum_{k=1}^{J}\exp((h(0)+\theta_k)/\sigma)};$$

5. among all posted patient payment schedules that implement the efficient assignment, the planner's optimum is unique up to an additive constant and is given by:

$$c_j^*(\lambda) \equiv \bar{c} - \theta_j, \qquad p^* = \mathbb{E}[\lambda] - \bar{c}.$$

   All patients who choose the same doctor pay the same out-of-pocket amount, with no within-doctor variation across λ, and cross-doctor differences in copays equal productivity differences:

$$c_j^* - c_{j'}^* = -(\theta_j - \theta_{j'}).$$

Proof. The proof has four steps.

Step 1: First-best effort within each doctor. Fix a patient (λ, e) assigned to doctor j. The planner

chooses effort to maximize:

$$h(m+\theta_j-\lambda)-m.$$

The taste term e_j is additive and does not affect effort. The first-order condition is:

$$h'(m+\theta_j-\lambda)=1,$$

whose unique solution is $m+\theta_j-\lambda=0$, giving:

$$m_j^*(\lambda,e)=\lambda-\theta_j.$$

This proves part (1). Substituting back gives $h(m_j^*+\theta_j-\lambda)=h(0)$, proving part (2).

In the first best, all doctors deliver the same health level $h(0)$, but higher-productivity doctors do so with less effort.

Step 2: Efficient allocation of patients across doctors. Substituting the first-best effort rule into the planner's objective, the social value of assigning patient (λ,e) to doctor j is:

$$h(0)+e_j-(\lambda-\theta_j)=h(0)-\lambda+(e_j+\theta_j).$$

Since $h(0)-\lambda$ is common across doctors, the planner assigns each patient to the doctor maximizing $e_j+\theta_j$, proving part (3).

Equivalently, doctor j' is preferred to j if and only if $e_{j'}-e_j\geq\theta_j-\theta_{j'}$: patients should choose the less productive doctor only when their idiosyncratic preference for that doctor outweighs the extra resource cost. Under i.i.d. Type-I extreme value shocks, this yields the logit market shares in part (4).

Step 3: Patient payment schedules that implement the efficient assignment. With first-best effort, every doctor delivers health $h(0)$, so patient (λ,e) choosing doctor j obtains utility $h(0)+e_j-c_j(\lambda)$. For decentralized choices to reproduce the efficient assignment $j^*=\arg\max_j\{e_j+\theta_j\}$, the payment schedule must satisfy, for each λ:

$$c_j(\lambda)=k(\lambda)-\theta_j,$$

for some scalar $k(\lambda)$. The cross-doctor copay difference is therefore pinned down:

$$c_j(\lambda)-c_{j'}(\lambda)=-(\theta_j-\theta_{j'}),$$

so more productive doctors must have lower copays by exactly their productivity advantage.

Step 4: The optimal schedule has no within-doctor variation across λ. With $c_j(\lambda) = k(\lambda) - \theta_j$, the patient's expected utility before paying the premium is:

$$h(0) - k(\lambda) + \mathbb{E}_e\left[\max_j \{e_j + \theta_j\}\right] \equiv h(0) - k(\lambda) + I(\theta).$$

Budget balance gives:

$$p = \mathbb{E}[\lambda] - \mathbb{E}[k(\lambda)],$$

so conditional utility at severity λ equals:

$$h(0) + I(\theta) - \mathbb{E}[\lambda] - \big(k(\lambda) - \mathbb{E}[k(\lambda)]\big).$$

The mean is independent of $k(\lambda)$; varying $k(\lambda)$ only redistributes utility across illness severities. Since the planner applies the strictly concave $v(\cdot)$, any variation in $k(\lambda)$ creates unnecessary risk. By Jensen's inequality, welfare is maximized at $k(\lambda) \equiv \bar{c}$, giving:

$$c_j^*(\lambda) \equiv \bar{c} - \theta_j, \qquad p^* = \mathbb{E}[\lambda] - \bar{c}.$$

This proves part (5). Under this contract, conditional utility is constant across λ, so the optimal payment system not only implements the efficient doctor allocation but also provides full insurance across severity states. □

Corollary (Link to the doctor-side implementation in the main text). Firm j maximizes market share times profit, $s_j(m_j) \cdot g_j(m_j)$. Using $\partial s_j / \partial m_j = \frac{1}{\sigma} s_j(1 - s_j) h'(m_j + \theta_j - \lambda)$, the first-order condition is:

$$\frac{1}{\sigma} s_j (1 - s_j)\, h'(m + \theta_j - \lambda)\, g_j(m) + s_j\, g_j'(m) = 0,$$

which simplifies (dividing by s_j) to:

$$h'(m + \theta_j - \lambda) = -\frac{\sigma\, g_j'(m)}{(1 - s_j)\, g_j(m)}.$$

Setting this equal to 1 (the first-best condition) and solving gives $g_j'/g_j = -(1-s_j^*)/\sigma$, so the insurer should set:

$$g_j(m) = A\exp\left(-\frac{1-s_j^*}{\sigma}\,m\right).$$

For a symmetric equilibrium with $s_j^* = 1/J$, this reduces to $g_j(m) = A\exp\left(-\frac{(1-1/J)}{\sigma}m\right)$, matching the formula in Proposition 2 (with $\phi = 1$). Combined with the patient-side copay $c_j^* = \bar{c} - \theta_j$, this decentralizes the planner's first best. The level parameter A shifts the doctor-side transfer and can be absorbed into the premium; it does not affect the marginal incentive.

F Monopolistic Limit

We show that, under a linear contract, the first-best allocation is achievable in the monopolistic limit ($\sigma \to \infty$) with vanishingly small firm cost-sharing and no consumer cost-sharing. For convenience, we drop the extensive margin decision and focus on the intensive margin. The results extend to the full model with an extensive margin: the intensive-margin contract shape derived here applies conditional on visiting, while the visit copay τ_v is separately pinned down by the Baily–Chetty condition in Proposition 2.

Consider a linear contract $g(m) = R - \tau_2 m$. Suppose the insurer sets $\tau_1 = 0$ and:

$$\tau_2 = \frac{(1-1/J)\,R}{\sigma}.$$

Substituting into the firm's first-order condition (equation 10),

$$h'(m-\lambda) - \tau_1 = \frac{\sigma\,\tau_2}{(1-1/J)} \cdot \frac{1}{R-\tau_2\,m},$$

the right-hand side becomes:

$$\frac{\sigma}{(1-1/J)} \cdot \frac{(1-1/J)\,R}{\sigma} \cdot \frac{1}{R-\tau_2\,m} = \frac{R}{R-\tau_2\,m}.$$

As $\sigma \to \infty$, $\tau_2 = (1-1/J)\,R/\sigma \to 0$, so $R - \tau_2\,m \to R$ and the right-hand side converges to 1. Since $\tau_1 = 0$, the first-order condition becomes $h'(m-\lambda) = 1$ for all λ, which implies $m^*(\lambda) = \lambda$—the first-best allocation.

Intuitively, as market power grows, even a vanishingly small cost-sharing rate suffices to align

firm incentives with social welfare. With $\tau_1 = 0$, consumers bear no out-of-pocket cost, so there is no risk-protection loss. The insurer achieves the first best by relying entirely on firm cost-sharing.

G Proofs for "Almost" Perfect Competition: $\sigma \to 0$; No Extensive Margin Costs

We characterize the optimal linear contract in the competitive limit ($\sigma \to 0$). We show that firm cost-sharing acts as a uniform expenditure cap on high-cost types, while consumer cost-sharing controls moral hazard for types below the cap. For convenience, we drop the extensive margin decision and focus on the intensive margin. As in Appendix F, the results extend to the full model: the cap-plus-coinsurance structure derived here applies conditional on visiting, while the visit copay τ_v is separately determined by the Baily–Chetty condition in Proposition 2.

Denote by τ_1' and τ_2' the limits of τ_1^* and τ_2^* respectively as $\sigma \to 0$ (we will prove these limits exist below by deriving them exactly).

We will first show that there is a cut-off type λ^* such that, if $\lambda < \lambda^*$, $\lim_{\sigma\to 0} m^*(\lambda, \tau_1^*, \tau_2^*) = m^*(\lambda, \tau_1', 0)$. Define $\lambda^* = \frac{R}{\tau_2'} - (h')^{-1}(\tau_1')$. If consumers were to choose m directly to maximize utility, consumers of this type would choose $m^* = \frac{R}{\tau_2'}$ (as they do at $\sigma = 0$). Note that this value of m^* would mean that firm margins were identically zero for that type. If $\lambda < \lambda^*$, $R - \tau_2 m^*(\lambda) > 0$ (since $m^*(\lambda) < R/\tau_2'$ because h is strictly increasing). The FOC of firm j is given by:

$$h'(m_j^*(\lambda) - \lambda) - \tau_1 = \frac{\sigma \cdot \tau_2}{(1 - s_j)} \frac{1}{(R - \tau_2 m_j^*)} \tag{52}$$

So, if $R - \tau_2 m_j > 0$ as $\sigma \to 0$, m^* must converge to $\lambda + (h')^{-1}(\tau_1') = m^*(\lambda, \tau_1', 0)$. We will show below that τ_1' is the same consumer cost-sharing the firm would choose if they had available only consumer cost-sharing and a uniform cap.

For $\lambda \geq \lambda^*$,

$$\lim_{\sigma\to 0} m^*\left(\lambda, \tau_1^*(\sigma), \tau_2^*(\sigma), \sigma\right) = \frac{R}{\tau_2'},$$

This is because $\lambda + (h')^{-1}\left(\tau_1'\right) > \frac{R}{\tau_2'}$, so $\lambda + (h')^{-1}\left(\tau_1'\right)$ cannot be a solution, as it would generate negative profit for the firm. Therefore, the best the firm can do is to provide $m^* = \frac{R}{\tau_2'}$.

Summarizing, in the limit, $\tau_1^*(\sigma)$ and $\tau_2^*(\sigma)$ converge to τ_1' and τ_2', and:

$$\lim_{\sigma \to 0} m^*\left(\lambda, \tau_1^*(\sigma), \tau_2^*(\sigma), \sigma\right) = \min\left\{\lambda + (h')^{-1}\left(\tau_1'\right),\ \frac{R}{\tau_2'}\right\}.$$

Now, we will characterize τ_1' and τ_2'. We will show that the limiting outcome is equivalent to a social planner's problem to choose τ_1 and a uniform cap for m at K when τ_2 is not available.

In this problem:

$$m = \min\left\{\lambda + (h')^{-1}(\tau_1),\ K\right\}.$$

Note first that any solution to this problem is also achievable by the insurer as $\sigma \to \infty$ by setting:

$$\tau_1'(\sigma) = \tau_1,$$
$$\tau_2'(\sigma) = \frac{K}{R}$$

in the limit as $\sigma \to 0$. Likewise, the solution already derived can be achieved by the insurer with only τ_1 and a cap by setting:

$$\tau_1 = \tau_1',$$
$$K = \frac{R}{\tau_2'}$$

given the solution already derived. So the social planner choosing τ_1 and K in the second problem is equivalent to the limiting case of the insurer choosing $\tau_1(\sigma)$ and $\tau_2(\sigma)$.

Therefore, as long as the insurer's solution in the second problem (where she can use both τ_1 and k at the same time) is not the same as in the insurer's problem when she can only use consumer cost-sharing, the limit of the first problem will not be the same as the insurer's problem when she can only use cost-sharing.

Once the insurer sets K and $\tau_1$in the second problem,

$$\lambda^* = K - h'^{-1}(\tau_1)$$

such that in the first problem,

$$\lim_{\sigma \to 0} m^*\big(\lambda, \tau_1^*(\sigma), \tau_2^*(\sigma), \sigma\big) = \begin{cases} \lambda + h'^{-1}(\tau_1) & \text{if } \lambda < k - h'^{-1}(\tau_1) \\ k & \text{if } \lambda \geq k - h'^{-1}(\tau_1) \end{cases}$$

which proves the statment 1 and 2.

H Information Asymmetry

ï»¿

Homogeneous altruism is not a rationale for consumer cost-sharing. We first show that when all firms share the same known altruism parameter $\beta \in (0,1)$—placing weight β on consumer utility and weight $1-\beta$ on profit—consumer cost-sharing is not necessary to achieve the first best.

Monopolist case. Following Ellis and McGuire (1990), suppose a single firm chooses m to maximize:

$$\beta\big(h(m-\lambda) - \tau_1 m\big) + (1-\beta)(R - \tau_2 m).$$

The first-order condition is:

$$h'(m-\lambda) = \tau_1 + \frac{1-\beta}{\beta}\tau_2.$$

Consumer cost-sharing τ_1 and scaled firm cost-sharing $\frac{1-\beta}{\beta}\tau_2$ enter the firm's decision symmetrically: the firm responds identically to a unit increase in τ_1 and a $\frac{\beta}{1-\beta}$ unit increase in τ_2. In particular, setting $\tau_1 = 0$ and $\tau_2 = \frac{\beta}{1-\beta}$ achieves $h'(m-\lambda) = 1$ for every λ—the first-best allocation—with no consumer out-of-pocket cost. Since $\tau_1 = 0$ provides strict risk protection, the planner always prefers $\tau_1^* = 0$. As altruism β rises, the required $\tau_2^* = \frac{\beta}{1-\beta}$ increases: more altruistic firms already internalize consumer welfare and therefore require stronger financial incentives to prevent overprovision.

Competitive case. In the competitive model, Proposition 2 shows that the insurer can achieve the first best with $\tau_1 = 0$ using an appropriate nonlinear payment schedule $g(m)$. Altruism does not change this conclusion. Because altruistic firms already place direct weight on consumer utility, their response to firm cost-sharing is attenuated; the optimal adjustment is a steeper $g(m)$, not positive τ_1. Formally, the firm's first-order condition continues to pin down $m(\lambda)$ uniquely for any given g, and the insurer can solve for the g that implements first-best effort at $\tau_1 = 0$, exactly as in the proof of Proposition 2.

Both arguments rely on altruism being *homogeneous* across firms and *observable* to the insurer,

so that $g(m)$ can be calibrated to β. When firms differ in their degree of altruism and the insurer cannot distinguish them, consumer cost-sharing becomes a distinct instrument. We now turn to this case.

Heterogeneous, unobservable altruism. We consider a setting in which the insurer cannot distinguish between doctors who differ in their degree of altruism toward patients. We present results for two doctor types with equal population weights; this assumption is made for simplicity only. The proposition and Corollary 4 both extend verbatim to N types with arbitrary positive weights $(\pi_1, \ldots, \pi_N)$: the necessary and sufficient condition for $\tau_1^* > 0$ is $\mathrm{Var}(\mathrm{MRS}) > 0$, i.e., that doctors are not all equally altruistic. The key argument is that asymmetric information creates a role for positive consumer coinsurance even when firm cost-sharing alone would otherwise be sufficient. Because firm cost-sharing affects less-altruistic doctors more strongly, replacing consumer cost-sharing with firm cost-sharing compresses the distribution of treatment intensities across doctor types at the cost of introducing consumption risk for patients, and this tradeoff favors a strictly positive consumer coinsurance rate. We then derive a closed-form expression for the magnitude of τ_1^* to leading order in the variance of doctor types (Corollary 4).

Assume there are two doctor types, β_1 and β_2, with equal population weight. Doctor i chooses treatment intensity according to:

$$m_i(\lambda) \equiv m(\lambda, \tau_1, \tau_2, \beta_i) \in \arg\max_m \left\{ \beta_i \big(h(m - \lambda) - \tau_1 m \big) - \tau_2 m \right\}.$$

Let pooled expected spending be:

$$\bar{m}(\tau_1, \tau_2) \equiv \frac{1}{2} \sum_{i=1}^{2} \mathbb{E}_\lambda [m_i(\lambda)],$$

and suppose the common premium is:

$$P(\tau_1, \tau_2) = (1 - \tau_1) \bar{m}(\tau_1, \tau_2),$$

so the insurance budget balances.

The planner chooses (τ_1, τ_2) to maximize:

$$\mathcal{W}(\tau_1, \tau_2) = \sum_{i=1}^{2} \mathbb{E}_\lambda \left[u\Big(W + h(m_i(\lambda) - \lambda) - \tau_1 m_i(\lambda) - P(\tau_1, \tau_2) \Big) \right].$$

Proposition 3. *Suppose u is strictly concave, h is increasing and strictly concave, and $h'(0) = 1$. Then:*

1. *if $\beta_1 > \beta_2$, the planner does not choose $\tau_1 = 0$; hence, on the domain $\tau_1 \geq 0$, any optimum satisfies $\tau_1^* > 0$;*

Proof. Let:

$$\mu \equiv \mathbb{E}[\lambda], \qquad x_i \equiv m_i(\lambda) - \lambda, \qquad \bar{x} \equiv \frac{x_1 + x_2}{2}.$$

Then $m_i(\lambda) = \lambda + x_i$, and the doctor's first-order condition is:

$$h'(x_i) = \tau_1 + \frac{\tau_2}{\beta_i}.$$

Since h is strictly concave, x_i is decreasing in:

$$q_i \equiv \tau_1 + \frac{\tau_2}{\beta_i}.$$

Using $P = (1 - \tau_1)(\mu + \bar{x})$, consumption for a patient treated by type i can be written as:

$$c_i(\lambda) = W - \mu + h(x_i) - \bar{x} - \tau_1 \big[(x_i - \bar{x}) + (\lambda - \mu) \big].$$

Consider first the case $\beta_1 > \beta_2$, and suppose toward a contradiction that $\tau_1 = 0$ is optimal. Then:

$$q_i = \frac{\tau_2}{\beta_i}.$$

Because $\beta_1 > \beta_2$, type 1 faces a lower effective marginal price, so:

$$x_1 > x_2.$$

Now increase τ_1 slightly and decrease τ_2 so as to keep aggregate spending fixed, i.e. keep $\bar{x}$ fixed. Since τ_1 affects both types equally but τ_2 affects type 2 more strongly (because $1/\beta_2 > 1/\beta_1$), this

perturbation compresses treatment choices:

$$x_1 \text{ falls}, \qquad x_2 \text{ rises}, \qquad \bar{x} \text{ stays fixed.}$$

This raises welfare for two reasons.

First, when $\tau_1 = 0$ we have:

$$c_i = W - \mu + h(x_i) - \bar{x}.$$

Hence, holding $\bar{x}$ fixed, welfare is the sum of:

$$g(x_i) \equiv u\big(W - \mu + h(x_i) - \bar{x}\big).$$

Because u is concave and h is concave, g is concave. Therefore compressing (x_1, x_2) while keeping their average fixed raises:

$$g(x_1) + g(x_2).$$

Second, the new term,

$$-\tau_1(x_i - \bar{x}),$$

is a transfer from the high-spending type to the low-spending type. Since utility is concave, this redistribution also raises welfare.

The only cost of increasing τ_1 is the mean-zero risk term,

$$-\tau_1(\lambda - \mu),$$

whose first-order effect at $\tau_1 = 0$ is zero, since:

$$\mathbb{E}[\lambda - \mu] = 0.$$

Therefore, starting from $\tau_1 = 0$, a small increase in τ_1 combined with a decrease in τ_2 that keeps aggregate spending fixed strictly increases welfare. This contradicts optimality of $\tau_1 = 0$. Hence, if $\beta_1 > \beta_2$, the optimum must satisfy:

$$\tau_1^* > 0.$$

□

Corollary 4 (Magnitude of optimal coinsurance). *Maintain the above setup. Write* $\mathrm{MRS}_i \equiv 1/\beta_i$. *Suppose that (i) the planner adjusts* τ_2 *as* τ_1 *varies so as to hold aggregate treatment* $\bar{x}$ *fixed, and (ii) altruism heterogeneity is small in the sense that*

$$R_A \cdot \frac{\tau_2^2 \,\mathrm{Var}(\mathrm{MRS})}{h''(\bar{x})^2} \;\ll\; 1, \tag{53}$$

where $R_A \equiv -u''(\bar{c})/u'(\bar{c})$ *is the Arrow–Pratt coefficient of absolute risk aversion evaluated at mean consumption* $\bar{c} \equiv W - \mu + h(\bar{x}) - \bar{x}$. *Then the welfare-maximising coinsurance rate satisfies, to leading order in* $\mathrm{Var}(\mathrm{MRS})$*:*

$$\boxed{\tau_1^* \;=\; \frac{\tau_2 \,\mathrm{Var}(\mathrm{MRS})}{R_A \,|h''(\bar{x})|\, E(\mathrm{MRS})\,\mathrm{Var}(\lambda)}} \tag{54}$$

Proof. We proceed in three steps. Throughout, write $\delta\mathrm{MRS}_i \equiv \mathrm{MRS}_i - E(\mathrm{MRS})$ and let $\sigma^2 \equiv \mathrm{Var}(\mathrm{MRS})$. The "leading order" qualifier means we retain terms through $O(\sigma^2)$ and neglect terms of order σ^2 multiplied by $R_A\tau_2^2\sigma^2/h''^2$, which is $o(\sigma^2)$ under assumption (53).

Step 1. Comparative statics holding $\bar{x}$ *fixed.*

Each doctor's first-order condition is:

$$h'(x_i) \;=\; \tau_1 + \tau_2 \,\mathrm{MRS}_i. \tag{55}$$

Differentiating with respect to τ_1, treating $\tau_2 = \tau_2(\tau_1)$:

$$h''(x_i)\,\frac{dx_i}{d\tau_1} \;=\; 1 + \mathrm{MRS}_i\,\frac{d\tau_2}{d\tau_1}. \tag{56}$$

We impose $\frac{d\bar{x}}{d\tau_1} = \frac{1}{2}\sum_i \frac{dx_i}{d\tau_1} = 0$. Approximating $h''(x_i) \approx h''(\bar{x})$ (valid to leading order in σ, since $x_i - \bar{x} = O(\sigma)$), we sum (56) over i:

$$h''(\bar{x})\cdot 0 \;=\; 2 + \Big(\textstyle\sum_i \mathrm{MRS}_i\Big)\frac{d\tau_2}{d\tau_1} \;=\; 2 + 2\,E(\mathrm{MRS})\,\frac{d\tau_2}{d\tau_1},$$

yielding:

$$\frac{d\tau_2}{d\tau_1} \;=\; -\frac{1}{E(\mathrm{MRS})}. \tag{57}$$

Substituting (57) back into (56):

$$\frac{dx_i}{d\tau_1} \;=\; \frac{1}{h''(\bar{x})}\Big(1 - \frac{\mathrm{MRS}_i}{E(\mathrm{MRS})}\Big) \;=\; \frac{-\delta\mathrm{MRS}_i}{h''(\bar{x})\,E(\mathrm{MRS})}. \tag{58}$$

Step 2. Welfare derivative.

The planner maximises $\mathcal{W}(\tau_1) = \frac{1}{2}\sum_i E_\lambda[u(c_i(\lambda))]$, where patient consumption when treated by doctor i is:

$$c_i(\lambda) \;=\; W - \mu + h(x_i) - \bar{x} - \tau_1\big[(x_i - \bar{x}) + (\lambda - \mu)\big]. \tag{59}$$

Write $\bar{c} \equiv W - \mu + h(\bar{x}) - \bar{x}$ for the consumption level when all heterogeneity vanishes. Differentiating (59) with respect to τ_1 (using $d\bar{x}/d\tau_1 = 0$):

$$\frac{dc_i}{d\tau_1} \;=\; \underbrace{\big(h'(x_i) - \tau_1\big)\frac{dx_i}{d\tau_1}}_{A_i} \;-\; \underbrace{(x_i - \bar{x})}_{B_i} \;-\; \underbrace{(\lambda - \mu)}_{C_i}. \tag{60}$$

From (55), $h'(x_i) - \tau_1 = \tau_2\,\mathrm{MRS}_i$, so:

$$A_i \;=\; \frac{-\tau_2\,\mathrm{MRS}_i\;\delta\mathrm{MRS}_i}{h''(\bar{x})\;E(\mathrm{MRS})}.$$

We evaluate $\frac{d\mathcal{W}}{d\tau_1} = \frac{1}{2}\sum_i E_\lambda\big[u'(c_i)\,\frac{dc_i}{d\tau_1}\big]$ by expanding $u'(c_i(\lambda))$ around $\bar{c}$. Note that:

$$c_i(\lambda) - \bar{c} \;=\; \underbrace{\big[h(x_i) - h(\bar{x}) - (x_i - \bar{x})\tau_1\big]}_{\approx\,\tau_2\,E(\mathrm{MRS})\,(x_i - \bar{x}) \;=\; O(\sigma)} \;-\; \tau_1(\lambda - \mu). \tag{61}$$

The first bracket is $O(\sigma)$ (from doctor heterogeneity) and the second is $O(\tau_1) = O(\sigma^2)$ (since τ_1^* will turn out to be $O(\sigma^2)$). Write $u'(c_i) \approx \bar{u}' + u''(\bar{c})(c_i - \bar{c})$.

Contribution of A_i. Since A_i does not depend on λ,

$$\tfrac{1}{2}\textstyle\sum_i E\,[u'(c_i)]\,A_i \;=\; \tfrac{1}{2}\sum_i \bar{u}'\,A_i \;+\; \tfrac{1}{2}\sum_i u''(\bar{c})\,E\,[c_i - \bar{c}]\,A_i.$$

For the *leading term*, A_i is proportional to $\mathrm{MRS}_i\,\delta\mathrm{MRS}_i$, so:

$$\begin{aligned}\tfrac{1}{2}\textstyle\sum_i \bar{u}'\,A_i \;&=\; \frac{-\bar{u}'\tau_2}{2\,h''(\bar{x})\,E(\mathrm{MRS})}\sum_i \mathrm{MRS}_i\;\delta\mathrm{MRS}_i \\ &=\; \frac{-\bar{u}'\tau_2}{2\,h''(\bar{x})\,E(\mathrm{MRS})}\cdot 2\sigma^2 \;=\; \frac{\bar{u}'\,\tau_2\,\sigma^2}{|h''|\;E(\mathrm{MRS})},\end{aligned} \tag{62}$$

where we used the identity (valid for any distribution with mean $E(\mathrm{MRS})$):

$$\tfrac{1}{2}\textstyle\sum_i \mathrm{MRS}_i\;\delta\mathrm{MRS}_i \;=\; \tfrac{1}{2}\sum_i\big(E(\mathrm{MRS}) + \delta\mathrm{MRS}_i\big)\delta\mathrm{MRS}_i \;=\; \sigma^2, \tag{63}$$

since $\sum_i \delta\text{MRS}_i = 0$ and $\frac{1}{2}\sum_i(\delta\text{MRS}_i)^2 = \sigma^2$. For the *correction term*, we have $E[c_i - \bar{c}] \approx \tau_2 E(\text{MRS})\,(x_i - \bar{x})$ (the λ-dependent part has mean zero). Since $x_i - \bar{x} = O(\sigma)$ and $A_i = O(\sigma)$, the product $E[c_i - \bar{c}] \cdot A_i = O(\sigma^3)$, and after multiplying by $u''(\bar{c})$, this correction contributes $O\big(R_A\,\tau_2^3\,E(\text{MRS})\,\sigma^2/h''^2\big)$ to the welfare derivative—which is $O(\sigma^2)$ times $R_A\tau_2^2\sigma^2/h''^2$, hence negligible under (53).[19]

Contribution of B_i. At leading order ($u' \approx \bar{u}'$), $\frac{1}{2}\sum_i \bar{u}'(x_i - \bar{x}) = 0$ because $\sum_i(x_i - \bar{x}) = 0$ by definition of $\bar{x}$. The next-order contribution is $-\frac{u''(\bar{c})}{2}\sum_i \tau_2 E(\text{MRS})(x_i - \bar{x})^2$. Since $(x_i - \bar{x})^2 = \tau_2^2(\delta\text{MRS}_i)^2/h''^2$, this equals $-u''(\bar{c})\tau_2^3 E(\text{MRS})\,\sigma^2/h''^2 = R_A\bar{u}'\tau_2^3 E(\text{MRS})\,\sigma^2/h''^2$, which is again negligible under (53).

Remark: cancellation at next order. For completeness, note that the corrections from A_i and B_i are both of order $R_A\bar{u}'\tau_2^3 E(\text{MRS})\,\sigma^2/|h''|^2$ and *carry the same sign*, so they reinforce rather than cancel. Under the small-heterogeneity condition (53), the combined correction is $o(\sigma^2)$ and drops out at leading order. Without this condition the formula for τ_1^* would acquire an additive correction proportional to $\tau_2^3 E(\text{MRS})\,\sigma^2/|h''|^2$.

Contribution of C_i. Here the randomness in λ is essential. We compute:

$$-\tfrac{1}{2}\sum_i E_\lambda\big[u'(c_i(\lambda))\,(\lambda - \mu)\big] \;=\; -\tfrac{1}{2}\sum_i \text{Cov}\big(u'(c_i), \lambda\big), \tag{64}$$

since $E[\lambda - \mu] = 0$. From (59), $\partial c_i/\partial\lambda = -\tau_1$ (holding x_i fixed, since doctor choice is ex ante). A first-order Taylor expansion of u' in λ gives:

$$u'(c_i(\lambda)) \;\approx\; u'(c_i(\mu)) + u''(\bar{c}) \cdot (-\tau_1)(\lambda - \mu), \tag{65}$$

where the approximation replaces $c_i(\mu)$ by $\bar{c}$ inside u''; this is valid since $c_i(\mu) - \bar{c} = O(\sigma)$ and the resulting error is $O(u'''\sigma\tau_1) = O(\sigma^3)$. Therefore:

$$\text{Cov}\big(u'(c_i), \lambda\big) \;\approx\; u''(\bar{c}) \cdot (-\tau_1) \cdot \text{Var}(\lambda).$$

Substituting into (64):

$$-\tfrac{1}{2} \cdot 2 \cdot u''(\bar{c})(-\tau_1)\text{Var}(\lambda) \;=\; u''(\bar{c})\,\tau_1\,\text{Var}(\lambda). \tag{66}$$

This is negative (welfare-reducing) since $u'' < 0$.

Step 3. First-order condition and solution.

[19] Explicitly, the correction from A_i equals $R_A\bar{u}'\tau_2^3 E(\text{MRS})\,\sigma^2/h''^2$.

Collecting (62) and (66), the leading-order welfare derivative is:

$$\frac{d\mathcal{W}}{d\tau_1} \approx \frac{\bar{u}'\,\tau_2\,\sigma^2}{|h''|\,E(\text{MRS})} + u''(\bar{c})\,\tau_1\,\text{Var}(\lambda).$$

Setting this to zero and writing $u''(\bar{c}) = -R_A\,\bar{u}'$:

$$\frac{\bar{u}'\,\tau_2\,\sigma^2}{|h''|\,E(\text{MRS})} = R_A\,\bar{u}'\,\tau_1^*\,\text{Var}(\lambda).$$

Dividing both sides by $R_A\,\bar{u}'\,\text{Var}(\lambda)$ yields the result:

$$\tau_1^* = \frac{\tau_2\,\text{Var}(\text{MRS})}{R_A\,|h''(\bar{x})|\,E(\text{MRS})\,\text{Var}(\lambda)}.$$ □

Interpretation. Equation (54) decomposes the optimal coinsurance rate into four economically interpretable components:

- Var(MRS): the benefit of coinsurance in screening doctor types. Greater heterogeneity in altruism (measured by the cross-sectional variance of $1/\beta_i$) raises the return to demand-side price signals.
- $|h''(\bar{x})|$: the curvature of the health production function. Steeper curvature implies that doctors' treatment choices are less responsive to price, reducing the sorting benefit of coinsurance and hence lowering τ_1^*.
- $R_A\,\text{Var}(\lambda)$: the insurance cost of coinsurance. Higher risk aversion or greater illness-severity dispersion makes cost-sharing more costly.
- $\tau_2/E(\text{MRS})$: the scale of the agency wedge. A larger supply-side copay τ_2 amplifies the treatment-intensity distortion across doctor types, increasing the screening benefit.

When altruism is homogeneous (Var(MRS) = 0), all doctors make identical treatment choices and there is nothing to screen: $\tau_1^* = 0$ and insurance is full (given that aggregate spending is held constant by τ_2).

I Risk Averse Firms

We show that under conditions (a)–(d) below, it is optimal to transfer expenditure risk to firms even when they are risk averse. These conditions need not all hold simultaneously; each provides a separate

sufficient argument for risk transfer. The optimal contract from Proposition 2 takes the form:

$$g(m) = A\exp\left(-\frac{\kappa}{\sigma}m\right), \qquad \kappa \equiv \left(1-\frac{1}{J}\right)\phi > 0,$$

where $A > 0$ is a scale parameter pinned down by the participation constraint; it shifts the firm's overall profit level without affecting the marginal incentive. We consider four conditions under which risk transfer to firms is optimal.

(a) Up-front payments. Suppose the firm can receive an upfront transfer X independent of patient volume or expenditure, and suppose the insurer must deliver total expected profit $\bar{\pi}$. The participation condition is:

$$X + \mathbb{E}[g(m)] = \bar{\pi}.$$

For any given σ, the insurer can choose A arbitrarily small and set:

$$X = \bar{\pi} - \mathbb{E}[g(m)].$$

Therefore $\mathbb{E}[g(m)]$ can be made arbitrarily close to zero, shifting compensation almost entirely into the fixed component X. The firm's exposure to expenditure risk through the patient-dependent component $g(m)$ becomes arbitrarily small.

(b) Sufficient market power (large σ). Now suppose front-loading is infeasible, so $X = 0$, and the participation condition becomes:

$$\mathbb{E}[g(m)] = \bar{\pi}.$$

Then A is pinned down by:

$$A = \frac{\bar{\pi}}{\mathbb{E}\left[\exp\left(-\frac{\kappa}{\sigma}m\right)\right]},$$

and therefore:

$$g(m) = \bar{\pi}\,\frac{\exp\left(-\frac{\kappa}{\sigma}m\right)}{\mathbb{E}\left[\exp\left(-\frac{\kappa}{\sigma}m\right)\right]}.$$

Assume $m \geq 0$ almost surely. Then:

$$0 \leq \exp\left(-\frac{\kappa}{\sigma}m\right) \leq 1, \qquad \exp\left(-\frac{\kappa}{\sigma}m\right) \to 1 \quad \text{pointwise as } \sigma \to \infty.$$

By dominated convergence:

$$\mathbb{E}\left[\exp\left(-\frac{\kappa}{\sigma}m\right)\right] \to 1,$$

so:

$$A \to \bar{\pi} \qquad \text{and} \qquad g(m) \to \bar{\pi} \quad \text{for every } m.$$

The profit schedule converges to a constant as $\sigma \to \infty$. Moreover:

$$\frac{\partial g(m)}{\partial m} = -\frac{\kappa}{\sigma} g(m) \to 0 \qquad \text{as } \sigma \to \infty.$$

When optimal incentives are weak (large σ), the payment schedule becomes nearly flat in m, so the risk borne by the firm vanishes even without front-loading.

(c) Law of large numbers. Suppose the firm treats N patients with i.i.d. severity draws $\lambda_1, \ldots, \lambda_N$. Total firm profit is:

$$\Pi_N = \sum_{n=1}^{N} g(m^*(\lambda_n)).$$

By the central limit theorem, $\Pi_N / N \approx \mathbb{E}[g] + N^{-1/2} Z$ for a mean-zero random variable Z, so $\mathrm{Var}(\Pi_N / N) = \mathrm{Var}(g)/N \to 0$. For any risk-averse firm, the Arrow–Pratt approximation gives a per-patient risk premium of order $O(1/N)$, which vanishes as $N \to \infty$. In the limit of a large firm facing many uncorrelated patients, the firm acts as if risk-neutral, and the risk-transfer argument for firm cost-sharing applies without qualification.

(d) Fee-for-service baseline. Suppose the baseline contract is fee-for-service, so the firm is paid more the more it does. In the paper's notation, $g(m) = R - \tau_2 m$ with $\tau_2^{\mathrm{FFS}} < 0$, so income increases with m. Since $m^*(\lambda)$ varies with patient severity, the firm bears income risk:

$$\mathrm{Var}\big(g_{\mathrm{FFS}}(m^*(\lambda))\big) = (\tau_2^{\mathrm{FFS}})^2 \,\mathrm{Var}(m^*(\lambda)).$$

Now suppose the insurer switches to a contract with $\tau_2 > \tau_2^{\mathrm{FFS}}$, adjusted to deliver the same mean profit. The firm's income variance under the new contract is:

$$\mathrm{Var}\big(g(m^*(\lambda))\big) = \tau_2^2 \,\mathrm{Var}(m^*(\lambda)).$$

For any $|\tau_2| < |\tau_2^{\mathrm{FFS}}|$, income variance falls relative to the FFS baseline. At $\tau_2 = 0$ (capitation), income variance is zero. The insurer can therefore reduce the firm's risk exposure by moving τ_2 toward zero—or introducing positive cost-sharing—without exceeding the risk already borne under FFS. This analysis treats $m^*(\lambda)$ as fixed; when the contract changes, the firm's optimal treatment also

changes, but this is a second-order effect that does not alter the qualitative conclusion.

J Behavioral Hazard

This appendix extends the baseline model to allow for *treatment-level* behavioral hazard: consumers systematically overvalue one treatment relative to another. This is the empirically relevant case for value-based insurance design (VBID), where the policy goal is to steer patients toward high-value treatments (Chernew et al., 2007; Baicker et al., 2015). We show that differential firm reimbursements can achieve the same steering as differential consumer copays, without exposing consumers to income risk.

Setup. Consider a single visit in which a provider chooses between two treatments, A and B, for a patient with illness severity λ. Treatment $t \in \{A, B\}$ costs the provider c_t and delivers true health benefit $h_t(\lambda)$ to the patient, with $h_A(\lambda) > h_B(\lambda)$ for all λ: treatment A is socially preferred. Due to behavioral hazard, however, patients perceive the health benefits as $\tilde{h}_A(\lambda)$ and $\tilde{h}_B(\lambda)$ with $\tilde{h}_B(\lambda) > \tilde{h}_A(\lambda)$—patients overvalue the inferior treatment, for example demanding a branded drug over a bioequivalent generic, or requesting antibiotics when none are indicated.

The insurer reimburses the provider g_t for delivering treatment t and charges the patient a treatment-specific copay τ_t. As in the main model, patients choose among J symmetric providers via logit demand with scale parameter σ.

Provider problem. Each provider j chooses a probability $\rho_j \in [0, 1]$ of delivering treatment A (and $1 - \rho_j$ for B). Conditional on ρ_j, patient i's perceived utility from provider j is:

$$\tilde{U}_{ij}(\lambda) = W - p - \rho_j \tau_A - (1 - \rho_j)\tau_B + \rho_j \tilde{h}_A(\lambda) + (1 - \rho_j)\tilde{h}_B(\lambda) + \varepsilon_{ij}.$$

The part of perceived utility that varies with ρ_j is:

$$\Delta_j \equiv \rho_j \left[\tilde{h}_A(\lambda) - \tilde{h}_B(\lambda) - (\tau_A - \tau_B)\right].$$

Denote $\tilde{\delta} \equiv \tilde{h}_A(\lambda) - \tilde{h}_B(\lambda) - (\tau_A - \tau_B)$, the perceived net benefit of treatment A over B inclusive of copays. The logit market share of provider j among visitors is:

$$s_j = \frac{\exp(\Delta_j/\sigma)}{\sum_{j'} \exp(\Delta_{j'}/\sigma)}.$$

Provider j's expected profit per visitor is $\rho_j(g_A - c_A) + (1-\rho_j)(g_B - c_B)$. The provider solves:

$$\max_{\rho_j \in [0,1]} s_j(\rho_j) \cdot \left[\rho_j(g_A - c_A) + (1-\rho_j)(g_B - c_B)\right].$$

Taking the first-order condition and evaluating at a symmetric equilibrium where $\rho_j = \rho$ and $s_j = 1/J$:

$$\frac{J-1}{J} \frac{\tilde{\delta}}{\sigma} \bar{\pi}(\rho) + \Delta g = 0, \tag{67}$$

where $\bar{\pi}(\rho) \equiv \rho(g_A - c_A) + (1-\rho)(g_B - c_B)$ is expected profit per patient and $\Delta g \equiv (g_A - c_A) - (g_B - c_B)$ is the profit differential. The first term captures the demand effect: since patients overvalue B ($\tilde{\delta} < 0$ when copays are equal), choosing A costs market share, and providers must be compensated by a positive profit differential $\Delta g > 0$.

Implementing the optimal treatment mix. For any target allocation ρ^*, the planner can use either instrument to satisfy (J1).

Consumer copays alone ($g_A - c_A = g_B - c_B$). With $\Delta g = 0$, equation (J1) requires $\tilde{\delta} = 0$, so:

$$\tau_B - \tau_A = \tilde{h}_B(\lambda) - \tilde{h}_A(\lambda).$$

The copay differential must exactly offset the perceived benefit gap. This is the standard VBID recommendation: lower copays for the high-value treatment.

Firm reimbursements alone ($\tau_A = \tau_B$). With equal copays, $\tilde{\delta} = \tilde{h}_A(\lambda) - \tilde{h}_B(\lambda) < 0$ is fixed. Solving (J1) for Δg:

$$\Delta g = -\frac{J-1}{J} \frac{\tilde{\delta}}{\sigma} \bar{\pi}(\rho^*). \tag{68}$$

Since $\tilde{\delta} < 0$, the right-hand side is positive: the reimbursement gap $\Delta g > 0$ compensates providers for the market share they sacrifice by steering patients to the less popular treatment. More generally, for any target $\rho^* \in (0,1)$, equations (J1) and the budget constraint define a one-parameter family of equivalent policies $(\tau_A, \tau_B, g_A, g_B)$ that all implement the same treatment allocation. The two instruments are substitutes.

Welfare ranking. The two equivalent-allocation policies deliver identical health outcomes and—holding the insurer's budget fixed—identical premiums. They differ only in risk exposure. Under differential copays, out-of-pocket spending varies with treatment received: patients who receive B pay $\tau_B > \tau_A$. Since treatment assignment may depend on factors patients cannot fully control ex

ante, differential copays expose risk-averse consumers to income risk. Under differential reimbursements with a uniform copay, out-of-pocket spending is independent of treatment. By the same risk-protection argument as in Proposition 2, the reimbursement policy is strictly preferred:

$$V^{\text{reimb}} > V^{\text{copay}}.$$