Insurance with Multiple Actors

Jason Abaluck
Oren Sarig
Jintaek Song*

April 2026

Abstract

We analyze how insurers should split cost-sharing between consumers and firms (e.g. patients and doctors, or drivers and car mechanics). With no contracting frictions, firm cost-sharing can replace consumer cost-sharing other than "visit copays." Optimal firm contracts depend on the marginal rate of substitution between consumer and firm cost-sharing. For Medicare physician services, paying physicians more up front but less as spending increases reduces risk for both patients and providers; net benefits are twice those from eliminating Medigap externality. For prescription drugs, a dollar of physician paperwork costs is 33–120 times more impactful than a dollar of consumer cost-sharing.

*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 physician services, 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 if care were free but they internalized premium impacts.

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.

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 physician services and prescription drugs. For Medicare physician services, we calibrate our model using existing estimates of utilization effects from Cabral and Mahoney (2019) as well as the impact of physician reimbursement on expenditures from Clemens and Gottlieb (2014). The first step is to identify physician profit margins in the status quo given their (unobserved) effort costs; these are inferred from physician behavioral response to changes in revenue. In the status quo, we estimate that 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. Shifting to a schedule where physicians are paid more up front but make lower profits the more they spend and then rebating premiums from reduced moral hazard would increase welfare by more than twice as much as eliminating the Medigap externality. 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 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.

[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.

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: a constant marginal rate of substitution between firm and consumer cost-sharing. 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 optimal cost-sharing under both transferability and non-transferability, Section 4 calibrates the optimal physician contract for Medicare physician services, Section 5 estimates optimal cost-sharing for prescription drugs in Medicare Part D, and Section 6 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 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 draw firm-specific 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:

$$u_{ij} = V_j + \varepsilon_{ij} \tag{1}$$

where:

$$V_j \equiv h(m_j(\lambda), \lambda) + v\big(W - \alpha_j(\lambda) - \tau_1(m_j(\lambda)) - p\big)$$

is the deterministic component of utility from firm j. Below, we suppress the i subscripts in most cases to ease notation. Here $h(m, \lambda)$ is the health benefit from services m given illness severity λ, $v(\cdot)$ is a strictly increasing, strictly concave utility defined over financial consumption normalized so that $v'(\bar{c}) = 1$ at some reference level of consumption, W is wealth, $\alpha_j(\lambda)$ is the side payment required of firm j for consumers of type λ (this can be positive or negative), τ_1 is consumer cost-sharing (a potentially nonlinear function of service costs), and p are premiums (explained below as part of the firm problem). We define financial consumption as $c \equiv W - \alpha - \tau_1(m) - p$. The health benefit function $h(m, \lambda)$ is strictly concave in m with $h_{m\lambda} > 0$: the marginal benefit of care is increasing in illness severity, so that optimal care is increasing in λ (a conventional single-crossing condition).[3] 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 taste shocks $\varepsilon_i \equiv (\varepsilon_{i1}, \ldots, \varepsilon_{iJ})$ are drawn from a smooth exchangeable distribution and represent the consumer's idiosyncratic preference for firms (e.g. convenience or taste for a given doctor).

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 $U^0(\lambda) = h(0, \lambda) + v(W - p)$. If they visit a firm, they receive expected utility given by:

$$\begin{aligned} U^{visit}(\lambda) &= E_\varepsilon\left(\max_j h(m_j, \lambda) + v(W - \alpha_j - \tau_1(m_j) - p) + \varepsilon_j\right) - K_{visit} \\ &\equiv \mathcal{S}(V_1, \ldots, V_J) - K_{visit}, \end{aligned} \tag{2}$$

where $\mathcal{S}$ is the expected-surplus function induced by the exchangeable error distribution, and K_{visit} includes transportation and hassle costs of engaging with firms. If they decide to patronize a firm, the ε_j shocks are realized and consumers choose the firm that maximizes equation 1.

[3] We also assume h is twice continuously differentiable.

We summarize the local competitiveness of this demand system by:

$$\kappa \equiv \left. \frac{s_j}{\partial s_j / \partial V_j} \right|_{V_1 = \cdots = V_J},$$

where s_j are the choice probabilities for firm j among consumers who decide to visit. In the symmetric case, κ is a constant. Absent income effects, price elasticities (a perhaps more familiar measure of competitiveness) are proportional to κ. Under i.i.d. Type-I extreme-value shocks, $\mathcal{S}(V) = \sigma \log \sum_{j=1}^{J} \exp(V_j/\sigma)$ and $\kappa = \sigma/(1 - 1/J)$, where σ is the logit scale parameter.

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

$$U(\lambda) = \max\{U^{visit}(\lambda), U^0(\lambda)\} \tag{3}$$

Finally, in Stage 0 (where no active decision is made by consumers), ex ante utility is given by:

$$E_\lambda \left[U(\lambda) \right] \tag{4}$$

where the expectation is taken over $F(\lambda)$, the distribution of the λ types. Because $v(\cdot)$ is concave, consumers are risk-averse over financial consumption and 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 ε_{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:

$$\begin{aligned} m_j(\lambda), \alpha_j(\lambda) &= \text{Argmax}_{m,\alpha}\, \pi(\lambda) = \text{Argmax}_{m,\alpha}\, s_j(m, \alpha, \lambda) \cdot (R - \tau_2(m) + \alpha) \\ &\text{s.t.} \quad m \geq k \end{aligned} \tag{5}$$

where R is the fixed component of firm reimbursement,[4] τ_2 are firm cost-sharing rates and $s_j(m, \alpha, \lambda)$

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

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 the form of visit copays.

In our baseline model, we assume that firms maximize profits. An alternative specification, especially common in healthcare, allows providers to be altruistic and choose service level to maximize a weighted average of profits (given by equation 5) 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:

$$E_\lambda \left[U(\lambda, \{\alpha_j^*(\lambda, \tau_1, \tau_2), m_j^*(\lambda, \tau_1, \tau_2)\}, p(\tau_1, \tau_2), \tau_1, \tau_2) \right] \tag{6}$$

where α_j^* and m_j^* are the optimal side payments and service levels which solve equation 5, p are premiums (defined below), and consumers solve equation 3 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.[5] Since firm profits are exogenously fixed, insurers that maximize

[5]This assumption is not impactful for our empirical exercises below, but it would be important to relax when studying

consumer surplus are also maximizing total surplus.

We further define equilibrium 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:

$$p(\tau_1, \tau_2) = E(m_i^*) - E(\tau_1(m_i^*)) + J\bar{\pi} - E(\alpha_i^*) \tag{7}$$

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).

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 ($\alpha = 0$) and no consumer cost-sharing. 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. Suppose the average consumer in this market requires $\$1,500$ worth of repairs in equilibrium. This cost-sharing tax averages \$30 per consumer ($\$1,500 \cdot 0.02$), so the mechanic nets $\$230 - 0.02m$ for a type m consumer, which averages to \$200. If a type λ_h consumer requires engine repairs which (in equilibrium) cost $m = \$2,000$, the mechanic will net only \$190 (230-0.02 $\cdot$ 2000) for that consumer (they would receive gross revenue of $\$2,190$ from the insurer, \$2,000 of which reimbursed for repair costs). Due to the incentive contract, higher type consumers are less profitable on average, counteracting moral hazard.

Our baseline model assumes that once consumers visit firms, they have no further bargaining ability; in fact, there is an equivalence between our model and bargaining models. In Appendix S1, we show that in a special case of our model, the equilibrium service-level condition can equivalently be written as the first-order condition of a Nash bargaining problem in which consumers are randomly

the empirical relationship between competition and contracts, since competition would impact contracts both via $\bar{p}i$ and via the modeled impact of σ.

assigned a firm and bargain against the outside option of switching to an alternative (symmetric) firm. The firm's bargaining weight is an increasing function of the demand-system market-power summary κ.[6]

3 Optimal Cost-Sharing

We first show 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. We then argue that insurers will often prefer to enforce non-transferability and that the solution in that case is decidedly not neutral, generally preferring firm over consumer cost-sharing. Finally, we extend our model to consider good and bad rationales for consumer cost-sharing in the non-transferable case.

3.1 Transferability Implies Neutrality

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.

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 3, choose among firms according to Equation 1, and firms choose service levels and side payments according to Equation 5. 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.

[6] 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, which the isomorphic representation in our model would interpret as arising from market power. 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 in which nonlinear contracts might require unrealistically steep penalties.

We prove this result in Appendix S2. 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 S3.[7] 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.

3.3 Theory with Non-Transferability

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

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. Firm cost-sharing implements this allocation with zero intensive-margin out-of-pocket spending. In this setting, optimal firm cost-sharing among visitors leads to the same expenditures a social planner would have chosen with full information on

[7] 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.

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.[8] We formally characterize the optimal contract and its equivalence to a "constrained first best" in Proposition 2. A direct corollary is that, if it is optimal for everyone to visit (for example, if $\lambda > k$ for all individuals), the constrained first best coincides with the full-information first best.

Proposition 2: Define the constrained first-best service levels $m^{cfb}(\lambda)$ as the service levels the social planner would choose if they could directly choose service levels and a visit copay (τ_v), but visit decisions are made endogenously according to equation 3. Let τ_v denote the optimal visit copay (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 5). Suppose firm demand is generated by a symmetric smooth random-utility system, summarized at the symmetric allocation by κ as defined in Section 2. When $\bar{\pi} > 0$, then $m^{cfb}(\lambda)$ is achievable for a given level of profits with coinsurances given by: $(\tau_v, \tau_1, \tau_2) = (\tau_v, 0, \tau_2^*(m))$. τ_v and $\tau_2^*(m)$ are characterized by the following equations:

$$[\tau_v]: \; P(\text{visit}) \cdot \frac{v'(c^{\text{vis}}) - \mu}{\mu} = -\left(m^{cfb}(\lambda^*) - \tau_v\right) \frac{\partial P(\text{visit})}{\partial \tau_v},$$

$$[\tau_2(m)]: \; g(m) = A \cdot \exp\left(-\frac{\mu}{\kappa} m\right),$$

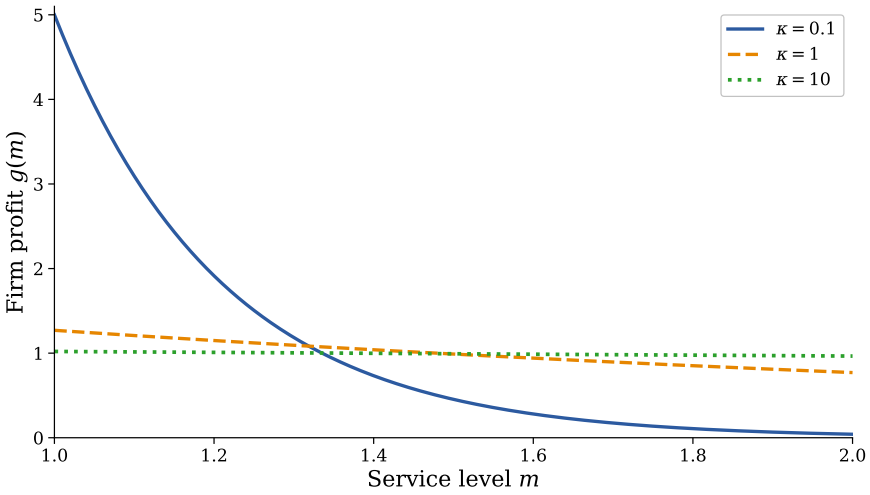
where λ^* is the marginal visiting type (indifferent between visiting and staying home) and $m^{cfb}(\lambda^*) - \tau_v$ is the net insurer cost of that marginal visitor and where $\mu \equiv \mu(\tau_v)$ is the common shadow value of a premium dollar at the optimum,

$$\mu = P(\text{visit})\, v'(W - p - \tau_v) + (1 - P(\text{visit}))\, v'(W - p),$$

and $c^{\text{vis}} = W - p - \tau_v$ is the financial consumption of visitors. The constrained-first-best service rule

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

Figure 1: Optimal firm profit schedule $g(m)$ by market competitiveness κ



Notes: Optimal contract from Proposition 2 for three values of the demand parameter κ governing local firm market power. Each curve is normalized so that average profits equal 1. Higher κ (more market power, or weaker demand sensitivity) yields a flatter, more linear schedule; lower κ (more competition) yields a steeper, convex schedule requiring stronger incentives at low service levels. In the logit case, $\kappa = \sigma/(1-1/J)$.

for interior visitors is:

$$h_m(m^{cfb}(\lambda), \lambda) = \mu \qquad \text{for all interior visitors.}$$

In general, $m^{cfb}(\lambda)$ is defined implicitly by $h_m(m, \lambda) = \mu$ for each λ; its shape depends on the chosen functional form for h. $U(\lambda)$ is defined as in equation 3 at the policy regime $(\tau_v, 0, \tau_2^*)$. μ is the common shadow value of a premium dollar.

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 (the gap between the marginal utility of visitor consumption and the shadow value of a premium dollar), and the right-hand side gives the moral hazard benefits.

$g(m)$ 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 contract $g(m)$ looks very different depending on the degree of competition, summarized by κ. Figure 1 graphs this contract for high and low κ. When κ is small and firms are 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 are necessary to offset this. When κ 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 constrained 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.[9] We next discuss these considerations.

3.4 Reasons for Consumer Cost-Sharing

Several extensions to our model restore some role for intensive margin consumer cost-sharing. We will discuss at greater length the case of linear contracts because of its relevance to our applications and briefly mention the others. We summarize these in Table 1 and discuss them at greater length in Appendix S4.

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, highly nonlinear contracts with more than 1-1 penalties for spending as our model calls for in the competitive case may be unrealistic in practice.

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.

In general, linear contracts cannot achieve constrained first-best service levels without some consumer cost-sharing. A given amount of firm cost-sharing will disproportionately impact service levels

[9]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).

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

Setting	**Key mechanism**	**Role for consumer cost-sharing**
Valid rationales (contracting frictions)		
Firm 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 when up-front payments infeasible, market power low, and pooling insufficient
Insufficient rationales		
Loading	Administrative costs proportional to service levels require steeper firm penalties	None
Altruism	Altruistic firms need steeper incentives but still respond to firm cost-sharing	None
Behavioral hazard	Consumers over- or under-use services; differential firm cost-sharing can correct this	None
Non-adherence	Consumers fail to follow prescribed treatment	Only can justify *negative* consumer cost-sharing

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 on them and would prefer expanding margins to attracting more consumers.[10] 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 B).

The behavior of the model in the perfectly competitive limit is also notable. In the case of exactly perfect competition $\kappa = 0$, 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 $\kappa \to 0$ does not generally converge to the case of perfect competition. When $\kappa = \epsilon > 0$ firm cost-sharing is effective, but disproportionately so for the highest expenditure types.[11] In the logit case, the

[10] They earn low margins on high spending types precisely because firm cost-sharing implies $\tau_2 > 0$, deliberately reducing margins for high spending types.

[11] 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

solution as $\kappa \to 0$ is exactly the same as if insurers could impose a uniform cap on expenditures (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 this in appendix S5 for the logit error case). 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—the practice of paying firms a fixed amount per case regardless of services rendered—corresponds to the 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. The monopolistic limit ($\kappa \to \infty$) implies that a linear contract achieves first-best service levels at a much smaller τ_2, making prospective payment unnecessarily aggressive. Conversely, in the competitive limit ($\kappa \to 0$), the optimality of prospective payment depends on target profits. If target profits $\bar{\pi}$ 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 ($\tau_2 \approx 0$). In Section 4 we estimate that the status-quo $\hat{\tau}_2 \approx -0.016$ (a small marginal subsidy on episode spending) and find that the optimal contract substantially improves welfare relative to either extreme: prospective payment is far from optimal, but pure fee-for-service also goes too far in the opposite direction.

In Appendix S4, we consider several other rationales for consumer cost-sharing. When firms are heterogeneous in their productivity, this creates a new source of extensive margin moral hazard in the choice of firms, but this should be addressed with firm-specific copays rather than intensive mar-

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 in equilibrium (i.e. if $R - \tau_2 m(\lambda) = 0$ and $m(\lambda') < m(\lambda_{\text{cap}})$, then markups will be strictly positive for λ').

gin coinsurances (Appendix S6). Asymmetric information about firms can cause firm cost-sharing to disproportionately impact patients of firms who respond more, creating a role for consumer cost-sharing to create stability. We develop a model of asymmetric information about firm altruism in Appendix S7, where the optimal coinsurance rate depends to first-order on the variance in the marginal rate of substitution between consumer and firm cost-sharing. Firm risk aversion can create a rationale for consumer cost-sharing although it can be mitigated if: 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.

Appendix S4 also discusses some tempting rationales for consumer cost-sharing including loading (from the Arrow (1963) case for consumer deductibles), altruism and behavioral hazard and shows that these can in general be dealt with using appropriate firm contracts.

4 Application #1: Medicare Physician Services

This section evaluates the feasibility of replacing Medicare's current patient coinsurance with the optimal physician fee schedule characterized in Section 3. 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 five objects: the distribution of patient types $F(\lambda)$, the physician market-power parameter κ from the theory (which corresponds to the logit scale parameter σ under the parametrization below), the minimum effort level k, the health-benefit curvature (which we parametrize with ω), and the current implicit penalty (or reward) for physician spending τ_2. We recover $F(\lambda)$ by inverting the physician first-order condition from observed episode revenue for patients with interior spending, and using existing estimates of how deductibles change visit decisions for patients who visit at the minimum spending level. We calibrate σ from quasi-experimental estimates of the impact of copays on PCP choice. We consider a range of assumptions about physician marginal costs for minimal visits (k) relative to the observed revenue for minimal visits in the data.

We identify (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 and visits to the elimination of patient coinsurance (Cabral and Mahoney, 2019; Brot-Goldberg et al., 2017) 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.

4.1 Data and Sample Construction

MarketScan episode data. 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.[12] We use data for 2014–2018 and restrict to Level 1 (CPT-coded) outpatient claims, the physician-fee-schedule services where physicians have a meaningful intensity choice. This excludes Level 2/3 HCPCS claims for ambulance, transportation, medical supplies, and durable medical equipment, which have zero physician-work RVUs and do not correspond to the model's continuous-intensity decision. 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, with and without offsetting physician incentives.

Revenue as the data primitive. 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.[13] We assume a linear physician revenue schedule $\text{Rev}(m) = R + (1-\tau_2)\, m$ that approximates Medicare's resource-based fee schedule (although we consider optimal nonlinear counterfactual schedules below). Dollar-equivalent physician effort m is a latent variable, and all moment conditions are expressed entirely in terms of the observed revenue distribution by inverting the revenue equation to substitute out for m. We model coinsurances τ_1 as a fraction of observed physician revenue; this is just a relabeling of patient and firm coinsurances relative to the model, where τ_1 is a fraction of m.[14]

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

[13] Rev_i equals the sum of COB (Medicare FFS payment), NETPAY (employer supplemental), DEDUCT, COPAY, and COINS in MarketScan.

[14] Specifically, $\tau_1 \text{Rev}(m) = \tau_1 R + \tau_1(1-\tau_2)m$. The term $\tau_1 R$ is a fixed lump sum per visit that does not affect the physician's intensive-margin first-order condition (though it does effect profits); the effective copay rate on m is exactly $\tau_1(1-\tau_2)$. A model that applied τ_1 directly to m would thus be equivalent to the revenue formulation with the substitution $\tau_1 \to \tau_1/(1-\tau_2)$.

Episode construction. 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.

Table 2: Observed revenue distribution from MarketScan episode data

Decile (above-kink)	Mean revenue Rev ($)
1	77
2	105
3	121
4	153
5	202
6	272
7	397
8	602
9	1,080
10	4,405
Kink revenue Rev_k	73
Episodes at kink	16.9%
Mean episode revenue (all)	637
Mean episode revenue (above kink)	741

Notes: Revenue is the total allowed amount per episode (Medicare FFS payment plus supplemental payer contributions), observed directly from MarketScan 2014–2018. Decile means are computed over episodes with revenue above the kink Rev_k = \$73, the modal episode revenue in the data (the point of mass bunching); 16.9% 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.

4.2 Structural Model and Identifying Equations

We adopt the health benefit function,

$$h(m, \lambda) = \left(1 - \frac{1}{\omega}\right)(m - \lambda) + \frac{\lambda + 1}{\omega} \ln \frac{m + 1}{\lambda + 1}, \tag{8}$$

which satisfies the conditions assumed in the theory (h strictly concave in m, $h_{m\lambda} > 0$). We assume CARA preferences over financial consumption, $v(c) = -\frac{1}{\gamma}\exp(-\gamma c)$, calibrating reference consumption $\bar{c}$ using existing consumption observed in the MarketScan data (at this $\bar{c}$, $v'(\bar{c}) = 1$). Ignoring income effects from premium discounts, we can interpret λ as approximately the first best service level that would be chosen by a social planner who internalized premium impacts of m (since $h_m(\lambda, \lambda) = 1$).[15] With this h function, marginal benefits are given by: $h_m(m, \lambda) = (1 - 1/\omega) + (\lambda + 1)/(\omega(m + 1))$,

[15] Approximately because, With this choice of $\bar{c}$, λ formally only corresponds to the first-best at status quo consumption levels due to income effects.

so the patient's bliss point (where $h_m = 0$) which patients would choose with zero cost-sharing is $m^{\text{bliss}} = (\lambda+1)/(1-\omega) - 1$, strictly above the first-best level λ.[16]

These assumptions, combined with the physician's first-order condition (equation 44), yield a closed-form mapping from observed revenue to patient type. In Appendix S8, we show how these assumptions, combined with a distributional assumption about types bunching at the kink (Assumption 1), suffice to identify the full distribution of λ_i from the observed distribution of revenue.

4.3 External Calibration

We assume that the ε error terms are i.i.d. logit and the number of physicians per market are large, which implies that $\kappa = \sigma$, the episode-level logit scale parameter. We calibrate this using evidence on the responsiveness of physician choice to patient cost-sharing (Sinaiko and Rosenthal, 2014). Details of the calibration are given in Appendix S9.

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. In the MarketScan data, the kink point is identified as the modal episode revenue among low expenditure visits (the value at which mass bunching occurs), which equals $\text{Rev}_k = \$73$.[17] The minimum effort level k itself is not directly observed; we calibrate it as a fraction of Rev_k. Our baseline assumes $k = 0.60 \cdot \text{Rev}_k = \$43.80$, reflecting physician practice-expense overhead of roughly 40% of visit revenue (consistent with AMA and MGMA benchmark data for office-based practice), and we consider values for k ranging from 50% to 90% of observed revenue (Appendix Table 1).

4.4 Identification of (ω, τ_2)

Given the externally calibrated (σ, k), the remaining structural parameters (ω, τ_2) are identified by matching two empirical moments derived from the physician's FOC (equation 45); the distribution of types for non-visitors is identified by matching an external moment of how visit probabilities respond to copays.

[16] The "+1" here is a regularization which ensures finite utility at $h(0, \lambda)$, so that the visit decision is well-defined. In practice, we estimate that $\lambda > 20$ for all visiting enrollees so this regularization has little impact for visitors. Likewise, net utility at the extensive margin is also determined by the parameter K_{visit}, so the functional form $h(0, \lambda)$ only determines the relative utility of visiting for different λ types, not the average utility of visiting relative to not visiting.

[17] The mode is computed as the most common value in the range \$0–\$150 of the episode revenue distribution. This value corresponds closely to the Medicare Physician Fee Schedule payment for CPT 99213 (a standard established-patient office visit), which is the most commonly billed evaluation and management code and had a national average payment of approximately \$73 during 2014–2018 (Centers for Medicare & Medicaid Services, 2024). This is consistent with the interpretation that the kink represents the minimum-effort visit type. Sensitivity to alternative kink values $\text{Rev}_k \in \{\$50, \$100, \$150, \$200\}$ is reported in Appendix Table 1.

Moment 1: Revenue change from Medigap (identifies ω). Cabral and Mahoney (2019) 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%, implying that its removal would reduce revenue by $1/1.172 - 1 \approx 14.7\%$. Our MarketScan sample consists of beneficiaries with employer-sponsored supplemental insurance, so the status quo is $\tau_1 = 0$ and the counterfactual is removing supplemental coverage ($\tau_1 = 0.20$). We assume based on Brot-Goldberg et al. (2017) that this copay change would induce 14.5% of beneficiaries to exit entirely, meaning that revenue would decline by 12.6% among non-exiting beneficiaries, among whom we calibrate the model.[18]

$$\frac{\text{Rev}^{\text{cf}} - \text{Rev}^{\text{data}}}{\text{Rev}^{\text{data}}} \approx \left(\frac{1}{1.172} - 1\right)(1 - 0.145) \approx -0.126, \tag{9}$$

where Rev^{data} is the observed mean Medigap revenue, and the counterfactual averages Rev^{cf} are computed by re-solving the physician FOC at $\tau_1 = 0.20$ for each patient type recovered from equation (46).

Moment 2: Revenue elasticity with respect to reimbursement (identifies τ_2). Let A denote a scalar that multiplies the entire physician revenue schedule: $\text{Rev}_A(m) \equiv A \cdot \text{Rev}(m)$. A change in A corresponds to an across-the-board shift in Medicare's conversion factor, multiplying total RVUs and thus raising or lowering reimbursement for every service proportionally while leaving relative prices across services unchanged. Profits are given by $A \cdot \text{Rev}(m) - m$ so increasing A increases the marginal return to providers of increasing m without increasing costs.

Clemens and Gottlieb (2014) exploit precisely this kind of variation: geographic differences in payment rates generated by the 1997 fee-schedule consolidation shift A differentially across areas, and they estimate a long-run (2001–2005) elasticity of total physician revenue with respect to the conversion factor of 1.45. The model analog is:

$$\left.\frac{d \log \text{Rev}(A)}{dA}\right|_{A=1} = 1.45, \tag{10}$$

where $\text{Rev}(A)$ is mean revenue when physicians re-optimize their service choices at the scaled schedule $A \cdot \text{Rev}(m)$, but evaluated at $A = 1$. Following Clemens and Gottlieb (2014), evaluating at $A = 1$

[18] Brot-Goldberg et al. (2017) study a large firm that switched all employees from a completely free plan (zero deductible, zero coinsurance) to a high-deductible plan with an effective consumer price of approximately 22% of total spending, analogous to our Medigap counterfactual of raising τ_1 from 0 to 0.20. They find that office-visit spending fell by 13–18% (Table 5), with an anticipation-adjusted estimate of 13–16%. The spending reductions are "entirely due to outright reductions in quantity" rather than price shopping (Table 6), so the spending decline approximates the quantity decline which is the operative force in Medicare. We use 14.5%, the midpoint of their range, for our calibration. One caveat is that their sample is a non-elderly employed population rather than Medicare beneficiaries.

shuts down any mechanical impact of changing A on revenue and isolates only the behavioral impact through changes in service levels m. Since Clemens and Gottlieb (2014) estimate this elasticity on non-Medigap Medicare patients, we evaluate $\text{Rev}(A)$ at $\tau_1 = 0.20$ for each type recovered from equation (46), excluding types who exit the market at $\tau_1 = 0.20$.

Identification of ω and τ_2. Ignoring for the moment the extensive margin, and ignoring terms of $O(\tau_2^2)$,[19] we show in Appendix S10 that Moment 1 can be rewritten as:

$$\frac{\text{Rev}^{\text{cf}} - \text{Rev}^{\text{data}}}{\text{Rev}^{\text{data}}} \approx -0.20\,(1-\tau_2)\,\frac{\omega}{1-\omega}\left(1 - \frac{R}{\text{Rev}^{\text{data}}}\right), \tag{11}$$

Similarly ignoring terms of $O(\tau_2^2)$, Moment 2 can be approximately rewritten as:

$$\left.\frac{d\log \text{Rev}(A)}{dA}\right|_{A=1} \approx \frac{\omega}{1-\omega} \cdot \frac{\sigma R}{(R-\tau_2 \bar{m})^2}\left(1 - \frac{R}{\text{Rev}^{\text{cf}}}\right), \tag{12}$$

The factor $\omega/(1-\omega)$ is common to both equations and determines the magnitude of the revenue responses. τ_2 determines the relative magnitude of the responses. When margins are fat ($R - \tau_2 \bar{m}$ large, so τ_2 is 0 or negative), then consumer coinsurances will be relatively more impactful than the revenue shifter. Alternatively, when margins are thin (e.g., because $\tau_2 > 0$), the revenue shifter will be relatively more impactful. The relative magnitude of the two moments above identifies margins, and thus τ_2.

Moment 3: Extensive-margin visit response (identifies f and λ_2). Given the calibrated $(\hat{\omega}, \hat{\tau}_2)$ and the recovered above-kink type distribution, we identify the visit margin using Brot-Goldberg et al. (2017), whose estimates imply that eliminating Medigap supplemental coverage (raising τ_1 from 0 to 0.20) causes approximately 14.5% of office-visit episodes to disappear on the extensive margin.

Under the assumed h, the visit surplus is strictly increasing in λ, so the 14.5% of visitors who exit are the lowest-λ types, all drawn from the bunched region (which comprises 16.9% of visitors). Combined with the uniform density assumption (Assumption 1), this pins down the bunching-region density f, the lower visit cutoff λ_2, and the nonfinancial visit cost K_{visit}; the detailed derivation is in Appendix S8.

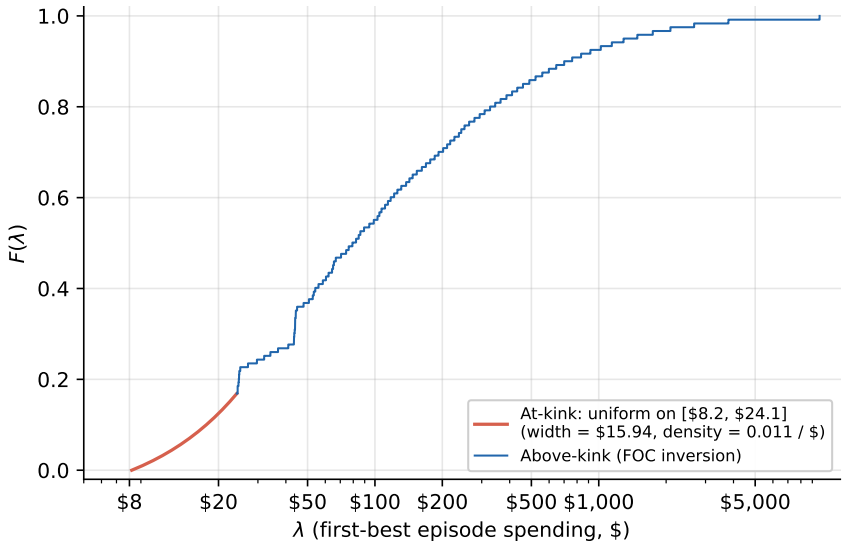
[19] We estimate the model using the exact equations.

4.5 Estimation Results

We solve the two-equation system (9)–(10) for (ω, τ_2) via a nonlinear equation solver with multiple starting values, and then recover f, λ_2, and K_{visit} from Moment 3. At the baseline specification ($\text{Rev}_k = \$73$, $k/\text{Rev}_k = 0.60$), $\hat{\tau}_2 = -0.016$ and $\hat{\omega} \approx 0.385$. Appendix Table 1 reports structural estimates across specifications; across the 25 alternative specifications, $\hat{\tau}_2$ ranges from -0.035 to $+0.001$ while $\hat{\omega}$ ranges from 0.349 to 0.390.

Distribution of $\hat{F}(\lambda)$. Figure 2 shows the recovered distribution of patient types λ at the baseline calibration. The above-kink portion (blue) is identified non-parametrically, while the at-kink portion (red) is identified by a parametric assumption on bunched types (Appendix S8). The distribution is heavily right-skewed: the median visiting type has first-best spending around \$83, while the top decile exceeds \$700 and the top centile exceeds \$3,500.

Figure 2: Recovered distribution of patient types $\hat{F}(\lambda)$



Notes: Baseline calibration: $\text{Rev}_k = \$73$, $k/\text{Rev}_k = 0.60$, $\hat{\omega} = 0.385$, $\hat{\tau}_2 = -0.016$. Blue step function: above-kink types identified via FOC inversion (equation 46), each centile carrying equal mass. Red segment: at-kink (bunched) types, assumed uniform on $[\lambda_2, \lambda_1] \approx [\$8.2, \$24.2]$ under Assumption 1. Horizontal axis on log scale. The excess mass around $\lambda \approx \$45$ reflects the second most common evaluation & management code (after our kink), for moderate-complexity office visits (revenue between \$108 and \$110).

Interpretation of $\hat{\omega}$. The estimated $\hat{\omega} = 0.385$ implies that physicians spend approximately 63% above the first-best level when patients have zero cost-sharing, or equivalently that a 10 percentage point increase in patient coinsurance reduces intensive-margin spending by approximately 6.3% (ra-

tionalizing the 12.6% reduction in Cabral and Mahoney (2019) from a 20% coinsurance).[20] Appendix Table 1 shows that $\hat{\omega}$ is remarkably stable across specifications, varying by less than five percentage points across the full $\text{Rev}_k \times k/\text{Rev}_k$ grid.

Interpretation of $\hat{\tau}_2$. The negative estimate $\hat{\tau}_2 = -0.016$ means that physicians currently face a marginal *subsidy* on episode spending: each additional dollar of episode expenditure raises per-patient profit by approximately 1.6 cents. In practice, the Clemens and Gottlieb elasticity is modest given ω and σ, meaning that margins must be relatively fat to rationalize this modest response, so τ_2 must be negative.

The model interprets the insensitivity of revenue to the revenue shifter as evidence of fat margins (induced by $\tau_2 < 0$). A prima facie alternative explanation is that physicians are altruistic, and so do not respond to revenue incentives despite having thin margins. We show in Appendix S11 that altruism also implies that physicians should be very responsive to consumer copays; in fact, altruism requires an even *more negative* τ_2 to rationalize the observed moments because it strengthens the response to copays more than revenues, so the copay to revenue ratio can only be restored by even fatter margins.

A negative τ_2 implies that the status-quo fee schedule rewards physicians for delivering more intensive episodes. Medicare's physician fee schedule 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 increase the revenue they receive for a given amount of effort, leading to a marginal subsidy (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.

[20] Taking the implicit derivative of $m(\lambda;\tau_1) \approx (\lambda+1)/(1-\omega(1-\tau_1)) - 1$ gives $d\ln m/d\tau_1 \approx -\omega/(1-\omega)$ at $\tau_2 \approx 0$. With $\hat{\omega} = 0.385$, a 10 percentage point increase in τ_1 therefore reduces log spending by $0.10 \times 0.385/0.615 \approx 0.063$, i.e. a 6.3% reduction; the corresponding overspending ratio under full insurance is $1/(1-\omega) \approx 1.63$.

Optimal Physician Contract Given the calibrated structural parameters $(\sigma, k, \hat{\omega}, \hat{\tau}_2)$ and the distribution of patient types, the optimal physician profit schedule from Proposition 2 is:

$$g(m) = C \cdot \exp\left(-\frac{\mu}{\sigma} m\right), \tag{13}$$

where μ is the shadow value of a premium dollar from Proposition 2, and the constant C is chosen so that the expected per-patient payment to physicians equals the per-patient rent R.[21] Computational details (the closed form for C, the shadow-value μ under CARA preferences, and the cutoff $\times$ ratio search) are deferred to Appendix S12.

The resulting profit schedule $g(m) = C\exp(-\mu m/\sigma)$ declines with episode expenditure and is structurally analogous to a physician-side deductible, although the implicit tax is less steep than we would expect under a deductible or prospective payment. Table 3 illustrates this. Under the status quo, per-episode profit is \$29 at the kink ($m = k =$ \$43.80), \$61 at $m =$ \$2,000, and \$125 for a 99th-percentile patient ($m \approx$ \$6,000), rising with m because $\hat{\tau}_2 < 0$. Under the optimal contract, the same episodes yield profits of \$66 at the kink, falling to just \$2 by $m =$ \$1,000: an average tax of \$0.07 per dollar of spending over that range. The economic logic, however, differs fundamentally from consumer deductibles. 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 profit decline is necessary because low-cost episodes are profitable to attract under the optimal schedule.

Table 4 reports how visit copays and welfare benefits vary with the CARA risk aversion parameters considered in Handel et al. (2015). Under this contract, the optimal visit copay is positive and decreasing in risk aversion, as we would expect from the standard Chetty-Baily trade-off between risk protection and moral hazard. Table 4 also compares the optimal policy (visit copay, zero coinsurances, optimal $g(m)$) to the policy of eliminating Medigap (0 visit copay, 20% coinsurances, status quo τ_2). Eliminating Medigap generates benefits relative to the status quo due to the externality mechanism in Cabral and Mahoney (2019).

The baseline model assumes physician risk-neutrality. In Appendix S13, we show that optimal contract *reduces* physician risk exposure relative to the baseline contract. Due to pooling across many patients, the overall impacts on physician risk are only a tiny fraction of impacts on patients and insurers.

[21] We have also assumed that the number of physicians in the choice set of each patient is large, so $1 - 1/J \approx 1$

Table 3: Status-Quo vs. Optimal Physician Profit Schedule

	Status Quo ($\hat{\tau}_2 = -0.016$)		Optimal contract $g(m)$		Episode share (%)	
Episode spending (m)	Revenue ($)	Profit ($)	Revenue ($)	Profit ($)	SQ	Opt
\$43.80 ($m = k$, kink)	73	29	110	66	40.1	47.8
\$200	232	32	237	37	27.4	26.6
\$400	435	35	418	18	10.0	8.5
\$600	638	38	609	9	7.5	6.6
\$1,000	1,045	45	1,002	2	6.7	4.7
\$2,000	2,061	61	2,000	0	5.0	3.8
\$5,925 (99th percentile)	6,050	125	5,925	0	3.3	1.9
	Average: \$38.09		Average: \$38.09			

Notes: Status-quo columns show revenue and profits given the revenue observed in the data and the corresponding profit inferred from the calibrated Rev_k and estimated $\hat{\tau}_2$; Profit = $R - \hat{\tau}_2 m$. Optimal columns show revenue and profit under the optimal contract given by equation (13) at the calibrated parameter values. The two "Episode share" columns report the percentage of episode mass in the non-overlapping bin containing each sample m value: [< \$100), [\$100, \$300), [\$300, \$500), [\$500, \$800), [\$800, \$1,500), [\$1,500, \$3,000), ≥ \$3,000, evaluated on a 200-type grid. The SQ column is normalized over all status-quo visitors and the Opt column is normalized over optimal-contract visitors (the planner deters 12.5% of status-quo visitors under τ_v = \$39.45), so both columns sum to 100% and are directly comparable as the *within-visitor* distribution of spending.

The optimal contract reduces direct Medicare outlay by \$208–\$225 per episode and total third-party (Medicare plus Medigap) spending by \$260–\$281 per episode, versus \$55 and \$180 respectively under Medigap elimination. The financial savings to Medicare are thus almost 4x larger than those associated with eliminating Medigap for physician services. The change in total welfare (taking into account the health and risk impacts of the service level changes) ranges from \$139-\$151 per episode, 2.3-2.4 as large as eliminating Medigap for these services.[22] In Appendix S14 we estimate that for outpatient physician services this would amount to annual Medicare savings of approximately \$5.5 billion from Medigap enrollees only. If we (somewhat heroically) extrapolate the benefits to all fee-for-service Medicare enrollees (assuming our estimated λ distribution is also representative of non-Medigap enrollees), the annual Medicare savings for outpatient physician services would be approximately \$11 billion. In principle, similar reforms could be extended to all Medicare outpatient services, although we focus on outpatient services that are within the purview of PCP decision-making.

[22] In Appendix S14 we compare these numbers with Cabral and Mahoney (2019). Their headline estimate, a Medigap externality on Medicare of approximately \$363/bene/year for all Part B physician services, corresponds by construction to our Panel A Δ Medicare column. Translated to our units at the observed episode rate of 1.84 episodes per visiting Medicare beneficiary per year, our Panel A externality is roughly \$101/bene/year for outpatient physician services, about 28% of CM's all-Part B figure. The gap reflects our restriction to outpatient physician services (excluding inpatient physician billing) and possible differences in sample composition; our SQ Medicare baseline of \$925/bene/year is similarly 44% of CM's all-Part B baseline of \$2,112/bene/year. Panel B's \$213/episode decomposes into a \$55 "externality" channel that a Medigap tax would also recover and a \$158 "beyond-externality" channel that the optimal physician contract recovers in addition, by reaching the constrained first-best moral-hazard reduction without imposing patient financial risk beyond small copays.

Table 4: Sensitivity of optimal contract to risk aversion

CARA γ	τ_v	Δ Medicare (\$/ep)	Δ Total cost (\$/ep)	ΔW (\$/ep)
Panel A: Remove Medigap				
0 (quasi-linear)	0	55	180	65
1×10^{-4}	0	55	180	64
2×10^{-4}	0	55	180	64
4.39×10^{-4} (mean)	0	55	180	61
6.12×10^{-4} (p90)	0	55	180	57
Panel B: Optimal physician contract				
0 (quasi-linear)	\$43.81	225	281	151
1×10^{-4}	\$42.80	222	278	149
2×10^{-4}	\$41.80	219	274	147
4.39×10^{-4} (mean)	\$39.45	213	266	142
6.12×10^{-4} (p90)	\$37.94	208	260	139

Notes: τ_v is the visit copay in dollars, matching Proposition 2 and Appendix A. Δ Medicare is the per-episode reduction in Medicare's direct outlay relative to the SQ Medicare+Medigap baseline, assuming Medicare retains an 80% share of allowed charges in both regimes (SQ Medicare baseline = \$502.93/ep). Δ Total cost is the per-episode reduction in total third-party insurance outlay (Medicare plus Medigap), with SQ baseline \$628.66/ep; the gap relative to Δ Medicare reflects reduced Medigap outlay (in Panel A, the Medigap 20% share becomes patient out-of-pocket rather than disappearing). ΔW is the per-episode equivalent variation relative to the status-quo Medigap baseline: the lump-sum wealth transfer that would leave an ex-ante consumer on the identified visiting support indifferent between the baseline and the alternative regime, re-solving the regime equilibrium (visit set, premium, and—in Panel A—physician treatment $m(\lambda; p, x)$) at each candidate transfer. *Panel A* holds the physician contract at the status-quo $\hat{\tau}_2 = -0.016$ and raises consumer cost-sharing to $\tau_1 = 0.20$. *Panel B* replaces the physician contract with the optimal nonlinear $g(m)$ from Proposition 2 and chooses the visit copay τ_v via the associated Chetty-Bailey equation.

Implementing Optimal and Simplified Contracts We show in Appendix S15 that both the optimal nonlinear contract and a simpler piecewise linear alternative can be reframed as functions of the existing RVU-based fee schedule, under an assumption that the mapping between RVUs and m is stable under the policy change. The piecewise linear contract pays more for a minimal visit, applies an approximately 9% markdown on the current schedule through ~\$900 in spending, and then reimburses physicians at marginal cost thereafter (via a 1.6% markdown from the existing schedule). This contract achieves roughly half the welfare benefits of the optimal nonlinear contract, with most of the losses coming because the marginal penalty for low-spending patients is too small.

Note that in principle, even our "optimal" nonlinear contract could be improved upon were Medicare to use a differently sloped contract conditioning on patient observable characteristics. Our contracts here are a lower bound on what is achievable by these even more flexible (but more complicated) contracts.

5 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 authorization 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.[23]

5.1 Data

We use data from the Center for Medicare and Medicaid Services (CMS) for 2011–2019.[24] 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 (Medicare Board of Trustees, 2016).

Drugs covered by Medicare Part D include most prescription drugs sold at retail pharmacies,

[23]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 5.3).

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

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.[25] This translates the expenditure-weighted PA share τ_2^* into a dollar cost per unit of expenditure:

$$\tau_2 = \tau_2^* \times \frac{\text{PA cost per drug}}{\bar{P}}, \tag{14}$$

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

Table 5 reports summary statistics for the sample.

[25]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.

5.2 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:

$$\Delta M_{it} = \beta_{\tau_1} \Delta \tau_{1,it} + \beta_{\tau_2^*} \Delta \tau^*_{2,it} + X'_{it} \alpha + e_{it}, \tag{15}$$

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 $t-1$:

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

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.

5.3 Reduced-Form Results

Table 6 reports IV estimates of equation (15), progressively adding controls across columns.

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

	(1)	(2)
	ΔM_{it}	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.

5.4 Structural Identification and Parameter Estimates

Under the ratio-normalized quadratic $h(m,\lambda) = m - \lambda - (m-\lambda)^2/(2\omega\lambda)$ and the local approximation $\tau_1 v'(c) \approx \tau_1$ (i.e. $v'(\bar{c}) = 1$), the firm FOC implies that at the observed (τ_1, τ_2):

$$\frac{\partial E(m)}{\partial \tau_1} \approx -\omega\bar{\lambda}, \tag{16}$$

$$\frac{\partial E(m)}{\partial \tau_2} \approx -\frac{\sigma\bar{\lambda}}{\omega R}, \tag{17}$$

where R is the per-patient physician rent and $\bar{\lambda} = E[\lambda]$ is mean illness severity. Note that in the ratio-normalized quadratic, $\partial m/\partial\tau_1 = -\omega\lambda$ is proportional to λ (rather than constant as in the additive specification), so the aggregate response depends on $\omega\bar{\lambda}$. In practice, the product $\omega\bar{\lambda}$ is identified as a single composite parameter from the data; the interpretation of ω changes (it now measures curvature relative to spending levels) but the identification structure is preserved.[26] The linearized first-difference specification (15) 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 $\hat{\beta}_{\tau_2^*} = -3{,}058$:

$$\widehat{\omega\bar{\lambda}} = |\hat{\beta}_{\tau_1}| \approx \$987 \text{ per year} \quad \text{(composite curvature-severity parameter)}, \tag{18}$$

$$\hat{\sigma} = \hat{\omega} \cdot R \cdot \frac{|\hat{\beta}_{\tau_2^*}|}{|\hat{\beta}_{\tau_1}|} \cdot \zeta, \tag{19}$$

where ζ is the conversion factor from PA share to τ_2 in model units (equation (14)).

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 4.

These two estimates are independent of R and of each other. $\widehat{\omega\bar{\lambda}} = \$987$ per year is identified directly from $\hat{\beta}_{\tau_1}$ and does not depend on the PA cost assumption. The coinsurance response identifies ω jointly with $\bar{\lambda}$ (and with $v'(\bar{c})$), requiring a normalization such as $v'(\bar{c}) = 1$. $\hat{\sigma}/R = |\hat{\beta}_{\tau_2^*}| \times \text{PA cost}/\bar{P} = 3{,}058 \times \text{PA cost}/\$2{,}300$, yielding $\hat{\sigma}/R = 66.5$ under the baseline PA cost of \$50 and $\hat{\sigma}/R = 18.6$ under

[26] Equations (16)–(17) 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 (14)) 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 4, where $\hat{\tau}_2$ is recovered from data and the condition $\tau_2 \cdot \bar{m} \ll R$ must be checked ex post.

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 5.5, a large $\hat{\sigma}/R$ also has strong implications for optimal policy.

5.5 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 3). 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.

For each assumed γ, we compute $v'(c) = \exp(-\gamma c)$ at the status-quo consumption level and solve for the optimal (τ_1^*, τ_2^*) jointly with the equilibrium shadow value $\mu(\gamma)$. The optimal consumer coinsurance τ_1^* remains near zero across all γ values because the fundamental mechanism—physician cost-sharing is far more effective per dollar—does not depend on the level of risk aversion. What changes with γ is the welfare gain from the policy switch and the exact shape of the optimal physician contract.

Table 8: Counterfactual Welfare Simulations

Panel A: Optimal coinsurance rate and PA (by CARA γ)			
	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} ($832)	0.3%	35.0%	3.6%
Panel B: Welfare gain (by CARA γ)			
		Welfare gain ($)	
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 coin-

surance 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.

6 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 physician services, we find physicians are currently earning higher economic profits when they spend more; compensating them so that true profits are declining in expenditures could reduce physician risk while dramatically reducing moral hazard. Small visit copays could prevent excess patient visits given such a schedule. The net benefits for physician services are more than twice as large as correcting the Medigap externality. 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.

The model does not take into account either pay for performance or the design of contracts in settings with physician misreporting of effort (Ma and McGuire, 1997). A natural extension of our work

is to accommodate models where physicians might undersupply effort due to asymmetric information between patients and physician, creating a role for quality contracts.

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.

Acquatella, Angelique and Victoria Marone, "The Risk Protection Value of Moral Hazard," NBER Working Paper 34156, National Bureau of Economic Research August 2025. Revised December 2025.

Arrow, Kenneth J., "Uncertainty and the welfare economics of medical care," *American Economic Review*, 1963, *53* (5), 941–973.

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 and Neale Mahoney, "Externalities and Taxation of Supplemental Insurance: A Study of Medicare and Medigap," *American Economic Journal: Applied Economics*, 2019, *11* (2), 310–344.

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.

_ , "Medicare Physician Fee Schedule: Overview of the Resource-Based Relative Value Scale," https://www.cms.gov/medicare/payment/fee-schedules/physician 2024. Accessed: 2026-03-05.

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

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.

_ , "Moral hazard versus liquidity and optimal unemployment insurance," *Journal of Political Economy*, 2008, *116* (2), 173–234.

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.

Prager, Elena, "Healthcare Demand under Simple Prices: Evidence from Tiered Hospital Networks," *American Economic Journal: Applied Economics*, 2020, *12* (4), 196–223.

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

Sinaiko, Anna D., "How Do Quality Information and Cost Affect Patient Choice of Provider in a Tiered Network Setting? Results from a Survey," *Health Affairs*, 2011, *30* (5), 902–910.

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

to Albert Ma, Ching and Thomas G. McGuire, "Optimal health insurance and provider payment," *American Economic Review*, 1997, *87* (4), 685–704.

A Proof of Proposition 2

Timing and the planner's objective. The constrained first best is a nested, time-consistent problem. At date 0 the planner chooses a common visit copay τ_v. Given τ_v, after observing λ, the consumer decides whether to visit, correctly anticipating the continuation allocation she would receive if she visited and the common premium p implied by budget balance. Let $S(\tau_v)$ denote the resulting rational-expectations visiting set. At date 2, taking $S(\tau_v)$ as fixed, the planner chooses treatment $m(\lambda)$ and intensive-margin out-of-pocket spending $x(\lambda)$ for realized visitors to maximize total welfare of visitors and nonvisitors, with the premium pinned down residually by budget balance. Equivalently, for any fixed set S and copay τ_v, the continuation problem is

$$\mathcal{W}^c(S;\tau_v) = \max_{\{m(\lambda),x(\lambda)\}_{\lambda\in S},\, p} \int_{\lambda\in S} U^{visit}(\lambda; m(\lambda), x(\lambda), \tau_v, p)\, dF + \int_{\lambda\notin S} U^0(\lambda; p)\, dF,$$

subject to budget balance. The date-0 planner then chooses τ_v to maximize ex-ante welfare,

$$\tau_v^{cfb} \in \arg\max_{\tau_v} \mathcal{W}(\tau_v) \equiv \mathcal{W}^c(S(\tau_v);\tau_v).$$

Because treatment is chosen only after the visiting set is fixed, the continuation planner allocates treatment efficiently conditional on visit; the only screening instrument is the visit copay τ_v.

Proof of Proposition 2. Fix τ_v and the induced visiting set $S = S(\tau_v)$, and solve the date-2 continuation problem. Suppressing constants that depend only on S (such as expected taste-shock surplus and hassle costs), the planner solves

$$\max_{\{m(\lambda),x(\lambda)\}_{\lambda\in S}} \int_{\lambda\in S} \big[h(m(\lambda),\lambda) + v\big(W - p - \tau_v - x(\lambda)\big)\big]\, dF + \int_{\lambda\notin S} \big[h(0,\lambda) + v(W-p)\big]\, dF,$$

subject to

$$p = J\bar{\pi} + \int_{\lambda\in S} \big(m(\lambda) - \tau_v - x(\lambda)\big)\, dF.$$

Let μ denote the shadow value of one premium dollar in this continuation problem. Since the visiting set is fixed at date 2, nonvisitors matter only through this common shadow value.

Take a visitor with $m(\lambda) > k$. Raising $m(\lambda)$ by one dollar increases health by $h_m(m(\lambda),\lambda)$ and

raises the premium by one dollar, whose social cost is μ. Therefore every interior visitor satisfies

$$h_m\big(m(\lambda),\lambda\big) = \mu.$$

Now perturb $(m(\lambda),x(\lambda))$ by increasing both by the same small amount. Net insurer cost, and hence the premium, is unchanged. Optimality therefore requires

$$h_m\big(m(\lambda),\lambda\big) \le v'\big(W - p - \tau_v - x(\lambda)\big).$$

Combining the two displays gives

$$v'\big(W - p - \tau_v - x(\lambda)\big) \ge \mu.$$

So a visiting state has weakly higher marginal utility of consumption than the average premium dollar. Raising $x(\lambda)$ alone therefore takes a dollar from a high-marginal-utility visitor and rebates it through the premium pool at value μ, which weakly lowers welfare. Hence

$$x(\lambda) = 0 \qquad \text{for every visitor.}$$

If $m(\lambda) = k$, feasibility already implies $x(\lambda) = 0$, so the conclusion holds for all visitors.

Once $x(\lambda) = 0$, all visitors have the same financial consumption

$$c^{vis} = W - p - \tau_v,$$

while nonvisitors consume $W - p$. The continuation problem therefore reduces to an Arrow–Debreu allocation across realized visitor states:

$$m^{cfb}(\lambda) \in \arg\max_{m \ge k} \{h(m,\lambda) - \mu m\}.$$

Thus every interior visitor satisfies

$$h_m\big(m^{cfb}(\lambda),\lambda\big) = \mu,$$

and the corner is $m^{cfb}(\lambda) = k$ whenever the unconstrained solution lies below k. This is the sense in which the continuation allocation is first-best conditional on the visiting set: once entry is sunk, the planner treats each realized visitor efficiently and smooths consumption fully across visitor states.

To characterize entry, note that for any fixed m, the gain from visiting rather than staying home is $h(m,\lambda) - h(0,\lambda)$, which has increasing differences in (m,λ) because $h_{m\lambda} > 0$. Since $m^{cfb}(\lambda)$ is increasing in λ, the actual gain from visiting under the continuation rule,

$$G(\lambda) = h\big(m^{cfb}(\lambda),\lambda\big) - h(0,\lambda),$$

is increasing in λ. Hence the visiting set is an upper tail, $S = [\lambda^*, \bar{\lambda}]$, and the cutoff type is pinned down by actual-treatment indifference,

$$U^{visit}\big(\lambda^*; m^{cfb}(\lambda^*), 0, \tau_v\big) = U^0(\lambda^*).$$

The date-0 planner then chooses τ_v to maximize ex-ante welfare $\mathcal{W}(\tau_v) = \mathcal{W}^c(S(\tau_v);\tau_v)$. A marginal increase in τ_v has two first-order effects. First, it worsens insurance for inframarginal visitors but lowers the premium for everyone; that direct risk-protection effect is

$$P(\text{visit})\big[\mu - v'(c^{vis})\big].$$

Second, it deters marginal visitors. Because the cutoff type is privately indifferent, changing whether the marginal type visits has no first-order utility effect except through insurer spending. Thus the welfare gain from deterring marginal visitors is just the premium saving from not insuring a visitor whose net insurer cost is $m^{cfb}(\lambda^*) - \tau_v$:

$$-\mu\big(m^{cfb}(\lambda^*) - \tau_v\big)\frac{\partial P(\text{visit})}{\partial \tau_v}.$$

Setting the sum equal to zero and using

$$\mu = P(\text{visit})\, v'(W - p - \tau_v) + (1 - P(\text{visit}))\, v'(W - p)$$

yields exactly the Baily–Chetty equation in Proposition 2,

$$P(\text{visit}) \cdot \frac{v'(c^{vis}) - \mu}{\mu} = -\big(m^{cfb}(\lambda^*) - \tau_v\big)\frac{\partial P(\text{visit})}{\partial \tau_v}.$$

So the constrained first best uses the visit copay only to screen the extensive margin, while the continuation allocation fully smooths consumption across visitor states.

Implementation. Note that $\tau_2(m)$ is *not* a planner object; it is the market schedule used to implement the planner allocation. Set $\tau_1 = 0$ and $\alpha_j = 0$, and write firm profit per treated consumer as $g(m) = R - \tau_2(m)$. A firm chooses m to maximize $s_j(m, \lambda)\, g(m)$. At a symmetric allocation,

$$\frac{\partial s_j / \partial V_j}{s_j} = \frac{1}{\kappa},$$

so the firm's first-order condition is

$$h_m(m(\lambda), \lambda) = -\kappa \frac{g'(m)}{g(m)}.$$

Choose

$$\frac{g'(m)}{g(m)} = -\frac{\mu}{\kappa}, \qquad \text{equivalently} \qquad g(m) = A \exp\left(-\frac{\mu}{\kappa} m\right),$$

with A chosen to hit the required average profit $\bar{\pi}$. Then the firm's condition coincides exactly with the planner's condition $h_m = \mu$, and the lower bound $m \geq k$ reproduces the same corner at k. Because consumers face zero intensive-margin OOP and the same visit copay τ_v, they anticipate the same continuation rule when deciding whether to visit. Hence the same cutoff, premium, and allocation are reproduced in equilibrium. □

B Monopolistic Limit

We show that, under a linear contract, the constrained-first-best intensive-margin allocation is achievable in the logit case in the monopolistic limit ($\kappa \to \infty$) with vanishingly small firm cost-sharing and no consumer cost-sharing. We drop the extensive margin and focus on the intensive margin. The results extend to the full model: the intensive-margin contract shape derived here applies conditional on visiting, while the visit copay τ_v is pinned down separately 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:

$$\tau_2(\kappa) = \frac{\mu R}{\kappa},$$

where μ is the planner's shadow value of a resource dollar (as in Proposition 2 and Appendix A).

Substituting into the firm's symmetric-equilibrium first-order condition (equation 44):

$$h_m(m,\lambda) - \tau_1 v'(c) = \frac{\kappa \tau_2}{R - \tau_2 m},$$

the right-hand side becomes:

$$\frac{\kappa \cdot (\mu R/\kappa)}{R - \tau_2 m} = \frac{\mu R}{R - \tau_2 m}.$$

As $\kappa \to \infty$, $\tau_2(\kappa) \to 0$, so $R - \tau_2 m \to R$ and the right-hand side converges to μ. Since $\tau_1 = 0$, the first-order condition therefore converges to:

$$h_m(m,\lambda) = \mu,$$

which is exactly the constrained-first-best intensive-margin condition from Proposition 2 and Appendix A.

Intuitively, as market power grows, even a vanishingly small cost-sharing rate suffices to align firm incentives with the planner's service-level target. With $\tau_1 = 0$, consumers bear no out-of-pocket cost on the intensive margin, so there is no risk-protection loss; the insurer achieves the constrained first best by relying entirely on firm cost-sharing.

Supplemental Appendices

S1 Equivalence to Bargaining

This appendix shows that, in a special case of the main model, the equilibrium service-level condition can equivalently be written as the first-order condition of a Nash bargaining problem between a consumer and a firm. The result is derived under four simplifying assumptions:

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.* We drop the constraint $m \geq k$ and restrict attention to parameter values for which the bargaining solution is interior.

Suppose there are $J \geq 2$ 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:

$$u_C(m,\lambda) = h(m,\lambda) + v(W - \tau_1 m - p), \qquad u_F(m) = R - \tau_2 m, \tag{20}$$

where p is the common premium defined in the text. If bargaining breaks down, the firm's outside option is zero. The consumer can instead switch to an alternative firm. Let $m^{\text{alt}}(\lambda)$ denote the service level the consumer would receive from an alternative firm, and suppose that switching entails a utility cost $\delta > 0$. Hence the consumer's outside option is:

$$U^{\text{out}}(\lambda) = h\Big(m^{\text{alt}}(\lambda), \lambda\Big) + v\Big(W - \tau_1 m^{\text{alt}}(\lambda) - p\Big) - \delta. \tag{21}$$

We restrict attention to parameters for which the bargaining surplus is strictly positive on the relevant domain, i.e. $u_C(m,\lambda) - U^{\text{out}}(\lambda) > 0$ and $u_F(m) = R - \tau_2 m > 0$, so that the Nash program below is

well-defined.

Taking $U^{\text{out}}(\lambda)$ as given in the bilateral bargaining problem, the matched consumer-firm pair solves the Nash program:

$$m^*(\lambda) \in \arg\max_m \left[u_C(m,\lambda) - U^{\text{out}}(\lambda)\right]^{1-\varphi} \left[u_F(m)\right]^{\varphi}, \tag{22}$$

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

$$m^*(\lambda) \in \arg\max_m (1-\varphi)\ln\left(u_C(m,\lambda) - U^{\text{out}}(\lambda)\right) + \varphi \ln(u_F(m)). \tag{23}$$

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

$$(1-\varphi)\frac{h_m(m^*(\lambda),\lambda) - \tau_1 v'(W - \tau_1 m^*(\lambda) - p)}{u_C(m^*(\lambda),\lambda) - U^{\text{out}}(\lambda)} - \varphi \frac{\tau_2}{R - \tau_2 m^*(\lambda)} = 0. \tag{24}$$

Rearranging gives:

$$h_m(m^*(\lambda),\lambda) - \tau_1 v'(W - \tau_1 m^*(\lambda) - p) = \frac{\varphi}{1-\varphi}\left(u_C(m^*(\lambda),\lambda) - U^{\text{out}}(\lambda)\right)\frac{\tau_2}{R - \tau_2 m^*(\lambda)}. \tag{25}$$

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 $m^{\text{alt}}(\lambda) = m^*(\lambda)$. In a symmetric equilibrium, the health and consumption terms cancel between the matched and outside options:

$$\begin{aligned} u_C(m^*(\lambda),\lambda) - U^{\text{out}}(\lambda) &= \left[h(m^*(\lambda),\lambda) + v(W - \tau_1 m^*(\lambda) - p)\right] \\ &\quad - \left[h(m^*(\lambda),\lambda) + v(W - \tau_1 m^*(\lambda) - p) - \delta\right] \\ &= \delta. \end{aligned} \tag{26}$$

Substituting this into the first-order condition yields:

$$h_m(m^*(\lambda),\lambda) - \tau_1 v'(W - \tau_1 m^*(\lambda) - p) = \frac{\varphi}{1-\varphi}\,\delta\,\frac{\tau_2}{R - \tau_2 m^*(\lambda)}. \tag{27}$$

In the main model, the firm's symmetric-equilibrium first-order condition with linear cost-sharing is (equation 44):

$$h_m(m,\lambda) - \tau_1 v'(W - \tau_1 m - p) = \frac{\kappa \tau_2}{R - \tau_2 m}. \tag{28}$$

Equations (27) and (28) coincide exactly when:

$$\frac{\varphi}{1-\varphi}\,\delta = \kappa, \qquad \text{i.e.} \qquad \varphi \;=\; \frac{\kappa}{\kappa+\delta}.$$

The firm's bargaining weight is therefore an increasing function of the main-text market-power summary κ: more concentrated markets (larger κ) give firms more leverage, and more competitive markets (smaller κ) give consumers more leverage.

S2 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). Throughout, $\Pi_j(\lambda)$ denotes firm j's per-beneficiary margin conditional on a type-λ consumer visiting firm j; the share-weighted profit object appearing in the main-text insurer problem is $\pi_j(\lambda) = s_j(\lambda)\,\Pi_j(\lambda)$. The side payment α_j may be of either sign: positive α_j raises consumer out-of-pocket spending and firm profit in parallel.

Fix a candidate symmetric premium p and fixed rate R. For a type-λ consumer who visits, firm j solves (Equation 5):

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

where s_j is the market share induced by the random-utility system in Equation 1. Let:

$$V_j \equiv h(m_j,\lambda) + v(W-\tau_1(m_j)-\alpha_j-p)$$

denote firm j's deterministic utility index; then $s_j = s_j(V_1,\ldots,V_J)$ depends on (m_j,α_j) only through V_j.

Define the consumer's net out-of-pocket cost and the firm's per-beneficiary margin 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 $\mathrm{OOP}_j - \Pi_j = T(m_j) - R$, where $T(m) \equiv \tau_1(m) + \tau_2(m)$ is the combined cost-

sharing schedule. Assume T is strictly increasing on $[k,\infty)$, so that T^{-1} is well-defined on the range $T([k,\infty))$. Then:

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

and the floor constraint $m_j \geq k$ becomes the range restriction $R + \mathrm{OOP}_j - \Pi_j \in T([k,\infty))$.

Given a particular decomposition (τ_1, τ_2) of T, the map $(m_j, \alpha_j) \leftrightarrow (\mathrm{OOP}_j, \Pi_j)$ is one-to-one: from (OOP_j, Π_j) we recover m_j as above and then $\alpha_j = \mathrm{OOP}_j - \tau_1(m_j)$. Note that the inverse requires knowledge of τ_1 separately, not just of T: across decompositions of the same T, α is exactly what adjusts to keep (OOP, Π, m) invariant.

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

$$s_j(\lambda) = s_j\Big(\big\{h\big(T^{-1}(R + \mathrm{OOP}_{j'} - \Pi_{j'}), \lambda\big) + v(W - \mathrm{OOP}_{j'} - p)\big\}_{j'=1}^{J}\Big),$$

so the firm's problem can be written as:

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

This problem depends on the primitive cost-sharing schedules only through the combined schedule T and the premium p. Therefore, at any symmetric equilibrium supporting a common premium p, the equilibrium margin, out-of-pocket spending, and service level:

$$\big(\mathrm{OOP}^*(\lambda; p),\ \Pi^*(\lambda; p),\ m^*(\lambda; p)\big)$$

depend on cost-sharing only through T, where $m^*(\lambda; p) = T^{-1}\big(R + \mathrm{OOP}^*(\lambda; p) - \Pi^*(\lambda; p)\big)$.

The argument now shows that the extensive-margin visit decision, the symmetric-equilibrium premium p, and the profit-target reimbursement level R are all determined jointly by T alone. We avoid circularity by treating p as a parameter until the final fixed-point step.

Fix p. By Step 1, the equilibrium objects $\mathrm{OOP}^*(\lambda; p)$, $\Pi^*(\lambda; p)$, and $m^*(\lambda; p)$ depend only on T. The type-λ consumer's visit-margin utilities,

$$U^{\mathrm{visit}}(\lambda; p) = \mathcal{S}\Big(\big\{h(m_j^*(\lambda; p), \lambda) + v(W - \mathrm{OOP}_j^*(\lambda; p) - p)\big\}_{j=1}^{J}\Big) - c_{\mathrm{visit}}, \qquad U^0(\lambda; p) = h(0, \lambda) + v(W - p),$$

therefore depend on cost-sharing only through T at fixed p. So does the visit indicator (cf. Equa-

tion 3):

$$I(\lambda; p) \equiv \mathbf{1}\{U^{\text{visit}}(\lambda; p) \geq U^0(\lambda; p)\}.$$

For each visiting consumer i, define the inclusive objects $\text{OOP}_i^*(p) = I(\lambda_i; p)\,\text{OOP}_{j(i)}^*(\lambda_i; p)$ and $m_i^*(p) = I(\lambda_i; p)\, m_{j(i)}^*(\lambda_i; p)$, where $j(i)$ is the firm chosen by i; both depend on cost-sharing only through T at fixed p.

The symmetric-equilibrium premium is pinned down by budget balance (from the main-text premium formula):

$$p \;=\; J\bar{\pi} + E\left[m_i^*(p) - \text{OOP}_i^*(p)\right], \tag{29}$$

and R is pinned down by the target-profit condition $E[\pi(\lambda)] = \bar{\pi}$, i.e.,

$$E[s^*(\lambda; p, R)\,\Pi^*(\lambda; p, R)] \;=\; \bar{\pi}, \tag{30}$$

where we have made the dependence of the Step-1 objects on R explicit. The right-hand sides of (29) and (30) depend on cost-sharing only through T (at fixed p, R), so the joint fixed-point map $(p, R) \mapsto \big(F_p(p, R; T),\, F_R(p, R; T)\big)$ itself depends on cost-sharing only through T. Any pair (p^*, R^*) solving this system â and hence the associated $(m^*, \text{OOP}^*, \Pi^*, U^{\text{visit}}, U^0, I)$ evaluated at (p^*, R^*) â depends on cost-sharing only through T.

Fix any combined schedule T and any two decompositions (τ_1, τ_2) and $(\tilde{\tau}_1, \tilde{\tau}_2)$ with $\tau_1 + \tau_2 = \tilde{\tau}_1 + \tilde{\tau}_2 = T$. Any symmetric equilibrium under (τ_1, τ_2) induces a symmetric equilibrium under $(\tilde{\tau}_1, \tilde{\tau}_2)$ with identical $(m^*, \text{OOP}^*, \Pi^*, p^*, R^*)$ and the same visiting population $\{\lambda : I(\lambda) = 1\}$; the only object that adjusts is the side payment, via $\tilde{\alpha}^*(\lambda) = \text{OOP}^*(\lambda) - \tilde{\tau}_1(m^*(\lambda))$. All consumer and firm outcomes are therefore invariant across decompositions of the same T. □

S3 Insurers Prefer Non-transferability

This appendix proves that, in a special-case of our model, insurers strictly prefer non-transferability.

Relative to the full model, impose the following additional assumptions:

1. *All consumers visit.* The same logic can alternatively be read as holding conditional on a fixed set of visiting types.

2. *Regularity of the transferable optimum.* The transferable optimum induces an interior service rule $m^T(\lambda)$ that is strictly increasing in λ on its interior range, with inverse Λ^T, and the com-

posite marginal-benefit object $h_m(m, \Lambda^T(m))$ is integrable on that range.

3. *Positive target profits.* The negotiated average per-firm profit satisfies $\bar{\pi} > 0$, so the construction below produces an implementing firm-profit schedule in the same positive-profit region as Proposition 2.

4. *Nontrivial risk exposure in the transferable optimum.* The transferable optimum induces net out-of-pocket payments:
$$\mathrm{OOP}^T(\lambda) \equiv \tau_1^T\big(m^T(\lambda)\big) + \alpha^T(\lambda)$$
that are non-constant on a set of positive measure.

Let $\mathcal{W}^T$ denote maximal consumer welfare in the transferable world and $\mathcal{W}^{NT}$ maximal consumer welfare in the non-transferable world.

Proposition. *Under the assumptions above,*

$$\mathcal{W}^{NT} > \mathcal{W}^T.$$

If $\mathrm{OOP}^T(\lambda)$ *is constant almost everywhere, the inequality weakens to* $\mathcal{W}^{NT} \geq \mathcal{W}^T$.

Proof. Let:
$$\big(m^T(\lambda), \mathrm{OOP}^T(\lambda), \Pi^T(\lambda), p^T\big)$$
denote the transferable optimum, where:

$$\Pi^T(\lambda) \equiv R - \tau_2^T\big(m^T(\lambda)\big) + \alpha^T(\lambda)$$

is the firm's net profit per beneficiary. By Proposition 1 (proved in Appendix S2), transferable-world equilibrium outcomes depend only on the combined cost-sharing schedule $T(m) \equiv \tau_1(m) + \tau_2(m)$, not on its decomposition between consumer and firm sides. In particular, the transferable optimum pins down a feasible service rule $m^T(\lambda)$ together with associated net out-of-pocket payments and firm profits per beneficiary.

We proceed by constructing a candidate non-transferable policy that replicates the transferable service levels but exposes consumers to less financial risk. Set:

$$\alpha^{NT}(\lambda) \equiv 0, \qquad \tau_1^{NT}(m) \equiv 0,$$

and define the non-transferable firm-profit schedule:

$$g^{NT}(m) \equiv R^{NT} - \tau_2^{NT}(m)$$

on the interior range of m^T by the differential equation:

$$\frac{g^{NT\prime}(m)}{g^{NT}(m)} = -\frac{1}{\kappa}\, h_m\big(m, \Lambda^T(m)\big).$$

Equivalently,

$$g^{NT}(m) = A \exp\left(-\frac{1}{\kappa}\int_{m_0}^{m} h_z\big(z, \Lambda^T(z)\big)\, dz\right),$$

where $A > 0$ is chosen so that expected per-firm profits equal the exogenous target $\bar{\pi}$.

Under the same symmetric smooth random-utility system summarized by κ, the interior implementation condition used in Proposition 2 implies that with $\tau_1 = 0$,

$$h_m(m, \lambda) = -\kappa \frac{g'(m)}{g(m)}.$$

By construction, the non-transferable schedule $g^{NT}(m)$ therefore implements exactly the same service rule $m^T(\lambda)$: plugging g^{NT} into the firm's FOC yields $h_m(m, \lambda) = h_m(m, \Lambda^T(m))$, whose unique solution in m is $m^T(\lambda)$ by strict monotonicity of h_m in λ (Assumption $h_{m\lambda} > 0$).

Now compare consumer financial consumption. Because all consumers visit, the transferable optimum satisfies the premium identity:

$$p^T = E\big[m^T(\lambda)\big] - E\big[\mathrm{OOP}^T(\lambda)\big] + J\bar{\pi},$$

while under the non-transferable replica,

$$p^{NT} = E\big[m^T(\lambda)\big] + J\bar{\pi} = p^T + E\big[\mathrm{OOP}^T(\lambda)\big].$$

Therefore financial consumption in the transferable optimum is:

$$c^T(\lambda) = W - p^T - \mathrm{OOP}^T(\lambda),$$

whereas under the non-transferable replica it is constant across λ:

$$c^{NT} = W - p^{NT} = W - p^{T} - E\big[\mathrm{OOP}^{T}(\lambda)\big].$$

These two allocations have the same mean:

$$E\big[c^{T}(\lambda)\big] = c^{NT}.$$

Health utility is identical under the two policies because the service rule $m^{T}(\lambda)$ is identical. Because firms remain symmetric and implement the same service rule under both policies, the discrete-choice surplus term from firm taste shocks is also unchanged. The welfare comparison therefore reduces to identical health utility plus a Jensen comparison over financial consumption: the non-transferable replica replaces the state-dependent payment $\mathrm{OOP}^{T}(\lambda)$ with a common premium-financed payment of the same mean.

By strict concavity of v and Jensen's inequality,

$$E\big[v(c^{T}(\lambda))\big] < v\Big(E\big[c^{T}(\lambda)\big]\Big) = v(c^{NT})$$

whenever $\mathrm{OOP}^{T}(\lambda)$ is non-constant on a set of positive measure. Therefore the non-transferable replica yields strictly higher consumer welfare than the transferable optimum.

Finally, $\mathcal{W}^{NT}$ is the maximum welfare attainable over all non-transferable policies, so it must weakly exceed the welfare delivered by this replica. Hence $\mathcal{W}^{NT} > \mathcal{W}^{T}$. If $\mathrm{OOP}^{T}(\lambda)$ is constant almost everywhere, Jensen gives only weak inequality, so the conclusion weakens to $\mathcal{W}^{NT} \geq \mathcal{W}^{T}$.

□

S4 Rationales for Consumer Cost-Sharing: Detailed Discussion

This appendix expands on the rationales for intensive-margin consumer cost-sharing summarized in Section 3.4 and Table 1.

Firm Heterogeneity

In the baseline model a visit copay handles extensive-margin heterogeneity across consumers while firm contracts handle intensive-margin heterogeneity within visits. When firms themselves vary in

productivity, a second extensive-margin problem arises: fully insured consumers have no incentive to choose more cost-effective firms. Appendix S6 shows that firm-specific copays resolve this, while (now firm-specific) firm contracts continue to handle intensive-margin moral hazard.

Asymmetric Information

If doctors vary in altruism but the insurer cannot distinguish between them, firm cost-sharing disproportionately reduces service levels for patients of less altruistic firms, who are already under-treated relative to patients of more altruistic firms. Consumer cost-sharing affects patients more uniformly across firm types and partially undoes this misallocation. Appendix S7 develops the model formally and shows that, to first order, the optimal coinsurance rate rises in the variance of firms' marginal rate of substitution between consumer and firm cost-sharing.

Firm Risk Aversion

Large hospitals can reasonably be treated as risk-neutral with respect to any individual patient; individual physicians, auto mechanics, and contractors typically cannot. We give four separate sufficient conditions under which the zero-intensive-margin-consumer-copay logic of Proposition 2 survives firm risk aversion. Each shows when firm risk aversion does not overturn that logic; they are sufficient rather than jointly necessary. Let firm Bernoulli utility over profit be $u_F(\pi)$, where $u_F' > 0$ and $u_F'' < 0$, and let $\mathrm{CE}(X) \equiv u_F^{-1}(\mathbb{E}[u_F(X)])$ denote the certainty equivalent of a random profit stream X. The insurer must deliver reservation certainty equivalent $\bar{\pi}^{CE} \equiv u_F^{-1}(\bar{U})$. The optimal incentive schedule from Proposition 2 has the form $g(m) = A\exp(-\mu m/\kappa)$.

(a) Up-front payments. Suppose the insurer can add a deterministic per-patient prospective payment B independent of realized service level, so firm profit per patient is $B + g(m)$. For any fixed small $A > 0$ the random component $g(m)$ is uniformly small, and there is a unique $B(A)$ satisfying $\mathbb{E}[u_F(B(A) + g(m))] = \bar{U}$. As $A \to 0$ the risky component vanishes, so firm risk can be made arbitrarily small.

(b) Sufficient market power (large κ). Now suppose up-front payments are infeasible. Let $A(\kappa)$ satisfy $\mathbb{E}[u_F(A(\kappa)\exp(-\mu m/\kappa))] = \bar{U}$. As $\kappa \to \infty$, $\exp(-\mu m/\kappa) \to 1$ pointwise, and dominated convergence gives $A(\kappa) \to \bar{\pi}^{CE}$ and $g(m) \to \bar{\pi}^{CE}$ for every m. The schedule becomes asymptotically flat, so firm risk vanishes even without front-loading.

(c) Law of large numbers. Suppose the firm treats N patients with i.i.d. severity draws. Average profit is $\bar{\Pi}_N = N^{-1}\sum_{n=1}^{N} g(m^*(\lambda_n))$, and the law of large numbers gives $\mathrm{CE}(\bar{\Pi}_N) \to \mathbb{E}[g(m^*(\lambda))]$, so the

per-patient risk premium converges to zero.

(d) Baseline linear contract with a marginal subsidy. Holding the induced distribution of $m^*(\lambda)$ fixed, suppose the benchmark is a linear contract $g_0(m) = R_0 - \tau_2^0 m$ with $\tau_2^0 < 0$ (matching the status-quo Medicare fee schedule). Any alternative linear contract with the same mean profit and $|\tau_2| < |\tau_2^0|$ has strictly lower profit variance. By concavity of u_F its certainty equivalent is weakly higher, so moving τ_2 toward zero reduces firm risk relative to the baseline. At $\tau_2 = 0$ the intensive-margin profit risk disappears entirely. In the Medicare application in Section 4 we assume salary-based compensation is infeasible and gauge residual risk exposure through market power and patient pooling.

Bad Reasons for Consumer Cost-Sharing

Loading. Proportional administrative loading does not rationalize intensive-margin consumer cost-sharing once firm cost-sharing is available. Adding a proportional loading factor $\ell > 0$ raises the marginal resource cost of service from 1 to $1+\ell$. The planner's interior first-order condition becomes $h_m(m_\ell^{\text{cfb}}(\lambda), \lambda) = (1+\ell)\,\mu_\ell$, which is Proposition 2 with the marginal resource cost scaled up. Running the Proposition 2 implementation argument yields the firm profit schedule $g_\ell(m) = A_\ell \exp(-(1+\ell)\mu_\ell m/\kappa)$, steeper than the no-loading schedule by a factor of $1+\ell$ in the exponent. Consumer consumption conditional on visiting remains $W - p - \tau_v^\ell$, independent of m, so the same risk-protection argument gives $\tau_1 = 0$. Proportional loading rescales the firm schedule but creates no new rationale for intensive-margin consumer cost-sharing, paralleling Ma and McGuire (1997).

Altruism. Ellis and McGuire (1990) analyzes an altruistic monopolist and finds that linear firm cost-sharing achieves the first best, with more altruistic firms requiring stronger incentives (more negative τ_2). Embedding altruism in our competitive model preserves this conclusion: steeper firm incentives are needed, but consumer cost-sharing is not.

Behavioral Hazard. We extend the baseline model to allow for treatment-level behavioral hazard: consumers systematically overvalue one treatment relative to another (Chernew et al., 2007; Baicker et al., 2015). Consider a single visit in which a provider chooses between treatments A and B for a patient with severity λ, where A is socially preferred ($h_A(\lambda) > h_B(\lambda)$) but patients overvalue B ($\tilde{h}_B(\lambda) > \tilde{h}_A(\lambda)$). The insurer reimburses the provider q_t for treatment t and charges the patient τ_t. Under logit demand with market-power parameter κ, the provider's symmetric-equilibrium first-order condition is

$$\frac{\tilde{\delta}}{\kappa}\,\bar{\pi}(\rho) + \Delta\pi \;=\; 0, \tag{31}$$

where $\tilde{\delta} \equiv \tilde{h}_A - \tilde{h}_B - (\tau_A - \tau_B)$ is the perceived net benefit of A over B, $\bar{\pi}(\rho)$ is expected profit per patient, and $\Delta\pi \equiv \pi_A - \pi_B$. Since $\tilde{\delta} < 0$ (patients overvalue B), steering toward A costs market share and requires a positive profit differential:

$$\Delta\pi \;=\; -\frac{\tilde{\delta}}{\kappa}\bar{\pi}(\rho^*). \tag{32}$$

A copay-only policy ($\Delta\pi = 0$) sets $\tilde{\delta} = 0$, leaving providers indifferent over all $\rho \in [0,1]$ rather than selecting a unique interior target. Among implementations that induce the same treatment lottery ρ^*, the reimbursement version (equal copays, differential q_t) weakly dominates: patient financial consumption is independent of the realized treatment, whereas any implementation with $\tau_A \neq \tau_B$ adds mean-preserving risk. The case for consumer-side instruments therefore rests on contracting frictions rather than on behavioral hazard per se. Patient non-adherence to prescribed drugs can justify *negative* consumer cost-sharing, since positive cost-sharing exacerbates non-adherence.

S5 Competitive Limit without the Extensive Margin

We characterize the optimal linear contract in the logit case in the competitive limit ($\kappa \to 0$). 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. We drop the extensive margin and focus on the intensive margin. As in Appendix B, 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 determined separately by the Baily–Chetty condition in Proposition 2.

Consider any convergent subsequence of optimal linear rates $(\tau_1^*(\kappa), \tau_2^*(\kappa))$ as $\kappa \to 0$ with strictly positive firm-side limit $\tau_2' > 0$, and let (τ_1', τ_2') denote its limit. The assumption $\tau_2' > 0$ is what delivers a finite positive implied expenditure cap

$$M_{\text{cap}} \equiv \frac{R}{\tau_2'};$$

the argument below characterizes the limiting allocation along any such subsequence, and we do not claim that every optimal subsequence has $\tau_2' > 0$. Let $m_c(\lambda; \tau_1')$ denote the consumer-side allocation, defined as the solution to the consumer first-order condition:

$$h_m(m, \lambda) = \tau_1' \, v'(W - \tau_1' m - p).$$

Define the cap-binding threshold λ_{cap} implicitly by $m_c(\lambda_{\text{cap}}; \tau_1') = M_{\text{cap}}$.

We characterize the limit of the equilibrium allocation $m^*(\lambda, \tau_1^*(\kappa), \tau_2^*(\kappa), \kappa)$ for each type, using the firm's symmetric-equilibrium first-order condition:

$$h_m\big(m^*(\lambda), \lambda\big) - \tau_1^*(\kappa)\, v'(c) = \frac{\kappa\, \tau_2^*(\kappa)}{R - \tau_2^*(\kappa)\, m^*(\lambda)}. \tag{33}$$

Types with $m_c(\lambda; \tau_1') < M_{\text{cap}}$ (below the cap). For these types, $R - \tau_2^*(\kappa)\, m^*(\lambda)$ stays bounded away from zero along the subsequence, so the right-hand side of (33) vanishes as $\kappa \to 0$. In the limit, the firm first-order condition collapses to the consumer first-order condition, and therefore:

$$\lim_{\kappa \to 0} m^*\big(\lambda, \tau_1^*(\kappa), \tau_2^*(\kappa), \kappa\big) = m_c(\lambda; \tau_1').$$

Types with $m_c(\lambda; \tau_1') \geq M_{\text{cap}}$ (above the cap): a limit-point argument. We show that the only possible limit point of $m^*(\lambda, \tau_1^*(\kappa), \tau_2^*(\kappa), \kappa)$ for such λ is M_{cap}. Suppose first that a candidate limit point satisfies $\lim m^*(\lambda) < M_{\text{cap}}$. Then $R - \tau_2^*(\kappa)\, m^*(\lambda)$ stays bounded away from zero along the relevant subsequence, so the right-hand side of (33) vanishes; the limit therefore solves the consumer first-order condition, which gives $m_c(\lambda; \tau_1') \geq M_{\text{cap}}$, contradicting the assumed limit below M_{cap}. Suppose instead that a candidate limit point satisfies $\lim m^*(\lambda) > M_{\text{cap}}$. Then eventually along the subsequence $\tau_2^*(\kappa)\, m^*(\lambda) > R$, so the firm earns strictly negative per-patient margins $R - \tau_2^*(\kappa)\, m^*(\lambda) < 0$; reducing m^* toward M_{cap} strictly increases firm profits, contradicting firm optimality. The only remaining possibility is:

$$\lim_{\kappa \to 0} m^*\big(\lambda, \tau_1^*(\kappa), \tau_2^*(\kappa), \kappa\big) = M_{\text{cap}}.$$

Limiting allocation. Combining the two cases:

$$m_\infty(\lambda) \;\equiv\; \lim_{\kappa \to 0} m^*\big(\lambda, \tau_1^*(\kappa), \tau_2^*(\kappa), \kappa\big) \;=\; \min\big\{m_c(\lambda; \tau_1'),\, M_{\text{cap}}\big\}.$$

Equivalence with a cap-plus-coinsurance problem. Consider now a planner's problem in which the only available instruments are a linear consumer coinsurance rate τ_1 and a uniform expenditure cap M_{cap} (with τ_2 unavailable). The induced allocation is:

$$m(\lambda; \tau_1, M_{\text{cap}}) \;=\; \min\big\{m_c(\lambda; \tau_1),\, M_{\text{cap}}\big\}.$$

Any (τ_1, M_{cap}) in this second problem is achievable as a $\kappa \to 0$ limit of the original problem by taking $\tau_1^*(\kappa) \equiv \tau_1$ and $\tau_2^*(\kappa) \equiv R/M_{\text{cap}}$; conversely, the limit of any optimal subsequence with $\tau_2' > 0$ corresponds to the choice $(\tau_1, M_{\text{cap}}) = (\tau_1', R/\tau_2')$ in the cap-plus-coinsurance problem and yields the same limiting allocation $m_\infty(\lambda)$. The two problems are therefore equivalent at the level of limiting allocations, and any such limit point $(\tau_1', M_{\text{cap}})$ solves the corresponding cap-plus-coinsurance problem. This proves statements (1) and (2) in the main text.

S6 Firm Heterogeneity

We characterize the first-best allocation when firms differ in productivity and the social planner can choose treatment effort for each patient, the allocation of patients to firms, and out-of-pocket payments. We then show that the same outcome can be decentralized with firm-specific patient copays and a firm-side contract $g_j(m)$. Throughout, μ denotes the planner's shadow value of a resource dollar in the problem below (the analogue of the μ that appears in Proposition 2 and Appendix A).

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

$$h(m+\theta_j, \lambda) + v\big(W - p - \text{OOP}_j(\lambda)\big) + e_j.$$

Consider first the direct planner who observes (λ, e) and each firm's productivity, and chooses an assignment rule $j(\lambda, e) \in \{1, \ldots, J\}$, effort $m_j(\lambda, e)$, a patient out-of-pocket schedule $\text{OOP}_j(\lambda)$, and a common premium p, to maximize expected patient welfare subject to budget balance. Let $\mu > 0$ denote the multiplier on the budget constraint.

For a patient assigned to firm j, the effort first-order condition is:

$$h_m(m+\theta_j, \lambda) = \mu.$$

Assume h is strictly concave in its first argument with inverse $h_m^{-1}(\mu, \lambda)$, and that the solution is

interior in the sense that $h_m^{-1}(\mu,\lambda) > \theta_j$ for all relevant (j,λ). Then the first-best effort rule is:

$$m_j^{FB}(\lambda) = h_m^{-1}(\mu,\lambda) - \theta_j.$$

More productive firms use less effort to deliver the same effective treatment $h_m^{-1}(\mu,\lambda)$, which is the same across firms and depends only on λ. If m_j^{FB} would be negative for some (j,λ), the rule generalizes to $m_j^{FB}(\lambda) = \max\{0,\, h_m^{-1}(\mu,\lambda) - \theta_j\}$; we assume interiority throughout the remainder of the appendix.

The direct planner also chooses $\mathrm{OOP}_j(\lambda)$ directly, without any firm-choice incentive constraint. The associated first-order condition $v'\big(W - p - \mathrm{OOP}_j(\lambda)\big) = \mu$ pins down financial consumption at the common level satisfying $v'(c^*) = \mu$ for every (j,λ), so at the direct planner optimum $\mathrm{OOP}_j(\lambda)$ is equalized across firms and across types. The consumption term therefore drops out of the firm-assignment comparison, leaving a comparison in the remaining firm-specific terms only.

Substituting the first-best effort rule into the planner's objective, welfare at the assignment stage depends on firm j only through the sum of the patient's taste shock and the resource savings from higher productivity:

$$e_j + \mu\theta_j.$$

Therefore the efficient assignment rule is:

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

If the $\{e_j\}$ are i.i.d. Type-I extreme value with scale $(1 - 1/J)\kappa$ (so that the main-text market-power summary κ applies), the efficient market share of firm j is:

$$s_j^{FB} = \frac{\exp\big(\mu\theta_j / [(1-1/J)\kappa]\big)}{\sum_{k=1}^{J} \exp\big(\mu\theta_k / [(1-1/J)\kappa]\big)}.$$

Higher-productivity firms receive proportionally larger shares, and the sensitivity of shares to productivity differences scales with μ/κ.

Suppose now that the planner cannot directly assign patients to firms and must instead induce the efficient assignment through a firm-specific out-of-pocket schedule $\mathrm{OOP}_j(\lambda)$. Under first-best effort, every firm delivers effective treatment $h_m^{-1}(\mu,\lambda)$, so patient (λ, e) choosing firm j obtains utility:

$$h\big(h_m^{-1}(\mu,\lambda),\lambda\big) + v\big(W - p - \mathrm{OOP}_j(\lambda)\big) + e_j.$$

For decentralized choice to reproduce $j^*(\lambda, e) = \arg\max_j \{e_j + \mu\theta_j\}$, the deterministic component of the patient's utility index must differ across firms by exactly $\mu\theta_j$. The implementing schedule therefore satisfies:

$$v\big(W - p - \mathrm{OOP}_j(\lambda)\big) = K(\lambda) + \mu\theta_j \tag{34}$$

for some scalar shift $K(\lambda)$ that is common across firms for a given λ. Equivalently:

$$\mathrm{OOP}_j(\lambda) = W - p - v^{-1}\big(K(\lambda) + \mu\theta_j\big).$$

Under this schedule, patient utility from choosing firm j becomes:

$$h\big(h_m^{-1}(\mu, \lambda), \lambda\big) + K(\lambda) + \mu\theta_j + e_j,$$

so decentralized choice exactly reproduces the planner's efficient assignment.

Given the assignment rule and effort rule, the only role of $K(\lambda)$ is to shift consumption across illness states without affecting which firm the patient picks. The planner therefore chooses $K(\lambda)$ only to manage consumer risk. Because v is strictly concave, v^{-1} is strictly convex, so the resource cost of delivering a given expected consumer surplus from the implementing schedule is strictly convex in $K(\lambda)$. By Jensen's inequality, welfare is maximized at $K(\lambda) \equiv \bar{K}$, a constant. Hence the optimal implementing consumer-payment schedule is:

$$\mathrm{OOP}_j^* = W - p^* - v^{-1}\big(\bar{K} + \mu\theta_j\big),$$

where p^* is the equilibrium premium at the optimum. All patients who choose the same firm pay the same out-of-pocket amount, with no within-firm variation across λ. More productive firms face lower patient copays, because patients would otherwise not internalize the resource savings from choosing them.

A first-order expansion of v^{-1} around reference consumption $\bar{c}$ gives:

$$\mathrm{OOP}_j^* - \mathrm{OOP}_{j'}^* \approx -\frac{\mu(\theta_j - \theta_{j'})}{v'(\bar{c})}.$$

Under the paper's normalization $v'(\bar{c}) = 1$, this becomes $\mathrm{OOP}_j^* - \mathrm{OOP}_{j'}^* \approx -\mu(\theta_j - \theta_{j'})$; in the additional special case $\mu = 1$, cross-firm copay differences are approximately equal to productivity differences, $\mathrm{OOP}_j^* - \mathrm{OOP}_{j'}^* \approx -(\theta_j - \theta_{j'})$.

Conditional on the patient choosing firm j, the insurer still needs a firm-side contract to implement the first-best effort rule. Let $g_j(m) \equiv R_j - \tau_{2,j}(m)$ denote firm j's per-patient profit. Firm j maximizes $s_j(m_j)\, g_j(m_j)$, where under the logit demand system summarized at the symmetric allocation by κ,

$$\frac{\partial s_j}{\partial m_j} = \frac{1}{(1-1/J)\kappa}\, s_j(1-s_j)\, h_m(m_j+\theta_j, \lambda).$$

Dividing the first-order condition $(\partial s_j/\partial m_j)\, g_j + s_j\, g_j' = 0$ by s_j yields:

$$h_m(m+\theta_j, \lambda) = -\kappa_j \frac{g_j'(m)}{g_j(m)}, \qquad \kappa_j \equiv \frac{(1-1/J)\kappa}{1-s_j^{FB}}.$$

Setting this equal to the first-best condition $h_m(m+\theta_j, \lambda) = \mu$ and solving gives:

$$\frac{g_j'(m)}{g_j(m)} = -\frac{\mu}{\kappa_j}, \qquad g_j(m) = A_j \exp\left(-\frac{\mu}{\kappa_j}\, m\right).$$

The slope μ/κ_j is firm-specific because κ_j depends on the efficient share s_j^{FB}; the level A_j is pinned down firm-by-firm by participation or transfer requirements. In the symmetric special case $s_j^{FB} = 1/J$ and $\kappa_j = \kappa$, the schedule collapses to the Appendix A / Proposition 2 form $g(m) = A\exp(-\mu m/\kappa)$. Combined with the patient-side copay rule above, this decentralizes the planner's first best. □

S7 Information Asymmetry

We first note 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—Proposition 2 still applies and the insurer can still achieve the constrained first best with $\tau_1 = 0$ using an appropriate nonlinear firm-profit schedule $g(m)$. The only change relative to the profit-maximizing baseline is that the altruistic firm's first-order condition is a weighted average of the consumer's and the profit-maximizer's, rescaled by $\beta/(1-\beta)$. Matching this rescaled condition to the constrained-first-best condition $h_m(m, \lambda) = \mu$ still pins down $m(\lambda)$ uniquely for any given g, and the insurer solves for the g that implements first-best effort at $\tau_1 = 0$ exactly as in the proof of Proposition 2. The resulting $g(m)$ is steeper than in the profit-maximizing case: more altruistic firms already internalize consumer welfare and therefore require stronger financial incentives to prevent overprovision. Since $\tau_1 = 0$ also provides strict risk protection, the planner always prefers $\tau_1^* = 0$ in this case.

This reasoning relies 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, developing a local approximation result for small heterogeneity in a two-type model; the same logic extends to N types with arbitrary positive weights by replacing averages with weighted averages.

Assume there are two doctor types $\beta_1, \beta_2 \in (0,1)$ with equal population weight. Doctor i chooses treatment intensity to maximize the altruistic objective,

$$m_i(\lambda) \in \arg\max_m \left\{ \beta_i \left[h(m,\lambda) + v(W - \tau_1 m - p) \right] + (1-\beta_i)(R - \tau_2 m) \right\}.$$

Let $\bar{m}(\tau_1,\tau_2) \equiv \frac{1}{2}\sum_i \mathbb{E}_\lambda[m_i(\lambda)]$ denote pooled expected spending, and suppose the common premium $p(\tau_1,\tau_2) = (1-\tau_1)\bar{m}(\tau_1,\tau_2)$ so the insurer's budget balances. The planner chooses (τ_1,τ_2) to maximize expected patient welfare,

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

Write $\bar{\lambda} \equiv \mathbb{E}[\lambda]$ for mean illness severity.

Doctor i's first-order condition is $\beta_i[h_m(m_i,\lambda) - \tau_1 v'(c_i)] = (1-\beta_i)\tau_2$. Under the local approximation $v'(\bar{c}) \approx 1$ (where $\bar{c}$ is the mean consumption level in the fully insured allocation), this simplifies to

$$h_m(m_i,\lambda) \;\approx\; \tau_1 + \tau_2 \,\mathrm{MRS}_i, \qquad \mathrm{MRS}_i \equiv \frac{1-\beta_i}{\beta_i}. \tag{35}$$

We refer to MRS_i as a marginal rate of substitution because it is precisely the rate at which doctor i substitutes consumer-side for firm-side cost-sharing in their locally linearized first-order condition: $\partial m_i/\partial \tau_2 = \mathrm{MRS}_i \, \partial m_i/\partial \tau_1$. Less altruistic doctors have higher MRS_i and respond more strongly to τ_2.

The remainder of the appendix works in the ratio-normalized quadratic specification $h_m(m,\lambda) = 1 - (m-\lambda)/(\omega\lambda)$ evaluated at $\lambda = \bar{\lambda}$, so that $h_{mm} \approx -1/(\omega\bar{\lambda})$ is common across types. If the insurer could observe β_i, Proposition 2 would deliver the constrained first best with $\tau_1 = 0$. Under unobservable altruism the insurer must offer a common (τ_1,τ_2) pair and doctors self-select through (35). When doctors differ in MRS_i, firm cost-sharing widens treatment dispersion because less-altruistic doctors cut back more aggressively; introducing some consumer cost-sharing at the expense of firm cost-sharing compresses that dispersion, at the cost of exposing patients to extra consumption risk. The proposition below derives the local welfare derivative and, as an immediate consequence, the

sign and magnitude of the optimal consumer coinsurance rate.

Proposition 1. *In the two-type model above, under the local approximation $v'(\bar{c}) \approx 1$ and the ratio-normalized quadratic h evaluated at mean severity, suppose the supply-side agency wedge is strictly positive ($\tau_2 > 0$), suppose the planner adjusts τ_2 as τ_1 varies so as to hold pooled expected spending $\bar{m}$ fixed, and suppose altruism heterogeneity is small in the sense that $R_A \tau_2^2 \operatorname{Var}(\mathrm{MRS})/h_{mm}^2 \ll 1$, where $R_A \equiv -v''(\bar{c})/v'(\bar{c})$ is the Arrow–Pratt coefficient of absolute risk aversion at mean consumption. Then the leading-order welfare derivative with respect to τ_1 is*

$$\frac{d\mathcal{W}}{d\tau_1} \;=\; \frac{\bar{v}'\,\tau_2 \operatorname{Var}(\mathrm{MRS})}{|h_{mm}|\,\mathbb{E}(\mathrm{MRS})} \;-\; R_A\,\bar{v}'\,\tau_1 \operatorname{Var}(\lambda) \;+\; o\big(\operatorname{Var}(\mathrm{MRS})\big), \tag{36}$$

where $\bar{v}' \equiv v'(\bar{c})$. Consequently, whenever $\beta_1 \neq \beta_2$ (so that $\operatorname{Var}(\mathrm{MRS}) > 0$), the planner's optimum on the domain $\tau_1 \geq 0$ is strictly interior, and is given to leading order by

$$\tau_1^* \;\approx\; \frac{\tau_2 \operatorname{Var}(\mathrm{MRS})}{R_A\,|h_{mm}|\,\mathbb{E}(\mathrm{MRS})\operatorname{Var}(\lambda)} \;=\; \frac{\omega\bar{\lambda}\,\tau_2 \operatorname{Var}(\mathrm{MRS})}{R_A\,\mathbb{E}(\mathrm{MRS})\operatorname{Var}(\lambda)}. \tag{37}$$

The two terms in (36) have a clean interpretation: the first is the screening benefit of consumer coinsurance (positive and proportional to cross-doctor heterogeneity in MRS), and the second is the insurance cost (negative and proportional to the risk-aversion-weighted severity variance). Equation (37) then follows by setting the derivative to zero and solving for τ_1^*. The sign result is a one-line consequence: the screening benefit is strictly positive at $\tau_1 = 0$ whenever $\operatorname{Var}(\mathrm{MRS}) > 0$ and $\tau_2 > 0$, while the insurance cost vanishes at $\tau_1 = 0$, so a small positive τ_1 strictly improves welfare.

Proof. Write $\delta\mathrm{MRS}_i \equiv \mathrm{MRS}_i - \mathbb{E}(\mathrm{MRS})$ and $\sigma^2 \equiv \operatorname{Var}(\mathrm{MRS})$. Throughout, "leading order" retains terms through $O(\sigma^2)$ and neglects terms of order $R_A\,\tau_2^2\,\sigma^2/h_{mm}^2 \cdot \sigma^2$, which is $o(\sigma^2)$ under the small-heterogeneity assumption. Let $x_i \equiv m_i - \bar{\lambda}$ and $\bar{x} \equiv (x_1 + x_2)/2$; evaluating (35) at $\lambda = \bar{\lambda}$ gives $x_i \approx \omega\bar{\lambda}(1 - q_i)$ with $q_i \equiv \tau_1 + \tau_2\,\mathrm{MRS}_i$.

Differentiating the doctor FOC (35) with respect to τ_1, treating $\tau_2 = \tau_2(\tau_1)$, gives $h_{mm}\,dx_i/d\tau_1 = 1 + \mathrm{MRS}_i\,d\tau_2/d\tau_1$. Imposing $d\bar{x}/d\tau_1 = \frac{1}{2}\sum_i dx_i/d\tau_1 = 0$ and summing across i yields $d\tau_2/d\tau_1 = -1/\mathbb{E}(\mathrm{MRS})$, and substituting back gives

$$\frac{dx_i}{d\tau_1} \;=\; \frac{-\delta\mathrm{MRS}_i}{h_{mm}\,\mathbb{E}(\mathrm{MRS})}. \tag{38}$$

Let $c_i^{\mathrm{eq}}(\lambda)$ denote the utility-equivalent scalar that summarizes, after the $v'(\bar{c}) \approx 1$ expansion, the

patient's welfare from treatment by doctor i. To leading order,

$$c_i^{\mathrm{eq}}(\lambda) \;=\; \bar{c} + \big[h(m_i,\bar{\lambda}) - h(\bar{m},\bar{\lambda}) - (x_i - \bar{x})\,\tau_1\big] - \tau_1(\lambda - \bar{\lambda}). \tag{39}$$

The object c_i^{eq} is not literal financial consumption; it absorbs the linearized health term $h(m_i,\bar{\lambda})$ so that welfare can be written as a single composition with v. Under the maintained approximation, the planner's first-order welfare derivative equals $\frac{1}{2}\sum_i \mathbb{E}_\lambda[v'(c_i^{\mathrm{eq}})\,dc_i^{\mathrm{eq}}/d\tau_1]$, and $c_i^{\mathrm{eq}} - \bar{c} = O(\sigma) + O(\tau_1)$. Differentiating (39) with respect to τ_1 (using $d\bar{x}/d\tau_1 = 0$) yields

$$\frac{dc_i^{\mathrm{eq}}}{d\tau_1} \;=\; A_i - (x_i - \bar{x}) - (\lambda - \bar{\lambda}), \qquad A_i \;\equiv\; \big(h_m(m_i,\bar{\lambda}) - \tau_1\big)\frac{dx_i}{d\tau_1} \;\approx\; \frac{-\tau_2\,\mathrm{MRS}_i\,\delta\mathrm{MRS}_i}{h_{mm}\,\mathbb{E}(\mathrm{MRS})}, \tag{40}$$

where we used $h_m(m_i,\bar{\lambda}) - \tau_1 \approx \tau_2\,\mathrm{MRS}_i$ from (35).

We evaluate $d\mathcal{W}/d\tau_1 = \frac{1}{2}\sum_i \mathbb{E}_\lambda[v'(c_i^{\mathrm{eq}})\,dc_i^{\mathrm{eq}}/d\tau_1]$ by expanding $v'(c_i^{\mathrm{eq}}) \approx \bar{v}' + v''(\bar{c})(c_i^{\mathrm{eq}} - \bar{c})$, where $\bar{v}' \equiv v'(\bar{c})$. The three terms in (40) contribute as follows.

The A_i term. Since A_i does not depend on λ, its leading contribution is $\frac{1}{2}\sum_i \bar{v}'\,A_i$. Using the identity $\frac{1}{2}\sum_i \mathrm{MRS}_i\,\delta\mathrm{MRS}_i = \sigma^2$ (valid because $\sum_i \delta\mathrm{MRS}_i = 0$ and $\frac{1}{2}\sum_i(\delta\mathrm{MRS}_i)^2 = \sigma^2$), this gives

$$\frac{1}{2}\sum_i \bar{v}'\,A_i \;=\; \frac{\bar{v}'\,\tau_2\,\sigma^2}{|h_{mm}|\,\mathbb{E}(\mathrm{MRS})}. \tag{41}$$

The subleading correction involves $v''(\bar{c})\,\mathbb{E}[c_i^{\mathrm{eq}} - \bar{c}]\,A_i$. Since $\mathbb{E}[c_i^{\mathrm{eq}} - \bar{c}] = O(\sigma)$ and $A_i = O(\sigma)$, the product is $O(\sigma^3)$ and contributes $O(R_A\,\tau_2^3\,\mathbb{E}(\mathrm{MRS})\,\sigma^2/h_{mm}^2)$ to the welfare derivative—$o(\sigma^2)$ under the small-heterogeneity assumption.

The $(x_i - \bar{x})$ term. At leading order this vanishes because $\sum_i (x_i - \bar{x}) = 0$. The next-order contribution is $-\frac{v''(\bar{c})}{2}\sum_i \tau_2\,\mathbb{E}(\mathrm{MRS})\,(x_i - \bar{x})^2$. Since $(x_i - \bar{x})^2 = \tau_2^2\,(\delta\mathrm{MRS}_i)^2/h_{mm}^2$, this equals $R_A\,\bar{v}'\,\tau_2^3\,\mathbb{E}(\mathrm{MRS})\,\sigma^2/h_{mm}^2$, the same order as the A_i correction and with the same sign; both are $o(\sigma^2)$ under the small-heterogeneity assumption and drop out at leading order. (Without that assumption, the formula for τ_1^* would pick up an additive correction proportional to $\tau_2^3\,\mathbb{E}(\mathrm{MRS})\,\sigma^2/|h_{mm}|^2$.)

The $(\lambda - \bar{\lambda})$ term. Here randomness in λ matters. Since $\partial c_i^{\mathrm{eq}}/\partial\lambda = -\tau_1$ (the doctor's choice is ex ante with respect to the patient's realized λ), a first-order Taylor expansion of v' in λ gives

$$-\frac{1}{2}\sum_i \mathbb{E}_\lambda\big[v'(c_i^{\mathrm{eq}}(\lambda))\,(\lambda - \bar{\lambda})\big] \;=\; -\frac{1}{2}\sum_i \mathrm{Cov}\big(v'(c_i^{\mathrm{eq}}),\lambda\big) \;\approx\; v''(\bar{c})\,\tau_1\,\mathrm{Var}(\lambda),$$

which is negative (welfare-reducing) since $v'' < 0$.

Collecting the two leading-order contributions and writing $v''(\bar{c}) = -R_A \bar{v}'$ gives the welfare derivative (36). When $\mathrm{Var}(\mathrm{MRS}) > 0$ and $\tau_2 > 0$, the first term of (36) is strictly positive at $\tau_1 = 0$ while the second vanishes, so $d\mathcal{W}/d\tau_1|_{\tau_1=0} > 0$ and the optimum is interior. Setting $d\mathcal{W}/d\tau_1 = 0$ and solving for τ_1^* yields (37). □

The formula (37) decomposes the optimal coinsurance rate into four interpretable components. The numerator scales with $\mathrm{Var}(\mathrm{MRS})$, which captures the screening benefit of consumer cost-sharing: greater cross-doctor heterogeneity in altruism raises the return to demand-side price signals. In the denominator, $|h_{mm}| = 1/(\omega\bar{\lambda})$ measures the curvature of the health production function at mean severity—steeper curvature means doctors' treatment choices are less price-responsive, which reduces the screening benefit and hence τ_1^*. The factor $R_A \mathrm{Var}(\lambda)$ captures the insurance cost of coinsurance: higher absolute risk aversion or greater illness-severity dispersion makes consumer cost-sharing more costly in welfare terms. Finally, the factor $\tau_2/\mathbb{E}(\mathrm{MRS})$ scales the agency wedge that consumer cost-sharing is screening: a larger supply-side copay amplifies the treatment-intensity distortion across doctor types and therefore raises the screening benefit. When altruism is homogeneous, $\mathrm{Var}(\mathrm{MRS}) = 0$ and the formula collapses to $\tau_1^* = 0$, matching the homogeneous-altruism result above.

S8 Calibration Details for Section 4

Under the logit demand model of Section 3, the physician's symmetric-equilibrium problem (equation 5) can equivalently be represented as a behavioral objective:

$$\max_{m \geq k} \quad h(m, \lambda) + v(W - p - \tau_1 m) + \sigma \ln\big[A \cdot \mathrm{Rev}(m) - m\big], \tag{42}$$

where $A \cdot \mathrm{Rev}(m) - m$ is per-patient profit. 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. A change in A scales the revenue function proportionally, leaving its slope unchanged. A enters only when computing the Clemens–Gottlieb elasticity for Moment 2 (Section 4.4).

The linear revenue schedule $\mathrm{Rev}(m) = R + (1 - \tau_2)\, m$ 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. Per-patient profit is $\pi(m) =$

$R - \tau_2 m$. Given structural parameters (R, τ_2), latent effort is recovered from observed revenue as:

$$m_i = \frac{\text{Rev}_i - R}{1 - \tau_2}. \tag{43}$$

Under a linear contract $g(m) = R - \tau_2 m$, the physician's first-order condition in a symmetric equilibrium is:

$$h_m(m, \lambda) - \tau_1 v'(c) = \frac{\tau_2 \kappa}{R - \tau_2 m}, \tag{44}$$

where $c = W - p - \tau_1 m$ is financial consumption.

Specializing equation (44) to the log-ratio h of equation (8) and evaluating at baseline $A = 1$, the FOC for an interior optimum ($m > k$) becomes:

$$\left(1 - \frac{1}{\omega}\right) + \frac{\lambda + 1}{\omega\,(m+1)} - \tau_1 v'(c) - \frac{\sigma\,\tau_2(1-\tau_2)}{R - \tau_2\,\text{Rev}} = 0, \tag{45}$$

where $m = (\text{Rev} - R)/(1 - \tau_2)$. Inverting the FOC at $\tau_1 = 0$ (the Medigap status quo, at which $v'(c)$ drops out and the inversion is independent of risk aversion γ) yields a closed-form mapping from observed revenue to patient type for each episode with $\text{Rev}_i > \text{Rev}_k$:

$$\lambda_i = \left(\frac{\text{Rev}_i - R}{1 - \tau_2} + 1\right)\left[1 - \omega\left(1 - \frac{\sigma\,\tau_2(1-\tau_2)}{R - \tau_2\,\text{Rev}_i}\right)\right] - 1. \tag{46}$$

The remaining 16.9% of episodes are bunched at the lower bound $m = k$; their types fall in an interval $[\lambda_2, \lambda_1]$ that cannot be identified from revenue alone, so we impose:

Assumption 1 (Uniform bunched types). *The density $f(\lambda)$ is constant on $[\lambda_2, \lambda_1]$.*

We adopt CARA preferences, $v(c) = -\frac{1}{\gamma}\exp(-\gamma\, c)$, so that $v'(c) = \exp(-\gamma\, c)$. At a reference consumption level $\bar{c}$, the effective patient price in the FOC (45) is $\tau_1 \exp(-\gamma\, \bar{c})$ rather than τ_1. We normalize $v'(\bar{c}) = 1$ (i.e. $\bar{c}$ corresponds to the certainty-equivalent consumption under the status-quo contract).

The model in principle accommodates income effects via $v'(\cdot)$ term (Acquatella and Marone, 2025; Chetty, 2008), with income effects identified by the degree of risk aversion. When $\tau_1 = 0$ (as in the status quo for Medigap enrollees), the $v'(c)$ term drops out entirely and the FOC is independent of risk aversion. For that moment, income effects are absent . For the Cabral–Mahoney counterfactual ($\tau_1 = 0.20$), the calibration uses the approximation $v'(c) \approx v'(\bar{c}) = 1$, so that the structural parameters (ω, τ_2) are approximately invariant to the assumed level of risk aversion γ. Risk aversion enters

primarily through the welfare evaluation and through the shadow value μ, which determines the constrained-first-best treatment rule and the optimal contract.

The upper cutoff λ_1 is the type at which the physician's interior choice equals k—types above λ_1 receive $m > k$, while types below are constrained to $m = k$:

$$\lambda_1 = (k+1)\left[1 - \omega\left(1 - \tau_1(1-\tau_2) - \frac{\sigma\,\tau_2(1-\tau_2)}{R - \tau_2\,\mathrm{Rev}_k}\right)\right] - 1, \tag{47}$$

where Rev_k is the revenue at minimum effort. The lower cutoff λ_2 is the marginal type who is just willing to visit at $m = k$. The visit condition requires $h(k,\lambda) - h(0,\lambda) \geq \tau_v + K_{\text{visit}}$ (where $h(0,\lambda)$ is equation (8) evaluated at $m = 0$), and $K_{\text{visit}} \geq 0$ is a nonfinancial visit cost (e.g. travel time, hassle of scheduling).

A type λ visits if the incremental health surplus from a visit, net of out-of-pocket spending and the visit copay, exceeds the visit cost: $h(m^\star(\lambda),\lambda) - h(0,\lambda) - \tau_1\,\mathrm{Rev}(\lambda) \geq \tau_v + K_{\text{visit}}$, where $m^\star(\lambda)$ is the physician's chosen treatment level and K_{visit} is a reduced-form intercept that absorbs the logit convenience term and other nonfinancial visit costs. Under the assumed h this visit surplus is strictly increasing in λ, so the 14.5% of visitors who exit when τ_1 moves from 0 to 0.20 are the lowest-λ types, all drawn from the bunched region (which comprises 16.9% of visitors).

We identify f, K_{visit}, λ_2, and the no-Medigap dropout cutoff λ_{drop} in four steps:

1. Under Assumption 1 the type density is constant at f on $[\lambda_2, \lambda_1]$. By continuity of the density at λ_1, the empirical density of the first few above-kink centiles (types just above λ_1) identifies f.

2. Let $\varepsilon_k = 0.169$ denote the at-kink episode share and ext $= 0.145$ the Brot-Goldberg et al. (2017) dropout share. Under uniform density f, the marginal dropout type λ_{drop} satisfies $f \cdot (\lambda_1 - \lambda_{\text{drop}}) = \varepsilon_k - \text{ext}$, so:

$$\lambda_{\text{drop}} = \lambda_1 - \frac{\varepsilon_k - \text{ext}}{f}. \tag{48}$$

   Types in $[\lambda_{\text{drop}}, \lambda_1]$ continue to visit under the no-Medigap counterfactual; types in $[\lambda_2, \lambda_{\text{drop}})$ drop out.

3. The type λ_{drop} is indifferent about visiting under the physician's no-Medigap treatment rule ($\tau_1^{\text{cf}} = 0.20$):

$$K_{\text{visit}} = h\big(m^{\text{cf}}(\lambda_{\text{drop}}), \lambda_{\text{drop}}\big) - h\big(0, \lambda_{\text{drop}}\big) - \tau_1^{\text{cf}} \cdot \mathrm{Rev}^{\text{cf}}\big(\lambda_{\text{drop}}\big), \tag{49}$$

   where m^{cf} and Rev^{cf} are the physician's no-Medigap best response.

4. The total bunching mass pins down λ_2:

$$\lambda_2 = \lambda_1 - \frac{\varepsilon_k}{f}. \tag{50}$$

Appendix Table 1: Sensitivity of structural estimates to Rev_k and k/Rev_k

Rev_k (\$)	k/Rev_k	k (\$)	R (\$)	$\hat{\tau}_2$	$\hat{\omega}$
Panel A: Rev_k = \$50 *(100 above-kink centiles, 10.7% at kink)*					
50	0.50	25.0	24.5	−0.020	0.389
50	0.60	30.0	19.3	−0.024	0.385
50	0.70	35.0	14.0	−0.029	0.380
50	0.80	40.0	8.6	−0.034	0.376
50	0.90	45.0	3.4	−0.035	0.374
Panel B: Rev_k = \$73 *(100 above-kink centiles, 16.9% at kink)* [baseline]					
73	0.50	36.5	36.0	−0.013	0.390
73	0.60	43.8	28.5	−0.016	0.385
73	0.70	51.1	20.8	−0.021	0.379
73	0.80	58.4	13.0	−0.027	0.373
73	0.90	65.7	5.2	−0.032	0.368
Panel C: Rev_k = \$100 *(88 above-kink centiles, 28.8% at kink)*					
100	0.50	50.0	49.7	−0.007	0.387
100	0.60	60.0	39.4	−0.010	0.381
100	0.70	70.0	29.0	−0.014	0.376
100	0.80	80.0	18.5	−0.019	0.369
100	0.90	90.0	7.8	−0.025	0.363
Panel D: Rev_k = \$150 *(64 above-kink centiles, 47.8% at kink)*					
150	0.50	75.0	74.9	−0.001	0.383
150	0.60	90.0	59.7	−0.004	0.377
150	0.70	105.0	44.3	−0.007	0.372
150	0.80	120.0	28.6	−0.012	0.365
150	0.90	135.0	12.5	−0.018	0.358
Panel E: Rev_k = \$200 *(55 above-kink centiles, 55.2% at kink)*					
200	0.50	100.0	100.1	+0.001	0.375
200	0.60	120.0	80.0	−0.000	0.370
200	0.70	140.0	59.6	−0.003	0.364
200	0.80	160.0	38.9	−0.007	0.357
200	0.90	180.0	17.5	−0.014	0.349

Notes: Each row solves the moment equations described in Section 4.4. At-kink fractions are the share of MarketScan episodes with reported revenues at or below Rev_k. Above-kink centiles use the full 100-centile distribution (MarketScan, above \$50) for Panels A–B and the subset above Rev_k for Panels C–E.

S9 Calibration of σ

We calibrate σ from evidence on the market share of new physician episodes. Sinaiko and Rosenthal (2014) study tiered physician networks in the Massachusetts Group Insurance Commission (GIC). Because 53% of physicians appeared in at least two GIC plans with different tier rankings, regres-

sions that include physician fixed effects together with plan, year, and specialty controls isolate the within-physician response to tier-induced copay variation, largely removing quality and reputation as confounders. The within-physician estimates imply approximately a 12% reduction in new-patient market share for physicians facing \$10–\$20 higher office-visit copays, which translates to \$30–\$60 higher episode copays with an average of 3 visits per episode. Mapping this through the logit formula $\Delta \ln s = -\Delta c/\sigma$ gives:

$$\sigma \approx \frac{\Delta c_{\text{episode}}}{|\Delta \ln s|} = \frac{30\text{–}60}{|\ln 0.88|} \approx \$235\text{–}\$470 \text{ per episode.}$$

We use $\sigma = \$255$ per episode as our baseline, corresponding to the low end of this range, since this already implies a substantial amount of physician market power.[27]

Two independent sources corroborate this calibration using hospital choice or hypothetical physician choices. Prager (2020) estimates a discrete-choice model of hospital demand with hospital fixed effects and quality controls, finding demand elasticities of -0.04 to -0.16 with respect to tiered copays—an order of magnitude consistent with σ in the hundreds of dollars, although the setting is hospitals rather than physicians. Sinaiko (2011) reports a complementary stated-preference experiment: a \$35 copay differential increased the share choosing a lower-copay physician by 3.5–11.7%. Because the scenario is experimentally assigned, it is clean on the quality-held-fixed dimension, but as stated rather than revealed preference we treat it as an auxiliary bound.

S10 Derivation of Approximate Moment Conditions

We derive the approximate closed-form expressions for Moments 1 and 2 stated in the identification paragraph.

Notation and setup

Under the health benefit function (8), the marginal benefit is $h_m(m, \lambda) = (1 - 1/\omega) + (\lambda + 1)/(\omega(m + 1))$. Substituting into the physician's first-order condition (45) and solving yields the equilibrium

[27] This dollar-denominated calibration is consistent with the structural model under the normalization $v'(\bar{c}) = 1$: with CARA utility, the logit formula becomes $\Delta \ln s = -v'(\bar{c})\,\Delta c/\sigma$, which reduces to $-\Delta c/\sigma$ at the normalized reference consumption.

service level for interior types ($m > k$):

$$m(\lambda) \;=\; \frac{\lambda+1}{1-\omega\big(1-\tau_1(1-\tau_2)-P\big)} - 1, \tag{51}$$

where:

$$P \;\equiv\; \frac{\sigma\,\tau_2(1-\tau_2)}{(1-1/J)\,(R-\tau_2\,\mathrm{Rev}(m))}$$

encodes how fiscal parameters affect physician behavior through the profit margin, and $\tau_1(1-\tau_2)$ is the effective consumer cost-sharing rate on effort m (since τ_1 is applied to revenue). Revenue for an interior type is $\mathrm{Rev}(m) = R + (1-\tau_2)m$.

Derivation of (51). The FOC states:

$$\left(1-\frac{1}{\omega}\right) + \frac{\lambda+1}{\omega\,(m+1)} \;=\; \tau_1(1-\tau_2) + \frac{\sigma\,\tau_2(1-\tau_2)}{(1-1/J)(R-\tau_2\,\mathrm{Rev})}.$$

Isolating $(\lambda+1)/(\omega(m+1))$:

$$\frac{\lambda+1}{\omega\,(m+1)} \;=\; \frac{1}{\omega} - 1 + \tau_1(1-\tau_2) + P \;=\; \frac{1}{\omega}\big[1-\omega\big(1-\tau_1(1-\tau_2)-P\big)\big],$$

so $m+1 = (\lambda+1)/[1-\omega(1-\tau_1(1-\tau_2)-P)]$, yielding (51). Note that P itself depends on m through $R-\tau_2\,\mathrm{Rev}(m)$; for small $|\tau_2|$, P varies only weakly across types. Since m and λ are denominated in dollars, the "+1" shift is negligible in the calibration-relevant range: $(m+1)/m$ and $(\lambda+1)/\lambda$ differ from 1 by less than 4% even at the smallest types.

Moment 1: Revenue change from raising τ_1

In the data, Medigap enrollees face $\tau_1 = 0$; the counterfactual removes supplemental insurance, setting $\tau_1 = 0.20$. We seek the intensive-margin percentage change in mean revenue.

From (51), the ratio of counterfactual to observed shifted service levels for a given type λ is:

$$\frac{m^{\mathrm{cf}}(\lambda)+1}{m^{\mathrm{data}}(\lambda)+1} \;=\; \frac{1-\omega(1-P^{\mathrm{data}})}{1-\omega(1-0.20(1-\tau_2)-P^{\mathrm{cf}})},$$

since $\lambda+1$ cancels. Since m is in dollars and $m \gg 1$, $(m+1)/m \approx 1$ and the ratio of shifted levels approximates the ratio of levels themselves. The key approximation is that P is approximately constant across the change in τ_1: $P^{\mathrm{cf}} \approx P^{\mathrm{data}}$. This holds because P depends on τ_1 only indirectly through m, and the shift in m has a second-order effect on P when $|\tau_2|$ is small (since P is proportional to τ_2).

Writing $\delta \equiv 0.20(1-\tau_2)\,\omega$ and $D \equiv 1-\omega(1-P)$, the ratio simplifies to:

$$\frac{m^{\text{cf}}}{m^{\text{data}}} \;\approx\; \frac{D}{D+\delta} \;=\; \frac{1}{1+\delta/D} \;\approx\; 1-\frac{\delta}{D}.$$

For small P, $D \approx 1-\omega$, so the proportional change in m is:

$$\frac{m^{\text{cf}}}{m^{\text{data}}} - 1 \;\approx\; -\frac{0.20(1-\tau_2)\,\omega}{1-\omega}. \tag{52}$$

The proportional scaling is approximately the same for every interior type λ.

Mean revenue among interior types changes as follows. Since $\text{Rev}(m) = R + (1-\tau_2)m$,

$$\begin{aligned}\overline{\text{Rev}}^{\text{cf}} - \overline{\text{Rev}}^{\text{data}} \;&=\; (1-\tau_2)\left(\bar{m}^{\text{cf}} - \bar{m}^{\text{data}}\right) \\ &\approx\; -(1-\tau_2)\,\bar{m}^{\text{data}} \cdot \frac{0.20(1-\tau_2)\,\omega}{1-\omega}.\end{aligned}$$

Dividing by $\overline{\text{Rev}}^{\text{data}} = R + (1-\tau_2)\bar{m}^{\text{data}}$:

$$\frac{\overline{\text{Rev}}^{\text{cf}} - \overline{\text{Rev}}^{\text{data}}}{\overline{\text{Rev}}^{\text{data}}} \;\approx\; -\frac{0.20(1-\tau_2)\,\omega}{1-\omega}\left(1 - \frac{R}{\overline{\text{Rev}}^{\text{data}}}\right), \tag{53}$$

where the last step uses $1 - R/\text{Rev} = (1-\tau_2)m/\text{Rev}$, i.e., the variable component of revenue as a share of total revenue.

Expression (53) is exact when $\tau_2 = 0$ (so that $P = 0$ regardless of τ_1). For $|\tau_2|$ small, the error is of order $\tau_2 \times \Delta\tau_1$: the change in P induced by the shift in m is proportional to τ_2 times the change in m, which is itself proportional to $\Delta\tau_1$. At our estimates ($\hat{\tau}_2 \approx -0.016$, $\Delta\tau_1 = 0.20$), the approximation error is on the order of 0.3%, negligible relative to the $\sim 13\%$ moment.

Moment 2: Revenue elasticity with respect to conversion factor A

The conversion factor A scales the entire revenue schedule: a uniform percentage increase in A raises all payment components proportionally, so that:

$$\text{Rev}_A(m) \;=\; A \cdot \text{Rev}(m) \;=\; A\left[R + (1-\tau_2)m\right].$$

This corresponds to the Medicare fee schedule experiment in **?**. Physician profit per patient at scale A is $\pi_A(m) = A\,\text{Rev}(m) - m = A\left[R + (1-\tau_2)m\right] - m$, which at $A = 1$ reduces to $R - \tau_2 m$ as in the base

model. The marginal profit with respect to m at scale A is $A(1-\tau_2)-1$; at $A=1$ this equals $-\tau_2$. We evaluate the elasticity at $A=1$ and $\tau_1 = 0.20$ (since the Clemens–Gottlieb estimate is identified on a predominantly non-Medigap population).

Define the effective cost-sharing rate at scale A:

$$\tau_2^A \;\equiv\; 1 - A(1-\tau_2),$$

so that $\tau_2^{A=1} = \tau_2$ and the marginal profit is $-\tau_2^A$. The physician's first-order condition at scale A takes the same form as in the base model with τ_2 replaced by τ_2^A and profit $R - \tau_2 m$ replaced by $\pi_A = A\,\mathrm{Rev}(m) - m$. Thus the parameter P generalizes to:

$$P_A \;=\; \frac{\sigma\,\tau_2^A}{(1-1/J)\,\pi_A},$$

and the equilibrium service level is $m(\lambda; A) = (\lambda+1)/[1-\omega(1-\tau_1(1-\tau_2)-P_A)] - 1$.

Differentiating $\ln m$ with respect to $\ln A$ at $A=1$:

$$\left.\frac{d\ln m}{d\ln A}\right|_{A=1} \;=\; \frac{\omega}{1-\omega(1-\tau_1(1-\tau_2)-P)}\left.\frac{dP_A}{d\ln A}\right|_{A=1}. \tag{54}$$

To evaluate $dP_A/d\ln A$, we apply the quotient rule holding m fixed (the indirect effect of A through m contributes a higher-order correction proportional to τ_2^2). We need two derivatives at $A=1$:

$$\left.\frac{d\tau_2^A}{d\ln A}\right|_{A=1} \;=\; A\left.\frac{d\tau_2^A}{dA}\right|_{A=1} \;=\; -(1-\tau_2) \;\approx\; -1,$$

$$\left.\frac{d\pi_A}{d\ln A}\right|_{A=1} \;=\; A\left.\frac{d}{dA}\big[A\,\mathrm{Rev}(m)-m\big]\right|_{A=1} \;=\; \mathrm{Rev}(m) \;=\; R+(1-\tau_2)m \;\approx\; R+m,$$

where the final approximations drop terms of order τ_2. Applying the quotient rule to $P_A = \sigma\,\tau_2^A/[(1-1/J)\pi_A]$:

$$\begin{aligned}\left.\frac{dP_A}{d\ln A}\right|_{A=1} &= \frac{\sigma}{(1-1/J)}\cdot\frac{\frac{d\tau_2^A}{d\ln A}\cdot\pi_1 \;-\; \tau_2\cdot\frac{d\pi_A}{d\ln A}}{\pi_1^2}\\ &\approx \frac{\sigma}{(1-1/J)}\cdot\frac{(-1)(R-\tau_2 m) \;-\; \tau_2(R+m)}{(R-\tau_2 m)^2}.\end{aligned}$$

The numerator simplifies: $-(R-\tau_2 m)-\tau_2(R+m) = -R+\tau_2 m-\tau_2 R-\tau_2 m = -R(1+\tau_2) \approx -R$, so:

$$\left.\frac{dP_A}{d\ln A}\right|_{A=1} \approx -\frac{\sigma R}{(1-1/J)(R-\tau_2 m)^2}.$$

Substituting into (54) and writing $D_\tau \equiv 1-\omega(1-\tau_1(1-\tau_2)-P)$ for the denominator of the equilibrium service level:

$$\left.\frac{d\ln m}{d\ln A}\right|_{A=1} \approx \frac{\omega}{D_\tau}\cdot\frac{\sigma R}{(1-1/J)(R-\tau_2 m)^2}. \tag{55}$$

For small τ_1 and P, $D_\tau \approx 1-\omega$; in the main text we use this further simplification. At the exact evaluation point $\tau_1 = 0.20$ (non-Medigap population), $D_\tau \approx 1-\omega(1-0.20) = 1-0.8\omega$, which differs from $1-\omega$ by $0.2\omega \approx 0.08$ at $\hat{\omega} = 0.385$.

Mean revenue is $\overline{\text{Rev}}_A = A\left[R+(1-\tau_2)\bar{m}(A)\right]$, so:

$$\left.\frac{d\ln \overline{\text{Rev}}_A}{d\ln A}\right|_{A=1} = 1 + \frac{(1-\tau_2)\bar{m}}{\overline{\text{Rev}}}\cdot\frac{d\ln\bar{m}}{d\ln A}.$$

The first term (= 1) is the mechanical effect of the fee increase holding behavior fixed: since A scales the entire revenue schedule, a 1% increase in A raises revenue by 1% mechanically. The second term is the behavioral response. Using (55) and $\frac{(1-\tau_2)\bar{m}}{\overline{\text{Rev}}} = 1-\frac{R}{\overline{\text{Rev}}}$:

$$\left.\frac{d\ln \overline{\text{Rev}}}{d\ln A}\right|_{A=1} \approx 1 + \left(1-\frac{R}{\overline{\text{Rev}}^{\text{cf}}}\right)\frac{\omega}{D_\tau}\cdot\frac{\sigma R}{(1-1/J)(R-\tau_2\bar{m})^2}, \tag{56}$$

where $\overline{\text{Rev}}^{\text{cf}}$ denotes mean revenue in the non-Medigap population ($\tau_1 = 0.20$).

At $\tau_2 = 0$, the denominator equals R^2, $D_\tau = 1-0.8\omega$ at $\tau_1 = 0.20$, and the behavioral term is:

$$\left(1-\frac{R}{\overline{\text{Rev}}^{\text{cf}}}\right)\frac{\omega}{1-0.8\omega}\cdot\frac{\sigma}{(1-1/J)R} \approx \frac{41}{114}\cdot\frac{0.385}{0.692}\cdot\frac{250}{73} \approx 0.36\times 0.556\times 3.42 \approx 0.68,$$

giving a total elasticity of approximately $1+0.68 = 1.68$. This exceeds the **?** estimate of 1.45, so the behavioral response must be attenuated. Fat margins ($\tau_2 < 0$) accomplish this: when $\tau_2 < 0$, each dollar of spending generates profit even at the margin, which reduces the physician's incentive to respond to the conversion factor. Formally, the denominator $(R-\tau_2\bar{m})^2$ increases (since $-\tau_2\bar{m} > 0$), lowering the behavioral term and bringing the total elasticity toward 1.45. This is what pins $\hat{\tau}_2 < 0$.

Expression (56) is exact in the limit $J \to \infty$ (competitive limit, so $1-1/J \to 1$) and when P_A is constant across types (which holds approximately when $|\tau_2\bar{m}/R|$ is small). At our estimates ($\hat{\tau}_2 \approx$

-0.016, $\bar{m} \approx 41$, $R \approx 28.5$), $|\tau_2 \bar{m}/R| \approx 0.023$, so the approximation is tight. The neglected indirect effect of A on m in the quotient-rule computation of $dP_A/d\ln A$ contributes terms of order $\tau_2(R+m)/(R-\tau_2 m)^2 \approx \tau_2/R$; the resulting error in the elasticity is of order τ_2^2, which is less than 0.03%.

S11 Altruism and the Identification of τ_2

The baseline model assumes that physicians maximize profit. A natural alternative is that physicians are partially altruistic, placing weight on patient welfare in addition to own profit. Because revenue enters only through the profit component of the altruistic objective, altruism attenuates the revenue elasticity (Moment 2) and could in principle mimic fat margins ($\tau_2 < 0$). However, patient cost-sharing τ_1 enters *both* the profit and patient-welfare components, so altruism simultaneously *amplifies* the copay elasticity (Moment 1). The net effect is to inflate the Moment 1/Moment 2 ratio, which can only be offset by making τ_2 *more* negative—not less.

An altruistic physician maximizes:

$$V_\alpha(m) \;=\; (1-\alpha)\, s_j(m,\lambda)\,(R-\tau_2 m) \;+\; \alpha\left[h(m,\lambda) - \tau_1 \,\mathrm{Rev}(m)\right], \tag{57}$$

where $\alpha \in [0,1]$ is the altruism weight. Evaluating the first-order condition at the symmetric equilibrium using the logit demand structure (Section 2) and defining:

$$\Gamma \;\equiv\; (1-\alpha)\,\frac{(1-1/J)}{\sigma}(R-\tau_2 m) \;+\; \alpha,$$

the FOC can be written as:

$$\left[h_m - \tau_1(1-\tau_2)\right]\Gamma \;=\; (1-\alpha)\,\tau_2. \tag{58}$$

The revenue shifter A (with $R \to AR$) enters Γ only through the profit-side term $(1-\alpha)\frac{(1-1/J)}{\sigma}(R-\tau_2 m)$. Implicitly differentiating (58):

$$\left.\frac{dm}{dA}\right|_{\alpha} \;=\; \left.\frac{dm}{dA}\right|_{\alpha=0} \times \frac{(1-\alpha)\,\frac{(1-1/J)}{\sigma}(R-\tau_2 m)}{\Gamma}, \tag{59}$$

which is strictly less than the selfish response for $\alpha > 0$: altruism attenuates Moment 2. By contrast, τ_1 enters *both* Γ components—profit (via demand) and patient welfare (via the out-of-pocket cost

τ_1 Rev). The analogous expression is:

$$\left.\frac{dm}{d\tau_1}\right|_{\alpha} = \left.\frac{dm}{d\tau_1}\right|_{\alpha=0} \times \frac{\Gamma}{(1-\alpha)\,\frac{(1-1/J)}{\sigma}(R-\tau_2 m)}, \tag{60}$$

the *reciprocal* of the Moment 2 factor, exceeding one for $\alpha > 0$: altruism amplifies Moment 1.

Define the amplification factor:

$$F(\alpha,\tau_2) \equiv \frac{\Gamma}{(1-\alpha)\,\frac{(1-1/J)}{\sigma}(R-\tau_2 m)} = 1 + \frac{\alpha\,\sigma}{(1-\alpha)(1-1/J)(R-\tau_2 m)}. \tag{61}$$

Moment 1 is scaled up by F and Moment 2 is scaled down by F, so their ratio is scaled by $F^2 > 1$ for any $\alpha > 0$. Since F is decreasing in $(R-\tau_2 m)$, the only way to restore the observed ratio under altruism is to raise $R-\tau_2 m$—i.e., to make τ_2 more negative (fatter margins). Altruism therefore *reinforces* the case for fat margins rather than substituting for them.

At baseline estimates ($R \approx \$28.49$, $\hat{\tau}_2 \approx -0.016$, $\sigma = \$250$, J large), $F \approx 1+\alpha\,\sigma/[(1-\alpha)(R-\tau_2\bar{m})]$. At a representative above-kink spending level $\bar{m} \approx 170$, $R-\tau_2\bar{m} \approx 31$ and $F \approx 1+8.1\,\alpha/(1-\alpha)$. Even $\alpha = 0.10$ gives $F \approx 1.90$ and inflates the moment ratio by $F^2 \approx 3.6$; restoring fit requires substantially more negative τ_2.

S12 Computation of the Optimal Physician Contract

This appendix records computational details for the optimal nonlinear contract $g(m)$ from Section 4.5.

The constant C in equation (13) is chosen so that the expected per-patient profit to physicians equals the status-quo average profit $\bar{\pi}$ (which equals $R-\tau_2\bar{m}$ under the linear contract), yielding the closed form:

$$C = \frac{\bar{\pi}}{\mathbb{E}[\exp(-m/\sigma_{\text{eff}}) \mid \text{visit}]}, \qquad \sigma_{\text{eff}} = \frac{\sigma}{\mu}, \tag{62}$$

where the expectation is taken over the equilibrium distribution of m among visiting patients under the optimal contract.

Under CARA utility $v(c) = -\gamma^{-1}\exp(-\gamma c)$ with the normalization $v'(\bar{c}) = 1$, the shadow value of a premium dollar from Proposition 2 is:

$$\mu = P(\text{visit})\,\exp\big(-\gamma\,(W-p-\tau_v)\big) + \big(1-P(\text{visit})\big)\,\exp\big(-\gamma\,(W-p)\big).$$

When risk aversion is small ($\gamma \to 0$), $\mu \to 1$ and the optimal contract coincides with the quasi-linear

benchmark. At higher γ the shadow value falls below one, flattening the profit schedule and shifting the optimal m^{opt}/λ ratio above one (Table 4).

For each γ we compute the optimum by a joint search over (a) the visit-margin cutoff j_{cut} on the ordered support of the 200-type discretization (100 bunching grid points plus 100 above-kink centiles), (b) the intensive-margin ratio ratio $= m^{\text{opt}}(\lambda)/\lambda$ above the kink, and (c) the visit copay τ_v, grid-searched within the IC-feasible interval $[\tau_v^{\min}(j_{\text{cut}}, \text{ratio}), \tau_v^{\max}(j_{\text{cut}}, \text{ratio})]$ defined by the exclusion IC (lower bound) and the participation IC (upper bound). Welfare is evaluated on the status-quo visiting support $\{\lambda \geq \lambda_2\}$, since types below λ_2—non-visitors under the status quo—are not separately identified from the episode-level moments used in calibration. The welfare expression subtracts the nonfinancial visit cost K_{visit} from each visiting type (as a real time/hassle cost of visiting a physician), so deterring a visitor saves the insurer $k - \tau_v$ per deterred visit at zero first-order welfare cost to the marginal deterred type. ΔW in Table 4 is therefore the per-episode equivalent variation relative to status-quo Medigap, conditional on the status-quo visiting support rather than full-population ex ante welfare. At our baseline calibration the unconstrained optimum deters roughly 12.5% of status-quo visitors—the lowest-λ types in the bunched region—at a visit copay of $\tau_v \approx \$39.45$ at the mean Handel-Kolstad CARA coefficient.

S13 Physician Risk Exposure Under the Optimal Contract

In the status quo, physician contracts expose them to earnings risk from most patients being healthy; in the optimal contract, the risk they face is that patients will be sick. 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.

To compare these risks, we simulate physician income distributions under both regimes. For each physician–patient–episode in the Medicare data, we compute actual revenue in the status quo and counterfactual compensation under $g^*(m)$ (rescaled to be revenue-neutral on average). We then construct year-level income distributions by bootstrapping episodes within physician, scaling up to the full patient panel using the ratio of average Medicare caseload to observed MarketScan caseload. We restrict to physicians with at least 20 observed episodes per year. Expected utility is computed under CARA utility for each risk-aversion parameter, and we report the certainty-equivalent income difference between the two regimes.

Appendix Table 2: Physician certainty-equivalent income change under optimal contract

CARA coefficient r	Total Δ CE ($/yr)	Total Δ CE ($/ep)	Risk-only Δ CE ($/yr)	Risk-only Δ CE ($/ep)
$r = 0$ (risk-neutral)	−5	+4.58	+0	+0.00
$r = 1 \times 10^{-4}$	+153	+4.78	+158	+0.20
$r = 3 \times 10^{-4}$	+308	+4.96	+313	+0.39
$r = 4 \times 10^{-4}$	+344	+5.01	+349	+0.43
$r = 1 \times 10^{-3}$	+398	+5.04	+403	+0.46

Notes: Certainty-equivalent income change is the dollar amount by which a physician under the optimal contract $g^*(m)$ is better (positive) or worse (negative) off relative to the status quo, under CARA utility with coefficient r. Both regimes are normalized to the same mean physician income. The *total* columns report the CE change on raw yearly income; the *risk-only* columns report the CE change after demeaning each physician's bootstrap distribution within regime, so they isolate the pure risk-premium channel from cross-physician redistribution (translation invariance of CARA makes this an exact decomposition: $\Delta^{\text{total}}_{j,r} = \Delta^{\text{mean}}_{j} + \Delta^{\text{risk}}_{j,r}$). The per-episode columns divide each physician's yearly CE change by their scaled episode count before averaging across physicians. Episodes are bootstrapped 100 times within physician; observed caseload is scaled up by the ratio of average Medicare new outpatient visits per physician to the MarketScan sample rate. Physicians with fewer than 20 observed episodes per year are excluded.

Table 2 reports results for five values of the CARA coefficient. The bottom-line is that the optimal contract *reduces* physician risk relative to the status quo, but the gains on a per episode basis are small relative to the benefits for insurers and patients in Table 4.[28]

S14 Comparison with Cabral-Mahoney Externality Estimates

This appendix relates the Medicare cost-saving numbers in Table 4 to the externality estimates in Cabral and Mahoney (2019) (henceforth CM), and uses the comparison to construct rough aggregate Medicare savings under each policy.

CM define the Medigap externality on Medicare as the additional Medicare outlay caused by Medigap-induced over-utilization of Part B services, equivalently the Medicare savings that would result if Medigap were eliminated. In the notation of our model this is exactly Δ Medicare in Panel A of Table 4: the reduction in Medicare's 80% share of allowed charges when τ_1 moves from 0 (the Medigap baseline) to 0.20 and physicians re-optimize treatment intensity at the status-quo $\hat{\tau}_2$. Our calibration is disciplined by CM's −14.7% revenue moment (equation (9)), so Panel A is, by construction, our model's estimate of the CM externality. The other 20% of the Δ Total cost column, the gap between Δ Medicare and Δ Total cost, is the mechanical shift of Medigap's 20% share onto patient out-of-pocket spending, which is not an externality on Medicare in CM's sense.

Cabral and Mahoney (2019) report that Medigap raises Part B physician spending by $454 per

[28] A fully optimal contract in the presence of physician risk aversion would attenuate τ_2 slightly from the value in Table 2, but the small physician risk exposure differences above suggest this attenuation can be ignored. The optimal contract also redistributes across physicians relative to the status quo, leaving RRR% of physicians better-off.

beneficiary per year, 17.2% of the beneficiary mean of $2,640 (1999–2004 dollars). Medicare pays 80% of allowed charges, so the externality on Medicare is $0.80 \times \$454 \approx \$363$ per Medigap beneficiary per year in their data window. CM use this estimate to argue for a tax on Medigap of approximately the same magnitude, which would internalize the moral-hazard externality.

Translating per-episode dollars to per-bene/year requires an assumption about Medicare's episode rate per beneficiary. Our calibration sample (IBM MarketScan Medicare Supplemental) consists of traditional Medicare fee-for-service beneficiaries aged 65+ with employer-sponsored supplemental coverage, restricted to Level 1 (CPT-coded) outpatient claims, the physician-fee-schedule services where physicians have a meaningful intensity choice. In the analytic sample we observe $\bar{e} \approx 1.84$ episodes per visiting patient per year (computed as total episodes divided by unique enrollee identifiers). Under this rate Panel A Δ Medicare equals $\$55/\text{ep} \times 1.84\,\text{ep/bene/yr} \approx \$101/\text{bene/yr}$. This is roughly 28% of CM's $363/bene/yr.

The remaining gap with CM reflects two differences. First, our sample covers outpatient physician services only, while CM's $2,640/bene/yr baseline aggregates over all Part B physician services including physician billing in inpatient settings (hospital visits, surgical assists, consultations). Second, our employer-sponsored Medigap sample may differ in composition from CM's broader Medicare population. As a check, the SQ Medicare baseline in our model is $\$502.93/\text{ep} \times 1.84 \approx \$925$ per visiting Medicare beneficiary per year, which is 44% of CM's all-Part-B Medicare baseline of $0.80 \times \$2{,}640 \approx \$2{,}112$. As a percentage of own SQ Medicare baseline the two estimates are similar: our Panel A reduces Medicare outlay by $\$55/\$503 \approx 10.9\%$, versus CM's $\$363/\$2{,}112 \approx 17.2\%/(1+0.172) \approx 14.7\%$.

Panel B's Δ Medicare of $213/episode is roughly 4× Panel A's $55, but it is not four times the CM externality. It decomposes additively into two distinct channels:

$$\underbrace{\$55/\text{ep}}_{\text{externality channel: same as Panel A}} + \underbrace{\$158/\text{ep}}_{\text{beyond-externality channel}} = \$213/\text{ep}.$$

The first piece is the moral-hazard reduction that a 20% patient coinsurance can induce, equivalently what CM's Medigap tax would recover. The second piece is the additional Medicare savings from implementing the constrained-first-best moral-hazard reduction via the optimal physician contract $g(m)$ while keeping $\tau_1 = 0$. Two forces drive the second piece. First, the optimal τ_1 in a Baily-Chetty problem is bounded above by the patient risk-protection margin, so consumer cost-sharing alone cannot reach first-best. Second, the status-quo $\hat{\tau}_2 = -0.016$ is a marginal subsidy on provider

effort, and replacing it with the optimal $g(m)$ removes this subsidy in addition to addressing Medigap-induced over-utilization.

In CM's framing the policy implication is sharp. A Medigap tax recovers the externality; an optimally redesigned physician contract recovers the externality plus roughly three times more in additional Medicare savings, without imposing any of the patient financial risk that motivated Medigap purchases in the first place.

Translating the per-episode numbers to aggregate Medicare savings requires multiplying by the episode rate ($\bar{e} \approx 1.84$ ep/bene/yr) and the affected beneficiary population. Two scope choices are natural. The conservative scope is current Medigap enrollees only (approximately 14 million beneficiaries in 2024, about 42% of fee-for-service Medicare). A broader extrapolation, which assumes our estimated λ distribution is representative of non-Medigap fee-for-service enrollees as well, extends to all fee-for-service Medicare beneficiaries (approximately 33 million). We exclude Medicare Advantage enrollees (about 32 million more) because they are already on capitated arrangements with different physician-incentive structures.

For Medigap enrollees (starting from $\tau_1 = 0$), the per-episode Medicare savings from adopting the optimal physician contract are \$213/ep, as reported in Panel B. For non-Medigap fee-for-service enrollees (starting from $\tau_1 = 0.20$), the calculation differs. Their status-quo physician behavior already incorporates the 20% patient coinsurance, so moving to $(\tau_1 = 0, g^*(m))$ involves *both* eliminating cost-sharing (which raises moral hazard by \$55/ep, reversing Panel A) *and* implementing the optimal contract (which saves \$213/ep in total relative to the Medigap SQ). The net savings for non-Medigap enrollees is therefore

$$\underbrace{\$213/\text{ep}}_{\text{savings relative to Medigap SQ}} \quad - \quad \underbrace{\$55/\text{ep}}_{\text{extra spending from } \tau_1: 0.20 \to 0} \quad = \$158/\text{ep},$$

exactly the beyond-externality channel. This is not a coincidence: non-Medigap enrollees have already internalized the \$55 externality reduction via their 20% coinsurance, so the contract reform can only recover the portion of Medicare savings that consumer cost-sharing cannot reach.

Applying these per-episode figures to each population:

$$\begin{aligned}
&\text{Medigap enrollees (14M):} && \$213 \times 1.84 \times 14\text{M} \approx \$5.5 \text{ billion/yr} \\
&\text{Non-Medigap FFS enrollees (19M):} && \$158 \times 1.84 \times 19\text{M} \approx \$5.5 \text{ billion/yr} \\
&\text{All fee-for-service Medicare (33M):} && \approx \$11 \text{ billion/yr.}
\end{aligned}$$

The naïve extrapolation of applying $213/ep to all 33M FFS enrollees would overstate aggregate savings by roughly $2 billion/yr ($13 vs. $11 billion/yr), because it ignores that non-Medigap enrollees start from a higher-coinsurance baseline. The corresponding Medigap-elimination (Panel A) aggregate is $1.4 billion/yr (Medigap-only, the policy-relevant figure under CM's framing); Panel A does not apply to non-Medigap enrollees who already face 20% cost-sharing. Even under the corrected aggregate, the optimal-contract reform generates roughly four times the Medicare savings of Medigap elimination for the Medigap population alone, and an additional comparable gain for non-Medigap enrollees.

These figures are for outpatient physician services only. Extending the contract reform to all Part B physician services (including inpatient physician billing) would scale the aggregates upward roughly in proportion to the all-Part-B vs. outpatient baseline ratio (about 2×, judging from the 44% baseline ratio above), giving Panel B aggregates on the order of $11 billion/yr (Medigap only) or $22 billion/yr (all fee-for-service Medicare). This extrapolation assumes the moral-hazard elasticity for inpatient physician services is similar to that for outpatient, and abstracts from any general-equilibrium adjustments to physician participation, fee-schedule politics, or patient sorting across contract regimes.

S15 Implementing a Near-Optimal Contract via Modified Medicare Billing

This appendix describes how a limited liability contract — a contract in which physician profits decline linearly with service intensity and are floored at zero — can be implemented as a simple modification of Medicare's existing RVU-based fee schedule, without requiring the payer to observe physician effort m directly.

Appendix Table 3 compares the optimal limited-liability linear contract to the status quo and the optimal nonlinear contract from Table 3.

Medicare's physician fee schedule reimburses each CPT code at a rate equal to its total RVU weight times a uniform conversion factor CF:

$$\text{Rev}_{\text{SQ}} = CF \times \left(\text{wRVU} + \text{PE\_RVU} + \text{MP\_RVU}\right).$$

Appendix Table 3: Limited-Liability Linear Contract vs. Status Quo and Optimal

	Status Quo ($\hat{\tau}_2 = -0.016$)		Optimal contract $g(m)$		LL linear contract	
Episode spending (m)	Rev. ($)	Profit ($)	Rev. ($)	Profit ($)	Rev. ($)	Profit ($)
\$43.80 ($m = k$, kink)	73	29	117	74	108	64
\$200	232	32	242	42	253	53
\$400	435	35	420	20	438	38
\$600	638	38	610	10	624	24
\$1,000	1,045	45	1,002	2	1,001	1
\$2,000	2,061	61	2,000	0	2,001	1
\$5,925 (99th percentile)	6,050	125	5,925	0	5,926	1
	Average: \$38.09		Average: \$38.09		Average: \$38.09	

Notes: Status-quo and optimal columns are identical to Table 3. The LL (limited liability) linear columns use $g_{\text{LL}}(m) = \max(\varepsilon, A_{\text{LL}} - s \cdot m)$ with $A_{\text{LL}} = \$67.00$, $s = 0.072$ (\$0.072 per dollar of spending), and breakpoint $m^* = (A_{\text{LL}} - \varepsilon)/s \approx \$918$ (above which physicians are reimbursed at cost plus a nominal margin $\varepsilon = \$1$). This contract uses the same visit copay $\tau_v = \$39.45$ and exclusion set as the optimal nonlinear contract and is revenue-neutral (average profit = \$38.09). Physicians re-optimize treatment intensity under the LL schedule; above the breakpoint, physicians are reimbursed at cost and provide first-best treatment. The linear contract recovers roughly half of the optimal nonlinear contract's welfare gain relative to the status quo.

In the model this is the linear schedule:

$$\text{Rev}_{\text{SQ}}(m) = \text{Rev}_k + (1 - \tau_2)(m - k),$$

where $\text{Rev}_k = \$73$ is the observed kink revenue (the modal episode revenue), $k = k_{\text{frac}} \cdot \text{Rev}_k$ is the corresponding effort level, and $\tau_2 \approx -0.016$ captures the marginal over-reimbursement relative to cost: physicians currently earn a $|\tau_2| \approx 1.6\%$ mark-up per additional unit of effort above the kink.

The near-optimal contract takes the form:

$$\pi_{\text{NEW}}(m) = \max\{\text{Rev}'_k - k - \tau'_2(m - k),\ 0\}, \qquad \tau'_2 > 0,$$

where Rev'_k is the policy-chosen kink revenue under the new scheme and $\tau'_2 > 0$ introduces a declining-profit margin above the kink. The corresponding revenue schedule is:

$$\text{Rev}_{\text{NEW}}(m) = \begin{cases} \text{Rev}'_k + (1 - \tau'_2)(m - k) & m \le m^*, \\ m & m > m^*, \end{cases} \tag{63}$$

where $m^* = k + (\text{Rev}'_k - k)/\tau'_2$ is the effort level at which profits reach zero (the limited liability constraint binds). For $m > m^*$ the physician is reimbursed exactly at cost.

On the linear portion $m \le m^*$, the target revenue schedule is achieved by the affine transformation $\text{Rev}_{\text{NEW}} = A + B \cdot \text{Rev}_{\text{SQ}}$. Substituting $\text{Rev}_{\text{SQ}}(m) = \text{Rev}_k + (1 - \tau_2)(m - k)$ and matching the value at

$m = k$ and the slope in m separately gives:

$$B = \frac{1-\tau_2'}{1-\tau_2}, \qquad A = \text{Rev}_k' - B \cdot \text{Rev}_k. \tag{64}$$

Since $\tau_2 < 0$ and $\tau_2' > 0$, we have $B < 1$: the reform applies a fractional markdown of $(1-B)\times 100\%$ to each additional dollar of fee-for-service revenue. The intercept $A = \text{Rev}_k' - B \cdot \text{Rev}_k$ is the change in the base payment at the kink revenue. Because $|\tau_2| \approx 0$, these approximate to $B \approx 1 - \tau_2'$ and $A \approx \text{Rev}_k' - \text{Rev}_k$: the reform is approximately a flat base-payment adjustment plus a proportional markdown of the existing fee schedule.

For $m > m^*$, the physician should be paid at cost m. Since m is not directly observed, we express cost reimbursement in terms of observable billing. Under the stable-mapping assumption, effort is recoverable as:

$$m = k + \frac{\text{Rev}_{\text{SQ}} - \text{Rev}_k}{1-\tau_2}.$$

The zero-profit threshold translates to the observable billing threshold:

$$\text{Rev}_{\text{SQ}}^* = \text{Rev}_k + \frac{(1-\tau_2)(\text{Rev}_k' - k)}{\tau_2'},$$

which is computable from the calibrated parameters and the policy choices. The complete revenue schedule (63) can therefore be written entirely in terms of observable billing as:

$$\text{Rev}_{\text{NEW}} = \min\left\{A + B \cdot \text{Rev}_{\text{SQ}},\ k + \frac{\text{Rev}_{\text{SQ}} - \text{Rev}_k}{1-\tau_2}\right\}. \tag{65}$$

The first argument is the declining-profit linear piece; the second is cost reimbursement. Note that the naive proposal $\max\{A + B \cdot \text{Rev}_{\text{SQ}}, 0\}$ would floor *revenue* at zero rather than profits, leaving the physician with negative profits $(-m)$ for high-effort episodes.

Two conditions are required.

1. The implementation does not require the payer to observe m directly, nor to know R separately from Rev_k. It requires only that the relationship between RVU-based billing and physician effort — summarized by the calibrated slope $(1-\tau_2)$ and observable anchor Rev_k — remain stable under the policy change. This accommodates stable upcoding: if physicians systematically bill above true effort, this is embedded in the calibrated $(1-\tau_2)$ and the implementation remains valid. What would violate this assumption is a *change* in coding behavior induced by

the reform itself — for example, more aggressive upcoding when $B < 1$ to offset the marginal clawback, or strategic CPT-code switching near the billing threshold Rev^*_{SQ}.

2. For full participation, Rev'_k must be set high enough that the marginal visiting type (the lowest illness-severity type λ_2 who visits under the new scheme) earns non-negative rents. This requires $\text{Rev}'_k \geq k + \tau'_2(\lambda_2 - k)$.