The Generosity Paradox:

When Less Generous Insurance Raises Spending

Iris SooJin Park*

March 16, 2026

Abstract

This paper shows that standard models of moral hazard predict a *generosity paradox*: less generous insurance can increase rather than decrease medical spending. Higher cost sharing makes it easier to reach the out-of-pocket maximum, where the marginal price is zero. Forward-looking consumers near that threshold therefore optimally increase their spending to reach the maximum. Using empirical estimates of the health need distribution and moral hazard responsiveness, I find that in many realistic scenarios, decreases in generosity lead to aggregate increases in spending and welfare losses. I discuss the practical implications of this for plan designers.

Keywords: nonlinear cost sharing, health insurance, moral hazard

* University of Wisconsin–Madison (email: park487@wisc.edu). I am grateful to Daniel Sacks and Justin Sydnor for invaluable advice and guidance, and to seminar participants at the University of Wisconsin–Madison for helpful comments. All errors are my own.

1 Introduction

A central premise in insurance design is that raising cost sharing reduces spending. Higher cost sharing increases the price of care, so standard moral hazard logic predicts lower spending in response. The core insight of this paper is that this reasoning is incomplete in nonlinear contracts with an out-of-pocket maximum. I show that reducing generosity can increase rather than decrease optimal spending, a result I call the "generosity paradox", with consequences for aggregate spending. The mechanism operates through rational responses to contract design, not through behavioral biases. Higher cost sharing lowers the total spending required to reach the out-of-pocket maximum, where the marginal price of care falls to zero. More broadly, the logic extends to any setting where marginal prices fall with cumulative use and use depends on marginal prices.

To illustrate, consider a plan with 20% coinsurance and a $2,000 out-of-pocket maximum. With no deductible, a household reaches the out-of-pocket maximum with $10,000 in total spending. With a $1,000 deductible, however, they reach that maximum with only $6,000 in total spending. A household that would have optimally chosen $7,000 in spending under the original plan, while still in the coinsurance region, may under the less generous plan instead find it optimal to reach the out-of-pocket maximum, where the marginal price falls to zero, and therefore spend more.

In this paper, I show that the generosity paradox is a general feature of standard models of health spending under nonlinear contracts and arises for a nontrivial set of consumers. I first establish this result in a perfect-foresight framework (Cardon and Hendel, 2001; Einav et al., 2013; Ho and Lee, 2023), where consumers observe their annual health needs and choose spending over the full nonlinear contract. In this setting, the spending increase is driven by intermediate-need consumers who switch from the coinsurance region to the out-of-pocket maximum in response to higher cost sharing. I then extend the analysis to forward-looking consumers who face uncertainty about future health needs but account for how current spending affects future out-of-pocket costs (Ellis, 1986; Cronin, 2019; Diaz-Campo, 2022). The paradox persists in this setting, with spending increases arising both from consumers who switch to the out-of-pocket maximum and from consumers who increase spending within the coinsurance region in anticipation of reaching the out-of-pocket maximum.

To assess the empirical relevance of the paradox, I simulate aggregate spending responses across a range of plan designs using existing empirical estimates of the health need distribution and moral hazard parameter from prior literature. I calibrate the simulations using

estimates from Ho and Lee (2023), who use administrative data from a large employer to estimate the standard perfect-foresight model of plan choice and utilization. I consider a range of plan designs, varying deductibles, coinsurance rates, and out-of-pocket maximums over values observed in employer-sponsored insurance. The simulations show that the generosity paradox can arise in the aggregate, leading to higher average spending under less generous coverage. I also show that when this occurs, higher spending does not translate into higher welfare, because the additional spending is concentrated where the gains from extra care are small, while reduced generosity leaves consumers more exposed to financial risk. I further show that heterogeneity in health needs plays a central role, as shifting the health need distribution can either amplify the paradox or attenuate it so that average spending falls under less generous coverage. This implies that the net effect on aggregate spending depends on how that distribution aligns with the contract structure.

This paper establishes a previously underappreciated implication of standard models of nonlinear insurance contracts. Early work on optimal insurance design established the tradeoff between risk protection and moral hazard control (Arrow et al., 1974; Zeckhauser, 1970). A central insight from this early literature is that straight deductible contracts, though optimal for risk protection, fail to control moral hazard once the deductible is met. In the presence of moral hazard, it can be optimal to give up some risk protection and allow greater out-of-pocket exposure while keeping positive marginal prices over a broader range of utilization through coinsurance. The same nonlinear contract structure that balances risk protection against moral hazard also implies that changes in generosity do not uniformly affect spending incentives. A modern structural literature builds on this foundation by modeling how forward-looking consumers optimize over the full nonlinear contract (Cardon and Hendel, 2001; Einav et al., 2013; Marone and Sabety, 2022; Diaz-Campo, 2022; Ho and Lee, 2023).[1] The generosity paradox follows from the same forward-looking behavior emphasized in these models.

Together, these insights have practical implications for plan designers. If a plan designer started from a fully estimated structural model, these dynamics would be incorporated in determining the optimal contract (e.g. Einav et al., 2013; Ho and Lee, 2023). In practice, however, plan changes rarely begin from full optimization. Employers and insurers often adjust cost-sharing incrementally in response to rising premiums, budget pressures, or regulatory changes, taking the existing plan structure as given. Rising spending may itself prompt plan designers to reduce generosity, precisely when more consumers are likely to

[1] Related work studies responses to nonlinear contracts when consumers face imperfect or delayed price information; see, for example, Anderson, Hoagland and Zhu (2024).

be near the out-of-pocket maximum, and the paradox is most likely to arise. It is in these incremental adjustments that this mechanism is most likely to be overlooked. A firm that raises deductibles, expecting to reduce spending by limiting moral hazard, may instead increase spending. In evaluating such changes, what matters is how a proposed change affects the share of consumers whose effective marginal price falls to zero under the new contract. More broadly, the relevant issue is not any single plan design change but whether adjustments taken together lower the spending threshold at which cost control disappears for the population the firm insures.

These results follow from the standard assumption that forward-looking households internalize the effect of today's spending on future prices. Whether consumers actually account for future prices is an open empirical question, with direct implications for the real-world relevance of the generosity paradox. Some evidence is consistent with myopic behavior, with consumers responding primarily to spot prices (Brot-Goldberg et al., 2017; Guo and Zhang, 2019). Other work documents forward-looking behavior, finding that utilization responds to future prices (Aron-Dine et al., 2015; Klein, Salm and Upadhyay, 2022; Johansson et al., 2023). In the discussion, I offer some preliminary thoughts on how the generosity paradox dynamics I identify for the standard model may differ in models with myopic consumers who respond only to spot prices.

The remainder of the paper proceeds as follows. Section 2 presents the baseline model of consumer optimization over nonlinear cost-sharing contracts. Section 3 characterizes when raising cost sharing increases optimal spending. Section 4 simulates spending responses across a range of plan designs and examines the welfare implications of the generosity paradox. Section 5 concludes.

2 Baseline Model

2.1 Out-of-Pocket Spending Schedule

A consumer faces a nonlinear contract (D, s, M) with deductible D, coinsurance rate $s \in (0, 1)$, and maximum out-of-pocket (MOOP) M. Let $m \geq 0$ denote medical spending and define

$$\bar{m}(D, s, M) = D + \frac{(M - D)}{s} \tag{1}$$

as the spending level at which the MOOP binds. Out-of-pocket spending is

$$\text{OOP}(m) = \begin{cases} m & m \leq D \\ D + s(m - D) & D < m < \bar{m}(D, s, M) \\ M & m \geq \bar{m}(D, s, M), \end{cases} \tag{2}$$

with piecewise-constant marginal OOP price $c \in \{1, s, 0\}$, where $c = 1$ below the deductible, $c = s$ in the coinsurance region, and $c = 0$ above the MOOP.

2.2 Static Model with Perfect Foresight

The static setup in this section follows the standard models of medical spending under nonlinear insurance contracts, including Cardon and Hendel (2001); Einav et al. (2013); Ho and Lee (2023).

At the beginning of the coverage year, a consumer with health need $\lambda > 0$ chooses annual medical spending $m \geq 0$ and derives gross utility from medical care

$$h(m; \lambda, \omega) = (m - \lambda) - \frac{(m - \lambda)^2}{2\omega\lambda}, \tag{3}$$

where $\omega > 0$ represents the household's price sensitivity and captures the degree of moral hazard.[2] Conditional on enrollment in a given plan, the consumer chooses medical spending to maximize the utility from care minus OOP costs. Fixed components such as income and premiums are omitted because they do not vary with utilization within a plan. The consumer therefore solves

$$V(\lambda) \equiv \max_{m \geq 0} \{h(m; \lambda, \omega) - \text{OOP}(m)\}. \tag{4}$$

For an interior optimum, spending satisfies the standard first-order condition equating the marginal benefit from care to the marginal OOP price, $h_m(m^*; \lambda, \omega) = \text{OOP}'(m^*)$. Because the nonlinear contract implies a constant marginal OOP price within each pricing region, the perfect-foresight case implies the interior solution

$$m^*(c; \lambda, \omega) = \lambda[1 + \omega(1 - c)]. \tag{5}$$

[2]The simulation abstracts from a fixed care-entry hassle cost of the type used in Ho and Lee (2023). When such a cost enters as a common additive term, it does not affect within-care spending choices or the cutoff values governing movement across spending regions. The simulation therefore isolates the intensive-margin implications of nonlinear cost sharing and contract geometry. Supplemental Appendix Figure D.1 shows that incorporating such a common hassle cost leaves the simulated spending response unchanged.

This yields three interior candidates: $m^*_{\text{ded}} = \lambda$ below the deductible, $m^*_{\text{coin}} = \lambda[1+\omega(1-s)]$ in the coinsurance region, and $m^*_{\text{cap}} = \lambda(1+\omega)$ above the MOOP.[3] Because the OOP schedule is kinked at the deductible and the MOOP, the global optimum is obtained by comparing the values of these three candidates.

2.3 Dynamic Model with Uncertainty

I also consider a forward-looking model with uncertainty about future health needs, following Diaz-Campo (2022) and related work on dynamic spending under nonlinear contracts (Ellis, 1986; Cronin, 2019). Consumers choose spending sequentially over the coverage year, taking into account that current spending changes cumulative spending and therefore future OOP prices. The relevant marginal price is thus a shadow price that reflects both the current spot price and the effect of today's spending on future OOP prices. Supplemental Appendix C.1 provides the full dynamic formulation.

3 The Generosity Paradox in Nonlinear Contracts

3.1 How Contract Geometry Shapes Incentives

The mechanism underlying the generosity paradox is a simple feature of nonlinear contract design. Every plan with a deductible, coinsurance, and out-of-pocket maximum has a spending threshold $\bar{m}(D, s, M)$ at which the consumer reaches the MOOP and the marginal price of care drops to zero. Raising the deductible while holding the MOOP fixed lowers this threshold, thereby compressing the coinsurance region and bringing zero-price care within reach. A consumer who would not have found it optimal to reach the MOOP under a more generous plan may find it optimal to do so after the deductible increase.

To illustrate, consider a plan with 10% coinsurance and an MOOP of \$2,000, comparing a \$500 deductible to a \$1,500 deductible:

$$\bar{m}(500) = 500 + \frac{2,000 - 500}{0.1} = \$15,500, \quad \bar{m}(1,500) = 1,500 + \frac{2,000 - 1,500}{0.1} = \$6,500.$$

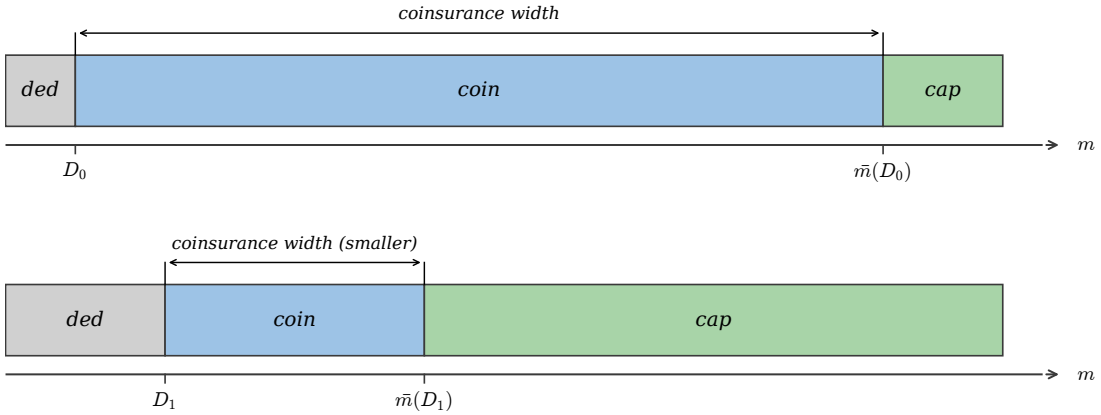
With the \$500 deductible, a household faces 10% coinsurance across \$15,000 of spending

[3] As in Ho and Lee (2023), this specification implies that higher-need consumers exhibit larger spending responses to plan changes, consistent with evidence from Ho and Lee (2023) and Brot-Goldberg et al. (2017). The same mechanism applies under the Einav et al. (2013) specification.

before reaching the MOOP. With the $1,500 deductible, the coinsurance region shrinks to $5,000, lowering the zero-marginal-price threshold by $9,000. The same compression mechanism arises when the coinsurance rate increases.

Figure 1 illustrates how a deductible increase changes the geometry of the contract. The top panel shows the three spending regions under deductible D_0, while the bottom panel shows how a higher deductible D_1 lowers the total spending required to reach the MOOP and compresses the coinsurance region. As a result, the effect of reduced generosity depends on where a consumer's health need places them along the contract. More generally,

Figure 1: Spending Regions under Different Deductible Levels



Note: Holding the coinsurance rate and MOOP fixed, the top panel shows a plan with deductible D_0 and coinsurance width $\bar{m}(D_0) - D_0$. The bottom panel shows a higher deductible $D_1 > D_0$, which shrinks the coinsurance region and lowers the spending threshold at which the out-of-pocket maximum binds from $\bar{m}(D_0)$ to $\bar{m}(D_1)$.

the generosity paradox requires both a nonempty coinsurance region and an MOOP. It can therefore arise in three-part contracts, and also in plans with coinsurance and an MOOP but no deductible, but not in straight deductible plans or pure coinsurance plans without a cap.

3.2 When Higher Deductibles Increase Spending

To formalize when reducing generosity raises spending, I characterize the ranges of realized health need over which a deductible increase moves consumers across pricing regions. A higher deductible can push some lower-need consumers from the coinsurance region into the full-price region, reducing spending, while pushing some intermediate-need consumers from the coinsurance region into the MOOP region, increasing spending.

To characterize which consumers switch regions, I first compare the values associated with each pricing region. Let V^{ded}, V^{coin}, and V^{cap} denote the values of stopping below

the deductible, remaining in the coinsurance region, and reaching the MOOP, respectively. Reaching the MOOP is optimal only if $V^{\text{cap}} \geq V^{\text{coin}}$ and $V^{\text{cap}} \geq V^{\text{ded}}$. Similarly, the coinsurance region is chosen only if it weakly dominates both neighboring options: $V^{\text{coin}} \geq V^{\text{ded}}$ and $V^{\text{coin}} \geq V^{\text{cap}}$. These pairwise comparisons define three cutoffs.

First, Appendix A.1 shows that

$$\Lambda_{\text{cap}} \equiv \frac{M}{1+\omega/2} \tag{6}$$

is the realized health need at which the consumer is indifferent between reaching the MOOP and stopping below the deductible, that is, $V^{\text{cap}} = V^{\text{ded}}$. Thus, for $\lambda < \Lambda_{\text{cap}}$, reaching the MOOP cannot be optimal, even if it weakly dominates the coinsurance region.

Next, define

$$\lambda_{\text{ded-coin}}(D) \equiv \frac{D}{1+\frac{\omega(1-s)}{2}}, \qquad \lambda_{\text{coin-cap}}(D) \equiv \frac{M-(1-s)D}{s(1+\omega\left(1-\frac{s}{2}\right))}. \tag{7}$$

The cutoff $\lambda_{\text{ded-coin}}(D)$ is the realized health need at which the consumer is indifferent between the full-price region and the coinsurance region, that is, $V^{\text{coin}} = V^{\text{ded}}$. The cutoff $\lambda_{\text{coin-cap}}(D)$ is the realized health need at which the consumer is indifferent between the coinsurance region and reaching the MOOP, that is, $V^{\text{cap}} = V^{\text{coin}}$. Derivations of these cutoffs are provided in Appendix A.2.

Putting these together, reaching the MOOP is optimal only when both conditions hold: $V^{\text{cap}} \geq V^{\text{coin}}$ and $V^{\text{cap}} \geq V^{\text{ded}}$. This implies the global cap-entry cutoff

$$\lambda_{\text{cap}}(D) \equiv \max\{\Lambda_{\text{cap}}, \lambda_{\text{coin-cap}}(D)\}. \tag{8}$$

That is, $\lambda_{\text{cap}}(D)$ is the minimum realized health need for which reaching the MOOP is optimal. By contrast, the coinsurance region is optimal when $V^{\text{coin}} \geq V^{\text{cap}}$ and $V^{\text{coin}} \geq V^{\text{ded}}$, so its boundaries are given directly by the two pairwise cutoffs $\lambda_{\text{ded-coin}}(D)$ and $\lambda_{\text{coin-cap}}(D)$. Thus the coinsurance region is optimal for $\lambda \in [\lambda_{\text{ded-coin}}(D), \lambda_{\text{cap}}(D))$, while the MOOP region is chosen for $\lambda \geq \lambda_{\text{cap}}(D)$.

A deductible increase therefore affects spending by moving some consumers out of the coinsurance region, either downward or upward, depending on which cutoff their realized health need crosses. Proposition 1 formalizes these responses by characterizing which consumers change their spending and by how much.

Proposition 1 (Heterogeneous Spending Responses to a Deductible Increase)**.** *Fix* (s, M, ω) *and consider a deductible increase from* D_0 *to* D_1*, where* $0 \le D_0 < D_1 < M$*. Assume that the coinsurance region is nonempty under both contracts.*

Let $\Delta m(\lambda) \equiv m^*(\lambda; D_1) - m^*(\lambda; D_0)$*. Then*

$$\Delta m(\lambda) = \begin{cases} -\omega(1-s)\lambda, & \text{if } \lambda \in \big[\lambda_{\text{ded-coin}}(D_0),\, \lambda_{\text{ded-coin}}(D_1)\big), \qquad (\textit{coin} \to \textit{ded}), \\ \omega s\, \lambda, & \text{if } \lambda \in \big[\lambda_{\text{cap}}(D_1),\, \lambda_{\text{cap}}(D_0)\big), \qquad (\textit{coin} \to \textit{cap}), \\ 0, & \textit{otherwise.} \end{cases} \tag{9}$$

Proof. See Appendix A.3.

Proposition 1 implies that lower-need consumers switch from coinsurance to the deductible region and reduce spending, while intermediate-need consumers switch from coinsurance to the MOOP and increase spending. In both cases, the magnitude of the response is proportional to λ, so higher-need consumers within each switching group experience larger spending changes. Consumers outside these switching ranges remain in the same region and do not change their spending.

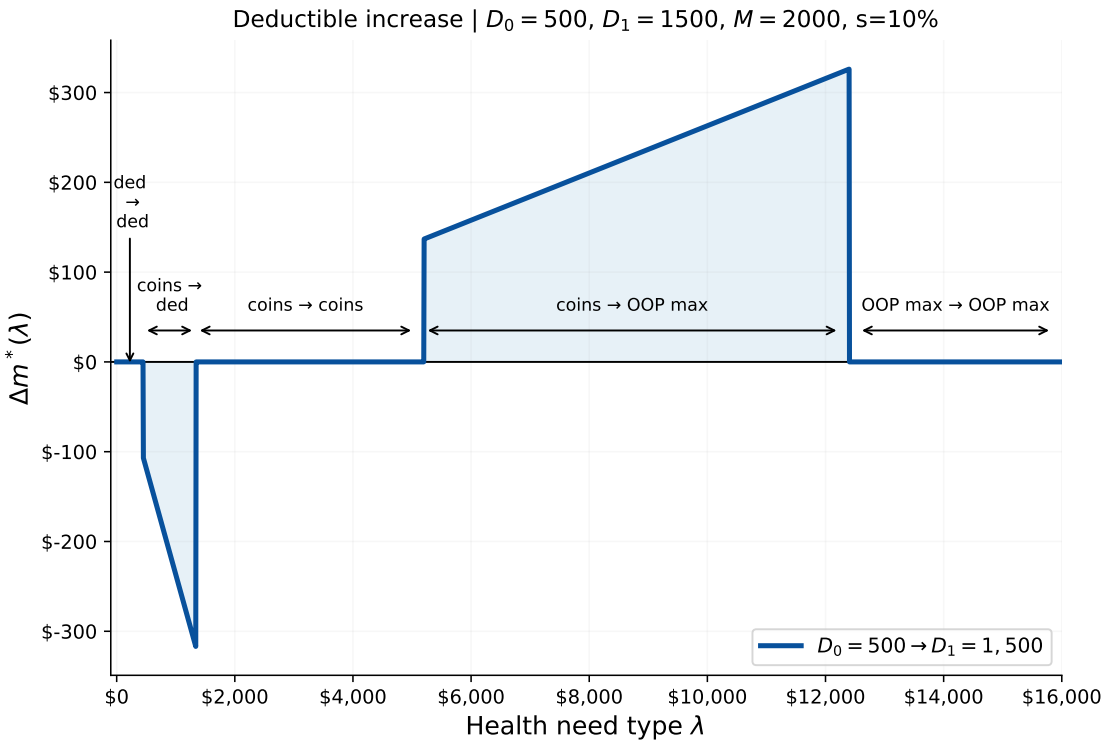
Figure 2 plots $\Delta m(\lambda)$ as a function of realized health need λ for a plan with a \$2,000 MOOP, a 10% coinsurance rate, and a deductible increase from \$500 to \$1,500. Throughout the paper, I set $\omega = 0.263$ following Ho and Lee (2023). Spending falls for lower-need consumers in $[447,\, 1,341)$, rises for intermediate-need consumers in $[5,201,\, 12,401)$, and is unchanged otherwise. In this example, the upward-response region is much wider than the downward-response region. Within each region, the magnitude of $\Delta m(\lambda)$ increases with λ.

Although Proposition 1 characterizes the up-switcher set, it does not yet establish when that set is nonempty. Proposition 2 provides this condition. Under a deductible increase, spending rises only for realized health needs in the interval $[\lambda_{\text{cap}}(D_1),\, \lambda_{\text{cap}}(D_0))$. Since $\lambda_{\text{coin-cap}}(D)$ is decreasing in D, a deductible increase from D_0 to D_1 implies $\lambda_{\text{coin-cap}}(D_1) < \lambda_{\text{coin-cap}}(D_0)$. Thus, the up-switcher set is empty only if the cap-relevance threshold binds already at the initial deductible, that is, if $\Lambda_{\text{cap}} \ge \lambda_{\text{coin-cap}}(D_0)$. Equivalently, consumers increase spending in response to a deductible increase only when $\Lambda_{\text{cap}} < \lambda_{\text{coin-cap}}(D_0)$. Let D_{crit} denote the deductible that solves

$$\Lambda_{\text{cap}} = \lambda_{\text{coin-cap}}(D_{\text{crit}}). \tag{10}$$

Then the up-switcher set is nonempty if and only if $D_0 < D_{\text{crit}}$.

Figure 2: Spending Responses to Deductible Increases by Health Need



Note: The figure plots the change in optimal spending, $\Delta m(\lambda) = m^*(\lambda; D_1) - m^*(\lambda; D_0)$, by realized health need λ, for a deductible increase from \$500 to \$1,500. The parameters are $M = 2{,}000$, $s = 0.10$, and $\omega = 0.263$, calibrated to estimates from Ho and Lee (2023). The annotations show how consumers move across regions of the contract as the deductible rises.

Proposition 2 (When the Up-Switcher Set Is Nonempty)**.** *Fix* (s, M, ω) *and define*

$$D_{\text{crit}} \equiv M \frac{\omega(1-s)+2}{\omega+2}. \tag{11}$$

Consider a deductible increase from D_0 *to* D_1*, where* $0 \leq D_0 < D_1 < M$*.*

The up-switcher set characterized in Proposition 1 is nonempty if and only if $D_0 < D_{\text{crit}}$*. In particular, if* $D_0 = 0$*, then this set is nonempty for every* $D_1 \in (0, M)$*.*

Proof. See Appendix A.3.

When the initial deductible is zero, the condition holds automatically, so introducing any positive deductible always generates up-switchers. More generally, if the initial deductible is positive and satisfies $D_0 < D_{\text{crit}}$, then any increase from the initial point generates a nonempty set of consumers who increase spending.

Figure 3 shows that the critical deductible D_{crit} is typically very close to the MOOP. Specifically, it plots how close the critical deductible is to the MOOP, measured by the ratio D_{crit}/M, against the coinsurance rate for different values of ω. At the average coinsurance rate of 19%,[4] D_{crit}/M ranges from 96% to 99% across the values of ω shown. Even at much higher coinsurance rates, D_{crit} remains a large share of the MOOP. Thus, in practice, the condition $D_0 < D_{\text{crit}}$ is satisfied for most realistic initial deductibles.

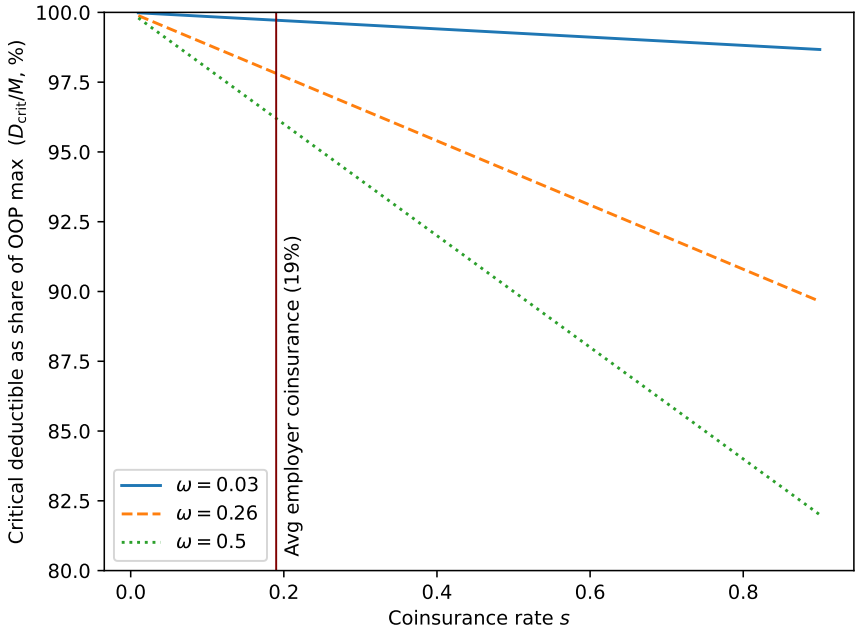
3.3 Extensions

Changes in coinsurance rates

The generosity paradox is not limited to deductibles and can also arise from higher coinsurance. As equation 1 shows, increasing the coinsurance rate lowers the spending threshold at which the consumer reaches the MOOP. A higher coinsurance rate affects spending through three channels. It can reduce spending by pushing some consumers from the coinsurance region below the deductible and by raising the marginal price of care for those who remain in the coinsurance region. At the same time, it lowers the total spending required to reach the MOOP, making the cap easier to reach and increasing spending for some consumers. In Supplemental Appendix B, I characterize the ranges of health need for which spending rises, falls, or remains unchanged after a coinsurance increase, and show that this case also generates a nonempty set of consumers who switch into the MOOP region and increase

[4] I use the average coinsurance rate reported by Kaiser Family Foundation (2025)

Figure 3: Critical Deductible as a Share of the MOOP



Note: The figure plots the critical deductible share D_{crit}/M against coinsurance s. Line styles correspond to $\omega \in \{0.03, 0.26, 0.5\}$. The vertical marker indicates the average coinsurance rate of 19% in employer-sponsored health insurance in 2025 as reported by Kaiser Family Foundation.

spending.

Dynamic Model with Uncertainty

The generosity paradox also arises when consumers choose spending sequentially under uncertainty. Following related models of dynamic spending under nonlinear contracts (Ellis, 1986; Cronin, 2019; Diaz-Campo, 2022), I consider a setting in which, unlike in the perfect-foresight case, annual spending is not chosen all at once. Instead, current spending affects not only current OOP costs but also future prices by moving the consumer closer to the MOOP. As a result, the relevant marginal price of care is not just the current spot price, but a shadow price that incorporates both today's spot price and the effect of current spending on future OOP prices. This preserves the same basic mechanism as in the static model. Spending today becomes more attractive when it moves the consumer closer to the MOOP, where future marginal prices are lower.

Supplemental Appendix C formalizes this forward-looking analogue under economically interpretable conditions on the continuation value.[5] Under these conditions, a higher deductible can increase spending through two channels. First, as in the perfect-foresight case,

[5]These conditions impose that the payoff to reaching the MOOP is weakly increasing in current health need and that the marginal value of moving closer to the MOOP is weakly higher when the deductible rises.

some consumers optimally switch into the MOOP region, generating a discrete jump in spending. Second, even consumers who remain in the coinsurance region may spend more because a higher deductible lowers the expected shadow price of current care by making it easier to reach the MOOP in the future.

4 Net Population Effects

I now turn to the aggregate spending effects of reduced generosity. Aggregating across realized health needs, the net effect on average spending can be written as

$$\mathbb{E}[\Delta m(\lambda)] = \mathbb{E}[\Delta m(\lambda)\mathbf{1}\{\Delta m(\lambda) > 0\}] + \mathbb{E}[\Delta m(\lambda)\mathbf{1}\{\Delta m(\lambda) < 0\}]\,. \tag{12}$$

Average utilization rises when the need-weighted mass shifted into the MOOP exceeds the need-weighted mass shifted into the full-price region.

To evaluate these aggregate effects, I assume that annual health needs λ are drawn from a lognormal distribution. Following Ho and Lee (2023), I calibrate the distribution of health needs to match mean annual needs of \$6,490 and a standard deviation of \$4,890, set the moral hazard parameter to $\omega = 0.263$, and set CARA risk aversion to $\psi = 0.0003$. Using this distribution, I simulate spending responses to deductible and coinsurance increases under multiple cost-sharing designs.

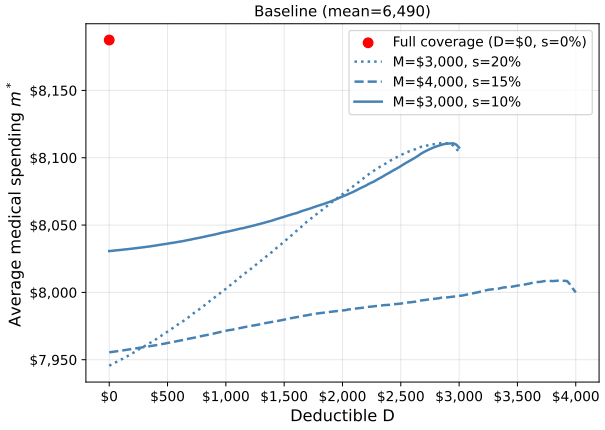
In addition to net spending responses, I examine how plan generosity affects consumer welfare, which depends on both the health benefits of care and exposure to financial risk. I summarize welfare using the certainty equivalent (CE), defined as the amount of certain resources that makes the consumer indifferent to enrolling in a contract $x = (D, s, M)$. Under CARA utility, the CE is given by

$$CE(x) = -\frac{1}{\psi}\log \mathbb{E}\left[\exp\left(-\psi\tilde{u}(x)\right)\right] - \text{premium}(x), \tag{13}$$

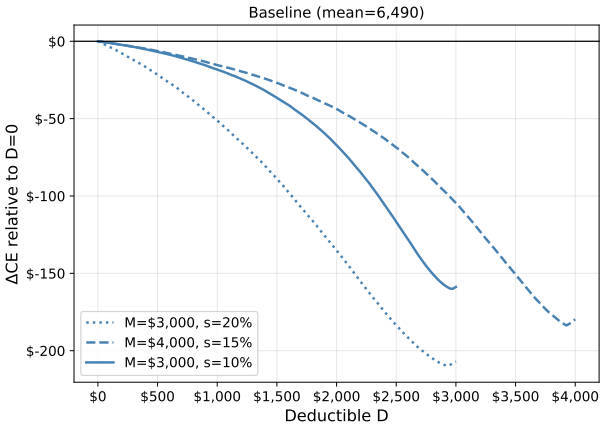
where $\tilde{u}(x) = h(m^*; \lambda, \omega) - \text{OOP}(m^*)$ captures utility from care after out-of-pocket costs. The premium is set to the actuarially fair value, defined as $\text{premium}(x) = \mathbb{E}[m^* - \text{OOP}(m^*)]$.

Figure 4 summarizes how average spending and consumer welfare vary with deductibles and coinsurance under the baseline calibration. Panels (a) and (b) vary the deductible for three plan designs: 10 and 20 percent coinsurance with a \$3,000 MOOP, and 15 percent coinsurance with a \$4,000 MOOP. Panels (c) and (d) turn to the variation in the coinsurance

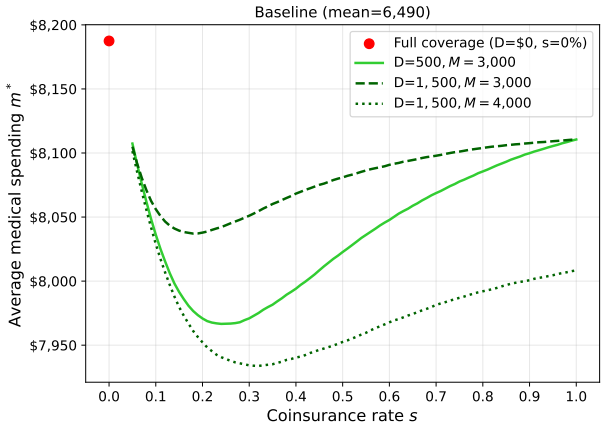
Figure 4: Average Spending and Consumer Welfare under Multiple Cost-Sharing Designs



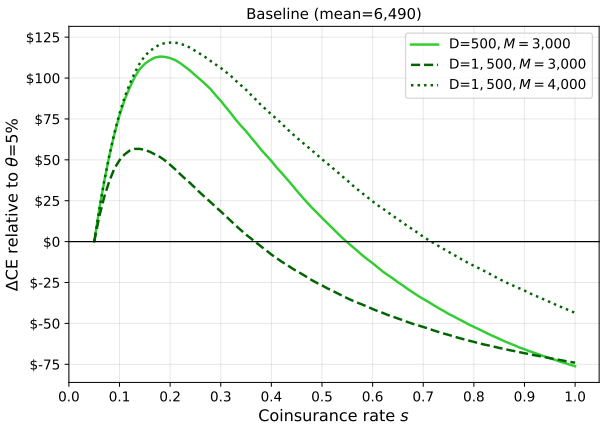
(a) Spending: deductibles



(b) Welfare: deductibles



(c) Spending: coinsurance



(d) Welfare: coinsurance

Note: The figure shows outcomes under the perfect-foresight model for several plan designs. Panels (a) and (b) vary the deductible for three contracts: 20 percent coinsurance with a $3,000 MOOP, 15 percent coinsurance with a $4,000 MOOP, and 10 percent coinsurance with a $3,000 MOOP. Panels (c) and (d) vary the coinsurance rate for three contracts: two with a $3,000 MOOP at deductibles of $500 and $1,500, and one with a $4,000 MOOP and a $1,500 deductible. Panels (a) and (c) plot average annual medical spending and Panels (b) and (d) plot the change in certainty equivalent (CE). Simulations use $N = 100{,}000$ draws of annual health needs. Following Ho and Lee (2023), mean annual health needs are calibrated to $6,490 with a standard deviation of $4,890, the moral hazard parameter is set to $\omega = 0.263$, and CARA risk aversion is set to $\psi = 0.0003$.

rate for three plan designs: two with a $3,000 MOOP at deductibles of $500 and $1,500, and one with a $4,000 MOOP and a $1,500 deductible.[6]

Panel (a) shows that average spending monotonically rises with the deductible across all three plan designs, consistent with the generosity paradox. The two plans with a $3,000 MOOP exhibit the strongest increase in spending as the deductible rises, but the increase is steeper under 20 percent coinsurance than under 10 percent coinsurance. Intuitively, when coinsurance is higher, consumers in the coinsurance region face a larger reduction in OOP spending once they reach the MOOP, so the incentive to push through to the cap becomes stronger. By contrast, the plan with a $4,000 MOOP and 15 percent coinsurance also shows an upward pattern, but the increase is more modest because a higher MOOP weakens the cap-reaching force.

Panel (b) shows that consumer welfare, measured by CE, declines as the deductible rises across all three plan designs. A higher deductible can raise total medical spending by pushing some consumers from the coinsurance region to the MOOP, but it still lowers insurer liability, so actuarially fair premiums fall (see Supplemental Appendix Figure D.2). The decline in CE reflects the fact that these premium reductions do not compensate for the loss of insurance protection. As the deductible rises, consumers bear the full marginal cost of care over a wider range of spending and face greater OOP exposure unless they reach the MOOP. Supplemental Appendix Table D.1 shows that the paradoxical spending increase is driven by consumers who switch from the coinsurance region to the MOOP, whereas the welfare decline is driven mainly by the much larger losses borne by consumers who remain in the coinsurance region and face higher OOP costs.

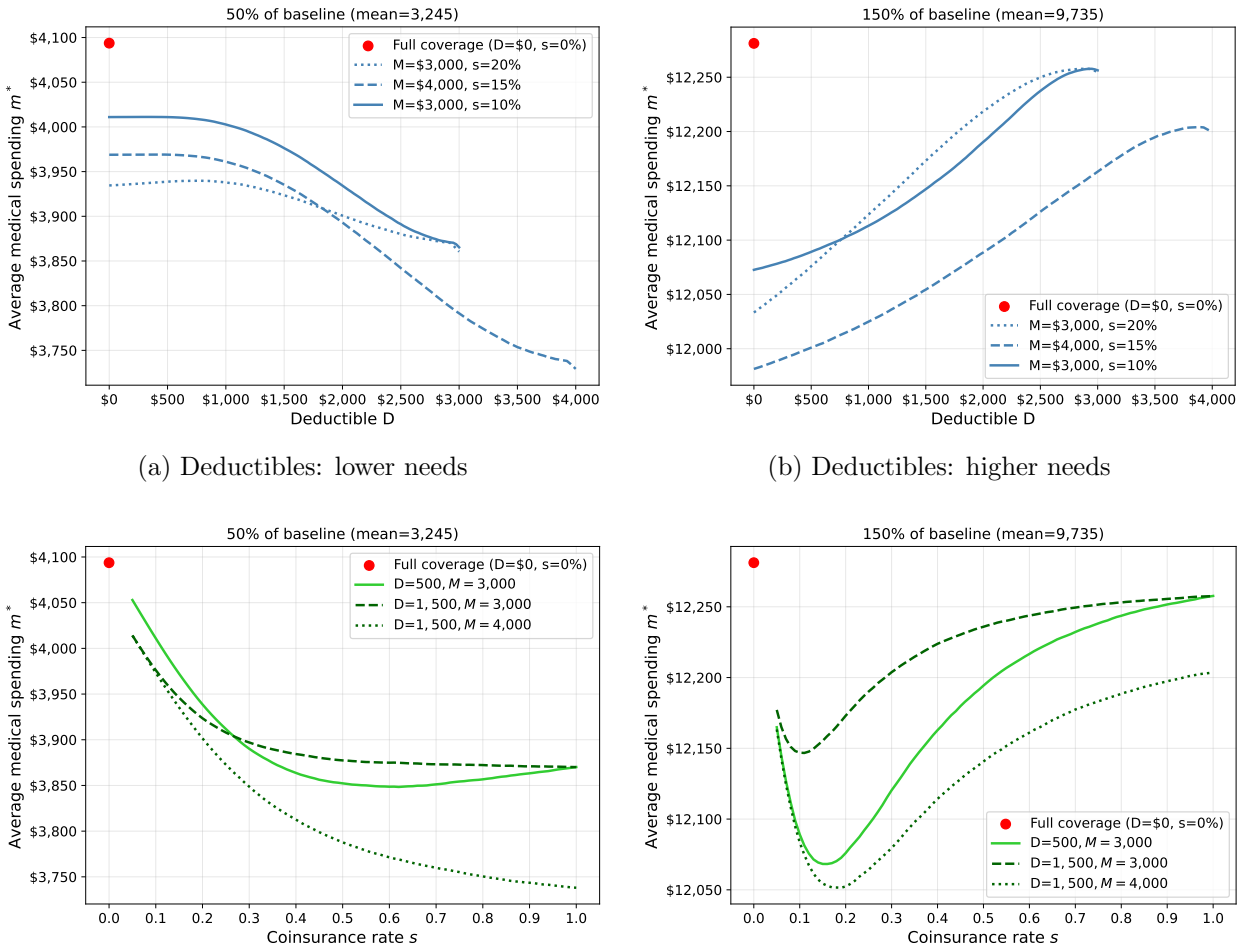
Panel (c) shows that average spending follows a U-shaped pattern as the coinsurance rate rises across all three plan designs, which arises from the model's two opposing forces. When the initial coinsurance rate is low, increasing coinsurance raises the marginal price of care for a broad set of consumers who remain below the MOOP, while generating relatively few new cap switchers, so average spending falls. As the coinsurance rate rises, however, the MOOP becomes substantially easier to reach, and the mass of consumers pushed into the zero-price region grows. The cap-reaching force then becomes strong enough to outweigh the standard price effect, causing average spending to rise.

Panel (d) shows that welfare is non-monotonic in coinsurance and need not move in the

[6]Appendix Figures D.3 and D.4 extend this analysis by mapping the spending response to small deductible and coinsurance increases over a broader range of contract parameters (D, s, M) and health-need distributions. Positive aggregate spending responses arise for a broad set of plan designs, indicating that the paradox is not confined to a small number of selected calibrations.

same direction as spending. Starting from a very low coinsurance rate, a moderate increase in coinsurance can raise CE for some plan designs even as spending falls. In this range, higher coinsurance reduces utilization and lowers premiums, and the resulting premium savings outweigh the welfare loss from less generous coverage. As coinsurance rises further, however, insurance protection deteriorates, and CE falls. Over the same range, spending can eventually rise because a higher coinsurance rate makes the MOOP easier to reach and pushes more consumers into the zero-price region.

Figure 5: Average Spending under Multiple Health-Need Distributions



(a) Deductibles: lower needs

(b) Deductibles: higher needs

(c) Coinsurance: lower needs

(d) Coinsurance: higher needs

Note: The figure shows how spending varies across lower- and higher-health-need populations under the same plan designs as in Figure 4. The low health-need distribution rescales the Ho and Lee (2023) baseline moments to 50 percent of their original values, with mean annual health needs of $3,245 and a standard deviation of $2,445. The high health-need distribution rescales the same moments to 150 percent of their original values, with mean annual health needs of $9,735 and a standard deviation of $7,335. The moral hazard parameter is set to $\omega = 0.263$, following Ho and Lee (2023).

Figure 5 shows that the aggregate spending response to cost sharing is highly sensitive

to the underlying distribution of health needs. To explore this heterogeneity, I construct two alternative populations by scaling the mean and standard deviation of annual health needs to 50 percent and 150 percent of the baseline values, respectively. The top row shows spending responses to deductible variation and the bottom row to coinsurance variation, with the lower-need population on the left and the higher-need population on the right.

For the lower-need population, the generosity paradox largely disappears for both deductible and coinsurance changes. With the lower-need distribution placing less mass near the MOOP, the cap-access effect is weak, and the standard price effect dominates, causing spending to fall as the deductible or the coinsurance rate rises. For the higher-need population, by contrast, the paradox becomes much stronger. Panel (b) shows a steeper increase in spending with the deductible, and panel (d) shows a less pronounced U-shape because the conventional price effect weakens relative to the cap-access effect. As a result, even modest increases in cost sharing push a larger share of consumers into the zero-price region, weakening the initial decline in spending.

The broader implication is that the aggregate spending effect of plan generosity cannot be inferred from the contract alone. The same contract change can reduce spending in a healthier population and increase it in a sicker one, as the net effect depends on how much mass the health-need distribution places near the relevant cost-sharing thresholds.

5 Conclusion

This paper shows that in standard models of moral hazard under nonlinear cost sharing, reducing generosity can increase rather than decrease optimal spending. Raising deductibles or coinsurance lowers the spending threshold at which the marginal price of care falls to zero. Consumers near that threshold may therefore find it optimal to push through to the out-of-pocket maximum, spending more despite less generous coverage. I show that the generosity paradox arises under both perfect foresight and forward-looking behavior with uncertainty and characterize the range of health needs for which it occurs. In many realistic plan designs, reductions in generosity lead to aggregate spending increases, and this additional spending does not improve welfare.

These results have practical implications for evaluating spending responses to changes in generosity within existing plan designs. Insurers and employers often change existing plan designs in response to rising spending, precisely when the paradox is most likely to arise. A change intended to reduce spending may instead increase it if it pushes more enrollees toward

the out-of-pocket maximum. Even without a full structural model, a plan designer can often infer the direction of the spending response by asking whether a proposed change moves a substantial share of enrollees closer to the cap. In nonlinear contracts, that interaction, rather than any single plan feature, determines whether a reduction in generosity reduces spending or paradoxically increases it.

The generosity paradox also speaks to how we interpret observed heterogeneous spending responses to changes in cost sharing. Empirical work typically attributes such heterogeneity to differences in price sensitivity across consumers. This paper shows that heterogeneous responses can arise even among consumers with identical price sensitivity. The same deductible increase reduces spending for consumers far from the out-of-pocket maximum and increases it for consumers near it. Observed heterogeneity in spending responses therefore reflects not only differences in moral hazard parameters but also differences in where health needs lie relative to the contract's thresholds.

More broadly, the generosity paradox depends on how consumers form expectations about future prices. Consumers near the out-of-pocket maximum spend more because they anticipate that crossing it drives the marginal cost of care to zero. To the extent consumers are myopic, the paradox is correspondingly attenuated, because reducing generosity raises or leaves unchanged the current spot price of care. Under myopia, for the paradox to arise, a consumer must move from the coinsurance region to the out-of-pocket maximum within a single period's spending. While this is theoretically possible, it is unlikely to have a meaningful aggregate impact. The degree to which consumers are forward-looking remains an open empirical question, with direct relevance given that cost-sharing increases remain the central tool for managing health care spending.

References

Anderson, David M., Alex Hoagland, and Ed Zhu. 2024. "Medical bill shock and imperfect moral hazard." *Journal of Public Economics*, 236: 105152.

Aron-Dine, Aviva, Liran Einav, Amy Finkelstein, and Mark Cullen. 2015. "Moral Hazard in Health Insurance: Do Dynamic Incentives Matter?" *The Review of Economics and Statistics*, 97(4): 725–741.

Arrow, Kenneth Joseph, et al. 1974. *Essays in the theory of risk-bearing.* North-Holland Amsterdam.

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

Cardon, James H, and Igal Hendel. 2001. "Asymmetric information in health insurance: evidence from the National Medical Expenditure Survey." *RAND Journal of Economics*, 408–427.

Cronin, Christopher J. 2019. "Insurance-induced moral hazard: a dynamic model of within-year medical care decision making under uncertainty." *International Economic Review*, 60(1): 187–218.

Diaz-Campo, Cecilia S. 2022. "Dynamic Moral Hazard in Nonlinear Health Insurance Contracts." Working paper, SSRN 4651013.

Einav, Liran, Amy Finkelstein, Stephen P. Ryan, Paul Schrimpf, and Mark R. Cullen. 2013. "Selection on Moral Hazard in Health Insurance." *American Economic Review*, 103(1): 178–219.

Ellis, Randall P. 1986. "Rational Behavior in the Presence of Coverage Ceilings and Deductibles." *The RAND Journal of Economics*, 17(2): 158–175.

Guo, Audrey, and Jonathan Zhang. 2019. "What to expect when you are expecting: Are health care consumers forward-looking?" *Journal of Health Economics*, 67: 102216.

Ho, Kate, and Robin S. Lee. 2023. "Health Insurance Menu Design for Large Employers." *The RAND Journal of Economics*, 54(4): 598–637.

Johansson, Naimi, Sonja C. de New, Johannes S. Kunz, Dennis Petrie, and Mikael Svensson. 2023. "Reductions in out-of-pocket prices and forward-looking moral hazard in health care demand." *Journal of Health Economics*, 87: 102710.

Kaiser Family Foundation. 2025. "2025 Employer Health Benefits Survey."

Klein, Tobias J., Martin Salm, and Suraj Upadhyay. 2022. "The response to dynamic incentives in insurance contracts with a deductible: Evidence from a differences-in-regression-discontinuities design." *Journal of Public Economics*, 210: 104660.

Marone, Victoria R, and Adrienne Sabety. 2022. "When should there be vertical choice in health insurance markets?" *American Economic Review*, 112(1): 304–342.

Zeckhauser, Richard. 1970. "Medical insurance: A case study of the tradeoff between risk spreading and appropriate incentives." *Journal of Economic theory*, 2(1): 10–26.

Appendix A Proofs of Section 3.2

Appendix A.1 Derivation of Λ_{cap}

The cap-reaching candidate is globally relevant only if it weakly dominates the full-price candidate. Using

$$V^{\text{ded}}(\lambda) = -\lambda, \qquad V^{\text{cap}}(\lambda) = \frac{\omega}{2}\lambda - M,$$

the cap weakly dominates full-price spending iff

$$V^{\text{cap}}(\lambda) \geq V^{\text{ded}}(\lambda) \iff \frac{\omega}{2}\lambda - M \geq -\lambda \iff \lambda \geq \frac{M}{1+\omega/2}.$$

Define

$$\Lambda_{\text{cap}} \equiv \frac{M}{1+\omega/2}.$$

For $\lambda < \Lambda_{\text{cap}}$, reaching the out-of-pocket maximum cannot be globally optimal. The comparison with the coinsurance candidate is handled separately by $\lambda_{\text{coin-cap}}(D)$.

Appendix A.2 Derivation of the cutoffs $\lambda_{\text{ded-coin}}(D)$ and $\lambda_{\text{coin-cap}}(D)$

Fix (s, M, ω) and deductible D. Using the reduced values

$$V^{\text{ded}}(\lambda) = -\lambda, \qquad V^{\text{coin}}(D;\lambda) = \frac{\omega\lambda}{2}(1-s)^2 - s\lambda - (1-s)D, \qquad V^{\text{cap}}(\lambda) = \frac{\omega}{2}\lambda - M,$$

the pairwise indifference cutoffs are obtained as follows.

Deductible–coinsurance cutoff. Setting $V^{\text{ded}}(\lambda) = V^{\text{coin}}(D;\lambda)$ gives

$$-\lambda = \frac{\omega\lambda}{2}(1-s)^2 - s\lambda - (1-s)D.$$

Rearranging,

$$(1-s)\Big(1 + \frac{\omega}{2}(1-s)\Big)\lambda = (1-s)D,$$

so

$$\lambda_{\text{ded-coin}}(D) = \frac{D}{1+\omega(1-s)/2}.$$

Coinsurance–cap cutoff. Setting $V^{\text{coin}}(D;\lambda) = V^{\text{cap}}(\lambda)$ gives

$$\frac{\omega\lambda}{2}(1-s)^2 - s\lambda - (1-s)D = \frac{\omega}{2}\lambda - M.$$

Thus

$$M - (1-s)D = s\Big(1 + \omega\left(1 - \frac{s}{2}\right)\Big)\lambda,$$

and hence

$$\lambda_{\text{coin-cap}}(D) = \frac{M - (1-s)D}{s\left(1 + \omega\left(1 - \frac{s}{2}\right)\right)}.$$

These are pairwise indifference thresholds. Global region assignment is characterized in the proof of Proposition 1.

Appendix A.3 Proofs for Propositions 1 and 2

Proof of Proposition 1. Define

$$\lambda_{\text{cap}}(D) \equiv \max\{\Lambda_{\text{cap}}, \lambda_{\text{coin-cap}}(D)\}.$$

Then at deductible D, the consumer is in the coinsurance region if $\lambda \in [\lambda_{\text{ded-coin}}(D), \lambda_{\text{cap}}(D))$ and above the MOOP if $\lambda \geq \lambda_{\text{cap}}(D)$.

Now fix $D_0 < D_1$. Since $\lambda_{\text{ded-coin}}(D)$ is strictly increasing in D, types in

$$[\lambda_{\text{ded-coin}}(D_0), \lambda_{\text{ded-coin}}(D_1))$$

switch from the coinsurance region to the deductible region. Since $\lambda_{\text{coin-cap}}(D)$ is strictly decreasing in D and Λ_{cap} is constant, $\lambda_{\text{cap}}(D)$ is weakly decreasing, so types in

$$[\lambda_{\text{cap}}(D_1), \lambda_{\text{cap}}(D_0))$$

switch from the coinsurance region to the MOOP region. No other switches occur, since the coinsurance region shrinks from both sides as D rises.

Using

$$m^*_{\text{ded}}(\lambda) = \lambda, \qquad m^*_{\text{coin}}(\lambda) = \lambda[1 + \omega(1-s)], \qquad m^*_{\text{cap}}(\lambda) = \lambda(1+\omega),$$

it follows that

$$\Delta m(\lambda) = \begin{cases} -\omega(1-s)\lambda, & \lambda \in [\lambda_{\text{ded-coin}}(D_0), \lambda_{\text{ded-coin}}(D_1)), \\ \omega s\,\lambda, & \lambda \in [\lambda_{\text{cap}}(D_1), \lambda_{\text{cap}}(D_0)), \\ 0, & \text{otherwise.} \end{cases}$$

□

Proof of Proposition 2. By Proposition 1, the up-switcher set is

$$[\lambda_{\text{cap}}(D_1), \lambda_{\text{cap}}(D_0)), \qquad \lambda_{\text{cap}}(D) = \max\{\Lambda_{\text{cap}}, \lambda_{\text{coin-cap}}(D)\}.$$

Since $\lambda_{\text{coin-cap}}(D)$ is strictly decreasing in D and $D_0 < D_1$, this set is nonempty iff $\Lambda_{\text{cap}} < \lambda_{\text{coin-cap}}(D_0)$.

Using

$$\lambda_{\text{coin-cap}}(D) = \frac{M-(1-s)D}{s\left(1+\omega\left(1-\frac{s}{2}\right)\right)} \qquad \text{and} \qquad \Lambda_{\text{cap}} = \frac{M}{1+\omega/2},$$

the threshold deductible satisfying $\lambda_{\text{coin-cap}}(D) = \Lambda_{\text{cap}}$ is

$$D_{\text{crit}} = M\,\frac{\omega(1-s)+2}{\omega+2}.$$

Because $\lambda_{\text{coin-cap}}(D)$ is strictly decreasing,

$$\lambda_{\text{coin-cap}}(D_0) > \Lambda_{\text{cap}} \iff D_0 < D_{\text{crit}}.$$

Hence, the up-switcher set is nonempty iff $D_0 < D_{\text{crit}}$. If $D_0 = 0$, then $0 < D_{\text{crit}}$, so the up-switcher set is nonempty for every $D_1 \in (0, M)$.

□

Supplemental Appendix for

The Generosity Paradox: When Less Generous Insurance Raises Spending

Iris SooJin Park

March 16, 2026

Appendix B Changes in Coinsurance Rates

As in the deductible case, let Λ_{cap} denote the smallest realized health need for which reaching the MOOP weakly dominates stopping below the deductible. In the coinsurance case, I also impose $D \leq \Lambda_{\text{cap}}$. This ensures that

$$\lambda_{\text{ded-coin}}(s) < \lambda_{\text{coin-cap}}(s).$$

As a result, consumers do not jump directly from stopping below the deductible to reaching the MOOP. Instead, there remains an intermediate range of health needs for which stopping in the coinsurance region is optimal. Appendix Figure B.1 plots this condition in $(\omega, D/M)$-space and shows that it holds for the vast majority of empirically relevant plan combinations.

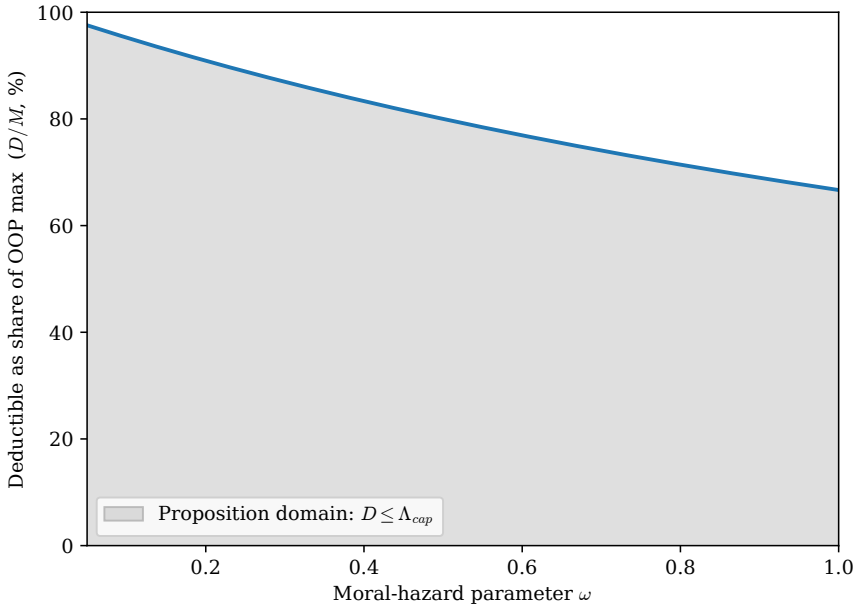
Proposition B.1 characterizes how a coinsurance increase from s_0 to s_1 partitions realized health needs into regions where spending rises, falls, or remains unchanged. Under the maintained condition $D \leq \Lambda_{\text{cap}}$, spending increases exactly for health needs in $\big[\lambda_{\text{coin-cap}}(s_1),\ \lambda_{\text{coin-cap}}(s_0)\big)$, with $\Delta m(\lambda) = \omega s_0 \lambda$. This interval is nonempty. Unlike in the deductible case, spending falls for two distinct groups. One group is pushed from the coinsurance region to below the deductible, generating $\Delta m(\lambda) = -\omega(1 - s_0)\lambda$. The other remains in the coinsurance region under both rates, with $\Delta m(\lambda) = -\omega(s_1 - s_0)\lambda$. Outside these intervals, optimal spending is unchanged.

Proposition B.1 (Heterogeneous Spending Responses to a Coinsurance Increase)**.** *Fix* (D, M, ω) *and assume*

$$D \leq \Lambda_{cap} \equiv \frac{M}{1 + \omega/2}.$$

Consider a coinsurance increase from s_0 *to* s_1 *with* $0 < s_0 < s_1 < 1$.

Figure B.1: Parameter Space Satisfying $D \leq \Lambda_{\text{cap}}$



Note: The figure plots the maintained-domain restriction for Proposition B.1 in $(\omega, D/M)$-space. The shaded region indicates parameter values satisfying $D \leq \Lambda_{\text{cap}}$, with the vertical axis reported as $100 \times (D/M)$.

Under the maintained assumption, $\lambda_{\text{coin-cap}}(s_1) < \lambda_{\text{coin-cap}}(s_0)$. *Thus, in particular, the up-switcher set*

$$\big[\lambda_{\text{coin-cap}}(s_1),\ \lambda_{\text{coin-cap}}(s_0)\big)$$

is nonempty.

Then the change in optimal spending satisfies

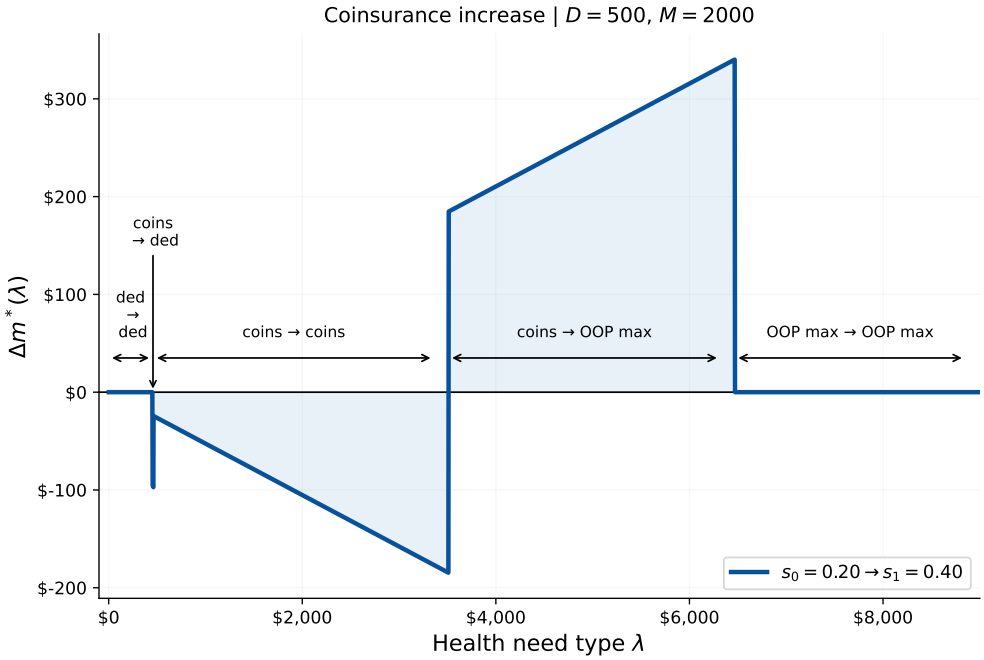
$$\Delta m(\lambda) = \begin{cases} -\omega(1-s_0)\lambda, & \text{if } \lambda \in \big[\lambda_{\text{ded-coin}}(s_0),\ \lambda_{\text{ded-coin}}(s_1)\big), \quad (\textit{coin} \to \textit{ded}), \\ -\omega(s_1-s_0)\lambda, & \text{if } \lambda \in \big[\lambda_{\text{ded-coin}}(s_1),\ \lambda_{\text{coin-cap}}(s_1)\big), \quad (\textit{coin} \to \textit{coin}), \\ \omega s_0 \lambda, & \text{if } \lambda \in \big[\lambda_{\text{coin-cap}}(s_1),\ \lambda_{\text{coin-cap}}(s_0)\big), \quad (\textit{coin} \to \textit{cap}), \\ 0, & \textit{otherwise.} \end{cases}$$

Proof. See below.

Figure B.2 plots $\Delta m(\lambda)$ for a coinsurance increase from 0.20 to 0.40, with $D = \$500$, $M = \$2,000$, and $\omega = 0.263$. Consistent with Proposition B.1, spending is unchanged for consumers below the deductible or at the MOOP under both coinsurance rates. Among consumers initially in the coinsurance region, spending falls for two groups and rises for one.

Lower-need consumers in $[452, 463)$ are pushed below the deductible, while intermediate-need consumers in $[463, 3,511)$ remain in the coinsurance region but reduce spending in response to the higher marginal price. Higher-need consumers in $[3,511, 6,468)$ are pushed to the MOOP and increase spending. As in Figure 2, the spending response is larger for higher-need consumers within each group, reflecting that $\Delta m(\lambda)$ is proportional to λ.

Figure B.2: Spending Responses to Coinsurance Increases by Health Need



Note: The figure plots the change in optimal spending, $\Delta m(\lambda) = m^*(\lambda; s_1) - m^*(\lambda; s_0)$, by realized health need λ, for a coinsurance increase from 0.20 to 0.40, holding the deductible fixed at $D = \$500$. The parameters are $M = \$2{,}000$ and $\omega = 0.263$, calibrated to estimates from Ho and Lee (2023). The annotations label how types are reassigned across regions of the contract as coinsurance rises. The spending response varies with health need because different types face different parts of the contract.

Proof of Proposition B.1. Fix (D, M, ω) and assume

$$D \leq \Lambda_{\text{cap}} \equiv \frac{M}{1 + \omega/2}.$$

For any $s \in (0, 1)$, define

$$\lambda_{\text{ded-coin}}(s) = \frac{D}{1 + \omega(1-s)/2}, \qquad \lambda_{\text{coin-cap}}(s) = \frac{M - (1-s)D}{s\left(1 + \omega - \frac{\omega}{2}s\right)}.$$

First,

$$\lambda_{\text{ded-coin}}(s) = \frac{D}{1 + \omega(1-s)/2} < D \leq \Lambda_{\text{cap}},$$

since $1 + \omega(1-s)/2 > 1$.

Next, define

$$\Delta(s;\lambda) \equiv V^{\text{coin}}(s;\lambda) - V^{\text{cap}}(\lambda).$$

Then $\lambda_{\text{coin-cap}}(s)$ is the unique solution to $\Delta(s;\lambda) = 0$, and

$$\Delta(s;\lambda) = \frac{\omega\lambda}{2}s^2 + \big(D - (1+\omega)\lambda\big)s + (M - D).$$

For fixed $s \in (0,1)$,

$$\frac{\partial\Delta(s;\lambda)}{\partial\lambda} = \frac{\omega}{2}s^2 - (1+\omega)s = s\Big(\frac{\omega}{2}s - (1+\omega)\Big) < 0,$$

so $\Delta(s;\lambda)$ is strictly decreasing in λ. Evaluating at $\Lambda_{\text{cap}} = M/(1+\omega/2)$ gives

$$\Delta(s;\Lambda_{\text{cap}}) = \frac{(1-s)\big(M(2+\omega(1-s)) - D(2+\omega)\big)}{2+\omega}.$$

Because $D \le \Lambda_{\text{cap}} = 2M/(2+\omega)$, the bracketed term is strictly positive for all $s \in (0,1)$, hence

$$\Delta(s;\Lambda_{\text{cap}}) > 0.$$

Since $\Delta(s;\lambda)$ is strictly decreasing in λ and vanishes at $\lambda = \lambda_{\text{coin-cap}}(s)$, it follows that

$$\lambda_{\text{coin-cap}}(s) > \Lambda_{\text{cap}} \qquad \text{for all } s \in (0,1).$$

Thus the coinsurance region is a genuine middle region.

Therefore, for each $s \in (0,1)$, the global region assignment is as follows: the consumer is below the deductible if $\lambda < \lambda_{\text{ded-coin}}(s)$, within the coinsurance region if

$$\lambda \in \big[\lambda_{\text{ded-coin}}(s), \lambda_{\text{coin-cap}}(s)\big),$$

and above the MOOP if

$$\lambda \ge \lambda_{\text{coin-cap}}(s).$$

Now compare two coinsurance rates $0 < s_0 < s_1 < 1$. Since

$$\lambda_{\text{ded-coin}}(s) = \frac{D}{1+\omega(1-s)/2},$$

the lower cutoff is strictly increasing in s, so

$$\lambda_{\text{ded-coin}}(s_0) < \lambda_{\text{ded-coin}}(s_1).$$

To study the upper cutoff, differentiate $\Delta(s;\lambda)$ with respect to s:

$$\frac{\partial \Delta(s;\lambda)}{\partial s} = \omega\lambda s + \big(D - (1+\omega)\lambda\big) = D - \lambda[1+\omega(1-s)].$$

Since

$$m^*_{\text{coin}}(s;\lambda) = \lambda[1+\omega(1-s)],$$

we have

$$\frac{\partial \Delta(s;\lambda)}{\partial s} = D - m^*_{\text{coin}}(s;\lambda).$$

For any $\lambda \geq \Lambda_{\text{cap}}$ and any $s \in (0,1)$,

$$m^*_{\text{coin}}(s;\lambda) > \lambda \geq \Lambda_{\text{cap}} \geq D,$$

so

$$\frac{\partial \Delta(s;\lambda)}{\partial s} < 0.$$

Hence, for each fixed $\lambda \geq \Lambda_{\text{cap}}$, $\Delta(s;\lambda)$ is strictly decreasing in s. Since $\lambda_{\text{coin-cap}}(s) > \Lambda_{\text{cap}}$, the equation $\Delta(s;\lambda) = 0$ is reached at a lower value of λ when s rises. Therefore

$$\lambda_{\text{coin-cap}}(s_1) < \lambda_{\text{coin-cap}}(s_0).$$

Combining the two cutoff movements yields

$$\lambda_{\text{ded-coin}}(s_0) < \lambda_{\text{ded-coin}}(s_1) < \lambda_{\text{coin-cap}}(s_1) < \lambda_{\text{coin-cap}}(s_0),$$

so the up-switcher set

$$\big[\lambda_{\text{coin-cap}}(s_1),\, \lambda_{\text{coin-cap}}(s_0)\big)$$

is nonempty.

It remains to compute spending changes. The region-specific spending levels are

$$m^*_{\text{ded}}(\lambda) = \lambda, \qquad m^*_{\text{coin}}(s;\lambda) = \lambda[1+\omega(1-s)], \qquad m^*_{\text{cap}}(\lambda) = \lambda(1+\omega).$$

Thus

$$\Delta m(\lambda) = \begin{cases} m^*_{\text{ded}}(\lambda) - m^*_{\text{coin}}(s_0;\lambda) = -\omega(1-s_0)\lambda, & \text{if } \lambda \in \big[\lambda_{\text{ded-coin}}(s_0),\ \lambda_{\text{ded-coin}}(s_1)\big), \\ m^*_{\text{coin}}(s_1;\lambda) - m^*_{\text{coin}}(s_0;\lambda) = -\omega(s_1-s_0)\lambda, & \text{if } \lambda \in \big[\lambda_{\text{ded-coin}}(s_1),\ \lambda_{\text{coin-cap}}(s_1)\big), \\ m^*_{\text{cap}}(\lambda) - m^*_{\text{coin}}(s_0;\lambda) = \omega s_0 \lambda, & \text{if } \lambda \in \big[\lambda_{\text{coin-cap}}(s_1),\ \lambda_{\text{coin-cap}}(s_0)\big), \\ 0, & \text{otherwise.} \end{cases}$$

No other switches occur, since the coinsurance region shrinks from both sides as s rises. □

Appendix C Dynamic Model with Uncertainty

Appendix C.1 Model setup

This section extends the model to forward-looking consumers who face uncertainty about future health needs, following the dynamic approach in Ellis (1986); Cronin (2019); Diaz-Campo (2022). The health benefit function $h(\cdot)$ remains as in equation 3, but is now applied period by period, so utility from current care in period t is written as $h(m_t;\lambda_t,\omega)$, with the same functional form in each period. Spending is therefore chosen over T periods rather than once at the start of the year. In each period t, the consumer observes current health need λ_t and cumulative spending through the previous period, C_{t-1}. These state variables determine both the current value of care and how additional spending affects future OOP costs. Given cumulative spending C_{t-1}, incremental OOP spending from current spending m_t is

$$\Delta\text{OOP}_t(C_{t-1}, m_t) = \text{OOP}(C_{t-1} + m_t) - \text{OOP}(C_{t-1}),$$

and cumulative spending evolves as $C_t = C_{t-1} + m_t$.

Because spending decisions are made sequentially under uncertainty, the consumer in each period must weigh the current health benefit of care against both its immediate OOP cost and its effect on future OOP costs through cumulative spending. Let $V_t(C_{t-1},\lambda_t)$ denote the consumer's value in period t after observing current health need λ_t and cumulative spending C_{t-1}. The value function satisfies

$$V_t(C_{t-1},\lambda_t) = \max_{m_t \geq 0}\Big\{h(m_t;\lambda_t,\omega) - \Delta\text{OOP}_t(C_{t-1},m_t) + \delta\,\mathbb{E}_t\big[V_{t+1}(C_t,\lambda_{t+1})\big]\Big\},$$

where $\delta \in (0,1)$ is the within-year discount factor and the expectation is taken over next period's health need λ_{t+1}. At the terminal period $t = T$, the continuation value vanishes, so the problem reduces to the static case.

For notational convenience, define the continuation value

$$W_{t+1}(C) \equiv \mathbb{E}_t[V_{t+1}(C, \lambda_{t+1})],$$

which gives the expected value of entering period $t+1$ with cumulative spending C. Then the period-t objective can be written as

$$G_t(m_t; \lambda_t, C_{t-1}) \equiv h(m_t; \lambda_t, \omega) - \Delta OOP_t(C_{t-1}, m_t) + \delta\, W_{t+1}(C_{t-1} + m_t).$$

The shadow price of current spending, given cumulative spending C_{t-1}, is defined as

$$p_t(C_{t-1}, m_t) \equiv \frac{\partial}{\partial m_t}\Delta\mathrm{OOP}_t(C_{t-1}, m_t) - \delta\, \frac{\partial}{\partial m_t} W_{t+1}(C_{t-1} + m_t).$$

The first term is the spot marginal OOP price of current spending. The second term is the marginal continuation-value benefit of spending more today. By raising cumulative spending C_t, current spending moves the consumer closer to the MOOP and can reduce future OOP costs.

For periods $t < T$, the consumer chooses spending so that the marginal health benefit of current spending equals its shadow price. The first-order condition for an interior optimum is $h_m(m_t^*; \lambda_t, \omega) = p_t(C_{t-1}, m_t^*)$. Under the quadratic h used above, this implies

$$m_t^* = \lambda_t + \omega\lambda_t\big(1 - p_t(C_{t-1}, m_t^*)\big).$$

Thus, the optimal spending rule takes the same form as in the static model, but the relevant marginal price is now the shadow price, which incorporates uncertainty about future health needs through the continuation value.

Appendix C.2 The Discrete Switching Channel

The first channel arises when a deductible increase makes it optimal for some consumers who would otherwise remain below the MOOP to push through and reach it instead. This occurs when the value of reaching the MOOP exceeds the value of remaining in the coinsurance region.

To formalize this, fix period t, state (C_{t-1}, λ_t), and deductible D. Let $V_t^{\text{below cap}}(D; \lambda_t, C_{t-1})$ denote the maximized value from remaining below the MOOP:

$$V_t^{\text{below cap}}(D; \lambda_t, C_{t-1}) \equiv \max_{\{m_t \geq 0:\, C_{t-1}+m_t < \bar{m}(D)\}} G_t(m_t; \lambda_t, C_{t-1}, D).$$

Likewise, let $V_t^{\text{above cap}}(D; \lambda_t, C_{t-1})$ denote the maximized value from choosing spending high enough to reach the MOOP:

$$V_t^{\text{above cap}}(D; \lambda_t, C_{t-1}) \equiv \max_{\{m_t \geq 0:\, C_{t-1}+m_t \geq \bar{m}(D)\}} G_t(m_t; \lambda_t, C_{t-1}, D).$$

Define $\Psi_t^{\text{cap}}(D; \lambda_t, C_{t-1})$ as the net value of pushing through to the MOOP rather than remaining below it:

$$\Psi_t^{\text{cap}}(D; \lambda_t, C_{t-1}) \equiv V_t^{\text{above cap}}(D; \lambda_t, C_{t-1}) - V_t^{\text{below cap}}(D; \lambda_t, C_{t-1}).$$

When $\Psi_t^{\text{cap}}(D; \lambda_t, C_{t-1}) \geq 0$, reaching the MOOP is weakly optimal.

Proposition C.1 characterizes the range of new cap-reaching consumers generated by a deductible increase from D_0 to D_1. Specifically, it focuses on consumers who begin period t in the coinsurance region ($D < C_{t-1} < \bar{m}(D)$), so cumulative spending at the start of the period is already above the deductible but still below the threshold for reaching the MOOP.

A natural regularity condition is that $\Psi_t^{\text{cap}}(D; \lambda_t, C_{t-1})$ is weakly increasing and continuous in λ_t. Economically, this means that as current health need rises, reaching the MOOP becomes weakly more attractive relative to remaining below it, and that this change occurs smoothly. Together with the conditions that $\Psi_t^{\text{cap}}(D; \lambda_t, C_{t-1}) < 0$ for sufficiently low λ_t and $\Psi_t^{\text{cap}}(D; \lambda_t, C_{t-1}) \geq 0$ for sufficiently high λ_t, this implies a well-defined cutoff level of health need above which reaching the MOOP is weakly optimal.

Proposition C.1 (Forward-looking cap switchers under sufficient conditions)**.** *Fix period t, prior cumulative spending C_{t-1}, and two deductibles $D_0 < D_1$. Assume*

$$D < C_{t-1} < \bar{m}(D) \qquad \text{for all } D \in [D_0, D_1].$$

For each $D \in [D_0, D_1]$, let

$$\lambda_{\text{cap}}^{FL}(D; C_{t-1}) \equiv \inf\{\lambda_t : \Psi_t^{\text{cap}}(D; \lambda_t, C_{t-1}) \geq 0\},$$

where $\Psi_t^{\text{cap}}(D; \lambda_t, C_{t-1})$ is continuous and weakly increasing in λ_t, negative for sufficiently

low λ_t, and nonnegative for sufficiently high λ_t.

Suppose further that for every fixed λ_t,

$$D \mapsto \Psi_t^{\text{cap}}(D; \lambda_t, C_{t-1})$$

is weakly increasing on $[D_0, D_1]$. Then

$$\lambda_{\text{cap}}^{FL}(D_1; C_{t-1}) \leq \lambda_{\text{cap}}^{FL}(D_0; C_{t-1}).$$

If strict, the new cap-reaching consumers are exactly

$$\big[\lambda_{\text{cap}}^{FL}(D_1; C_{t-1}), \lambda_{\text{cap}}^{FL}(D_0; C_{t-1})\big).$$

A higher deductible can make cap-reaching more attractive because, once the consumer is already in the coinsurance region, reaching the MOOP requires less additional OOP spending under a higher deductible. This direct gain is reinforced when the continuation-value loss from a higher deductible becomes weaker as cumulative spending rises. An economically interpretable sufficient condition for the monotonicity assumption in Proposition C.1 is therefore that $W_{t+1,D}(C; D)$ is weakly increasing in cumulative spending C. This means that the continuation-value effect of a higher deductible becomes less harmful as the consumer moves closer to the MOOP, since future OOP exposure is then less sensitive to the deductible. As a result, a deductible increase can lower the cap-entry cutoff and generate a new range of cap-reaching consumers, given by $[\lambda_{\text{cap}}^{\text{FL}}(D_1; C_{t-1}), \lambda_{\text{cap}}^{\text{FL}}(D_0; C_{t-1}))$.

Proof. For each deductible D, define

$$\mathcal{S}_{\text{cap}}(D; C_{t-1}) \equiv \{\lambda_t : \Psi_t^{\text{cap}}(D; \lambda_t, C_{t-1}) \geq 0\}.$$

By assumption, for each fixed D, the map

$$\lambda_t \mapsto \Psi_t^{\text{cap}}(D; \lambda_t, C_{t-1})$$

is continuous and weakly increasing, negative for sufficiently low λ_t and nonnegative for sufficiently high λ_t. Hence

$$\mathcal{S}_{\text{cap}}(D; C_{t-1}) = [\lambda_{\text{cap}}^{FL}(D; C_{t-1}), \infty).$$

Now fix λ_t. Since

$$D \mapsto \Psi_t^{\text{cap}}(D; \lambda_t, C_{t-1})$$

is weakly increasing on $[D_0, D_1]$, we have

$$\Psi_t^{\text{cap}}(D_1; \lambda_t, C_{t-1}) \geq \Psi_t^{\text{cap}}(D_0; \lambda_t, C_{t-1}).$$

Therefore any λ_t for which cap-reaching is weakly optimal under D_0 is also one for which cap-reaching is weakly optimal under D_1. Thus

$$\mathcal{S}_{\text{cap}}(D_0; C_{t-1}) \subseteq \mathcal{S}_{\text{cap}}(D_1; C_{t-1}).$$

Since both sets are upper intervals, it follows that

$$\lambda_{\text{cap}}^{FL}(D_1; C_{t-1}) \leq \lambda_{\text{cap}}^{FL}(D_0; C_{t-1}).$$

If strict, the newly induced cap-reaching consumers are exactly

$$\mathcal{S}_{\text{cap}}(D_1; C_{t-1}) \setminus \mathcal{S}_{\text{cap}}(D_0; C_{t-1}) = [\lambda_{\text{cap}}^{FL}(D_1; C_{t-1}), \lambda_{\text{cap}}^{FL}(D_0; C_{t-1})).$$

□

Appendix C.3 The Smooth Channel

The second channel is a smooth spending response within the coinsurance region. It captures how a deductible increase can raise spending even when the consumer remains strictly inside the coinsurance region. In this case, there is no discrete switch to the MOOP. The spot price of care is unchanged, so any increase in spending must come from the dynamic value of moving closer to the MOOP and thereby lowering expected future OOP costs.

Proposition C.2 (Forward-looking smooth increases)**.** *Fix period t, prior cumulative spending C_{t-1}, current health need λ_t, and deductible D_0. Suppose that for all D in a neighborhood of D_0,*

$$D < C_{t-1} \quad \textit{and} \quad D < C_{t-1} + m_t^*(D, \lambda_t) < \bar{m}(D).$$

Assume that $G_t(m_t; \lambda_t, C_{t-1}, D)$ is strictly concave in m_t and that $W_{t+1}(C; D)$ is differentiable in (C, D).

In particular, when

$$W_{t+1,CD}\big(C_t^*(D_0,\lambda_t);D_0\big) > 0,$$

we have

$$\frac{dm_t^*(D,\lambda_t)}{dD}\Big|_{D_0} > 0.$$

Therefore,

$$m_t^*(D_1,\lambda_t) > m_t^*(D_0,\lambda_t)$$

for all sufficiently small deductible increases $D_1 > D_0$.

Proposition C.2 characterizes when the smooth channel raises spending. Consider a consumer who has already passed the deductible by the start of period t ($D < C_{t-1}$) and remains below the MOOP after making the period-t spending choice ($D < C_{t-1}+m_t^*(D,\lambda_t) < \bar{m}(D)$). The choice therefore stays entirely within the coinsurance region. As a result, the current marginal price of care is unchanged, so any increase in spending must reflect the future value of moving closer to the MOOP.

The proposition also imposes two regularity conditions. First, G_t is strictly concave in m_t, a standard assumption that captures diminishing marginal value of additional utilization. As current spending rises, each additional dollar of care contributes less to the consumer's overall payoff, ensuring that the period-t problem has a unique interior optimum. Second, $W_{t+1}(C;D)$ is differentiable in (C,D), so small changes in cumulative spending and the deductible lead to smooth changes in continuation value.

Proposition C.2 gives a general sufficient condition for the smooth channel. A deductible increase raises spending within the coinsurance region when the continuation value satisfies

$$W_{t+1,CD}\big(C_t^*(D_0,\lambda_t);D_0\big) > 0.$$

The cross-partial $W_{t+1,CD}(C;D)$ measures how the marginal continuation value of cumulative spending changes when the deductible rises. If $W_{t+1,CD} > 0$, then an extra dollar of spending today becomes more valuable under a higher deductible because it moves the consumer closer to the MOOP. Under this condition, spending rises as the deductible increases from D_0 to D_1, so that $m_t^*(D_1,\lambda_t) > m_t^*(D_0,\lambda_t)$.

Proof of Proposition C.2. Consider a neighborhood of D_0 such that

$$D < C_{t-1} \quad \text{and} \quad D < C_{t-1} + m_t^*(D,\lambda_t) < \bar{m}(D).$$

The first inequality means the consumer enters period t already in the coinsurance region. The second means that, after choosing optimal current utilization, cumulative spending remains below the level required to reach the MOOP. Therefore, throughout the interval from C_{t-1} to $C_{t-1} + m_t^*(D, \lambda_t)$, the marginal out-of-pocket price is the coinsurance rate s. It follows that the period-t out-of-pocket increment is locally

$$\Delta OOP_t(C_{t-1}, m_t; D) = s\, m_t,$$

so it does not vary with D.

Recall that

$$W_{t+1}(C; D) \equiv \mathbb{E}_t[V_{t+1}(C, \lambda_{t+1}; D)]$$

and

$$G_t(m_t; \lambda_t, C_{t-1}, D) \equiv h(m_t; \lambda_t, \omega) - \Delta OOP_t(C_{t-1}, m_t; D) + \delta\, W_{t+1}(C_{t-1} + m_t; D).$$

Define

$$H(m_t, D) \equiv \frac{\partial G_t}{\partial m_t}(m_t; \lambda_t, C_{t-1}, D) = h_m(m_t; \lambda_t, \omega) - s + \delta\, W_{t+1,C}(C_{t-1} + m_t; D).$$

Since $m_t^*(D, \lambda_t)$ is an interior optimizer, it satisfies

$$H(m_t^*(D, \lambda_t), D) = 0.$$

By assumption, $W_{t+1}(C; D)$ is differentiable in (C, D), so $H(m_t, D)$ is continuously differentiable in (m_t, D). Moreover, strict concavity of G_t in m_t implies

$$H_m(m_t^*(D_0, \lambda_t), D_0) < 0.$$

Hence the first-order condition pins down the optimum locally as a differentiable function of D.

Next,

$$H_D(m_t, D) = \delta\, W_{t+1,CD}(C_{t-1} + m_t; D),$$

because on this neighborhood the current-period out-of-pocket increment is $s\, m_t$, so the only D-dependence in the first-order condition comes through the continuation value. The

Implicit Function Theorem therefore gives

$$\frac{dm_t^*(D,\lambda_t)}{dD}\Big|_{D_0} = -\frac{H_D}{H_m}\Big|_{(m_t^*(D_0,\lambda_t),D_0)} = -\frac{\delta\, W_{t+1,CD}\big(C_{t-1}+m_t^*(D_0,\lambda_t);D_0\big)}{H_m\big(m_t^*(D_0,\lambda_t),D_0\big)}.$$

Because the denominator is negative, the sign of the spending response is the sign of the cross-partial:

$$\operatorname{sign}\left(\frac{dm_t^*(D,\lambda_t)}{dD}\Big|_{D_0}\right) = \operatorname{sign}\big(W_{t+1,CD}\big(C_t^*(D_0,\lambda_t);D_0\big)\big)\,,$$

where

$$C_t^*(D_0,\lambda_t) = C_{t-1} + m_t^*(D_0,\lambda_t).$$

Hence, if $W_{t+1,CD}(C_t^*(D_0,\lambda_t);D_0) > 0$, then

$$\frac{dm_t^*(D,\lambda_t)}{dD}\Big|_{D_0} > 0.$$

Therefore $m_t^*(D_1,\lambda_t) > m_t^*(D_0,\lambda_t)$ for all sufficiently small deductible increases $D_1 > D_0$. □

Appendix C.3.1 Two-Period Verification under Lognormal Health Risk

To illustrate the sufficient condition in Proposition C.2 more concretely, consider the two-period model under the lognormal specification used in the existing literature. Because period 2 is final, the continuation value relevant for the period-1 choice simplifies to

$$W_2(C;D) \equiv \mathbb{E}[V_2(C,\lambda_2;D)].$$

Lemma C.1 shows that when period 2 begins in the coinsurance region and period-2 health need is lognormally distributed, the continuation value satisfies

$$W_{2,CD}(C;D) > 0.$$

Thus, in the two-period model with lognormal health risk, the sufficient condition in Proposition C.2 holds, providing a direct justification for the smooth spending response in period 1.

Lemma C.1 (Two-period continuation-value cross-partial)**.** *Consider the two-period model, with period 2 terminal, and suppose period-2 health need λ_2 is lognormally distributed with*

density f_2. Fix a period-2 state (C, D) such that

$$D < C < \bar{m}(D),$$

and define the remaining spending needed to reach the MOOP by

$$x(C, D) \equiv \bar{m}(D) - C.$$

Then the continuation value relevant for period 1, $W_2(C; D) \equiv \mathbb{E}[V_2(C, \lambda_2; D)]$, satisfies

$$W_{2,CD}(C; D) = \frac{1 - s}{1 + \omega\left(1 - \frac{s}{2}\right)} f_2\left(\frac{x(C, D)}{1 + \omega\left(1 - \frac{s}{2}\right)}\right) > 0.$$

Proof of Lemma C.1. Fix (C, D) with $D < C < \bar{m}(D)$ and define

$$x \equiv x(C, D) = \bar{m}(D) - C > 0.$$

Since the consumer enters period 2 already in the coinsurance region, the period-2 out-of-pocket increment is

$$\Delta OOP_2(C, m_2; D) = \begin{cases} s\, m_2, & m_2 < x, \\ s\, x, & m_2 \geq x. \end{cases}$$

Thus the terminal problem is

$$V_2(C, \lambda_2; D) = \max_{m_2 \geq 0} \{h(m_2; \lambda_2, \omega) - s \min\{m_2, x\}\}.$$

If the consumer remains below the MOOP, the optimal choice is

$$m_2^{\text{coin}}(\lambda_2) = (1 + \omega(1 - s))\lambda_2,$$

with value

$$V^{\text{coin}}(\lambda_2) = h((1 + \omega(1 - s))\lambda_2; \lambda_2, \omega) - s(1 + \omega(1 - s))\lambda_2.$$

If the consumer reaches the MOOP, the out-of-pocket cost is the fixed amount $s\, x$. Once the MOOP is reached, additional spending is free at the margin, so the optimal choice solves the full-coverage problem. Hence

$$m_2^{\text{cap}}(\lambda_2) = (1 + \omega)\lambda_2,$$

and the associated value is

$$h((1+\omega)\lambda_2;\lambda_2,\omega) - s\,x.$$

Define

$$V^{\text{cap}}(\lambda_2) \equiv h((1+\omega)\lambda_2;\lambda_2,\omega)\,,$$

so the cap-reaching value is

$$V^{\text{cap}}(\lambda_2) - s\,x.$$

Comparing the two values,

$$V^{\text{cap}}(\lambda_2) - s\,x - V^{\text{coin}}(\lambda_2) = s\left[\lambda_2\Big(1+\omega\Big(1-\frac{s}{2}\Big)\Big) - x\right].$$

Hence reaching the MOOP is optimal if and only if

$$\lambda_2 \geq \lambda_2^{\text{cap}}(x) \equiv \frac{x}{1+\omega\left(1-\frac{s}{2}\right)}.$$

Therefore

$$W_2(C;D) = \int_0^{\lambda_2^{\text{cap}}(x)} V^{\text{coin}}(\lambda)\,f_2(\lambda)\,d\lambda + \int_{\lambda_2^{\text{cap}}(x)}^{\infty} \big(V^{\text{cap}}(\lambda) - s\,x\big)\,f_2(\lambda)\,d\lambda.$$

Because the two branches coincide at the cutoff, differentiation with respect to x gives

$$W_{2,x}(x) = -s\left[1 - F_2(\lambda_2^{\text{cap}}(x))\right].$$

Differentiating again,

$$W_{2,xx}(x) = \frac{s}{1+\omega\left(1-\frac{s}{2}\right)}\,f_2(\lambda_2^{\text{cap}}(x))\,.$$

Since the terminal problem depends on (C,D) only through

$$x(C,D) = \bar{m}(D) - C,$$

the continuation value can be written as

$$W_2(C;D) = \widetilde{W}_2(x(C,D)).$$

By the chain rule,

$$W_{2,C}(C;D) = \widetilde{W}_2'(x)\,x_C$$

and

$$W_{2,CD}(C;D) = \widetilde{W}_2''(x)\, x_C x_D + \widetilde{W}_2'(x)\, x_{CD}.$$

Because

$$x(C,D) = \bar{m}(D) - C,$$

we have

$$x_C = -1, \qquad x_D = \bar{m}'(D) = 1 - \frac{1}{s} = -\frac{1-s}{s}, \qquad x_{CD} = 0.$$

Therefore

$$W_{2,CD}(C;D) = \widetilde{W}_2''(x)\, x_C x_D = W_{2,xx}(x)\, x_C x_D.$$

Substituting for $W_{2,xx}(x)$ yields

$$W_{2,CD}(C;D) = \frac{1-s}{1+\omega\left(1-\frac{s}{2}\right)}\, f_2\left(\frac{x(C,D)}{1+\omega\left(1-\frac{s}{2}\right)}\right).$$

Since the lognormal density is strictly positive on $(0,\infty)$ and $x(C,D) > 0$, it follows that

$$W_{2,CD}(C;D) > 0.$$

□

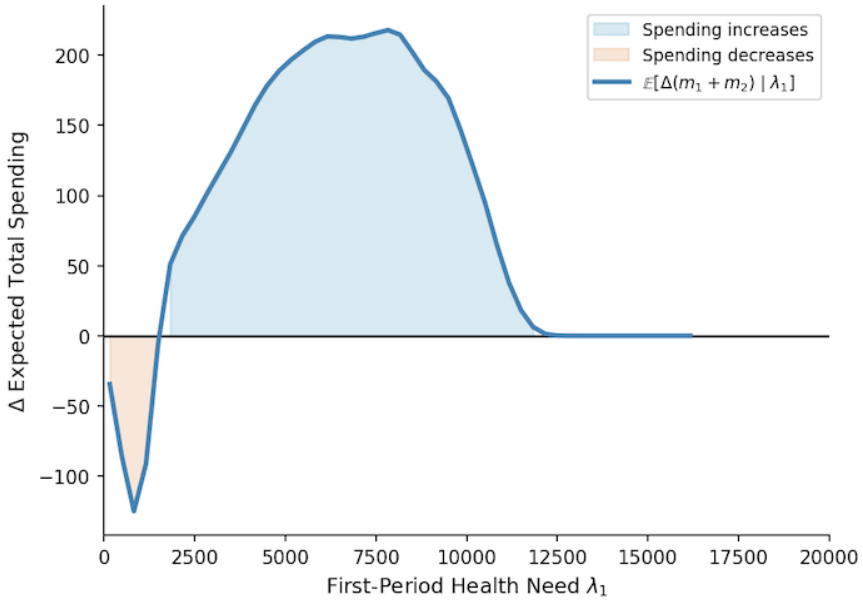
Appendix C.4 Simulation of the Dynamic Model with Uncertainty

Figure C.1 provides a simulation result of the forward-looking model in a two-period setting. It shows how expected total spending changes when the deductible increases from $D_0 = \$500$ to $D_1 = \$1,500$, holding fixed the same contract parameters as Figure 2 and using a period-specific health need distribution inferred from Ho and Lee.[7] The x-axis shows realized first-period health need λ_1. For each value of λ_1, the figure plots the expected change in total spending across both periods, integrating over uncertainty in second-period health need λ_2. In period 1, the consumer observes λ_1 and chooses m_1, taking into account that current spending affects the price of care in period 2. In the terminal period $t = 2$, the consumer observes λ_2 and chooses m_2 with no remaining uncertainty.

The figure shows that the spending response varies substantially with first-period health need. For low values of λ_1, expected total spending falls under the higher deductible. These

[7] The contract parameters are $M = \$2,000$, $s = 0.1$, $\delta = 1$, and $\omega = 0.263$. Health need in each period is drawn independently from a lognormal distribution with mean and standard deviation equal to half the annual values from Ho and Lee, reflecting the two-period structure.

Figure C.1: Spending Responses to a Deductible Increase in the Dynamic Model with Uncertainty

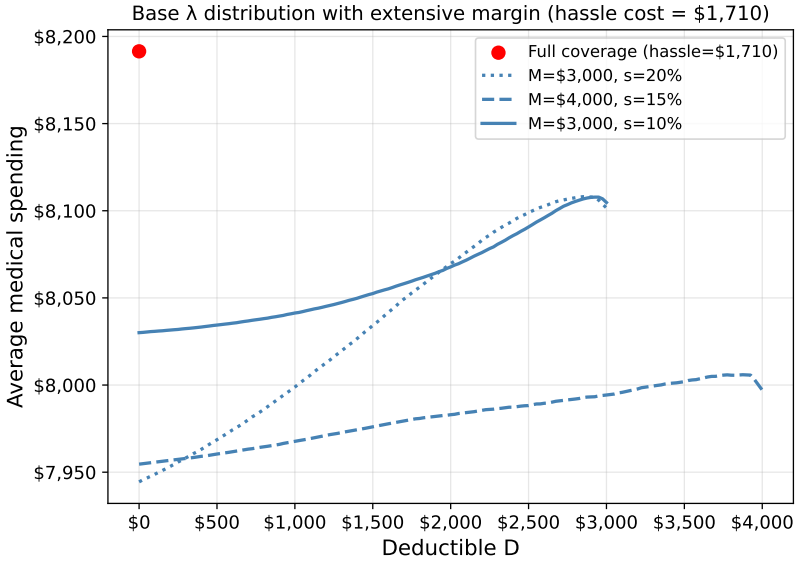


Note: The figure plots $\mathbb{E}[\Delta(m_1+m_2) \mid \lambda_1]$, the expected change in total two-period spending, as a function of realized first-period health need λ_1. The deductible increases from $D_0 = \$500$ to $D_1 = \$1{,}500$. Contract parameters are $M = \$2{,}000$, $s = 0.10$, $\delta = 1$, and $\omega = 0.263$. Health need in each period is drawn from a lognormal distribution with per-period mean \$3,245 and standard deviation \$2,434, equal to half the annual values from Ho and Lee. The plotted line is smoothed using a Gaussian filter for visual clarity.

consumers are pushed back toward the deductible region, where they face a higher current price and reduce spending. Over a broad intermediate range, expected total spending instead rises. This increase reflects both channels explained above. Some consumers raise spending while remaining in the coinsurance region, while others switch from the coinsurance region to the MOOP. For high values of λ_1, the response returns toward zero, since these consumers reach the MOOP under both contracts and the deductible increase no longer affects total spending.

Appendix D Supplementary Figures and Tables

Figure D.1: Average Spending under Multiple Cost-Sharing Designs with Extensive-Margin Hassle Cost



Note: The figure plots average annual medical spending m^* as the deductible varies across three plan designs: 20 percent coinsurance with a $3,000 out-of-pocket maximum, 15 percent coinsurance with a $4,000 out-of-pocket maximum, and 10 percent coinsurance with a $3,000 out-of-pocket maximum. The model includes an extensive-margin hassle cost of $z = \$1{,}710$, the highest value estimated by Ho and Lee (2023), so consumers seek care only when the net utility gain from spending exceeds the hassle cost; otherwise they optimally choose $m^* = 0$. Simulations use $N = 100{,}000$ draws of annual health needs calibrated following Ho and Lee (2023): mean health needs $6,490, standard deviation $4,868, moral hazard parameter $\omega = 0.263$, and CARA risk aversion $\psi = 0.0003$.

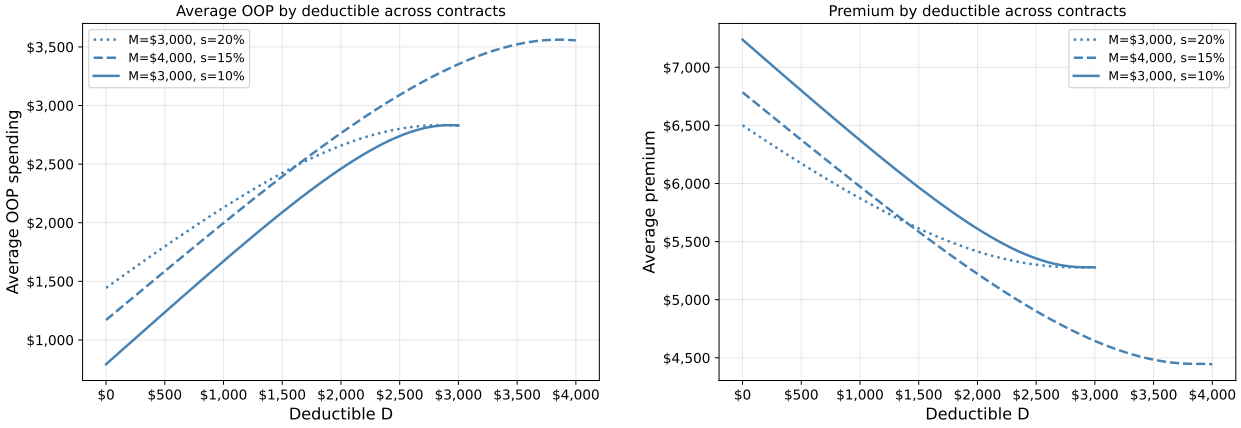
Table D.1: Decomposition of Spending and Welfare Changes from Raising the Deductible from $500 to $1,500

Panel A. Aggregate Changes	
$\Delta\mathbb{E}[m]$	75.97
$\Delta\mathbb{E}[\text{OOP}]$	678.03
ΔPremium	-602.06
ΔEU	-74.98

Panel B. Transition Decomposition				
Transition group	Share	$\Delta\mathbb{E}[m]$	$\Delta\mathbb{E}[\text{OOP}]$	ΔEU
Coinsurance $\rightarrow$ MOOP	0.397	81.60	231.99	11.24
Coinsurance $\rightarrow$ Coinsurance	0.484	0.00	435.21	-144.08
Coinsurance $\rightarrow$ Deductible	0.022	-5.63	10.83	-0.50
Cap $\rightarrow$ Cap	0.097	0.00	0.00	58.30
Deductible $\rightarrow$ Deductible	<0.001	0.00	0.00	0.05
Total	1.000	75.97	678.03	-74.98

Notes: The table reports a decomposition of the effects of increasing the deductible from $500 to $1,500, holding fixed M = $2,000 and $s = 0.10$. The simulation sets $\omega = 0.263$ and draws health needs λ from a lognormal distribution calibrated to have mean $6,490 and standard deviation $4,890, using 100,000 simulated draws. Panel A reports aggregate changes in average medical spending, average out-of-pocket spending, the actuarially fair premium, and expected utility. Panel B decomposes the aggregate changes into transition groups. The increase in spending is driven primarily by consumers switching from the coinsurance region into the MOOP region, while the decline in expected utility is driven mainly by consumers who remain in the coinsurance region and face higher out-of-pocket exposure. Because the premium falls in this exercise, the welfare decline is not due to a mechanical increase in premiums.

Figure D.2: Average OOP Spending and Premiums under Multiple Cost-Sharing Designs

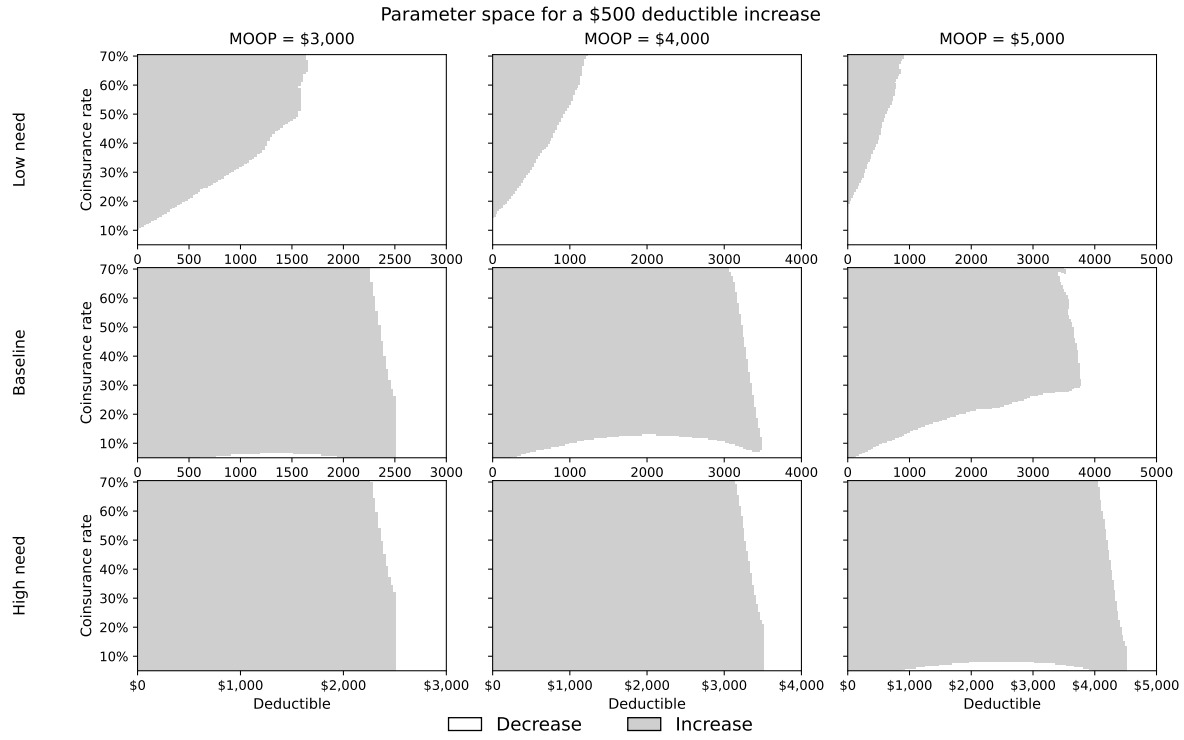


(a) Average out-of-pocket spending

(b) Average premium

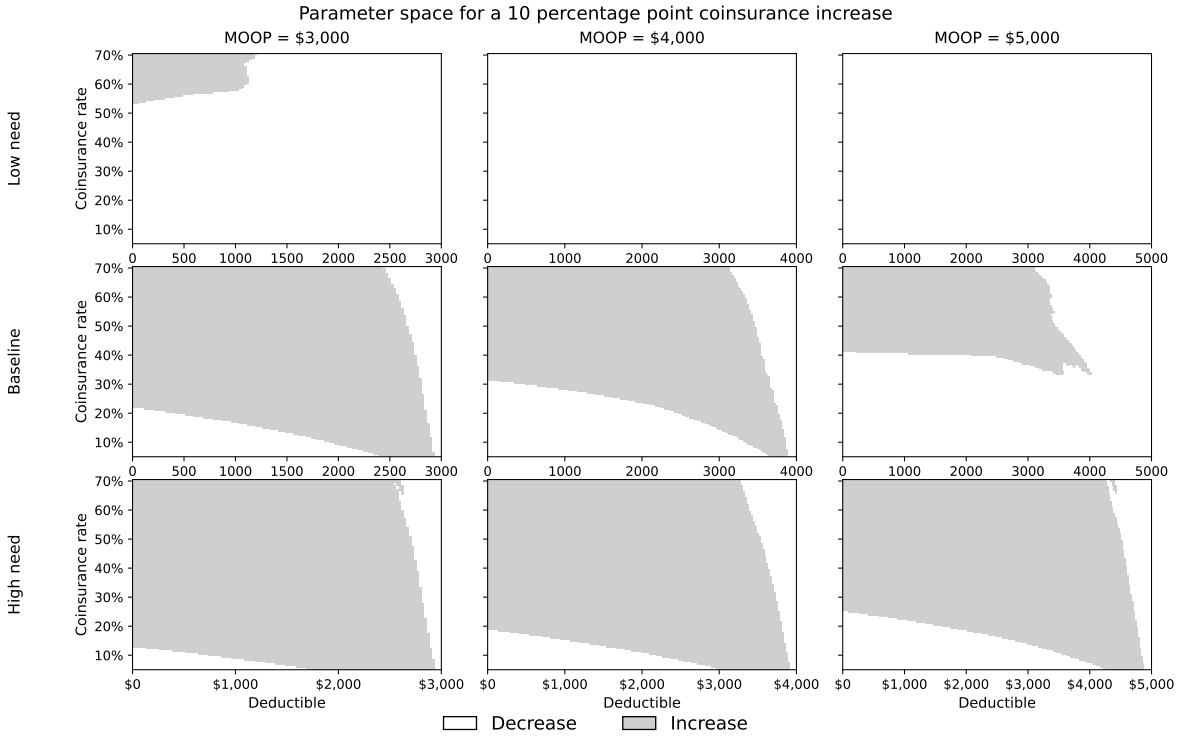
Note: The figure plots average out-of-pocket spending and average actuarially fair premiums as the deductible varies across three plan designs: 20 percent coinsurance with a $3,000 out-of-pocket maximum, 15 percent coinsurance with a $4,000 out-of-pocket maximum, and 10 percent coinsurance with a $3,000 out-of-pocket maximum. Premiums are actuarially fair and equal to average insurer liability under each contract. Simulations use $N = 100{,}000$ draws of annual health needs calibrated following Ho and Lee (2023), with mean $6,490, standard deviation $4,868, moral hazard parameter $\omega = 0.263$, and CARA risk aversion $\psi = 0.0003$.

Figure D.3: Parameter-Space Map of the Sign of the Average Spending Response to a $500 Deductible Increase



Note: Each panel maps the sign of the simulated change in average medical spending from a $500 increase in the deductible. The horizontal axis gives the initial deductible D, and the vertical axis gives the coinsurance rate s. Columns vary the out-of-pocket maximum, $M \in \{\$3{,}000, \$4{,}000, \$5{,}000\}$, and rows vary the lognormal distribution of health needs λ: (3,245, 2,434) in the top row, (6,490, 4,868) in the middle row, and (9,735, 7,302) in the bottom row, where each pair gives the mean and standard deviation. Shaded regions indicate parameter combinations for which a deductible increase raises average spending, and unshaded regions indicate parameter combinations for which spending falls. In all panels, the moral hazard parameter is fixed at $\omega = 0.263$, and results are based on 100,000 Monte Carlo draws.

Figure D.4: Parameter-Space Map of the Sign of the Average Spending Response to a 10 Percentage Point Coinsurance Increase



Note: Each panel maps the sign of the simulated change in average medical spending from increasing the coinsurance rate by 10 percentage points. The horizontal axis gives the deductible D, and the vertical axis gives the initial coinsurance rate s. Columns vary the out-of-pocket maximum, $M \in \{\$3{,}000, \$4{,}000, \$5{,}000\}$, and rows vary the lognormal distribution of health needs λ: (3,245, 2,434) in the top row, (6,490, 4,868) in the middle row, and (9,735, 7,302) in the bottom row, where each pair gives the mean and standard deviation. Shaded areas indicate that a coinsurance increase raises spending; unshaded areas indicate that spending falls. The moral hazard parameter is fixed at $\omega = 0.263$, and results are based on 100,000 Monte Carlo draws.