Northwestern | INSTITUTE FOR POLICY RESEARCH

IPR Working Paper Series

WP-22-42

Multidimensional Screening and Menu Design in Health Insurance Markets

Hector Chade
Arizona State University

Victoria Marone
University of Texas at Austin

Amanda Starc
Northwestern University and IPR

Jeroen Swinkels
Northwestern University

Version: October 14, 2022

Abstract

The researchers study a general screening model that encompasses a health insurance market in which consumers have multiple dimensions of private information and a price-setting insurer (e.g., a monopolist or a social planner) offers vertically differentiated contracts. They combine theory and empirics to provide three novel results: (i) optimal menus satisfy intuitive conditions that generalize the literature on multidimensional screening and shed light on insurer incentives; (ii) the insurer's problem with an unlimited number of contracts is well-approximated with only a small set of contracts; and (iii) under an additional assumption, the problem becomes dramatically simpler and can be solved using familiar graphical analysis. Calibrated numerical simulations validate assumptions, quantify the differential incentives of a monopolist and a social planner, and evaluate common policy interventions in a monopoly market.

The authors are grateful to Zack Cooper, Stuart Craig, Michael Dickstein, Eilidh Geddes, Ben Handel, Kate Ho, Amanda Kowalski, Tim Layton, Neale Mahoney, Alessandro Pavan, Mike Powell, Mark Shepard, Ashley Swanson, and seminar participants at Northwestern, the Utah Winter Business Economics Conference, Stanford, UCLA, Paris School of Economics, Michigan State, and University of Nevada-Reno for their helpful comments.

1 Introduction

Asymmetric information is central to our understanding of massive swaths of the economy. Regulators oversee firms that have private information about costs and consumer tastes. Investors evaluate entrepreneurs with differing abilities and project qualities. And in a particularly critical sector of the economy, health insurers sell insurance to consumers who know more about both their health status and their taste for insurance. These examples highlight settings in which there are multiple important dimensions of private information, and in which the space of possible contracts across which consumers can be screened is potentially vast. Until recently, however, the majority of both theoretical and empirical papers on screening have either considered a one-dimensional hidden information problem or else a multidimensional problem with a restricted contract space.

When both of these issues are considered together, solving screening problems becomes much harder. And as we explain below, even what is known about such problems does not apply directly in many settings of interest, including the canonical model of a health insurer's problem. We therefore have only a limited understanding of issues such as optimal exclusion, incentives to screen, and distortions in coverage once we take seriously both the fact that consumers can vary along multiple important dimensions and that insurers may introduce new contracts in response to this heterogeneity. While these issues have been explored in the empirical literature on health insurance markets, it remains unclear to what extent analytical properties can (or cannot) be derived.

In this paper, we combine theoretical and numerical analysis to address a number of challenging questions. What are the properties of an insurer's optimal menu of vertically differentiated health insurance contracts, and how do they change as, for example, the insurer's objective varies between that of a pure monopolist and a utilitarian social planner? What are the consequences of limiting the set of contracts the insurer can offer? Is there a way to recast this complex menu design problem in simpler terms, allowing one to solve it graphically? Importantly, while our theoretical framework is tailored to answering these questions in the context of a health insurance market, it is general enough that our results apply substantially more generally.

We begin by providing a set of theoretical results that substantially extends what is known about this class of multidimensional screening problems. Our framework is tailored to a health insurance market, but is general enough to inform other settings. We then combine these theoretical developments with a calibrated numerical analysis of a health insurance market. This combination of theoretical results and numerical analysis is central to our contribution. For one, our theoretical exploration provides a convergence result that gives a more solid foundation to standard numerical exercises commonly carried out in applied work. So, our theoretical results support empirical analysis. In addition, we show that under a simple additional assumption—quasiconcavity of the consumer's problem—the derivation of the insurer's optimal menu becomes dramatically more tractable, and our results more transparent. This assumption is hard to motivate from theoretical primitives, but is easy to check numerically, and we find that it is well supported in our calibrated numerical setting. Hence, numerical analysis supports theory as well.

Our model of a health insurance market is as follows. An insurer faces a population of consumers who have private information about their risk aversion, distribution of health states, and taste for healthcare utilization. The insurer designs a menu of vertically differentiated insurance contracts, where each contract has a premium and an out-of-pocket cost function that determines how much of a consumer's healthcare costs are covered by the insurer. The consumer's outside option is a base level of coverage provided by the government, which is exogenous. The insurer's payoff is a weighted average of consumer surplus, profits, and government spending. This general payoff function subsumes, for example, the case of a monopolist insurer as well as a utilitarian social planner. The timing is as follows: the insurer offers a menu of contracts. Consumers observe the menu, learn their type, and then choose a contract. Consumers then privately learn their health state and choose their healthcare utilization.

The fact that consumers only privately observe their health state allows for (ex-post) moral hazard in the model. While our theoretical analysis would still apply (and be somewhat simpler) absent moral hazard, we incorporate this complication because it is a first-order concern in real-world health insurance markets (Manning et al., 1987). And as is well known, the presence of moral hazard introduces interesting tradeoffs. Given the informational constraints, the only way to reduce consumers' exposure to financial loss under a bad health realization is to lower their marginal cost of healthcare utilization, thereby inducing them to use beyond the efficient level. Even for a social planner, the problem is therefore more complicated than simply pooling all consumers at full insurance (Arrow, 1965; Pauly, 1968; Zeckhauser, 1970).

Our numerical analysis is based on a population of consumers calibrated to match demographics of the under-65 US population and parameter estimates from Marone and Sabety (2022), allowing for an extremely flexible and empirically realistic distribution of consumer types. We implement the model using this population and a finite set of piecewise linear and concave insurance contracts. We maintain that the government provides a base level of coverage at a $10,000 deductible and out-of-pocket maximum contract, and that the government covers the cost associated with base coverage regardless of what coverage level a consumer ultimately selects.[1]

We interpret the insurer's menu design problem as one in which the insurer chooses a premium schedule and recommends an allocation of coverage to each consumer type, subject to the usual incentive compatibility constraints. We derive necessary conditions on optimal menus in two contexts: one in which there is a fixed set of contracts the insurer can offer, and one in which the insurer can offer a continuum of contracts. Our necessary conditions correspond to two different perturbations of a given menu. The main perturbation changes premiums for all contracts above a given coverage level. The second perturbation changes the quality of a given contract, while keeping prices the same.

[1] In this way, we implement "incremental pricing" as described by Weyl and Veiga (2017) and implemented in Einav et al. (2010). Our model is also flexible enough to capture "total pricing," as implemented in Handel et al. (2015).

Our optimality conditions are intuitive and generalize the well-known screening conditions in the case of one-dimensional private information. They also generalize the case where consumers are multidimensional, but the insurer is restricted to offering just one or two contracts. The results shed light on the insurer's incentives to exclude, screen, and distort coverage away from the efficient level. For example, we show that a monopolist insurer has more incentive to exclude than a utilitarian planner, and also distorts coverage downwards, below the efficient level.

We then provide a convergence result that links the insurer's problem with a finite set of contracts to the fully general problem with a continuum of them. We show that the latter is well-approximated by the former when the insurer can use a sufficiently rich set of contracts. We view this result as an important link between settings in which product characteristics can be "fully endogenous" (i.e., when the insurer can offer a continuum of contracts) to settings in which the contract space is limited by a regulator. It also provides the applied researcher solid theoretical ground on which to conduct analysis with a finite number of contracts, often a prerequisite for computational tractability. Indeed, this has been the approach in a number of recent applied papers (Azevedo and Gottlieb, 2017; Ho and Lee, 2021; Marone and Sabety, 2022). Our paper provides a unified framework for thinking about this approach, as well as the theoretical foundations that justify it. In our numerical application, convergence in the density of the contract space is remarkably fast: insurers can capture over 98 percent of the available payoff with as few as five contracts.

Motivated by the convergence result, we explore a simplified version of the insurer's problem when limited to a fixed set of contracts. We reframe the problem of setting the premium of each contract to one of setting the *incremental* premium of each marginal level of coverage. We show that under certain conditions, the insurer's problem "decouples," in the sense that the optimality condition for each incremental premium is entirely independent of all other premiums. This decoupling dramatically simplifies the problem, yielding a potentially powerful tool for analyzing it. The assumption that guarantees that the decoupled problem corresponds to the true problem is that consumers' payoffs are quasiconcave in coverage level at the optimal menu. This assumption is hard to justify from primitives, as it is an equilibrium outcome that depends on the distribution of consumer preferences and costs in the market. However, after solving for the optimal menu numerically in a given population, the assumption is easily checked. We find that in our empirical setting, the assumption is well supported. For a variety of insurer objective functions, the optimal price schedule is quasiconcave in coverage level for 99 percent of consumers, and the difference between the insurer's payoff from solving the simplified problem agrees with its payoff from solving the true problem within a margin of 1 percent.

The simplified version of the problem allows us to recast the analysis in familiar terms. Equilibrium outcomes depend on the demand curves for incremental coverage, the associated marginal revenue curves, and the marginal cost of providing incremental coverage. As the problem can be solved separately for each incremental level of coverage, each margin can be analyzed graphically. A monopolist sets marginal cost equal to marginal revenue, while a utilitarian planner sets marginal cost equal to price. A planner with an excess cost of public funds sets marginal cost equal to a weighted average of marginal revenue and price. A resulting comparative static is that if the insurer puts less weight on consumer surplus (and more weight on profits), then the entire premium schedule becomes steeper and higher, and consumers choose less coverage on every margin. A monopolist therefore serves fewer consumers than the planner at each incremental level of coverage. As this includes the first increment, it means that a monopolist always excludes a strictly positive mass of consumers from the market. We also show that a monopolist has more incentives to screen than a utilitarian planner, in the sense that the monopolist optimally uses weakly more contracts.

The numerical analysis confirms and quantifies several of our theoretical results. We solve for the optimal menu that would be offered by a social planner, a social planner facing an excess cost of public funds, and a monopolist. Consistent with our theoretical results, We find that a monopolist insurer excludes substantially more consumers than a social planner. Likewise, the monopolist screens more than the social planner, separating consumers across a range of coverage levels, and in the end offering much less coverage in the population. The monopolist's optimal menu reduces social welfare by $743 per household per year (equal to 7 percent of household average total healthcare spending) relative to what can be achieved by a social planner. Finally, as the cost of public funds rises, the social planner begins acting more like the monopolist, excluding more and more consumers from the market.

We then use our numerical framework to explore how a regulator might best intervene on behalf of consumers in a monopoly insurance market. We focus on three types of interventions. We first allow the regulator to mediate prices in the form of taxes or subsidies (to which the monopolist can strategically respond). We then restrict the set of contracts the monopolist can offer. Finally, we allow the regulator to adjust the base level of coverage. We find that the most effective policy tool is raising the base level of coverage, which in effect squeezes the monopolist out of the market entirely. Absent this possibility, we find that restricting the set of contracts the monopolist can offer and implementing a non-linear subsidy scheme are both reasonably effective at increasing coverage and consumer surplus in the market. By contrast, the monopolist's ability to strategically respond to subsidies makes the linear subsidy schemes perform poorly.

Our paper is related to an extensive theoretical literature on screening as well as a large empirical literature on health insurance markets. With respect to the empirical literature, our model of consumer demand for health insurance builds on a workhorse introduced by Cardon and Hendel (2001) used in several papers in the empirical literature (for example, Einav et al., 2013; Azevedo and Gottlieb, 2017; Ho and Lee, 2021; Marone and Sabety, 2022). Our formulation is general enough to capture a number of alternative institutional settings (for example, whether or not the government covers the cost of base coverage regardless of the consumer's contract choice), and we further enrich the model to allow a unified treatment of insurers with differing objective functions. Our graphical analysis of the insurer's problem builds on the foundational approach in Einav et al. (2010), who focus on competitive markets and two contracts. We show how to extend this approach to an arbitrary number of contracts by focusing on incremental premiums and coverage levels. Our focus on health insurance menu design for multidimensional consumers is also closely related to recent work by Marone and Sabety (2022) and Ho and Lee (2021), who each solve for the optimal menus of contracts that would be offered by a utilitarian social planner in their respective empirical settings. We build on these findings by asking to what extent various features of those solutions will hold in general, and how optimal menus would change with the insurer objective function.

Our theoretical approach is related to the seminal works by Stiglitz (1977) (insurance), Mussa and Rosen (1978) (quality provision), and Maskin and Riley (1984) (quantity provision), who similarly analyze a principal-agent problem with private information, but focus on only one-dimensional private information. There is a subsequent important literature on screening with multidimensional private information, including Wilson (1993), Armstrong (1996), Rochet and Choné (1998), and Manelli and Vincent (2006). This literature has been surveyed by Rochet and Stole (2003). Our class of problems belongs to their Section 5, which they call "the one-dimensional instrument" case. An early contribution to this class is the parametric example solved in Laffont et al. (1987), which is a special case of our general formulation. More recently, Deneckere and Severinov (2017) provide a solution for a class of problems with two-dimensional private information. The simplified version we analyze was pioneered by Wilson (1993) in his work on nonlinear pricing, but for a specific problem without common values. Veiga and Weyl (2016) conduct a similar exercise to ours with common values, but with just one possible contract choice (plus an outside option). The extension to many possible contracts substantially complicates the analysis.

Finally, our theoretical approach is tightly linked to the empirical setting. Health insurance has some particular features that prevent us from using much of the extensive technical apparatus developed in this literature. For example, existing incentive compatibility and optimal exclusion results rely on convexity and homogeneity assumptions on the consumer's utility as a function of their type. These assumptions fail in our setting, where the utility function is built up from more primitive objects, such as the certainty equivalent of a lottery over health outcomes. We therefore remove those assumptions and derive our results using the perturbation arguments described above. Ultimately, our results rely on a few, permissive properties of the agent's utility and the principal's cost, meaning that our model can be applied beyond just health insurance settings.

The paper has distinct theoretical and numerical analyses, which are targeted to different groups of readers. It is organized as follows. In Section 2, we describe the model. In Sections 3 and 4, we present theoretical results, including the optimality conditions and convergence. These sections can be skipped by the applied reader without loss of continuity. In Section 5, we present the simplified reformulation of the problem, provide a graphical analysis, and discuss the insurer's incentives to exclude and screen. In Section 6, we discuss our numerical application, assess key assumptions and solve for the optimal menus of various types of insurers. In Section 7, we apply our analysis to evaluate the impact of regulatory intervention. This section can be skipped by readers primarily interested in our theoretical results. Section 8 concludes.

2 The Model

We consider a model of a health insurance market in which an insurer chooses a set of vertically ordered contracts to offer and their associated premiums. Heterogeneous consumers then select a single contract, incur health shocks, and choose their subsequent healthcare utilization. Consumers have multidimensional private information at the time they choose an insurance contract. Realized health is also private information, allowing ex-post moral hazard. Selection—either adverse or advantageous—and moral hazard are thus intertwined. A government may also provide a base level of insurance coverage to all consumers.

While our application is to health insurance, our model is a general workhorse for settings with multidimensional screening. And given the limited number of general results in this literature (see Armstrong (1996), Rochet and Choné (1998), and Manelli and Vincent (2006) for notable exceptions), we substantially push the technical frontier. To help the reader who is more interested in the application, we will separate much of our discussion of the technical contribution into "Technical Remarks" and footnotes. These can be skipped without loss of continuity.

The Consumer. There is a strictly risk-averse consumer (or a continuum thereof). She has CARA preferences, and is privately informed about her taste for healthcare utilization ω, her coefficient of absolute risk aversion ψ, and her distribution F over potential health states l, which has density f on bounded support $[0, \bar{l}]$.[2] We denote the consumer's type by $\theta = (\omega, \psi, F)$. The distribution of θ is given by a joint cdf G on $\Theta = [0, \bar{\omega}] \times [0, \bar{\psi}] \times \Delta([0, \bar{l}])$. The support of G is some rectangular subset $supp\, G = [\underline{\omega}, \bar{\omega}] \times [\underline{\psi}, \bar{\psi}] \times \mathcal{F}$ of Θ.[3] We assume that G has a continuous density function g. For convenience, we assume there are $\overline{F}$ and $\underline{F} \in \mathcal{F}$ such that each F in $\mathcal{F}$ first-order-stochastically dominates $\underline{F}$ and is first-order-stochastically dominated by $\bar{F}$. That is, there is an unambiguously sickest and healthiest type in the population.

If the consumer chooses a dollar amount $a \in [0, \bar{a}]$ of healthcare utilization ("spending") when her health state is l and her taste for healthcare is ω, then she enjoys a utility level which in dollar terms is given by $b(a, l, \omega)$, where b is twice-continuously differentiable, strictly decreasing in l and strictly increasing in a.[4] That is, an agent is hurt by a worse health outcome, but helped by more healthcare spending. We assume $b_{aa} < 0$, $b_{a\omega} > 0$ and $b_{al} > 0$, such that the consumer has declining marginal utility for healthcare, but that marginal utility is higher when she has either worse health or a higher taste for healthcare.[5] A canonical example introduced by Einav et al. (2013) is $b(a, l, \omega) = (a - l) - (1/(2\omega))(a - l)^2$, which satisfies all the assumptions for $a \geq l$. This

[2]For the numerical exercises, we will take l unbounded and with an atom where the agent wants no healthcare. The formal analysis can accommodate these, but at the cost of more notation and less transparent analysis.

[3]Whenever we talk about $\Delta([0, \bar{l}])$, we implicitly endow it with the topology of weak convergence.

[4]We use increasing and decreasing in the weak sense of nondecreasing and nonincreasing, adding "strictly" when needed, and similarly with positive and negative, and concave and convex. Also, for any function f and argument x of f, we write $(f)_x$ for the total derivative of f with respect to x. We use the symbol $=_s$ to indicate that the objects on either side have strictly the same sign.

[5]We in fact only need these conditions to hold for a and l such that $b_a(a, l, \omega) \in [0, 1]$, because in our environment the consumer will optimally choose such an a given l and ω.

example belongs to a canonical class of b functions that satisfy $b(a, l, \omega) = \hat{b}(a - l, \omega)$, with $\hat{b}$ increasing in $a - l$.

Insurance Contracts. An insurance contract consists of an out-of-pocket cost function that specifies how much the consumer pays for each level of healthcare spending. There is an exogenously given set of potential contracts indexed by a scalar $x \in [0, 1]$. If a consumer chooses a contract x and healthcare spending level a, then her out-of-pocket cost is $c(a, x)$. We take c to be twice-continuously differentiable for almost all (a, x), with $0 \leq c_a \leq 1$, $c_{aa} \leq 0$, $c_x \leq 0$ for $a > 0$, and $c_{ax} < 0$.[6] That is, out-of-pocket costs are increasing and concave in the level of healthcare spending, and as x gets larger, the out-of-pocket cost function gets lower and shallower as a function of a. Contracts are thus vertically differentiated, with higher x corresponding to higher coverage.

Optimal Choice of Healthcare Spending. Given a contract x, a health state realization l, and taste for healthcare utilization ω, the consumer chooses an optimal level of healthcare spending a. Let $a^*(l, x, \omega) \equiv \arg\max_{a \in [0, \bar{a}]}(b(a, l, \omega) - c(a, x))$ be that optimum.[7] Let

$$z(l, x, \omega) \equiv b(a^*(l, x, \omega), l, \omega) - c(a^*(l, x, \omega), x) \tag{1}$$

be the consumer's income-equivalent payoff given (l, x, ω).

Optimal Choice of Insurance Contract. Let y be the initial wealth of the consumer. Since the consumer has CARA preferences, we can usefully simplify her problem by expressing her preferences in certainty-equivalent units. Consider a consumer of type θ who chooses contract x with premium p and out-of-pocket cost function $c(\cdot, x)$. Her expected utility is $\int \left(-e^{-\psi(y - p + z(l, x, \omega)}\right) dF(l)$, which has certainty equivalent $y - p + v(x, \theta)$, where

$$v(x, \theta) \equiv -\frac{1}{\psi} \log \int e^{-\psi z(l, x, \omega)} dF(l). \tag{2}$$

For any two contracts x and x', the consumer's willingness to pay for the discrete jump from x to x' is given by $v(x', \theta) - v(x, \theta)$, while her marginal willingness to pay for incremental coverage is given by $v_x(\cdot, \theta)$. Faced with a menu of (x, p) pairs, the consumer chooses the contract that maximizes the difference between the dollar value of her health activity and the premium: $\max_{x \in [0,1]} (v(x, \theta) - \rho(x))$.

The Government. The government provides a base level of insurance $x^0 \in [0, 1]$. If the consumer chooses healthcare spending level a, the cost to the government is $k(a, x^0) = a - c(a, x^0)$. The government is risk neutral, but may face an excess cost of public funds, reflecting dead weight losses in the tax system.

[6] We allow ourselves to consider cases with $c_{ax} = 0$ in our numerical exercise. Theoretically, this is tractable but creates technical complications without economic insight.

[7] The notation is justified since, under our assumptions, $a^*(\cdot, x, \omega)$ is unique for almost all l, and so, since F is atomless, it is irrelevant which optimal a is chosen when there is more than one such optimum.

The Insurer. The insurer is risk neutral and is a price-setter. Depending on the economic context, the insurer might be a monopolist, a social planner, or a firm designing insurance for its workers. Our model is flexible enough to cover all of these cases. The insurer chooses a premium schedule ρ specifying a premium $\rho(x)$ for each insurance contract.

We assume that ρ is left continuous in x, which will ensure that the consumer always has an optimal choice of insurance contract.[8] Without loss of generality, we take ρ to be increasing, since the consumer will never choose a contract for which some higher coverage level is available at a weakly lower premium. Let $\mathcal{P}$ be the set of such premium schedules. To reflect that the consumer always has an option of taking the government-provided insurance level x^0, we require that $\rho(x^0) = 0$. The insurer may also face other constraints on the set of premium schedules. We denote the set of allowable premium schedules as a closed set $\mathcal{P}^0 \subset \mathcal{P}$.

The insurer also makes a recommendation $\chi(\theta)$ of insurance contract to each type θ. A menu (ρ, χ) is incentive compatible if and only if, for all θ,

$$\text{(IC)} \qquad \chi(\theta) \in \arg\max_{x\in[0,1]} (v(x,\theta) - \rho(x))$$

If the consumer chooses contract x and healthcare spending a, then the cost to the insurer is $k(a,x) - k(a,x^0)$, reflecting that the first $k(a,x^0)$ of healthcare spending is covered by the government. We therefore implement "incremental pricing," as described by Weyl and Veiga (2017), meaning that the government covers the cost of base coverage regardless of which contract the consumer ultimately selects.[9]

Timing. The timing is as follows. At time 0, the government sets x^0. At time 1, the insurer chooses the premium schedule ρ and recommends an allocation χ, and the consumer learns her type θ. At time 2, facing ρ, and knowing θ (but not her health state realization l), the consumer chooses an insurance contract x and pays $\rho(x)$. At time 3, the consumer learns her health state l, chooses a level of healthcare spending a, and pays out-of-pocket cost $c(a,x)$.

Expected Insured Costs. A consumer of type θ enrolled in contract x incurs expected insured healthcare spending equal to

$$\gamma^I(x,\theta) \equiv \int k(a^*(l,x,\omega),x)dF(l).$$

[8] This follows since $v(\cdot,\theta)$ is continuous and since ρ left continuous implies that $-\rho$ is upper semicontinuous.

[9] Our model is also flexible enough to capture an alternative regime of "total pricing," under which the government would only pay for base coverage if the consumer selected base coverage. The insurer would then cover the full cost $k(a,x)$ of providing coverage above x^0. Note that this distinction does not matter when the insurer is the social planner, as in this case the government supplies both base and incremental coverage. But, as shown by Weyl and Veiga (2017) and discussed in Handel and Ho (2021), it may matter a great deal when the insurer is a private firm.

The portion paid by the government is equal to

$$\gamma^G(x, x^0, \theta) \equiv \int k(a^*(l, x, \omega), x^0) dF(l).$$

Note that as written, the government's portion of insured costs is tied to the consumer's choice of healthcare spending under her chosen contract x. It may alternatively be the case that the government's portion is determined by what the consumer would have done had she taken minimum coverage x^0, in which case we would have $\gamma^G(x^0, \theta) \equiv \int k(a^*(l, x^0, \omega), x^0) dF(l)$. The decision of how to set the government's share of insured costs is a regulatory one. We consider both cases in our analysis. Regardless, the net cost to the insurer of covering the consumer is $\gamma^I - \gamma^G$. We make the following assumption regarding γ^I and γ^G:

Assumption 1 (Marginal Costs) *The functions γ^I and γ^G are continuous. The derivatives γ^I_x and γ^G_x are defined for almost all θ, and are uniformly bounded where defined.*

See Online Appendix B.3 for primitives. These primitives subsume as a special case the canonical b and the case of c piecewise linear.

The Insurer's Objective Function. To cover a broad set of cases in a unified and parsimonious way, we model the insurer's objective using weights $w = (w^C, w^I, w^G) \geq 0$ on consumer surplus, profits, and government spending, respectively. Given a set of weights w, a base coverage level x^0, an insurance contract x, and a premium p, the insurer facing type θ has payoff

$$(3) \qquad S(p, x, \theta) = w^C(\underbrace{v(x, \theta) - p}_{\substack{\textit{Consumer}\\ \textit{surplus}}}) + w^I(\underbrace{p - \gamma^I(x, \theta) + \gamma^G(x, x^0, \theta)}_{\substack{\textit{Insurer profit on}\\ \textit{incremental coverage}}}) - w^G \underbrace{\gamma^G(x, x^0, \theta)}_{\substack{\textit{Govt. spending}\\ \textit{on base coverage}}}.$$

We suppress that S depends on w and x^0 as they will be fixed for the relevant portion of the analysis. Table 1 describes the weights that would correspond to different types of insurers in the context of our model. A monopolist corresponds to $w = (0, 1, 0)$, reflecting that it cares only about itself. A social planner with a cost of public funds τ (where typically $\tau > 1$) corresponds to $w = (1, \tau, \tau)$.

Table 1. Example Insurer Objective Functions

Insurer	w^C	w^I	w^G
Monopolist	0	1	0
Social planner	1	1	1
Social planner with cost of funds τ	1	τ	τ

Notes: The table shows the weights w that would correspond to different types of insurers.

Given weights w, we can now write each of these insurer's problems as simply

$$\text{(P)} \qquad \max_{\rho \in \mathcal{P}^0, \chi} \int_\Theta S(\rho(\chi(\theta)), \chi(\theta), \theta) dG(\theta)$$
$$s.t. \qquad IC \text{ and } \rho(x^0) = 0,$$

where recall that x^0 corresponds to the consumer's outside option, and so the IC constraint together with $\rho(x^0) = 0$ capture the participation constraint. The central contribution of this paper is to provide insight into the optimal structure of (ρ, χ).

In addition to providing flexibility in the insurer's objective function, this framework also easily permits regulatory intervention in the market. For example, a regulator may wish to restrict the insurer to a fixed set of allowable contracts, but allow the insurer to price those contract freely. In this case, $\mathcal{P}^0$ is the subset of $\mathcal{P}$ with steps at the pre-specified contracts. Since the consumer prefers the highest contract in an interval where ρ is flat to any other contract in that interval, this is equivalent to offering only the set of contracts corresponding to the step points of ρ. Alternatively, the regulator may wish to specify a restriction on price schedules $Q(\rho) \leq 0$ for some function Q (or several such functions), in which case $\mathcal{P}^0$ is the subset of $\mathcal{P}$ that satisfies these restrictions. The function Q could reflect a limit on the firm's expected profit, expected profit margin, or the expected cost to the government. It could also reflect a constraint on the fraction of consumers excluded from the market. In both cases, the set $\mathcal{P}^0$ is closed.

Technical Remark 1 (Role of Price Schedule) We work directly with the price schedule ρ as a function of the insurance contract x, rather than as a function of the type θ as is standard in the mechanism design literature. As Rochet (1985) argues, the two approaches are equivalent. And, as we discuss more fully below, because there will typically be many θ's choosing any given x, this is technically more natural since it automatically imposes that two types who choose the same contract pay the same price. More importantly, we proceed largely as if ρ alone is the design variable. This is because for any given ρ, our structure has enough single-crossing embedded in it that for almost all θ, the consumer has a unique optimal contract choice. See the proof of Lemma 2 in Appendix A.6.

Technical Remark 2 (Stochastic Menus) Stochastic mechanisms can be very useful to the principal when types are multidimensional (Manelli and Vincent, 2006), or when the type includes the agent's risk aversion (Kadan et al., 2017). For example, having the premium on the insurance contract targeted at types with low risk aversion be determined by a lottery would help dissuade more risk averse types from imitating the less risk averse types. We find it implausible that the insurer would be allowed to run such lotteries (indeed, many regulations prevent charging identical consumers different premiums), and so we rule them out here for reasons of economic realism.

Technical Remark 3 (Subsumed Screening Problems) If one takes v, γ^I and γ^G as primitives, rather than building them up as we did from a health insurance setting, then we have a substantially general model of multidimensional screening with product quality or quantity that lies in $\mathbb{R}$ (see Section 5 in Rochet and Stole, 2003). For example, our setting subsumes extensions to multidimensional private information of the one-good nonlinear pricing problem in Maskin and Riley (1984), or the quality-provision problem in Mussa and Rosen (1978), as well as optimal regulation settings in the tradition of Baron and Myerson (1982). For our analysis to go through, it is important that the model be vertical, in that for some dimension τ of the consumer's private information, there is strict single-crossing, $v_{x\tau} > 0$. We establish this property for $v_{x\psi}$ below.

3 Consumer Demand for Insurance

We first derive comparative statics of insurance demand as a function of the consumer's type. We will show that more risk averse consumers have a higher demand for insurance, as do sicker consumers. That is, demand always increases in the consumer's risk aversion parameter ψ given ω and F, and it also increases in the distribution of health states F when the distribution becomes worse (in a precise sense), given ψ and ω. The relationship between taste for healthcare utilization and demand for insurance (determining the direction of "selection on moral hazard") is more complex. Concavity in the out-of-pocket cost function $c(\cdot, x)$ makes the demand for insurance change in an ambiguous fashion when ω changes, given ψ and F. However, we can derive an unambiguous result describing relationship when contracts are linear.

We will find it useful to define the following "marginal-utility-adjusted" density of health states given x and θ:

$$m(l|x,\theta) = \frac{e^{-\psi z(l,x,\omega)} f(l)}{\int e^{-\psi z(l',x,\omega)} f(l') dl'}. \tag{4}$$

This is a transformed density of l where the weight on each health state l is updated by the marginal utility to the consumer of an extra dollar in that state. To see the role of m, note that by the Envelope Theorem, the derivative of the consumer's ex-post payoff with respect to coverage level is $z_x(l,x,\omega) = -c_x\left(a^*(l,x,\omega), x\right)$, since the effects on z via the associated change in the optimal level of healthcare utilization can be ignored. Hence from (2),

$$v_x(x,\theta) = -\frac{1}{\psi} \frac{\int e^{-\psi z(l,x,\omega)} (-\psi z_x(l,x,\omega)) f(l) dl}{\int e^{-\psi z(l,x,\omega)} f(l) dl} = -\int c_x(a^*(l,x,\omega),x) m(l|x,\theta) dl, \tag{5}$$

meaning that the marginal effect of higher coverage on a consumer's certainty equivalent payoff is the average under m of paying $-c_x$ less in each health state.

We are interested in the comparative statics of χ with respect to θ. We therefore analyze the behavior of v_x with respect to ω, ψ, and F, since this will pin down the behavior of χ.

Proposition 1 (Properties of Insurance Demand) *The consumer's demand for insurance sat-*

isfies the following properties for all x and θ: (i) $v_{x\psi} > 0$, and thus $\chi(\omega, \cdot, F)$ is increasing in ψ; (ii) if $\{f(\cdot|\tau)\}_{\tau\in[0,1]}$ is a parametrized family of densities ordered by strict monotone-likelihood-ratio property (MLRP), then $v_{x\tau} > 0$, and thus $\chi(\omega, \psi, \cdot)$ is increasing in τ;[10] *(iii) if $b(a, l, \omega) = \hat{b}(a-l, \omega)$ and c is convex in a (includes the linear case), then $v_{x\omega} > 0$, and thus $\chi(\cdot, \psi, F)$ is increasing in ω.*

Technical Remark 4 (Demand for Insurance and ω) One can also show that $\chi(\cdot, \psi, F)$ is increasing when F is dirac at 0.[11] Beyond these cases, the comparative statics in ω are complex. One can show that $v_{x\omega} > 0$ if and only if

$$0 < \left(\int b_\omega m dl\right)_x = \int b_{\omega a} a_x^* m dl + \int b_\omega m_x dl,$$

where the first term in the last expression is strictly positive. But, as x increases, the cdf M decreases (since as insurance improves, the marginal utility of income becomes more equal across states), and when $b(a, l, \omega) = \hat{b}(a-l, \omega)$, $(b_\omega)_l =_s (a^* - l)_l$. If $c(\cdot, x)$ is convex then sicker individuals face a higher marginal cost of care, and so $(a^* - l)_l \leq 0$, and $\int b_\omega m_x dl$ is positive. But, if $c(\cdot, x)$ is sufficiently concave, then the second term is negative and overwhelms the first term.

4 Optimal Menu Design

We now describe necessary conditions that an optimally designed menu satisfies. After some preliminaries, we consider two versions of the insurer's menu design problem. We first consider the case in which the insurer is restricted to offering a finite set of fixed contracts. We then turn to the case in which the insurer can offer a continuum of contracts, such that it can also control the qualities of the contracts offered. In both cases, we derive necessary conditions for optimality of ρ that substantially generalize the familiar screening conditions in Mussa and Rosen (1978) and Maskin and Riley (1984) for the one-dimensional case. We emphasize that these conditions are necessary only, since the problem does not have enough structure for us to show that the insurer's payoff is quasiconcave in ρ.

We then show that the optimal menu under a fixed set of contracts converges to the optimal menu under a continuum of contracts as the number of contracts in the fixed set grows large. Because it is substantially more tractable and can approximate the continuum case arbitrarily well, we view the case with a fixed set of contracts to be of primary importance. We close the section with a number of other results about the insurer's problem, including the incentive to exclude and screen consumers, and the existence of positive trade in the market.

[10] A family of densities $\{r(\cdot|t)\}_{t\in[0,1]}$ has the strict *MLRP* if $r(s|t_h)/r(s|t_l)$ is strictly increasing in s for all $t_h > t_l$. In this case we will say that the cdf R shifts in the strict *MLRP* sense.

[11] By (4) the cdf M is also degenerate at 0. Thus, $v_x(x, \theta) = -c_x(a^*(0, x, \omega), x)$ and hence $v_{x\omega} = -c_{xa} a_\omega^* > 0$.

4.1 Preliminaries

To simplify our analysis of the insurer's objective function S, we separate the portion that represents gains from trade from the portion that represents a transfer between the insurer and consumers. To this end, let

$$\mathcal{S}(x,\theta) \equiv w^I(v-\gamma^I) - (w^G - w^I)\gamma^G,$$

where the term $(v-\gamma^I)$ is the dollar value of the social surplus created by allocating a consumer of type θ to contract x, while $(w^G - w^I)\gamma^G$ captures the effect of government transfers to the insurer. We can then rewrite the insurer's payoff as $S(p,x,\theta) = \mathcal{S}(x,\theta) - (w^I - w^C)(v(x,\theta) - p)$, where the second term measures the value the insurer places on consumer surplus. It will be important in what follows that $\mathcal{S}$ does not depend on p.

We can interpret the marginal gains from trade from insurance in familiar terms. The derivative of social surplus $(v-\gamma^I)$ with respect to coverage level is given by

$$v_x - \gamma_x^I = \underbrace{\int(-c_x)mdl \;-\int(-c_x)fdl}_{\substack{\text{Marginal value of} \\ \text{risk protection}}} \;-\; \underbrace{\int(1-c_a)a_x^* fdl}_{\substack{\text{Marginal social cost} \\ \text{of moral hazard}}}\;.$$

Recall that m reflects health states weighted by marginal utilities. So, $\int(-c_x)mdl$ represents the benefit to the consumer of marginally more generous insurance, while $\int(-c_x)fdl$ is the cost to the insurer. The difference between the two represents the marginal value of risk protection provided by insurance. As coverage level increases, the additional healthcare spending a_x^* induced by insurance confers on the consumer a marginal benefit of b_a, which at an optimum level of spending equals its marginal out-of-pocket cost c_a. The full marginal social cost to the insurer, however, remains 1. Averaging across all health states, $\int(1-c_a)a_x^* fdl$ then represents the marginal social cost of spending induced by insurance.

As a final preliminary, for any θ, let $\bar{x}(\theta,\rho)$ be the largest best response to ρ and $\underline{x}(\theta,\rho)$ the smallest best response. It will simplify the derivations if for almost all θ, $\bar{x}(\theta,\rho)$ and $\underline{x}(\theta,\rho)$ (which may be equal) are the only best responses for θ. Formally, say that ρ has the *two-best-response property* (*2BRP*) if for almost all (ω, F), the best response correspondence $X(\omega,\cdot,F,\rho)$ has at most two elements for any ψ. We will also assume that F has a finite-dimensional parametrization $\tilde{F}(\cdot|\mathbf{t})$, where $\mathbf{t}\in[0,1]^n$, and $\tilde{F}(\cdot|\mathbf{t})$ is strictly *MLRP* increasing in the first coordinate of $\mathbf{t}$.[12] We will also say that two price schedules are close to each other if for a given contract available at a given price under one price schedule, something almost as good is available for only a slightly higher price under the other.[13]

[12]That is, there is $\tilde{G}$ a joint cdf on $[0,\bar{\omega}]\times[0,\bar{\psi}]\times[0,1]^n$ with density $\tilde{g}$ such that for all $Y\subset[0,\bar{\omega}]\times[0,\bar{\psi}]\times\Delta([0,\bar{l}])$, we have $G(Y) = \tilde{G}\{(\omega,\psi,\mathbf{t})|(\omega,\psi,\tilde{F}(\cdot|\tau))\in Y)\}$.

[13]That is, for two price schedules ρ' and ρ'', the distance $d(\rho',\rho'')$ is the smallest number such that for each x, there is $\hat{x}$ within $d(\rho',\rho'')$ to the left of x with $\rho''(\hat{x})\le\rho'(x)+d(\rho',\rho'')$, and vice versa. Formally,

$$d(\rho',\rho'') = \min\{\delta|\rho''(\max(x-\delta,0))\le\rho'(x)+\delta \text{ and } \rho'(\max(x-\delta,0))\le\rho''(x)+\delta \text{ for all } x\in[0,1]\}.$$

Technical Remark 5 (Genericity of *2BRP*) Our strong intuition is that *2BRP* holds generically. For any three $x' < x'' < x'''$, there is a locus of θ where the consumer is indifferent between x' and x'' and one where the consumer is indifferent between x' and x'''. It would be extremely surprising if these loci corresponded over any region, but v is sufficiently complicated that formalizing this is intractable beyond some special examples. We can do the analysis that follows without *2BRP*, but the notational load is extreme, and the economics less transparent.

4.2 Optimally Pricing a Fixed Set of Contracts

Suppose the insurer is restricted to offering a fixed set of contracts $\{x^k\}_{k=1}^K$, where $x^0 < x^1 < \cdots < x^K \leq 1$, but can freely set their associated premiums. Consider a candidate price schedule ρ, and a perturbation in which the insurer raises (or reduces) by a constant amount the premiums on all contracts more generous than a given contract x. As premiums increase, two things happen. First, the insurer makes more money on inframarginal consumers who continue to choose a contract above x. Second, some consumers who previously chose a contract above x will substitute to contract x (or below). The switchers will generate a different amount of surplus than previously. At the optimum, for either an increase or decrease in premiums, the insurer balances the two effects.

Formally, fix (ω, F) and some $0 \leq k < K$ and, suppressing them in what follows, let $\hat{\psi}$ be the boundary type such that types less risk averse than $\hat{\psi}$ choose x^k or below, while types more risk averse than $\hat{\psi}$ choose x^{k+1} or above. Now, raise the premiums for all contracts $k+1$ and above by a small amount ε and, abusing notation, let $\hat{\psi}(\varepsilon)$ be the new boundary type after the perturbation.

Consumers with risk aversion between $\hat{\psi}$ and $\hat{\psi}(\varepsilon)$ now substitute from their previous choice of contract to a lower contract. The size of this effect depends on (i) how thick the density of types is near $\hat{\psi}$ $(g(\hat{\psi}))$; (ii) how quickly the boundary moves $(\hat{\psi}_\varepsilon(0))$; and (iii) the per-consumer impact on the insurer of the induced change in contract choice measured by $\mathcal{S}$. When $\hat{\psi}$ is interior, *2BRP* implies that the boundary type $\hat{\psi}$ is indifferent between contact $\underline{x} = x^{\underline{k}}$ for some $\underline{k} \leq k$ and contract $\bar{x} = x^{\bar{k}}$ for some $\bar{k} > k$, and that these two contracts are the only two optimal choices. In this case, we can define a ratio

$$r = \frac{\mathcal{S}(\bar{x}, \hat{\psi}) - \mathcal{S}(\underline{x}, \hat{\psi})}{v_\psi(\bar{x}, \hat{\psi}) - v_\psi(\underline{x}, \hat{\psi})}, \tag{6}$$

where the denominator captures the speed at which the boundary type moves and the numerator captures the impact of that move on the insurer.[14] Multiplying r by $g(\hat{\psi})$ captures effects (i)–(iii).

The minimum is well-defined since ρ is left-continuous. It is straightforward to check that d is a metric. Indeed, d is the Levy metric (Billingsley (1995); Problem 14.5, p.198) adjusted to take account of the fact that x lies in a compact support, and we will refer to it as such henceforth.

[14]If $\hat{\psi}$ is not interior, then set $r = 0$, since in that case, $\hat{\psi}_\varepsilon(0) = 0$. In the proof, we show that with probability one there is some (ω, F)-type such that either $\hat{\psi}$ is interior or the consumer has a strict preference between his favorite contact below x^k and his favorite contract above x^{k+1}. In that event, $\hat{\psi}$ will equal either $\underline{\psi}$ or $\bar{\psi}$ as appropriate, and will remain that way even when the price vector is perturbed by a small amount.

The other effect of the perturbation is that the insurer now makes more money on infra-marginal consumers who continue to choose a contract above x^k. The size of this effect depends on the number of (ω, F)-type consumers who are more risk averse than $\hat{\psi}$ $(1 - G(\hat{\psi}))$. At the optimum, the insurer balances the expected value of all of these effects across (ω, F)-types. Reintroducing dependencies on (ω, F), the overall impact on the insurer's payoff when facing type (ω, F) is

$$\mathcal{V}(x^k, \omega, F) \equiv (w^I - w^C)(1 - G(\hat{\psi}(x^k, \omega, F)|\omega, F)) - r(x^k, \omega, F)g(\hat{\psi}(x^k, \omega, F)|\omega, F). \tag{7}$$

We can now state our optimality theorem. Write $G(\omega, F)$ for the marginal of G onto (ω, F).

Theorem 1 (Optimality Condition: Fixed Set of Contracts) *Let (ρ, χ) be optimal given $\{x^k\}_{k=0}^K$, and let ρ satisfy 2BRP. Then, $\int \mathcal{V}(x^k, \omega, F)dG(\omega, F) \leq 0$ for $k < K$ with equality if $\rho(x^k) < \rho(x^{k+1})$.*

The role of $\rho(x^k) < \rho(x^{k+1})$ is that when $\rho(x^k) = \rho(x^{k+1})$, the insurer cannot lower $\rho(x^{k+1})$ without also lowering $\rho(x^k)$, given that price schedules must be monotone. The proof is in Appendix A.2.

4.3 A Continuum of Contracts

We next consider what happens when the insurer is free to offer all coverage levels x in $[0, 1]$. As before, fix an x strictly above which we will raise the price by ε, and given x and (ω, F), let $\hat{\psi}$, $\hat{\psi}(\varepsilon)$, $\bar{x}$ and $\underline{x}$ be defined as before. If $\bar{x} > \underline{x}$, then r defined by equation (6) continues to capture the effect of types who flow from above x to below x when ε is raised. But, because we are in the continuum, it can easily be that the best contract choice correspondence is single valued at $\hat{\psi}$, so that $\bar{x} = x = \underline{x}$. In this case, it is useful think of r as reflecting a limit where $\bar{x} - \underline{x}$ is strictly positive but small, and note that Cauchy's Mean Value Theorem then tells us that

$$r = \frac{\mathcal{S}_x(x, \hat{\psi})}{v_{x\psi}(x, \hat{\psi})},$$

an intuition we formalize in Appendix A.4. With the definition of r modified in this way, we can again show that the value of the perturbation facing (ω, F) is $\mathcal{V}(x, \omega, F)$, and so Theorem 1 generalizes readily to the continuum. See Theorem 3 in Appendix A.4. There is also an additional necessary condition that must hold in the continuum case, related to the insurer's ability to adjust coverage levels of the contracts offered, in addition to their prices. This intermediate case of a finite number of contracts with endogenous qualities is discussed in Appendix A.3.

4.4 Some Relationships to the Literature

Theorem 1 (and Theorem 3 in Appendix A.3) generalize several results from the literature on principal-agent problems with private information. First, our optimality condition for the premium schedule $\int \mathcal{V} dG(\omega, F) = 0$ can in fact be interpreted in quite a familiar way. When the insurer is a monopolist, it has a *marginal revenue* = *marginal cost* interpretation, and when the insurer is a social planner, it has a *price* = *marginal cost* interpretation. We will make this point in more detail in Section 5, and so we defer the details.

Second, consider the monopoly case and a continuum of contracts, and assume that there is only one (ω, F), but that ψ is the consumer's private information. Then the setting reduces to a standard *one-dimensional principal-agent problem*, and it can be shown that our main condition given in equation (7) reduces to

$$\mathcal{S}_x g - v_{\psi x}(1 - G) = 0,$$

and so reflects the standard efficiency versus information-rents trade-off. That is, providing slightly more coverage to a type ψ changes efficiency by $\mathcal{S}_x g$, but also has an impact $v_{\psi x}(1-G)$ on the information rents that must be given to types higher than ψ. If instead we change a given quality x while leaving its premium unaltered, then the perturbation has bite only if χ is constant on some interval (ψ^l, ψ^h). Online Appendix B.6 shows that the ensuing condition reduces to the standard "ironing" condition (Fudenberg and Tirole, 1991, Chapter 7),

$$\int_{\psi^l}^{\psi^h} \left(\mathcal{S}_x - v_{x\psi} \frac{1-G}{g} \right) g dl = 0.$$

Third, return to multidimensional types, and assume that we restrict the monopolist to choosing a single contract, which is a special case of the setting with a finite number of contracts discussed in Appendix A.3. Online Appendix B.6 shows that in this case, our necessary conditions coincide with those of Veiga and Weyl (2016). Namely, the conditions can be combined to derive a single necessary condition on the optimal contract x with a term involving the covariance between the marginal benefit for the consumer, v_x, and the cost to the insurer γ^I, calculated using the density of types on the margin between choosing x and the outside option x^0.[15,16]

Finally, and in line with Technical Remark 3, if we start with functions v and γ^I as primitives (without the structure provided by the insurance problem), then our results provide the optimality conditions for suitable extensions of, say, Mussa and Rosen (1978) and Maskin and Riley (1984) with multidimensional types.

[15] Veiga and Weyl (2016) interpret a positive covariance as "adverse sorting" in that the marginal types are costly for the firm, and a negative covariance as "advantageous sorting."

[16] One can generalize this construction to any finite number of contracts (with two covariances in the resulting expression). We omit this development for several reasons. First, we find the interpretation of the resulting expression to be more involved than that driven directly by the two perturbations. Second, the covariance terms disappear in the limit as steps grow small. And third, we are skeptical that there are economically interesting primitives giving structure to these covariances.

4.5 Convergence

We now show that the solution to the problem where the insurer can offer a continuum of contracts is well-approximated by the problem with a finite set of contracts. Definition 1 defines convergence of sets of price schedules. Theorem 2 then shows that if $\mathcal{P}^n$ converges to $\mathcal{P}^0$, then the payoff to the insurer does as well, and that the limit of optimal solutions is optimal.

Definition 1 *Say that a sequence $(\mathcal{P}^n)$ of closed subsets of the closed subset $\mathcal{P}^0 \subseteq \mathcal{P}$ converges to $\mathcal{P}^0$ if for all $\rho \in \mathcal{P}^0$, there is a sequence (ρ^n) with each $\rho^n \in \mathcal{P}^n$ such that $\rho^n \to \rho$.*

Theorem 2 (Convergence) *Let $\mathcal{P}^0$ be closed, and let $\mathcal{P}^n \to \mathcal{P}^0$. Then, the payoff to the insurer under $\mathcal{P}^n$ converges to her payoff under $\mathcal{P}^0$. Further, if $\rho^n \to \hat{\rho}$ is any convergent sequence of optimal solutions for the insurer given $\mathcal{P}^n$, then $\hat{\rho}$ is optimal for the insurer in $\mathcal{P}^0$, and the payoffs to the consumer of each type converges to those under $\hat{\rho}$.*

Theorem 2 has two main implications. First, for numerical purposes, the modeler can use any reasonable set of fixed contracts, and be confident that they get a result that approximates what the insurer can achieve with a continuum of contracts. The details of how the sequence $\mathcal{P}^n$ is constructed simply do not matter, as long as the set of contracts grows dense.

Second, this result provides theoretical flexibility. If the insurer can offer a sufficiently rich set of fixed contracts, then there is a vanishing amount of value added by also allowing it to modify the coverage levels of those contracts, as in Appendix A.3. We can therefore work in the case of a (large) fixed set of contracts or in the continuum, whichever is more convenient. In the remainder of the paper, we focus on fixed set of contracts. Our numerical exploration suggests that the number of allowable contracts can be shockingly small and still closely approximate the continuum limit in terms of insurer and consumer payoffs.[17] We discuss the speed of convergence in Section 6.2.

4.6 Incentives to Exclude, Screen, and Trade

Besides being intuitive, the optimality condition $\int \mathcal{V} dG = 0$ also provides insight into the insurer's incentives to optimally exclude and screen types. For example, one can show that from the point of view of the social planner, a monopolist excludes too many consumers from insurance above x^0. One can also show that if ω was the only source of private information, a social planner would completely pool types, while a monopolist insurer may even completely sort types, an extreme example of differential incentives to screen. See Online Appendix B.7 for details. Since these insights will be much more transparent in Section 5, we postpone further discussion of their economic implications.

[17] We conjecture that if the contracts are relatively evenly spaced, then convergence is of the order $1/K^2$. If the insurer's profit had a Gateaux derivative everywhere, with bounds on the second derivative, then the rate of convergence result would follow as long as the optimal ρ is interior. Where *2BRP* holds, the Gateaux derivative as one moves from ρ linearly towards $\hat{\rho}$ is $\int\int(\hat{\rho}(x) - \rho(x))\mathcal{V} dG dx$. We do not know how to show that *2BRP* holds everywhere or how to tame the speed at which the derivative changes.

Another question of interest is whether a monopolist insurer always trades (that is, makes strictly positive profits).[18] Assume that the government's costs γ^G increase with consumers' chosen level of coverage, that $\bar{F}$ is non-degenerate (so that the worst risk type faces real risk), and that the outside option is strictly less than full insurance. Under these conditions, the monopolist insurer will always choose to sell to a strictly non-empty set of types.

Proposition 2 (Positive Trade) *Assume that the insurer is a monopolist, that there is a continuum of contracts, that the government's costs γ^G increase with consumers' chosen level of coverage, and that $x^0 < 1$. Then, because $\bar{F}$ is non-degenerate, any optimal menu for the insurer involves a strictly positive amount of trade. That is, the insurer sells contracts strictly greater than x^0 to a positive-measure set of types.*

The proof is in Appendix A.5, but the intuition is as follows. From the point of view of the monopolist, there is "no moral hazard" at x^0, since the government pays for any spending that occurs. Giving a little extra insurance to some types therefore has a first-order gain in terms of the insurance motive, but only a second-order cost in terms of medical spending that is viewed as wasteful from the point of view of the monopolist. If the government instead adjusts its policy so that the monopolist bears the full cost of all spending induced by higher coverage, then trade will only occur so long as there are positive gains from trade in the market for incremental coverage.

5 A Simplified Problem

We now reformulate the insurer's problem in a way that dramatically simplifies the derivations and allows the problem to be analyzed in a familiar graphical framework. Formally, we show that under a quasiconcavity assumption on the consumer's problem, the insurer's full problem of setting a price schedule on a fixed set of contracts can be reduced to a set of entirely independent problems of setting the marginal price of incremental coverage. A similar approach has been proposed by Wilson (1993) in a much simpler setting. The strengths and weaknesses of this approach are discussed in depth in Armstrong (2016).

5.1 The Reformulation

The key aspect of the simplification is to reformulate everything in terms of incremental levels of coverage. To that end, we wish to express the insurer's expected payoff on a given consumer in a given contract as the payoff the insurer obtains when the consumer takes the outside option *plus* all the incremental effects of moving the consumer from one coverage level to the next until the relevant contract is reached.

[18] This question has received attention both in the empirical insurance literature (see the no-trade result in Hendren, 2013) and in the theoretical insurance literature with either adverse selection or both adverse selection and moral hazard (see Chade and Schlee (2020) and Chade and Swinkels (2022) for no-trade and trade results).

Fix a set of allowable contracts $x^0 < x^1 < x^2 < \cdots < x^K \leq 1$. For a given premium schedule ρ, and for $k = 1, \ldots, K$, let $p^k = \rho(x^k) - \rho(x^{k-1})$ be the marginal premium between adjacent contracts. Similarly, let $v^k(\theta) = v(x^k, \theta) - v(x^{k-1}, \theta)$ be a type-θ consumer's marginal willingness to pay between adjacent contracts, $\gamma^{I,k}(\theta) = \gamma^I(x^k, \theta) - \gamma^I(x^{k-1}, \theta)$ be the marginal insured cost, and $\gamma^{G,k}(x^0, \theta) = \gamma^G(x^0, x^k, \theta) - \gamma^G(x^0, x^{k-1}, \theta)$ be the government's part of that cost. Note that since contracts are vertically differentiated and the premium schedule is increasing in coverage level, p^k, v^k, $\gamma^{I,k}$, and $\gamma^{G,k}$ are all weakly positive. The insurer's marginal payoff from charging type θ a marginal premium p^k to move up a coverage level is then

$$S^k(p^k, \theta) = w^C(v^k(\theta) - p^k) + w^I(p^k - \gamma^{I,k}(\theta) + \gamma^{G,k}(x^0, \theta)) - w^G\gamma^{G,k}(x^0, \theta).$$

Given this notation, the insurer's objective function can be re-expressed. Let $\tilde{k}(\theta, \rho)$ be the optimal contract chosen by a consumer of type θ facing price schedule ρ (that is, $\chi(\theta) = x^{\tilde{k}(\theta,\rho)}$). The payoff to the insurer on type θ given ρ is then

$$S^0(\theta) + \sum_{k=1}^{\tilde{k}(\theta,\rho)} S^k(p^k, \theta), \tag{8}$$

where $S^0(\theta) = w^C v(x^0, 0) + w^I\gamma^I(x^0, \theta) + w^G(x^0, x^0, \theta)$ is the payoff from putting type θ into contract x^0, and where we take the sum to be zero when $\tilde{k}(\theta, \rho) = 0$.

The function $\tilde{k}$ is still complicated, since the consumer's optimal contract is still defined by a set of non-local incentive constraints. But if the consumer's problem has some additional structure, we can substantially simplify $\tilde{k}$. Say that a price schedule ρ is *quasiconcave consistent (QC) for* a consumer of type θ if the consumer's payoff $v(x^k, \theta) - \rho(x^k)$ is single-peaked in k (that is, single-peaked in coverage level). When ρ is *QC* for θ, then the consumer's payoff reaches its peak at the last point where their marginal payoff $v^k(\theta) - p^k$ is positive. This is very useful, because now $\tilde{k}$ is defined by a local incentive constraint: $\tilde{k}(\theta, \rho) = \max\{k | v^k(\theta) \geq p^k\}$.

If ρ is *QC* for θ, then we can rewrite the insurer's payoff on type θ as

$$S(x^0, \theta) \ + \sum_{\{k | v^k(\theta) \geq p^k\}} S^k(p^k, \theta). \tag{9}$$

If ρ is *QC* for all θ, then we can write the insurer's expected payoff from providing increment k of insurance to all types who are willing to pay for the increment as

$$\tilde{\Pi}^k(p^k) \equiv \int_{\{\theta | v^k(\theta) \geq p^k\}} S^k(p^k, \theta) dG(\theta),$$

meaning the insurer's total payoff is given by

$$\tilde{\Pi}(\rho) \equiv \int S(x^0, \theta) dG(\theta) + \sum_{k=1}^{K} \tilde{\Pi}^k(p^k).$$

To see this, integrate (9) with respect to G and swap the order of integration and summation.[19]

So, consider the problem

$$(\tilde{P}) \qquad \max_{(p^1,\ldots,p^K)} \sum_{k=1}^{K} \tilde{\Pi}^k(p^k).$$

Since each of the insurer's incremental payoffs $\tilde{\Pi}^k(p^k)$ is a function only of the incremental price p^k, the solution $\tilde{\rho}$ to $\tilde{P}$ can be solved one contract at a time. That is, the optimal price schedule $\tilde{\rho}$ can be constructed from the set of optimal marginal prices $(\tilde{p}^1, \ldots, \tilde{p}^K)$, where on each margin, $\tilde{p}^k \in \arg\max_{p^k} \tilde{\Pi}^k(p^k)$. $\tilde{P}$ is a much simpler problem than P.

When can we use the solution to $\tilde{P}$ to understand the solution to P? Note first that if ρ is QC for all consumer types except a small set, then we will have $\Pi(\rho) \cong \tilde{\Pi}(\rho)$, with the approximation arbitrarily good as the measure of the set of types where ρ is not QC goes to zero. The two solutions will therefore be nearly the same if the solution to each is QC for all but a small measure of θ. If we knew *a priori* that any optimal solution to either P or $\tilde{P}$ was QC for all θ, then the two problems would be exactly the same.

We are unaware of *theoretical* primitives that justify QC (indeed, Deneckere and Severinov (2017) cast serious doubt on whether such primitives generally exist). However, we note three features of QC here. First, the simplification holds only for vertically differentiated contracts, since otherwise there is no natural order on contracts (our necessary conditions can be extended to the non-stacked case). Second, single-dimensional consumer heterogeneity is neither necessary nor sufficient for QC to hold. For example, if moral hazard is initially strong but is lower at high levels of spending, QC is unlikely to hold even if the only dimension of heterogeneity is risk aversion: the insurer will want to offer lower level of insurance, where there is significant incremental moral hazard, to few consumers, but higher level of insurance, where there is little incremental moral hazard, to many consumers. Third, in our empirical setting below, differences in risk types (that is, in F) drive a substantial portion of variation in willingness to pay (WTP), and this seems to help QC to hold. In the end, QC must be checked empirically, and our numerical results illustrate a method for doing

[19]Formally, letting $\mathbb{I}_A$ be the indicator function of the set A,

$$\begin{aligned}
\int \sum_{\{k|v^k(\theta)\geq p^k\}} S^k(p^k,\theta)dG(\theta) &= \int \sum_{k=1}^{K} \mathbb{I}_{\{v^k(\theta)\geq p^k\}} S^k(p^k,\theta)dG(\theta) = \sum_{k=1}^{K} \int \mathbb{I}_{\{v^k(\theta)\geq p^k\}} S^k(p^k,\theta)dG(\theta) \\
&= \sum_{k=1}^{K} \int_{\{\theta|v^k(\theta)\geq p^k\}} S^k(p^k,\theta)dG(\theta) = \sum_{k=1}^{K} \tilde{\Pi}^k(p^k).
\end{aligned}$$

so. In Section 6, we further discuss this condition and evaluate the extent to which optimal menus are QC in our calibrated population.

5.2 Analyzing the Simplified Problem

In the simplified problem, we can think about the insurer's optimal price increments p^k one at a time. In this section, we use this simplicity to analyze the solution to $\tilde{P}$, and show how to think of that solution in familiar terms.

To begin, rewrite the objective of the insurer in terms of quantities instead of prices. That is, instead of choosing incremental prices, we can think of the insurer as choosing the fraction of consumers that will purchase each incremental coverage level. When the incremental price is p^k, this fraction is equal to $Q^k(p^k) = \int_{\{\theta | v^k(\theta) > p^k\}} dG(\theta)$. When $Q^k \in (0,1)$, it is strictly decreasing in p^k, and thus has an inverse function P^k defined by $P^k(Q^k(p^k)) = p^k$ for every p^k.[20]

Let

$$C^{I,k}(q^k) = \int_{\{\theta | v^k(\theta) > P^k(q^k)\}} \gamma^{I,k}(\omega, F) dG(\theta)$$

be the insurer's cost of providing incremental coverage level k to the q^k consumers who purchase at price $P^k(q^k)$. Let the marginal cost $MC^{I,k}$ be the derivative of $C^{I,k}$. Similarly, let

$$C^{G,k}(q^k) = \int_{\{\theta | v^k(\theta) > P^k(q^k)\}} \gamma^{G,k}(x^0, \omega, F) dG(\theta)$$

be the government's cost of q^k, with associated marginal cost $MC^{G,k}$, and let

$$V^k(q^k) = \int_{\{\theta | v^k(\theta) > P^k(q^k)\}} v^k(\theta) dG(\theta)$$

be aggregate consumer utility when q^k consumers are served.

Note that $V_q^k(q^k) = P^k(q^k)$. It is now straightforward to verify that

$$\tilde{\Pi}^k(P^k(q^k)) = w^C[V^k(q^k) - P^k(q^k)q^k] + w^I[P^k(q^k)q^k - C^{I,k}(q^k) + C^{G,k}(q^k)] - w^G C^{G,k}(q^k), \tag{10}$$

and thus we can think of the insurer as solving $\max_{q^k} \tilde{\Pi}^k(P^k(q^k))$. We can also now usefully decompose the insurer's payoff into a "benefit" equal to $(w^I - w^C)P^k(q^k)q^k + w^C V^k(q^k)$ and a "cost" equal to $w^I C^{I,k}(q^k) - (w^I - w^G)C^{G,k}(q^k)$. In the case of a monopolist, when $(w^C, w^I, w^G) = (0, 1, 0)$, the benefit is simply revenue, $P^k(q^k)q^k$, and the cost is simply the expected insured cost of incremental coverage, $C^{I,k}(q^k) - C^{G,k}(q^k)$.

Denoting the price-elasticity of demand by ϵ, so that $1/\epsilon = P_{q^k}^k q^k / P^k$, we can write the derivative

[20]To see that Q^k is strictly decreasing where it is interior, recall that $v_{x\psi} > 0$ and so $v^k(\omega, \cdot, F)$ is strictly increasing. Hence, $\{\theta | v^k(\theta) > p^k\}$ is strictly shrinking in p^k.

of the insurer's objective function as

$$(11)\quad \left(\tilde{\Pi}^k(P^k(q^k))\right)_{q^k} = \underbrace{P^k(q^k)\left(w^I + (w^I - w^C)\tfrac{1}{\epsilon}\right)}_{Marginal\ benefit} - \underbrace{(w^I MC^{I,k}(q^k) - (w^I - w^G)MC^{G,k}(q^k))}_{Marginal\ cost}.$$

The first term is the insurer's marginal benefit of giving more consumers incremental coverage level k.[21] As quantity increases, the insurer receives $P^k(q^k)$ on the extra unit sold, but P^k is falling at rate $P^k(q^k)/\epsilon$, resulting in a transfer from the consumer to the insurer valued at $w^I - w^C$. The second term is the insurer's marginal cost, where $w^I MC^{I,k}(q^k)$ is the incremental insured cost of the marginal consumer, and $(w^I - w^G)MC^{G,k}(q^k)$ is the insurer's valuation of the associated government spending.

At an optimum, marginal benefit is equal to marginal cost, yielding a familiar markup equation. For example, when the insurer is a monopolist, the optimality condition reduces to $P^k(1+(1/\epsilon)) = MC^{I,k} - MC^{G,k}$. Furthermore, all the terms in $\tilde{\Pi}^k_{p^k} = 0$ can be unpacked to obtain an expression that is a direct analog of the optimality condition $\int \mathcal{V} dG = 0$.[22] This makes intuitive sense, as increasing the marginal premium p^k raises the price schedule ρ for all $x > x^{k-1}$ and so is effectively the perturbation discussed in Section 4.2. It should therefore not be a surprise that we could have interpreted the original optimality condition as a markup equation, as we do here.

5.3 Graphical Analysis

The simplified problem is composed of a set of independent two-contract problems along each potential coverage level margin. It can therefore be analyzed graphically in the spirit of Einav et al. (2010). Indeed, they suggest a similar "incremental" approach to generalize their model to more than two contracts in the case of perfect competition, and Geruso et al. (2019) take a first step in this direction by extending the graphical analysis to accommodate three contracts. Our analysis formalizes the assumptions necessary to carry out this approach. In addition, the flexibility of our insurer objective function allows our graphical analysis to nest both the case of a monopolist insurer (as in Mahoney and Weyl, 2017) and the case of a social planner (as in Marone and Sabety, 2022). For simplicity, we normalize the insurer's weight on its own profits w^I to 1, and suppose the government's cost of providing base coverage does not depend on the consumer's chosen contract ($\gamma^G_x = 0$), and hence $MC^{G,k} = 0$ for all k.

Figure 1 illustrates the insurer's problem for one marginal coverage level. It shows the inverse demand function P^k relevant on that margin, the associated marginal revenue function $MR^k = P^k(1 + (1/\epsilon))$, and the insurer's marginal cost curve $MC^k = MC^{I,k}$.[23] The insurer's marginal

[21] The marginal benefit also in principle includes a term $w^C(V^k_{q^k}(q^k) - P^k(q^k))$, but since the marginal consumer is indifferent about paying $P^k(q^k)$, this term is zero.

[22] See Lemma 3 in Appendix A.7.

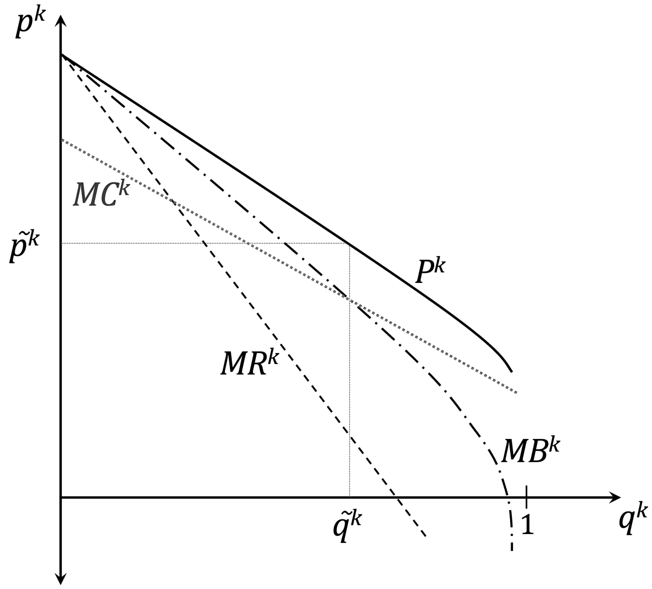
[23] Note that we have drawn the marginal cost as decreasing in the quantity of consumers that purchase the marginal coverage, reflecting an assumption that there is adverse selection (Einav et al., 2010).

benefit of serving more consumers is then $MB^k(q^k) = P^k(q^k)(1+(1-w^C)\frac{1}{\epsilon})$, which can be written as a convex combination of the marginal revenue and inverse demand curves, depending on the weight given to the consumer:

$$MB^k(q^k) = (1-w^C)MR^k(q^k) + w^C P^k(q^k).$$

The marginal benefit curve shown is an example for a case where $w^C \in (0,1)$.[24] The insurer's optimal quantity $\tilde{q}^k$ obtains where $MB^k = MC^k$.[25]

Figure 1. Insurer's Optimal Choice of q^k



Notes: The figure shows the inverse demand curve P^k, the marginal revenue curve MR^k, the insurer's marginal cost curve MC^k, and the insurer's marginal benefit curve MB^k in the market for incremental coverage amount k. The insurers optimal quantity $\tilde{q}^k$ obtains where the marginal benefit curve intersects the marginal cost curve.

Figure 1 subsumes a number of cases of interest. In the case of a monopolist ($w^C = w^G = 0$), MB^k coincides with the marginal revenue curve MR^i, and the optimal quantity solves $MR^k = MC^k$ (so long as the solution is interior). For a utilitarian social planner with no excess cost of public funds ($w^C = w^G = 1$), MB^k coincides with the inverse demand curve, and the optimal quantity solves $P^k = MC^k$. As drawn, the social planner chooses the corner $\tilde{q}^k = 1$. Finally, in

[24]Note that $w^C > 1$ is a viable possibility.

[25]Note we have drawn the marginal benefit curve as crossing the marginal cost curve from above. This is not guaranteed from our primitives. Indeed, there is the possibility of a single crossing from below or of no crossing at all, in which case the solution will not be interior, or of multiple crossings. One can similarly write the insurer's average benefit as $AB^k(q^k) = (1-w^C)P^k(q^k) + w^C(V^k(q^k)/q^k)$, and note that this is a convex combination of the average benefit function of the monopolist (their average revenue), P^k, and that of the social planner, V^k/q^k, and the insurer's average cost AC^k as $AC^k(q^k) = (1-w^G)((C^{I,k}(q^k)/q^k) - (C^{G,k}(q^k)/q^k)) + w^G(C^{I,k}(q^k)/q^k)$, which is also a convex combination. The insurer is better off not selling incremental insurance x^k to anyone if AB^k lies below AC^k at the optimal interior choice. In our example figure, where marginal cost crosses marginal benefit from below, this "participation" constraint is automatically satisfied, since there are no fixed costs, so that profits are zero at $q^k = 0$ and increasing as we move towards $\tilde{q}^k$.

the case of a planner with an excess cost of public funds ($w^C < 1 = w^G$), MB^k is the usual convex combination of the marginal revenue and inverse demand curves. As the cost of public funds rises (which, given our normalization, corresponds to w^C falling), we approach the monopoly solution.

5.4 Comparative Statics

The simplified problem yields some simple *monotone comparative statics* of economic interest, many of which can be read directly from Figure 1. First, when the consumer is weighted more heavily in the insurer's objective function, the marginal benefit curve rotates up towards the demand curve. The optimal quantity $\tilde{q}^k$ on every marginal coverage level is therefore increasing in w^C (and the optimal price $\tilde{p}^k$ is decreasing).[26] Since the price of x^0 is fixed at zero, an increase in w^C makes the premium schedule lower and flatter. Conversely, if the consumer is weighted less heavily relative to the insurer, for example if the insurer is a social planner facing a rising cost of public funds, the marginal benefit curve rotates down towards the marginal revenue curve. As the cost of public funds increases to infinity, the marginal benefit curve eventually coincides exactly with the marginal revenue curve (that is, the monopolist's and the social planner's solutions would coincide).

Thus far we have assumed that the government's cost of providing x^0 does not depend on the consumer's chosen contract ($\gamma_x^G = 0$). In this case, a change in w^G has no effect on Figure 1, since the government's cost of providing x^0 is simply a fixed sum. If $MC^{G,k}$ is instead strictly positive, then increasing w^G causes the marginal cost curve to go up, since the government's cost of providing x^0 would be increasing in coverage level due to moral hazard. In this case, an increase in w^G results in an increase in the optimal price, making the premium schedule higher and steeper.

5.5 Exclusion and Screening

In Section 4.6, we argued that a monopolist has stronger incentives than a social planner to both exclude and screen consumers. The simplified problem allows us to strengthen these results and visualize them graphically.

First note that as long as the insurer values profits more than consumer surplus, the marginal benefit curve will diverge to $-\infty$ as q^k goes to 1, as depicted in Figure 1. The reason for this is that q^k goes to 1, the reciprocal elasticity of demand $1/\epsilon$ goes to $-\infty$. So long as this term gets any weight in the insurer's objective, i.e., as long as profits are weighted at least slightly more heavily than consumer surplus, the optimal marginal quantity $\tilde{q}^k$ is therefore strictly less than one.

Proposition 3 (Optimal Exclusion at Every Level) *If $w^I > w^C$, then $\tilde{q}^k < 1$ for all k.*

The proof is in Appendix A.8. Proposition 3 applies at every marginal coverage level, including

[26] By standard monotone comparative statics results, this is true even if there are multiple crossings of marginal benefit and marginal cost, if there was a single crossing from below, or if there was originally no crossing.

the first increment of coverage. It thus implies that an insurer with $w^I > w^C$ optimally excludes a *strictly* positive measure set of consumers from the market for incremental coverage.[27]

Proposition 3 also sheds light on the differential incentives of the social planner and the monopolist to screen consumers. As drawn in Figure 1, the demand curve everywhere exceeds the marginal cost curve, and so, given QC, the social planner wishes to provide a coverage level weakly greater than k to all consumers.[28] But, consistent with Proposition 3, the monopolist optimally allocates incremental coverage k to some but not all consumers. Hence, again under QC, the monopolist is allocating some consumers to a contract k or below, and the monopolist uses more contracts than the social planner.

Finally, note that Proposition 3 also implies that due to the forces of both exclusion and screening, a monopolist will offer less coverage than a social planner. We investigate the numerical magnitude of these differences in Section 6.4.

6 Numerical Analysis

To build upon these theoretical predictions, we calibrate a model of a health insurance market and evaluate market outcomes under various scenarios. Beyond offering a numerical illustration of our key theoretical results, this approach also allows us to evaluate the magnitudes of the equilibrium impacts of various policy interventions of interest.

6.1 Description of the Calibrated Market

Consumers. We simulate a population of consumers using a distribution of demographics chosen to match the under-65 US population and parameter estimates reported in Marone and Sabety (2022).[29] Each consumer is a household composed of some number of individuals. Each household is characterized by type $\theta = (\psi, \omega, F)$, where F is assumed to have a shifted log-normal distribution such that $\log(l + \kappa) \sim N(\mu, \sigma^2)$. Consumer preferences feature constant absolute risk aversion, and we parameterize b such that $b(a, l, \omega) = (a - l) - \frac{1}{2\omega}(a - l)^2$.

Table 2 summarizes the characteristics of our simulated population. The average household would have total healthcare spending equal to \$12,170 under a full insurance contract, but only

[27] Optimal exclusion from coverage has precedent in the literature, but only without common values and with much more structure on the consumer's payoff function (Armstrong, 1996; Deneckere and Severinov, 2017).

[28] To see that the situation of Figure 1 can occur, with $MC^{I,k}(1) \leq P^k(1)$, let $\hat{\theta} = (\hat{\omega}, \underline{\psi}, \underline{F})$ be the type in the population with the lowest marginal willingness to pay for k. (We appeal to Proposition 1 to know that this type has the most favorable risk distribution $\underline{F}$ and the lowest risk aversion $\underline{\psi}$, but may have $\hat{\omega}$ interior, and we assume this type is unique for simplicity.) Then, $MC^{I,k}(1) = \gamma^{I,k}(\hat{\theta})$, and $P^k(1) = v^k(\hat{\theta})$, and so $MC^{I,k}(1) \leq P^k(1)$ if and only if $\gamma^{I,k}(\hat{\theta}) \leq v^k(\hat{\theta})$. Primitives for this are easily established. For example, if the healthiest type in the population still faces risk, then $\gamma^{I,k}(\hat{\theta}) \leq v^k(\hat{\theta})$ holds as long as the least risk-averse type in the population is sufficiently risk averse.

[29] Details of the simulation procedure are provided in Online Appendix B.1.

$10,684 under the null contract, reflecting moral hazard. Facing an equal odds gamble between $0 and $100, the average household would have a certainty equivalent of $48.9, reflecting risk aversion. Online Appendix Figure B.1 provides additional information on the joint distribution of types in the population.

Table 2. Population Summary Statistics

Sample demographic	Mean	Percentile 10	25	Median	75	90
Demographics						
Number of adults	1.9	2.0	2.0	2.0	2.0	2.0
Number of children	0.6	0.0	0.0	0.0	1.0	2.0
Average age of household adults	43.5	26.2	32.6	43.6	54.3	60.7
Dimensions of type θ						
Health state distribution parameter μ	1.6	0.3	0.9	1.6	2.3	2.8
σ	1.0	0.8	0.9	1.0	1.2	1.3
κ	0.6	0.1	0.3	0.5	0.9	1.3
Moral hazard parameter ω	1.4	0.8	1.0	1.3	1.7	1.9
Risk aversion parameter ψ	0.9	0.2	0.4	0.6	1.1	1.9
Resulting characteristics						
CE of equal odds gamble between $0 and $100 ($)	48.9	47.6	48.6	49.2	49.5	49.7
Expected total spending, null contract ($000)	10.6	3.0	4.4	7.8	13.8	22.3
full insurance ($000)	11.9	4.1	5.7	9.2	15.2	23.6

Notes: The table shows descriptive statistics for our simulated population of 10,000 households. Note that the moral hazard parameter and coefficient of absolute risk aversion reported are relative thousands of dollars.

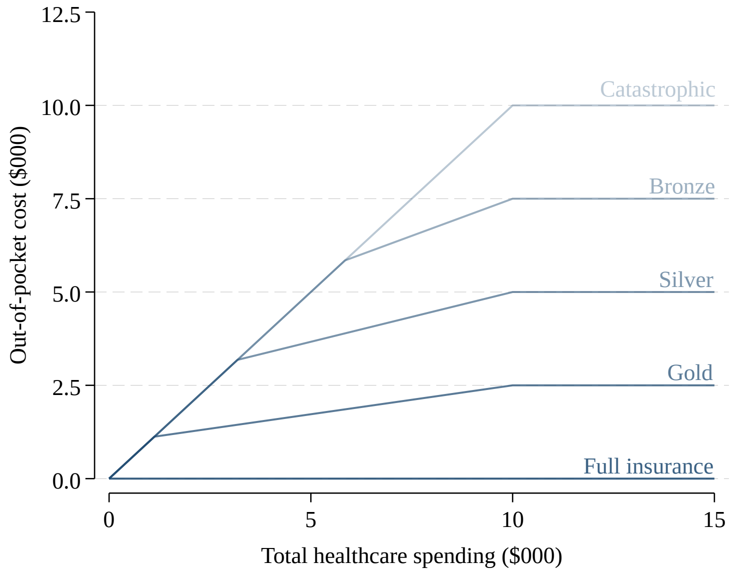
Insurance Contracts. We consider a set of contracts that are piecewise linear, with a deductible, coinsurance region, and out-of-pocket maximum design. We suppose that the base level of coverage x^0 is a "Catastrophic" contract with a deductible and out-of-pocket maximum of $10,000. Given the theoretical analysis, we restrict attention to a potentially large, but fixed set of contracts. Our baseline set of allowable contracts is depicted in Figure 2. Because they roughly correspond to the levels of coverage available on the Affordable Care Act exchanges, we refer to the contracts between Catastrophic and full insurance as Bronze, Silver, and Gold.[30] As will become clear, the returns to allowing an increasingly "dense" contract space are economically small.

6.2 Convergence

Theorem 2 predicts that an insurer's payoff when restricted to a finite set of contracts will converge to its unrestricted counterpart as the number of contracts grows. It is silent, however, on how quickly this may occur. We illustrate and investigate this result by computing optimal menus on an increasingly dense set of allowable contracts. Figure 2 depicts a set of five allowable contracts,

[30] The contracts' deductibles, coinsurance rates, and out-of-pocket maximums are: $5,846, 40%, $7,500 for Bronze; $3,182, 27%, $5,000 for Silver; and $1,125, 15%, $2,500 for Gold. The actuarial value of the five contracts in our population of consumers are: 0.40, 0.49, 0.61, 0.79, and 1.00.

Figure 2. Potential Contracts



Notes: The figure shows our focal set of allowable contracts. The base level of coverage x^0 provided by the government is the Catastrophic contract.

spaced at $2,500 out-of-pocket maximum intervals between the minimum and maximum levels of coverage. We increase (and decrease) the density of this contract space by varying the number of contracts used to span this range. We move from just two contracts (in which case there is just Catastrophic and full insurance) to 65 contracts (in which case 15 contracts are added between each of the five original contracts, meaning contracts are spaced at $156 out-of-pocket maximum intervals).[31]

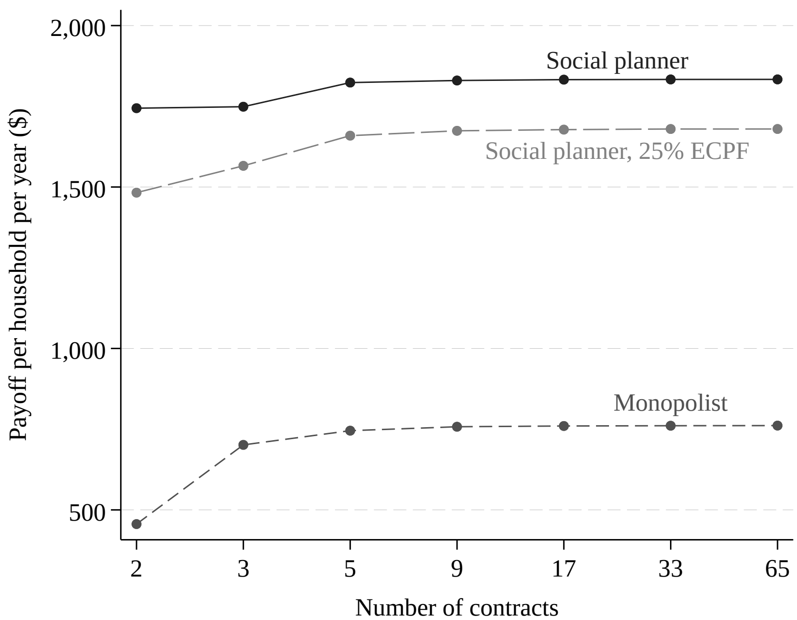
Figure 3 reports the results of this exercise. It plots insurer payoffs as a function of the number of contracts in the allowable contract space for three different insurers: a social planner with no excess cost of funds, a planner with a 25 percent excess cost of funds, and a monopolist. As predicted, insurer payoffs are increasing in the density of the contract space. But in practice, the returns to additional density are small. We find that after nine contracts (spaced at $1,125 out-of-pocket maximum intervals), the gains from moving to 65 contracts do not exceed $10 per household per year for any insurer. After five contracts, gains do not exceed $19. These results are consistent with both Marone and Sabety (2022) and Ho and Lee (2021), who find that only a limited number of contracts are sufficient to capture almost all the available surplus in their settings.

There are, however, economically meaningful gains from between two and five contracts. Over this range, the social planner facing an excess cost of funds can increase social surplus by $177 per household per year, and a monopolist can increase its profits by $289. For the social planner,

[31] We increase the set of allowable contracts by successively adding a contract between each pair of adjacent contracts. We proceed in this iterative manner so that under successively "dense" contract spaces, all previously allowable contracts remain allowable.

these gains reflect the ability to find a plan that more closely matches the tastes of consumers in the population. For the monopolist, these gains reflect this same increase in potential gains from trade, as well as the ability to more effectively screen consumers and thereby extract greater rents from the market. Our results suggest that while only a modest number of contracts are needed to closely approximate the limiting environment, there are potentially meaningful consequences of over-restricting the contract space, for example to only two contracts. Of course, the precise number of contracts at which payoffs flatten out may vary across settings, in particular with the size of the range between minimum and maximum allowable coverage.

Figure 3. Convergence



Notes: The figure shows optimal insurer payoffs as a function of the number of contracts used in the allowable contract space. Insurer payoffs are reported on a per-consumer per-year basis, and are measured relative to allocating all consumers to the Catastrophic contract.

Consistent with Theorem 2, we also find that the optimal premium schedules and therefore the optimal allocations themselves converge as the density of the contract space increases. In the case of the monopolist insurer, consumer surplus also converges alongside producer surplus. Online Appendix Figure B.2 depicts the convergence of allocations. As the density of the contract space increases, the insurers "fill in" in the neighborhood of the their desired allocation under a sparser contract space. The numerical results are thus quite robust to the density of the contract space.

6.3 Performance of the Simplified Problem

Armed with the convergence results, we proceed with the set of five fixed contracts. We next investigate how well the simplified version of the problem (presented in Section 5) approximates the true problem (presented in Section 2). Table 3 reports these results for our three focal insurers. For each insurer, the table reports the solution to the true problem P, as well as the solution to the simplified problem $\tilde{P}$. Specifically, it reports the optimal premium schedule, the associated allocations, and the associated insurer payoff when evaluated according to the objective function of each version of the problem.

Table 3. Performance of the Simplified Problem

	Premiums *\$000s*				Allocations *Pct. of households*					Insurer Payoff *\$000s*		
Insurer	Brnz.	Slvr.	Gold	Full	Cstr.	Brnz.	Slvr.	Gold	Full	Π	$\tilde{\Pi}$	Pct. QC
Social planner												
Solution to P	0.16	0.32	0.67	3.21	<0.01	–	<0.01	1.00	–	1.823	1.823	1.00
Solution to $\tilde{P}$	0.13	0.30	0.68	3.19	<0.01	–	<0.01	1.00	–	1.823	1.823	1.00
Social planner, 25% ECPF												
Solution to P	1.53	2.80	4.64	7.15	0.14	<0.01	0.13	0.74	–	1.659	1.631	0.91
Solution to $\tilde{P}$	1.32	2.85	4.73	7.23	0.13	0.03	0.13	0.71	–	1.655	1.654	0.99
Monopolist												
Solution to P	2.02	4.09	6.46	9.02	0.39	0.03	0.32	0.26	–	0.745	0.739	0.96
Solution to $\tilde{P}$	2.00	4.13	6.50	9.00	0.38	0.06	0.29	0.26	–	0.745	0.744	0.99

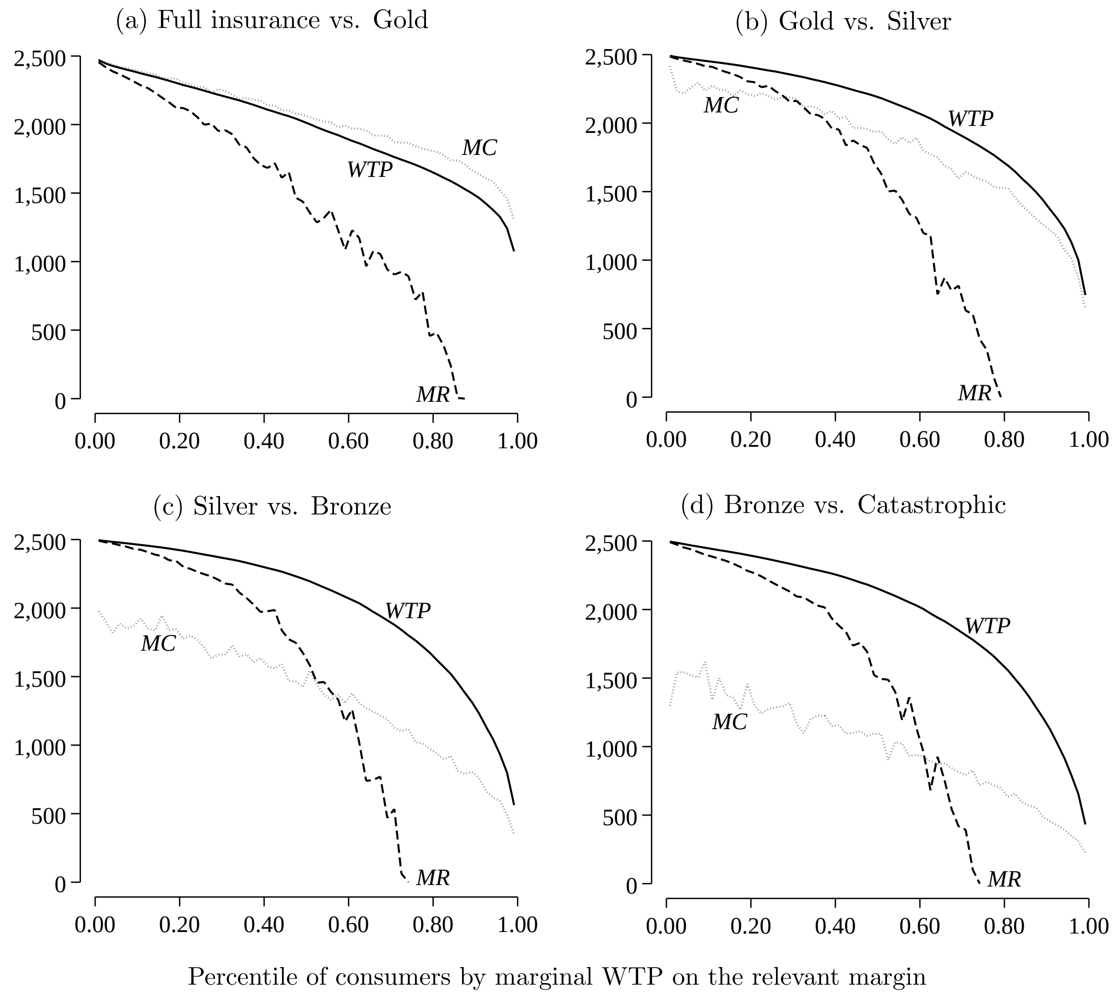
Notes: The table reports the premium schedules ρ chosen by insurers with different objective functions when solving the two formulations of the menu design problem: the original problem P and the simplified problem $\tilde{P}$. The table also reports the associated allocations and the insurer's payoff when evaluated according to the objective functions of each problem (Π and $\tilde{\Pi}$, respectively). Insurer payoffs are expressed on a per household per year basis, and are measured relative to the allocation of all consumers to the Catastrophic contract. The final column (Pct. QC) reports the percent of consumers for whom the premium schedule is quasiconcave consistent.

Recall that the key condition necessary for the two versions of the problem to coincide is that consumer payoffs are quasiconcave in coverage level at the optimal menu. The final column of Table 3 reports the fraction of consumers for whom the given price schedule fulfills this condition. We find that it holds for nearly all consumers in the population. That is, the allocation recommended by the insurer under the solution to the simplified problem is followed by nearly all consumers. Because the quasiconcavity assumption is so close to being universally satisfied, it is not surprising that insurer payoffs under the two versions of the problem are extremely close.

The quality of the approximation of the simplified problem means that it can be a very useful tool for understanding the solution to the true problem. The graphical analysis presented in Section 5.3 described how to solve the insurer's problem graphically on one marginal coverage level. When considering more than two potential contracts, there are more margins to consider. Figure 4 demonstrates how to carry out the graphical analysis on all margins simultaneously, in order to solve visually for the optimal menu across the full set of contracts. The four panels represents the "markets for incremental coverage" on each of the four margins between our five contracts. Each panel depicts the marginal willingness to pay curve WTP for the given incremental coverage amount, the associated marginal revenue curve MR, and the marginal cost curve MC associated with providing that coverage level increment.[32]

[32] Consistent with our baseline formulation of the model, we have implemented "incremental" pricing here in that the insurer's cost of providing Bronze coverage is simply the incremental cost over providing Catastrophic (and not the full cost of providing Bronze). If instead we implemented "total" pricing, the only change to Figure 4

Figure 4. Illustration of Graphical Analysis: Monopolist and Social Planner's Problems



Notes: The figure demonstrates the graphical analysis of the simplified problem. Each panel represents the "market for incremental coverage" between each pair of adjacent contracts. The vertical axes are measured in dollars. The horizontal axes report the percentage of consumers choosing a given marginal level of coverage. Consumers are ordered on the horizontal axes according to their marginal willingness to pay for the additional coverage offered on each margin. The solid line (WTP) represents consumers' willingness to pay on each margin, the dotted line (MC) represents the marginal cost curve, and the dashed line (MR) represents a monopolist's marginal revenue curve. The MC and MR curves are constructed as connected binned scatter plots using 100 points.

To solve the insurer's problem, one simply needs to find the intersection of the marginal benefit and marginal cost curves in each panel. As discussed in Section 4, a monopolist's marginal benefit curve is the marginal revenue curve. The quantities at which MR intersects MC in each panel therefore reveal the fraction of consumers to whom the monopolist wishes to provide that coverage level increment. For example, on the margin between Bronze and Catastrophic, marginal revenue

would be that the MC curve on the margin between Bronze and Catastrophic would shift up by an amount equal to the cost of supplying the Catastrophic contract, as depicted in Online Appendix Figure B.3. This would have the effect of substantially lowering the insurer's optimal quantity on that margin, likely introducing a violation of quasiconcavity.

exceeds marginal cost for about the 60 percent of consumers with the highest willingness to pay, consistent with the fact that we see the monopolist optimally allocating 61 percent of consumers to coverage above the Catastrophic contract (c.f. Table 3). The associated optimal marginal premium ($2,021) can then be read from the value of the willingness to pay curve at this quantity. On the margin between Gold and Silver coverage, marginal revenue exceeds marginal cost for about the first 25 percent of consumers, consistent with the fact that the monopolist optimally allocates roughly this fraction of consumers to Gold coverage or above. The same exercise can be repeated for a social planner with zero cost of funds using the intersections of the WTP and MC curves.

Figure 1 also provides a visual test of the quasiconcavity assumption that is critical for the graphical solution to coincide with the solution to the original problem. Recall that if a price schedule is QC for a given consumer, the consumer only purchases a given coverage level increment so long as they have also purchased *every* lower coverage level increment. They will not "skip" any coverage level increment. A price schedule that is QC for all consumers will therefore have two properties: (i) incremental quantities $\tilde{q}^k$ will be decreasing in coverage level, and more specifically, (ii) the set of consumers that purchase at higher coverage levels will be a subset of those that purchase at lower coverage levels. Property (i) can be assessed visually in Figure 1. For example for the monopolist, the intersection between MR and MC occurs further and further to the right as one progresses from Panel (a)–(d) (i.e., as coverage level decreases).

For any price schedule that satisfies property (i), property (ii) will hold so long as the position of consumers on the demand curve *does not change too much* across different coverage level margins.[33] Violations of property (ii) can arise when different consumers' willingness to pay are driven by different things—for example, the value of risk protection versus an expected reduction in out-of-pocket spending—because the rate of change of these components in coverage level can be quite different. The same consumer may therefore be located high on the demand curve for one coverage level increment, but low on the demand curve for another. With multidimensional consumer types, this type of reordering is sure to happen to some extent, but the extent to which it happens is ultimately an empirical question. In practice, we find that violations of property (ii) are rare (c.f. Table 3).

6.4 Exclusion, Screening, and Comparative Statics

Consistent with the theoretical results in Sections 5, Table 3 shows that the monopolist provides *less coverage* than the social planner, uses *more contracts* in its optimal menu, and *excludes more* consumers from the market for incremental coverage. Under the monopoly allocation, 39 percent of consumers are excluded from the market, and consumers are screened across four contracts. In the

[33] It is not necessary for consumers' position on the demand curve to be exactly consistent across coverage level margins because a given price schedule will only be screening consumers across (at most) the number of fixed contracts available. Whole sections of the demand curve will therefore choose the same contract, and consumers that have moved position slightly within that section will not cause a violation of property (ii).

social planner's allocation (without an excess cost of funds), no consumers are excluded, and nearly all consumers are pooled at the Gold contract (see Marone and Sabety (2022) for a full discussion of this result). The finding that the utilitarian social planner has little incentive to attempt to screen consumers is also echoed in Bundorf et al. (2012) and Ho and Lee (2021).

When the planner faces an excess cost of public funds, it begins behaving somewhat more like the monopolist, in that it now places more weight on profits than on consumer surplus. We can therefore explore comparative statics with respect to the insurer's objective function by comparing outcomes between our two social planners facing different costs of public funds.[34] Consistent with the results in Section 5.4, as τ increases from 1 to 1.25, the planner begins to both screen and exclude, and the optimal amount of coverage provided decreases.

We note that all of these numerical results persist qualitatively in a world without moral hazard. In a population of consumers identical to our focal population, but without moral hazard ($\omega \approx 0$ for all consumers), the social planner of course optimally pools all consumers in full insurance. Relative to the planner, the monopolist again provides relatively less coverage, screens consumers across multiple contracts, and excludes some consumers from the market entirely. However, given that there are larger gains from trade from insurance in a world without moral hazard, the monopolist (like the social planner) increases the amount of coverage provided under its optimal menu.[35]

6.5 Welfare

Unsurprisingly, social welfare is lower under monopoly than under the social planner's solution. We now quantify these welfare differences, and investigate the impacts of various policy interventions.

Table 4 reports outcomes under a number of benchmark cases, under the optimal menus chosen by each of our three focal insurers, and under a set of policy interventions. In each case, the table reports welfare outcomes, spending outcomes, and the percentage of consumers enrolled in each contract under the relevant allocation. The welfare outcomes are average per-household per-year social surplus, consumer surplus, and producer surplus, each measured relative to the allocation of all consumers to the Catastrophic contract. The spending outcomes are average per-household per-year government spending, premiums, and expected out-of-pocket spending.

Panel A first shows our benchmark outcomes, which will serve as useful points of comparison going forward. The four benchmarks are (i) the first best allocation of consumers to contracts (which can be achieved only with type-specific pricing), (ii) the allocation of all consumers to the full insurance contract, (iii) the allocation of all consumers to the Catastrophic contract, and (iv) the perfectly competitive outcome. We implement the first three benchmarks as if the insurer were

[34] Recall that in our model, a social planner with a cost of public funds τ corresponds to welfare weights of $(1, \tau, \tau)$. We can therefore think of an increase in the cost of public funds in the same way as a decrease in the weight on consumer surplus.

[35] Results available upon request.

Table 4. Welfare Outcomes and Policy Simulations

	Scenario	Welfare outcomes *\$000 per household*			Spending outcomes *\$000 per household*			Allocations *Pct. of households*				
		SS†	*CS*†	*PS*	*Gov*	*Prem*	*OOP*	Cstr.	Brnz.	Slvr.	Gold	Full
Panel A. Benchmarks												
*	First best	1.86	1.86	–	–	9.75	1.85	<0.01	0.01	0.23	0.56	0.20
	Full insurance for all	1.74	1.74	–	11.90	–	–	–	–	–	–	1.00
	Minimum coverage for all	–	–	–	5.64	–	5.36	1.00	–	–	–	–
	Competitive equilibrium	1.05	1.05	–	5.64	1.31	4.18	0.06	0.77	0.16	<0.01	<0.01
Panel B. Optimal menus												
	Social planner	1.82	1.82	–	9.09	0.77	1.80	<0.01	–	<0.01	1.00	–
	Social planner, 25% ECPF	1.66	1.66	–	5.60	3.80	2.08	0.14	<0.01	0.13	0.72	0.01
	Monopolist	1.08	0.33	0.74	5.64	3.07	3.26	0.39	0.05	0.28	0.28	–
Panel C. Policy interventions												
(i)	Linear taxes/subsidies	0.74	0.64	0.10	5.15	2.70	3.44	0.58	0.04	–	0.38	<0.01
(ii)	Nonlinear subsidies	1.30	0.48	0.82	5.78	3.55	2.79	0.30	0.01	0.28	0.42	–
(iii)	Restrict which plans allowed	0.94	0.48	0.46	5.64	4.49	1.70	0.46	–	–	–	0.54
(iv)	Adjust base coverage	1.82	1.82	–	9.86	–	1.80	–	–	–	1.00	–

Notes: The table shows welfare outcomes, spending outcomes, and allocations under various scenarios. The first set of columns reports social surplus (*SS*), consumer surplus (*CS*), and producer surplus (*PS*) in thousands of dollars per household per year. Note that consumer welfare is normalized to zero at the Catastrophic contract, and accounts for the tax burden associated with government spending. The second set of columns reports expected government spending (*Gov*), premium spending (*Prem*), and expected out-of-pocket spending (*OOP*), again in thousands of dollars per household per year. The final set of columns reports the percentage of households allocated to each contract. †Relative to allocating all consumers to the Catastrophic contract when there is no excess cost of public funds (ECPF).

the social planner facing a zero excess cost of funds, meaning all surplus accrues to consumers.[36] Allocating all consumers to full insurance results in social surplus that is \$1,743 per household per year higher than allocating all consumers to the Catastrophic contract. This is only slightly lower than the \$1,860 attainable under the first best allocation. The competitive outcome features substantial unravelling, but still generates a large amount of social surplus as very few consumers are excluded. Panel B then reports outcomes at the optimal menus chosen by our three focal insurers. Social surplus under a social planner facing no excess cost of public funds is equal to \$1,825 per household per year. Under the monopolist, social surplus falls to \$1,081, and consumer surplus falls to \$330.

Interestingly, social surplus improves slightly under the monopolist relative to a perfectly competitive market. This result is consistent with Veiga and Weyl (2016), who suggest that there may be an interior optimum level of competition in insurance markets. And as suggested by Diamond (1992), a regulator of a competitive market may prefer to auction off the right to serve the market as a monopolist instead of permitting free entry and competition to unravel the available gains from trade. Of course, the monopolist captures the majority (69 percent) of the surplus it generates. Consumers are better off under competition.

[36] We calculate the competitive equilibrium proposed by Azevedo and Gottlieb (2017).

7 Policy Analysis

In this section, we show how theory and empirical analysis can guide policymakers. Adverse selection is the canonical rationale for public policy intervention in insurance markets. Many policies – including taxes and subsidies – aim to counteract the pricing distortion created by selection. In this section, we formally analyze such policies. We begin by considering the local impact of taxes and subsidies contract by contract. The degree to which the government will want to subsidize (or tax) insurance will depend on (a) the degree of adverse selection (as measured by slope of the marginal cost curve), (b) consumer responsiveness to price changes (as measured by the concavity of the demand curve), and (c) the social cost of public funds (as measured by weight w^G.) After deriving the key metric that incorporates these three factors, we apply our results to our numerical setting. We then compare pricing regulation to benefit regulation and consider combinations of non-local perturbations.

Focusing on the case of a monopolist insurer, we note that subsidies lower the insurer's marginal costs $MC^{I,k}$. Fix a single marginal coverage level of interest and suppress k. To be general, let the subsidy scheme be parameterized by $s \in [-1, 1]$, where $\sigma(q|s)$ is the subsidy when the monopolist serves q consumers. Assume that $\sigma(q|0) \equiv 0$, so that $s = 0$ corresponds to no subsidies, and that $\sigma_s(q|s) \geq 0$ and $\sigma_{qs}(q|s) \geq 0$, so that a higher s corresponds to higher subsidies both as an absolute and at the margin. An important example is a linear subsidy scheme σ given by $\sigma(q|s) = sq$. A linear subsidy (that is, constant across consumers) will lower the marginal cost curve by a constant. With a decreasing marginal cost curve, this will raise the quantity served.

Let

$$q(s) = \arg\max_q \left(P(q)\, q - C(q) + \sigma(q|s)\right)$$

be the optimal quantity served by the monopolist facing subsidy scheme s, where we assume that σ has enough regularity that q is well-defined and continuously differentiable. Further assume that the government's problem of choosing s is characterized by the first-order condition. Let *bang for first buck* (BFFB) be defined as $\beta_s(0)/\sigma_s(q^m|0)$, so that BFFB is the benefit the government realizes on the first dollar spent on subsidies. Comparing BFFB to w^G then tells us whether the government wishes to tax or subsidize. If $w^G <$ BFFB, then the government sets $s > 0$ and subsidizes insurance. If $w^G \geq$ BFFB, then the government sets $s \leq 0$ and taxes insurance.

In Online Appendix B.8, we show that

$$\text{BFFB} = w^C q^m \frac{\sigma_{qs}}{\sigma_s} \frac{1}{2 + \left(\frac{MC_q}{-P_q} + q^m \frac{P_{qq}}{P_q}\right)} + w^I. \tag{12}$$

The fraction σ_{qs}/σ_s reflects the amount by which the first dollar spent on subsidies lowers the effective marginal cost of the monopolist. For example, with a linear subsidy of s on all consumers served, $q^m \sigma_{qs}/\sigma_s = 1$. The second fraction reflects the degree to which the monopolist's quantity served responds to changes in the marginal cost. When marginal cost is constant and demand is linear, the fraction is 1/2. A decreasing marginal cost curve increases BFFB, but concave demand decreases BFFB.[37] In our setting, the (continuous) joint distribution of consumer types and the (non-linear) structure of contracts lead to both decreasing marginal cost and substantial demand concavity, as in Figure 4. As a result, the sign and magnitude of BFFB are an empirical question.

Armed with this theoretical insight, Panel C of Table 4 reports the impact of common strategies a regulator might use to intervene on behalf of consumers in a monopoly market. These policies are: (i) linear taxes or subsidies on contracts supplied by the monopolist; (ii) nonlinear subsidies which the monopolist receives only on additional consumers that obtain coverage; (iii) banning the monopolist from offering certain contracts, and (iv) raising the minimum level of coverage x^0. In each case, we assume the regulator aims to maximize consumer surplus, taking into account the tax burden associated with government spending available at zero excess cost of public funds.

We first consider policy (i), as suggested by the theoretical exercise. Here, the regulator can intervene by implementing linear taxes or subsidies on inside-option contracts. That is, the regulator announces a vector of taxes/subsidies that the monopolist will receive on each consumer enrolled in each contract (except Catastrophic), and the monopolist then chooses its optimal price schedule.[38] Substantial concavity of the demand curve reduces BFFB to surprisingly low levels with linear subsidies. As a result, we find that the regulator's best course of action is to actually tax the monopolist, thereby lowering government expenditures (and the consumers' tax burden) but also lowering the average level of coverage in the market. Consumer surplus increases by \$315 per household per year relative to the unregulated monopolist outcome. The welfare gains are primarily due to lower government spending, rather than increased coverage (and risk protection). The revenue raised by the taxes implemented reduces the consumer's tax burden by \$493.

The results of policy (i) may not reflect the political incentives of regulator intervention in insurance markets, namely, that a popular policy goal to raise coverage levels. We next consider a policy (ii) in which the regulator can *only* subsidize (and not tax) coverage, and moreover, is able to announce that these subsidies are only available to the monopolist for marginal consumers. That is, the vector of subsidies announced only applies to consumers served *beyond* the number of consumers served under the monopolist's unregulated optimal allocation. Theory suggests that such a non-linear subsidy scheme may be more effective at increasing consumer surplus. The non-linearity can substantially lower the cost of implementing subsidies, as they need not be paid on inframarginal consumers. Indeed, we find that in this case the regulator raises the levels of coverage

[37]If $MC_q = 0$, $P_{qq} = 0$, and $w^C = w^I = 1$, then BFFB = 1.5 for a linear subsidy. The logic of concavity driving BFFB is similar to pass-through of both cost shocks and taxes.

[38]Computationally, this is quite a difficult problem to solve on behalf of the regulator, as it must calculate the monopolist's best response to every potential vector of taxes/subsidies it might consider. We gain traction by using the logic of the simplified problem from Section 5. Namely, for every permutation of potentially allowable contracts, we solve for the optimal marginal tax/subsidy that the regulator would choose to implement on each marginal level of coverage, integrate these to form an optimal vector of taxes/subsidies, and then choose which permutation of allowable contracts maximizes the regulator's objective.

obtained in the market while increasing government spending by only \$143 per household per year. Interestingly, however, the consumer gains from this higher coverage are only \$298 per household per year. The gains do not substantially outweigh the tax increase due to higher government spending: total consumer surplus increases by only \$155 per household per year relative to the unregulated monopolist outcome.

We also consider alternative policies that affect the menu of contracts available to the insurer. Under policy (iii), the regulator chooses which of the four incremental levels of coverage to allow the monopolist to offer. Given strategic pricing by the monopolist, the regulator needs to trade off the welfare losses from additional exclusion against the welfare gains from more pooling. We find that the optimal policy is to ban the Bronze, Silver, and Gold contracts, and let the monopolist offer only the full insurance contract. While this strategy is effective in shifting some consumers to higher coverage, it also shifts some consumers to lower coverage. Relative to the monopolist's unconstrained solution, an additional 7 percent of the population is excluded from the market. As suggested by Geruso et al. (2019), there is a trade-off between exclusion and the generosity of coverage: consumer surplus increases by \$156 per household per year (due to more generous coverage on the intensive margin) relative to the unregulated monopoly market.

Finally, policy (iv) considers the case in which the regulator can raise the minimum level of coverage x^0, allowing it to shrink the size of the market served by the monopolist. This strategy will not always allow the regulator to reach the socially optimal allocation (since the optimal menu may involve screening), but it will prevent the primary problem of under-insurance in the market. In our setting, given that the social planner's solution was to pool all consumers at Gold, the regulator can fully restore maximal consumer and social surplus by raising the level of the outside option to the Gold contract. The monopolist cannot profitably offer the full insurance contract when consumers receive the Gold contract for free (for the same reason the full insurance contract was not part of the social planner's optimal allocation), and therefore effectively exits the market.

Finally, note first that maximal consumer and social surplus can be restored by simply offering the monopolist a take-it-or-leave-it offer to either serve the market at prices specified by the regulator in exchange for a lump sum payment, or else to not be able to serve the market. While this is indeed the optimal regulatory intervention, it may be difficult to implement.[39] Our results instead highlight the roles of both theory and empirical analysis in guiding policy. Conditional

[39]As it need not share rents with the monopolist, the regulator will implement the same price schedule as a social planner. The regulator could also achieve this outcome by implementing quantity targets. In particular, letting $\Pi^M \geq 0$ be the monopolist's profits absent regulatory intervention, and $\{\hat{q}^k\}_{k=1}^K$ be the regulator's chosen allocation, then the regulator can offer a payment $B = \Pi^M - \sum \hat{\Pi}^k(\hat{q}^k)$ if at least $\hat{q}^k$ is implemented for all k, and zero otherwise. Such schemes require, for example, that the market is well-defined. It also assumes the presence of an ex-ante competitive insurance market to which the regulator can auction the right to serve the market, or else that the regulator can bargain (holding all the bargaining power) with the monopolist over a contract to service the whole market. In practice, we consider a finite number of combinations of the simplified problem. As is well known, more efficient but still simple instruments exist: subsidizing quantity only beyond some threshold $\hat{q} < q^m$ leaves σ_{qs} alone, but lowers σ_s, with BFFB increasing without bound as $\hat{q}$ approaches q^m. Of course, our graphical analysis is most useful for analyzing local changes in the marginal cost curve. Consideration of large taxes and subsidies (or, at the extreme, banning some contracts) may require a more complex and computationally intense analysis.

on an exogenous minimum level of coverage and a restriction that the government cannot supply additional coverage itself, regulator intervention has a valuable role in the private market. Both theoretical and numerical analysis suggests that the most effective strategy for raising coverage when facing a strategic private insurer is a nonlinear subsidy scheme. This is in stark contrast to regulating a perfectly competitive market, in which a linear subsidy can restore the optimal feasible allocation (Azevedo and Gottlieb, 2017). When the regulator is able to supply incremental coverage itself, this is clearly a better course of action. That said, we note these results rely on the assumption that the monopolist does not have an informational or cost advantage in supplying coverage relative to the regulator.

8 Conclusion

We develop a principal-agent model with multidimensional private information that is both general and well-suited to understanding optimal menu design in health insurance markets. Our model encompasses the problems facing a government, a private entity developing an insurance program on behalf of its employees, and a monopolist insurance provider. We develop theoretical tools to describe the optimal menu of insurance contracts, and use those tools to study positive trade, optimal exclusion, and incentives to screen. We further develop a simplified version of the problem that, in our numerical setting, is a very close approximation to the actual problem at hand. The simplified problem is amenable to familiar graphical analysis, and allows us to derive comparative statics and further results on exclusion and screening. We build a numerical framework that lets us quantify the magnitudes of theoretically identified effects, and use the numerical tools to shed light on a variety of policy questions.

There are a number of directions for further work. First, the CARA assumption suppresses income effects, and there is good reason to believe that these effects are important, particularly at lower levels of coverage than we permit in our numerical analysis. Relaxing this assumption would require a large technical leap. Second, while the results establish an important benchmark to understand insurance markets with market power, and while the assumption of a single principal is appropriate when thinking about a single insurer such as a government or a monopolist, there are many interesting settings where the market is oligopolistic. We hope that the tools developed here form a basis for theoretical exploration of such markets. Finally, while the simplified version of the problem works extremely well in our specific setting, it is as yet unknown how generally this approach is valid. Further theoretical and numerical exploration is of prime interest.

References

Armstrong, M. (1996). Multiproduct nonlinear pricing. *Econometrica: Journal of the Econometric Society*, 51–75.

Armstrong, M. (2016). Nonlinear pricing. *Annual Review of Economics 8*, 583–614.

Arrow, K. J. (1965). Uncertainty and the welfare economics of medical care: Reply (the implications of transaction costs and adjustment lags). *The American Economic Review 55*(1/2), 154–158.

Azevedo, E. M. and D. Gottlieb (2017). Perfect competition in markets with adverse selection. *Econometrica 85*(1), 67–105.

Baron, D. P. and R. B. Myerson (1982). Regulating a monopolist with unknown costs. *Econometrica: Journal of the Econometric Society*, 911–930.

Billingsley, P. (1995). *Probability and Measure.* New York: Wiley.

Bundorf, M. K., J. Levin, and N. Mahoney (2012). Pricing and welfare in health plan choice. *American Economic Review 102*(7), 3214–3248.

Cardon, J. H. and I. Hendel (2001). Asymmetric Information in Health Insurance: Evidence from the National Medical Expenditure Survey. *RAND Journal of Economics*, 408–427.

Chade, H. and E. Schlee (2020). Insurance as a lemons market: Coverage denials and pooling. *Journal of Economic Theory 189.*

Chade, H. and J. Swinkels (2022). Disentangling moral hazard and adverse selection. Technical report, Working Paper.

Deneckere, R. and S. Severinov (2017). A solution to a class of multi-dimensional screening problems: Isoquants and clustering. *Tech. rep.*.

Diamond, P. (1992). Organizing the health insurance market. *Econometrica: Journal of the Econometric Society*, 1233–1254.

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

Einav, L., A. Finkelstein, S. P. Ryan, P. Schrimpf, and M. R. Cullen (2013). Selection on moral hazard in health insurance. *American Economic Review 103*(1), 178–219.

Einav, L., A. N. Finkelstein, and M. Cullen (2010). Estimating Welfare in Insurance Markets Using Variation in Prices. *The Quarterly Journal of Economics CXV*(3).

Fudenberg, D. and J. Tirole (1991). *Game Theory.* MIT press.

Geruso, M., T. J. Layton, G. McCormack, and M. Shepard (2019, September). The two margin problem in insurance markets. Working Paper 26288, National Bureau of Economic Research.

Handel, B., I. Hendel, and M. D. Whinston (2015). Equilibria in health exchanges: Adverse selection versus reclassification risk. *Econometrica 83*(4), 1261–1313.

Handel, B. and K. Ho (2021). The industrial organization of health care markets. In *Handbook of Industrial Organization*, Volume 5, pp. 521–614. Elsevier.

Hendren, N. (2013). Private information and insurance rejections. *Econometrica 81*(5), 1713–1762.

Ho, K. and R. Lee (2021). Health insurance menu design for large employers. Working Paper 27868, National Bureau of Economic Research.

Kadan, O., P. J. Reny, and J. M. Swinkels (2017). Existence of optimal mechanisms in principal-agent problems. *Econometrica 85*(3), 769–823.

Laffont, J.-J., E. Maskin, and J.-C. Rochet (1987). Optimal nonlinear pricing with two-dimensional characteristics. *Information, Incentives and Economic Mechanisms*, 256–266.

Mahoney, N. and E. G. Weyl (2017). Imperfect competition in selection markets. *Review of Economics and Statistics 99*(4), 637–651.

Manelli, A. M. and D. R. Vincent (2006). Bundling as an optimal selling mechanism for a multiple-good monopolist. *Journal of Economic Theory 127*(1), 1–35.

Manning, W. G., J. P. Newhouse, N. Duan, E. B. Keeler, and A. Leibowitz (1987). Health insurance and the demand for medical care: Evidence from a randomized experiment. *The American Economic Review 77*(3), 251–277.

Marone, V. R. and A. Sabety (2022, January). When should there be vertical choice in health insurance markets? *American Economic Review 112*(1), 304–42.

Maskin, E. and J. Riley (1984). Monopoly with incomplete information. *The RAND Journal of Economics 15*(2), 171–196.

Mussa, M. and S. Rosen (1978). Monopoly and product quality. *Journal of Economic Theory 18*(2), 301–317.

Pauly, M. V. (1968). The economics of moral hazard: Comment. *American Economic Review 58*(3), 531–537.

Rochet, J.-C. (1985). The taxation principle and multi-time hamilton-jacobi equations. *Journal of Mathematical Economics 14*(2), 113–128.

Rochet, J.-C. and P. Choné (1998). Ironing, sweeping, and multidimensional screening. *Econometrica*, 783–826.

Rochet, J.-C. and L. A. Stole (2003). The economics of multidimensional screening. *Econometric Society Monographs 35*, 150–197.

Shannon, C. (1995). Weak and strong monotone comparative statics. *Economic Theory 5*, 209–227.

Stiglitz, J. (1977). Monopoly, non-linear pricing and imperfect information: The insurance market. *Review of Economic Studies 44*(3), 407–430.

Veiga, A. and E. G. Weyl (2016). Product design in selection markets. *The Quarterly Journal of Economics 131*(2), 1007–1056.

Weyl, E. G. and A. Veiga (2017). Pricing institutions and the welfare cost of adverse selection. *American Economic Journal: Microeconomics 9*(2), 139–48.

Wilson, R. B. (1993). *Nonlinear Pricing.* Oxford University Press on Demand.

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

Appendices

Appendix A Theory

The following lemma will be used repeatedly. Its proof is in Online Appendix B.2.

Lemma 1 *The best-response correspondence $X(\rho, \theta)$ is upper hemicontinuous in ρ and θ. The consumer's value function $V(\rho, \theta) \equiv \max_{x \in [0,1]}(v(x, \theta) - \rho(x))$ is continuous.*[40]

A.1 Proof of Proposition 1

In each case it suffices to prove the first assertion since the second follows from a standard monotone comparative statics result.

(*i*) Recall that $v_x = -\int c_x m dl$, where since $-c_x(a^*(\cdot, x, \omega), x)$ is increasing, it is sufficient to show that m satisfies strict MLRP in (l, ψ). But, from (4), for any x and ω, and for any $\psi_h > \psi_l$,

$$\frac{m\,(l|x, \psi_h)}{m\,(l|x, \psi_\ell)} = e^{(\psi_h - \psi_\ell)(-z(l,x,\omega))} \frac{\int e^{-\psi_\ell z(l',x,\omega)} f(l')dl'}{\int e^{-\psi_h z(l',x,\omega)} f(l')dl'},$$

and so, by the definition of MLRP, it would be sufficient to show that $z(\cdot, x, \omega)$ is strictly decreasing. But, from (1), and using the Envelope Theorem to ignore the effects on z via a, we have that $z_l(l, x, \omega) = b_l(l, a^*(l, x, \omega), \omega) < 0$.

(*ii*) It is sufficient to show that m satisfies strict MLRP in (l, τ). But, for $\tau_h > \tau_\ell$ we have

$$\frac{m(l|x, \tau_h)}{m(l|x, \tau_\ell)} = \frac{f(l|\tau_h)}{f(l|\tau_\ell)} \frac{\int e^{-\psi z(l',x,\omega)} f(l'|\tau_\ell)dl'}{\int e^{-\psi z(l',x,\omega)} f(l'|\tau_h)dl'},$$

since $\{f(\cdot|\tau)\}_{\tau \in [0,1]}$ is ordered by strict MLRP.

(*iii*) It is easy to show that $v_\omega = \int b_\omega m dl$ and thus $v_{x\omega} > 0$ if and only if

$$\int b_{\omega a} a_x^* m dl + \int b_\omega m_x dl > 0. \tag{13}$$

The first term is always strictly positive since $b_{a\omega} > 0$ and $a_x^* > 0$. The second term can be written as $\int b_\omega (m_x/m) m dl$. Now, differentiating (4) with respect to x yields $m_x/m = \psi(\mathbb{E}_m[z_x] - z_x)$, and since $z_x = -c_x$ and $(-c_x)_l = -c_{xa} a_l^* > 0$, it follows that m_x/m is strictly decreasing in l, single-crosses zero from above, and integrates to zero. But, when $b = \hat{b}$ and c is linear in a, $(b_\omega)_l = 0$. Hence, the second term in (13) is 0. □

[40] To see why it is not an immediate consequence of the Theorem of the Maximum, note that $v(x, \theta) - \rho(x)$ is not continuous in ρ.

A.2 Proof of Theorem 1

Recall that by Proposition 1, Part (i), for any given (ω, F) and ρ, $X(\rho, \omega, \cdot, F)$ is single-valued $G(\cdot|\omega, F)$-almost everywhere. Therefore, we can without ambiguity take any selection $\upsilon(\cdot)$ from $X(\rho, \cdot, \omega, F)$ and write

$$\Pi(\rho, \omega, F) = \int S(\rho(\upsilon(\psi)), \upsilon(\psi), \psi, \omega, F) dG(\psi|\omega, F).$$

as the expected payoff to the insurer from premium schedule ρ given (ω, F), so that the designer's problem is simply to maximize $\int \Pi(\rho, \omega, F) dG(\omega, F)$ by choice of ρ subject to $\rho(x^0) = 0$.

Let $x = x^k$, and let ρ^ε be the premium schedule in which $\rho^\varepsilon(x') = \rho(x')$ for $x' \leq x$, and $\rho^\varepsilon(x') = \rho(x') + \varepsilon$ for $x' > x$. Let

$$\Delta(\omega, \psi, F) = \max_{k' > k}(v(x^{k'}, \omega, \psi, F) - \rho(x^{k'})) - \max_{k' \leq k}(v(x^{k'}, \omega, \psi, F) - \rho(x^{k'})),$$

noting that Δ is strictly increasing in ψ and when F moves in an *MLRP* direction by Proposition 1 Parts (i) and (ii). Thus in particular, Δ strictly increases in the second coordinate of θ, and so since $\tilde{G}$ has a density $\tilde{g}$, it follows that it is either $\Delta(\omega, \underline{\psi}, F) = 0$ or $\Delta(\omega, \bar{\psi}, F) = 0$ only for a zero G-measure set of (ω, F). Fix (ω, F) such that *2BRP* holds and neither $\Delta(\omega, \underline{\psi}, F) = 0$ nor $\Delta(\omega, \bar{\psi}, F) = 0$, and suppress (x, ω, F) in what follows. Let us define $\hat{\psi}$ as the type dividing those who choose strictly above x facing ρ, and, in a small abuse of notation, write $\hat{\psi}(\varepsilon)$ as the dividing type facing ρ^ε. To formalize this, if $\Delta(\omega, \underline{\psi}, F) > 0$, so that even $\underline{\psi}$ strictly prefers an action strictly above x to any action at or below x, then $\hat{\psi} = \underline{\psi}$ and for ε small, $\hat{\psi}(\varepsilon) = \underline{\psi}$ as well. If $\Delta(\omega, \bar{\psi}, F) < 0$, so that even $\bar{\psi}$ strictly prefers an action below x to one strictly above x, then $\hat{\psi} = \bar{\psi}$ and for ε small, $\hat{\psi}(\varepsilon) = \bar{\psi}$ as well. Finally, if $\Delta(\omega, \underline{\psi}, F) < 0 < \Delta(\omega, \bar{\psi}, F)$, then $\hat{\psi}$ is given by $\Delta(\omega, \hat{\psi}, F) = 0$, and $\hat{\psi}$ is given by $\Delta(\omega, \hat{\psi}(\varepsilon), F) = \varepsilon$. Let $\bar{x} \equiv \bar{x}(\hat{\psi}, \rho)$ and $\underline{x} \equiv \underline{x}(\hat{\psi}, \rho)$.[41] By *2BRP*, these are the only best responses for $\hat{\psi}$ facing ρ. Thus, by upper hemicontinuity of the best response correspondence X in ρ, for small ε no type near $\hat{\psi}$ will choose anything other than $\bar{x}$ or $\underline{x}$ facing ρ^ε. Hence for ε small and positive, types between $\hat{\psi}$ and $\hat{\psi}(\varepsilon)$ will switch from $\bar{x}$ to $\underline{x}$ and for ε small and negative, types between $\hat{\psi}(\varepsilon)$ and $\hat{\psi}$ switch from $\underline{x}$ to $\bar{x}$, while other types will maintain their previous behavior.

Where $\hat{\psi}$ is interior, the defining condition for $\hat{\psi}(\varepsilon)$ for ε small (positive or negative) is thus

$$v(\underline{x}, \hat{\psi}(\varepsilon)) - \rho(\underline{x}) = v(\bar{x}, \hat{\psi}(\varepsilon)) - \rho(\bar{x}) - \varepsilon,$$

[41] Note that $\bar{x}$ or $\underline{x}$ need not be adjacent to x; it may be that the optimal choice jumps past multiple quality levels as ψ passes through $\hat{\psi}$.

and so by the Implicit Function Theorem,

$$\hat{\psi}_\varepsilon(\varepsilon) = \frac{1}{v_\psi(\bar{x}, \hat{\psi}(\varepsilon)) - v_\psi(\underline{x}, \hat{\psi}(\varepsilon))},$$

where since $\hat{\psi}(0) = \hat{\psi}$, we have $\hat{\psi}_\varepsilon(0) = 1/(v_\psi(\bar{x}, \hat{\psi}) - v_\psi(\underline{x}, \hat{\psi})) \in (0, \infty)$, using $v_{\psi x} > 0$. Of course, if $\Delta(\omega, \underline{\psi}, F) > 0$ or $\Delta(\omega, \bar{\psi}, F) < 0$, then $\hat{\psi}_\varepsilon(0) = 0$.

Now,

$$\Pi(\rho^\varepsilon) - \Pi(\rho) = (w^I - w^C)(1 - G(\hat{\psi}(\varepsilon)))\varepsilon + \int_{\hat{\psi}}^{\hat{\psi}(\varepsilon)} (S(\rho(\underline{x}), \underline{x}, \psi) - S(\rho(\bar{x}), \bar{x}, \psi))dG(\psi),$$

where this expression also makes sense when $\varepsilon < 0$ under the usual convention that $\int_a^b = -\int_b^a$ when $a > b$. Thus,

$$\Pi_\varepsilon(\rho^\varepsilon) = (w^I - w^C)(-g(\hat{\psi}(\varepsilon))\hat{\psi}_\varepsilon(\varepsilon)\varepsilon + (1 - G(\hat{\psi}(\varepsilon)))) + (S(\rho(\underline{x}), \underline{x}, \hat{\psi}(\varepsilon)) - S(\rho(\bar{x}), \bar{x}, \hat{\psi}(\varepsilon)))g(\hat{\psi}(\varepsilon))\hat{\psi}_\varepsilon(\varepsilon)$$

But, recall that $S(p, x, \theta) = \mathcal{S}(x, \theta) - (w^I - w^C)(v(x, \theta) - p)$, and so,

$$\begin{aligned} & S(\rho(\underline{x}), \underline{x}, \hat{\psi}(\varepsilon)) - S(\rho(\bar{x}), \bar{x}, \hat{\psi}(\varepsilon)) \\ = \; & \mathcal{S}(\underline{x}, \hat{\psi}(\varepsilon)) - \mathcal{S}(\bar{x}, \hat{\psi}(\varepsilon)) - (w^I - w^C)(v(\underline{x}, \hat{\psi}(\varepsilon)) - \rho(\underline{x}) - (v(\bar{x}, \hat{\psi}(\varepsilon)) - \rho(\bar{x}))) \\ = \; & \mathcal{S}(\underline{x}, \hat{\psi}(\varepsilon)) - \mathcal{S}(\bar{x}, \hat{\psi}(\varepsilon)) + (w^I - w^C)\varepsilon, \end{aligned}$$

where the second equality follows from the defining equation for $\hat{\psi}(\varepsilon)$.

But then, substituting and taking a limit, where $\hat{\psi}$ is interior,

$$\begin{aligned} \Pi_\varepsilon(\rho^\varepsilon)|_{\varepsilon=0} &= (w^I - w^C)(1 - G(\hat{\psi})) - \frac{\mathcal{S}(\bar{x}, \hat{\psi}) - \mathcal{S}(\underline{x}, \hat{\psi})}{v_\psi(\bar{x}, \hat{\psi}) - v_\psi(\underline{x}, \hat{\psi})} g(\hat{\psi}) \\ &= (w^I - w^C)(1 - G(\hat{\psi})) - rg(\hat{\psi}) \end{aligned}$$

and so, reinstating (x, ω, F), we have that $\Pi_\varepsilon(\rho^\varepsilon, \omega, F)|_{\varepsilon=0} = \mathcal{V}(x, \omega, F)$. Recall also that when $\Delta(\omega, \underline{\psi}, F) > 0$ or $\Delta(\omega, \bar{\psi}, F) < 0$, then $\hat{\psi}_\varepsilon(0) = 0$ and so, since we defined $r = 0$ in this case, we once again have $\Pi_\varepsilon(\rho^\varepsilon)|_{\varepsilon=0} = (w^I - w^C)(1 - G(\hat{\psi})) - rg(\hat{\psi})$.

Finally, note from the previous displayed equation that $(\Pi(\rho^\varepsilon, \omega, F))_\varepsilon|_{\varepsilon=0}$ is uniformly bounded as we vary (ω, F). In particular, by Cauchy's Mean Value Theorem (CMVT), when $\hat{\psi}$ is interior, r is of the form $\mathcal{S}_x/v_{x\psi}$ for some $x \in (\underline{x}, \bar{x})$ and so is uniformly bounded. Thus, by Lebesgue's Dominated Convergence Theorem (LDCT),

$$\left(\int \Pi(\rho^\varepsilon, \omega, F)dG(\omega, F)\right)_\varepsilon\bigg|_{\varepsilon=0} = \int \mathcal{V}(x, \omega, F)dG(\omega, F).$$

and we are done, noting that the perturbation with $\varepsilon > 0$ is always feasible, while the perturbation with $\varepsilon < 0$ is feasible as long as $\rho(x^k) < \rho(x^{k+1})$. $\square$

A.3 Endogenizing Quality: Another Optimality Condition

We now derive an additional necessary condition that must hold if the insurer can also vary the coverage levels of the contracts offered, in addition to their prices. This second condition becomes relevant when the insurer is constrained in the *number* of contracts it can offer, but can choose both their price and their generosity.

Consider the perturbation in which the insurer just raises (or reduces) the generosity of a single contract, x^k, replacing x^k by $x^k + \varepsilon$. Fix (ω, F), and assume x is chosen by ψ in some positive-lengthed interval (ψ^l, ψ^h). There are three effects. First, consumers who stick with x generate a different amount of surplus than they did before, changing the insurer's payoff by $\int_{\psi^l}^{\psi^h} (\mathcal{S}_x - (w^I - w^C)v_x)dG(\psi)$.

Second, some types immediately below ψ^l now choose the new contract $x + \varepsilon$ instead of their previous choice, which was $\underline{x}^l \equiv \underline{x}(\psi^l)$. This has value $v_x(x, \psi^l)r^l$ to the insurer, where if ψ^l is interior, we define

$$r^l = \frac{\mathcal{S}(x, \psi^l) - \mathcal{S}(\underline{x}^l, \psi^l)}{v_\psi(x, \psi^l) - v_\psi(\underline{x}^l, \psi^l)}, \tag{14}$$

while if $\psi^l = 0$, we take $r^l = 0$. In this expression, $\mathcal{S}(x, \psi^l) - \mathcal{S}(\underline{x}^l, \psi^l)$ reflects the change in the insurer's payoff when the agent switches from $\underline{x}^l$ to $x + \varepsilon$, with the utility of the switching consumer type disappearing from the calculation because they are by definition indifferent. We will show that the v_x term and denominator of r^l capture the speed at which the boundary between those who switch and those who do not is moving.

Third, some types immediately above ψ^h will switch their choice down from $\bar{x}^h \equiv \bar{x}(\psi^h)$ to $x + \varepsilon$, with net effect $-v_x(x, \psi^h)r^h$, where

$$r^h = \frac{\mathcal{S}(\bar{x}^h, \psi^h) - \mathcal{S}(x, \psi^h)}{v_\psi(\bar{x}^h, \psi^h) - v_\psi(x, \psi^h)}. \tag{15}$$

if ψ^h is interior, and zero otherwise. Reintroducing the dependence of the various objects on (x, ω, F), the overall impact of the perturbation on the insurer's payoff is is

$$\begin{aligned}
\mathcal{W}(x, \omega, F) \quad \equiv \quad & -v_x(x, \psi^h(x, \omega, F))r^h(x, \omega, F)g(\psi^h(x, \omega, F)|\omega, F) \\
& + \int_{\psi^l(x,\omega,F)}^{\psi^h(x,\omega,F)} (\mathcal{S}_x(x, \theta) - (w^I - w^C)v_x(x, \theta))g(\psi|\omega, F)d\psi \\
& + v_x(x, \psi^l(x, \omega, F))r^l(x, \omega, F)g(\psi^l(x, \omega, F)|\omega, F),
\end{aligned}$$

where if for given (ω, F), x is never chosen, then we take $\mathcal{W}(x, \omega, F) = 0$.

We can now state the optimality condition associated with this perturbation. The proof is in Online Appendix B.4.

Theorem 3 (Second Optimality Condition: Fixed Number of Contracts) *Let (ρ, χ) be optimal given $\{x^k\}_{k=0}^K$, and let ρ satisfy 2BRP. Then, $\int \mathcal{W}(x^k, \omega, F) dG(\omega, F) = 0$ for $k = 1, ..., K$.*

A.4 Optimality in the Continuum

In this section, we state and prove the analogs to Theorems 1 and 3 in the continuum. The proof is in Online Appendix B.5.

Theorem 4 (Optimality Conditions: Continuum of Contracts) *Let (ρ, χ) be optimal given $\mathcal{P}$, and let ρ satisfy 2BRP. Then, we have $\int \mathcal{W}(x, \omega, F) dG(\omega, F) = 0$ for all x, and $\int \mathcal{V}(x, \omega, F) dG(\omega, F) \leq 0$ except in a countable subset of $[x^0, 1]$ with equality if $\rho(x') > \rho(x)$ for $x' > x$.*

Technical Remark 6 (Main Perturbation in Continuum Case) To see why $\int \mathcal{V} dG = 0$ need not hold for *all* x in Theorem 4, assume that for a given (ω, F) there is ψ^J where $\underline{x}(\psi^J) < x = \bar{x}(\psi^J)$. If one *raises* the premium of all contracts strictly above x, types just to the right of ψ^J will shift their choice from a little above $\bar{x}(\psi^J)$ down to x, while if one *lowers* the premium of all contracts strictly above x, types just to the left of ψ^J will shift their choice from near $\underline{x}(\psi^J)$ to near $\bar{x}(\psi^J)$. The appropriate expression for r (see (24) in Online Appendix B.5) thus differs in the two cases, and if there is a positive-measure set of types having a jump ending at x, there can be a difference between the left- and right-hand derivatives of payoffs with respect to the perturbation. At the cost of significant extra notation, one can explicitly tie down these derivatives, but the additional economic insight is small, especially given that this issue can only occur for a countable set of x's.

A.5 Proof of Proposition 2

It suffices to show that strictly positive profit menus exist. Fix $\hat{\psi} \in (0, \bar{\psi})$. Consider the menu with a single item $x > x^0$ priced at

$$p(x) = v(x, \hat{\psi}, \bar{\omega}, \bar{F}) - v(x^0, \hat{\psi}, \bar{\omega}, \bar{F}).$$

This is accepted by all types in a neighborhood of $(\bar{\psi}, \bar{\omega}, \bar{F})$. Using that $c(a^*, x) - c(a^*, x^0)$ is increasing in l, the cost of serving each customer is no more than $\int (c(a^*(l, x, \bar{\omega}), x) - c(a^*(l, x, \bar{\omega}), x^0)) d\bar{F}$. So, the profit per customer is at least

$$J(x) \equiv v(x, \hat{\psi}, \bar{\omega}, \bar{F}) - v(x^0, \hat{\psi}, \bar{\omega}, \bar{F}) - \int c(a^*(l, x, \bar{\omega}), x) - c(a^*(l, x, \bar{\omega}), x^0)) d\bar{F}.$$

Trivially, $J(x^0) = 0$. But

$$J_x(x) = v_x(x, \hat{\psi}, \bar{\omega}, \bar{F}) - \int (-c_x(a^*(l, x, \bar{\omega}), x))) d\bar{F} - \int (c_a(a^*(l, x, \bar{\omega}), x^0) - c_a(a^*(l, x, \bar{\omega}), x))) a_x^*(l, x, \bar{\omega}) d\bar{F},$$

and so,

$$J_x(x^0) = v_x(x^0, \hat{\psi}, \bar{\omega}, \bar{F}) - \int (-c_x(a^*(l, x^0, \bar{\omega}), 0))) \bar{f}(l) dl.$$

We would thus be done if $v_x(x^0, \hat{\psi}, \bar{\omega}, \bar{F}) > \int (-c_x(a^*(l, x^0, \bar{\omega}), x^0))) d\bar{F}$, since then, $J(x) > 0$ for x just to the right of x^0. But, from (5)

$$v_x(x^0, \hat{\psi}, \bar{\omega}, \bar{F}) = \int (-c_x(a^*(l, x^0, \bar{\omega}), x^0))) m(l | x^0, \bar{\omega}, \hat{\psi}, \bar{F}) dl,$$

where $m(\cdot | x^0, \bar{\omega}, \hat{\psi}, \bar{F})$ strictly MLRP dominates $\bar{f}$. Hence, $-c_x(a^*(l, x^0, \bar{\omega}), x^0))$ is a strictly increasing function of l. □

A.6 Proof of Theorem 2

The following lemma tells us that for any given closed set $\mathcal{P}^0 \subseteq \mathcal{P}$, if we take a sequence $\mathcal{P}^n$ of increasingly fine approximation to $\mathcal{P}^0$ then anything the insurer can do in $\mathcal{P}^0$ can come arbitrarily close what can be done in $\mathcal{P}^n$.

Lemma 2 *The insurer's payoff* $\Pi(\rho)$ *is continuous in* ρ.

Proof We assert first that the set of θ where $X(\rho, \theta)$ is singleton valued has full G-measure. To see this, note that by Proposition 1 (i), for each (ω, F) the function v is strictly supermodular in x and ψ, and so for each pair ψ'' and ψ' with $\psi'' > \psi'$, the smallest best response at ψ'' is at least as large as the largest best response at ψ', or formally,

$$\inf X(\rho, \psi'', \omega, F) \geq \sup X(\rho, \psi', \omega, F).$$

But then, for each (ω, F), there is a countable set of values of ψ such that $X(\rho, \cdot, \omega, F)$ is unique except on this set (see Shannon, 1995). Since the distribution over ψ conditional on (ω, F) is atomless, it follows that with probability one conditional on (ω, F), $X(\rho, \cdot, \omega, F)$ is unique. Since (ω, F) was arbitrary, we are done.

Fix $\grave{\rho}$ and $\grave{\rho}^n \to \grave{\rho}$, and fix any measurable selection $\grave{\chi}(\cdot)$ from $X(\grave{\rho}, \cdot)$ and $\grave{\chi}^n(\cdot)$ from $X^n(\grave{\rho}, \cdot)$, so that

$$\Pi(\grave{\rho}^n) = \int S(\grave{\rho}^n(\grave{\chi}^n(\theta)), \grave{\chi}^n(\theta), \theta) dG(\theta),$$

and similarly for $\Pi(\grave{\rho})$. Let θ be any type for whom $X(\grave{\rho}, \theta)$ has unique element $\grave{x}$. Then, $\grave{\chi}^n(\theta) \to \grave{\chi}(\theta)$ by Lemma 1. But, also from Lemma 1, $V(\grave{\rho}^n, \theta) = v(\grave{\chi}^n(\theta), \theta) - \grave{\rho}^n(\grave{\chi}^n(\theta))$ converges to $V(\grave{\rho}, \theta)$, and since v is continuous, $v(\grave{\chi}^n(\theta), \theta)$ converges to $v(\grave{\chi}(\theta), \theta)$. But then, it follows that

$\grave{\rho}^n(\grave{\chi}^n(\theta)) \to \grave{\rho}(\grave{\chi}(\theta))$. Hence, since S is continuous, $S(\grave{\rho}^n(\grave{\chi}^n(\theta)), \grave{\chi}^n(\theta), \theta) \to S(\grave{\rho}(\grave{\chi}(\theta)), \grave{\chi}(\theta), \theta)$. But then, by *LDCT*, since S is bounded, and since the set of θ where $X(\rho, \theta)$ is singleton valued has full G-measure, $\Pi(\grave{\rho}^n) \to \Pi(\grave{\rho})$, and we are done. $\square$

Proof of Theorem 2 Immediate from Lemmas 1 and 2. $\square$

Theorem 2 provides an upper hemicontinuity result for the set of optimal solutions in our problem as the set of allowable premium schedules is varied. A natural question is whether the set of *optimal* solutions is lower hemi-continuous as well. Unfortunately, this is not true.

Example 1 *Assume that the insurer has exactly two distinct optima ρ^* and ρ^{**} in $\mathcal{P}^0$ and that $\mathcal{P}^n$ consists of some growing set of premium schedules where ρ^* is always an element of $\mathcal{P}^n$, but ρ^{**} is not. Then, the insurer has unique solution ρ^* in each approximation. If a regulator prefers ρ^{**} to ρ^*, then the regulator is strictly harmed by the restriction to $\mathcal{P}^n$, no matter how large is n.*

In this example, the insurer has two optima that imply different things in terms of, for example, the payoff to the consumer. If the insurer has only *one* optimum, then everything must converge. Our strong intuition is that only for very unusual examples will there be more than one optimum in $\mathcal{P}^0$. The intuition is that while it is not unusual that the payoff to the insurer has multiple peaks as one runs over $\mathcal{P}^0$, it would be surprising if for any given specification of G two of those peaks had exactly the same height. A proof eludes us.

A.7 Equation (11) and the Analog of $\int \mathcal{V} dG = 0$

Let $\tilde{\psi}^k(p^k, \omega, F) = \arg\min_\psi \left|v^k(\omega, \psi, F) - p^k\right|$. Because $v_{x\psi} > 0$, $\tilde{\psi}^k(p^k, \omega, F)$ is unique, and any given type has marginal willingness to pay for x^k greater than p^k if and only if ψ is above $\tilde{\psi}^k(p^k, \omega, F)$.[42] The following lemma uses this property to characterize $\tilde{\Pi}^k_{p^k}$.

Lemma 3 *We have*

$$\tilde{\Pi}^k_{p^k}(p^k) = \int \tilde{\mathcal{V}}^k(p^k, \omega, F) dG(\omega, F),$$

where

$$\begin{aligned}(16) \quad \tilde{\mathcal{V}}^k(p^k, \omega, F) = &(w^I - w^C)(1 - G(\tilde{\psi}^k(p^k, \omega, F)|\omega, F)) \\ &- \tilde{\psi}^k_{p^k}(p^k, \omega, F)(w^I(p^k - \gamma^{I,k}(\theta)) - (w^G - w^I)\gamma^{G,k}(x^0, \theta))g(\tilde{\psi}^k(p^k, \omega, F)|\omega, F).\end{aligned}$$

Note that $\tilde{\mathcal{V}}$ is the direct analog to $\mathcal{V}$ in this setting, since where $\tilde{\psi}^k_{p^k} \neq 0$,

$$\tilde{\psi}^k_{p^k}(p^k, \omega, F) = \frac{1}{v^k_\psi(\omega, \tilde{\psi}^k(p^k, \omega, F), F)} = \frac{1}{v_\psi(x^k, \omega, \tilde{\psi}^k(p^k, \omega, F), F) - v_\psi(x^{k-1}, \omega, \tilde{\psi}^k(p^k, \omega, F), F)}.$$

[42] Since $v_{x\psi} > 0$, $v^k = v(x^k, \omega, \psi, F) - v(x^{k-1}, \omega, \psi, F) = \int_{x^{k-1}}^{x^k} v_x(x, \omega, \psi, F) dx$ is strictly increasing in ψ.

Proof of Lemma 3 We have that

$$\left\{\theta | v^k(\theta) \geq p^k\right\} = \left\{(\omega, \psi, F) | \psi \geq \tilde{\psi}^k(p, \omega, F)\right\}$$

and so,

$$\tilde{\Pi}^k(p^k) = \int \int_{\tilde{\psi}^k(p^k,\omega,F)}^{\bar{\psi}} S^k(p^k, \theta) g(\psi|\omega, F) d\psi dG(\omega, F).$$

But then,

$$\tilde{\Pi}^k_{p^k}(p^k) = \int \left(\begin{array}{c} \int_{\tilde{\psi}^k(p^k,\omega,F)}^{\bar{\psi}} S^k_{p^k}(p^k, \theta) g(\psi|\omega, F) d\psi \\ -\tilde{\psi}^k_{p^k}(p^k, \omega, F) S^k(p^k, \omega, \tilde{\psi}^k(p^k, \omega, F), F) g(\tilde{\psi}^k(p^k, \omega, F)|\omega, F) \end{array} \right) dG(\omega, F).$$

But, $S^k_{p^k}(p^k, \theta) = w^I - w^C$, and if $\tilde{\psi}^k_{p^k}(p^k, \omega, F) \neq 0$, then $v^k(\omega, \tilde{\psi}^k(p^k, \omega, F), F) = p^k$ and so, evaluated at $\theta = (\omega, \tilde{\psi}^k(p^k, \omega, F), F)$,

$$S^k(p^k, \theta) = w^I(p^k - \gamma^{I,k}(\theta)) - (w^G - w^I)\gamma^{G,k}(x^0, \theta),$$

and the claimed expression follows. □

A.8 Proof of Proposition 3

Let $\underline{v}^k = \min_{\omega \in [\underline{\omega}, \bar{\omega}]} v^k(\omega, \underline{\psi}, \underline{F})$, noting that by Lemma 1, $v^k(\theta) \geq \underline{v}^k$ for all $\theta \in supp\, G$, with strict equality except on the $G(\omega, F)$-zero-measure set where $F = \underline{F}$. But, whenever $p^k < v^k(\omega, \underline{\psi}, F)$, $\tilde{\psi}^k(p^k, \omega, F) = \underline{\psi}$, and so, $\lim_{p^k \downarrow \underline{v}^k} \tilde{\psi}^k(p^k, \omega, F) = \underline{\psi}$ and $\lim_{p^k \downarrow \underline{v}^k} \tilde{\psi}^k_{p^k}(p^k, \omega, F) = 0$, and thus $\lim_{p^k \downarrow \underline{v}^k} \tilde{\mathcal{V}}^k(p^k, \omega, F) = w^I - w^C$. Further $\tilde{\mathcal{V}}^k(p^k, \omega, F)$ is bounded on the compact set $[\underline{v}^k, \underline{v}^k + 1] \times [\underline{\omega}, \bar{\omega}] \times \mathcal{F}$, since all components of it are uniformly bounded ($\tilde{\psi}^k_p$ in particular is either equal to 0 or to $1/((x^k - x^{k-1}) v_{x\psi}(x, \omega, \tilde{\psi}^k, F))$ for some $x \in [x^{k-1}, x^k]$). But then by the Lebesgue Dominated Convergence Theorem,

$$\lim_{p^k \downarrow \underline{v}^k} \int \tilde{\mathcal{V}}^k(p^k, \omega, F) dG(\omega, F) = \int \lim_{p^k \downarrow \underline{v}^k} \tilde{\mathcal{V}}^k(p^k, \omega, F) dG(\omega, F) = w^I - w^C > 0,$$

and so it follows that $\tilde{p}^k > \underline{v}^k$. Since $supp\, G$ is a rectangle, it follows that $\tilde{p}^k > v^k$ for a positive G-measure set of consumers, since this set includes a neighborhood of $(\tilde{\omega}, \underline{\psi}, \underline{F})$ for any $\tilde{\omega} \in \arg\min_{\omega \in [\underline{\omega}, \bar{\omega}]} v^k(\omega, \underline{\psi}, \underline{F})$. □

Appendix B Online Appendix

B.1 Computational Details and Additional Numerical Results

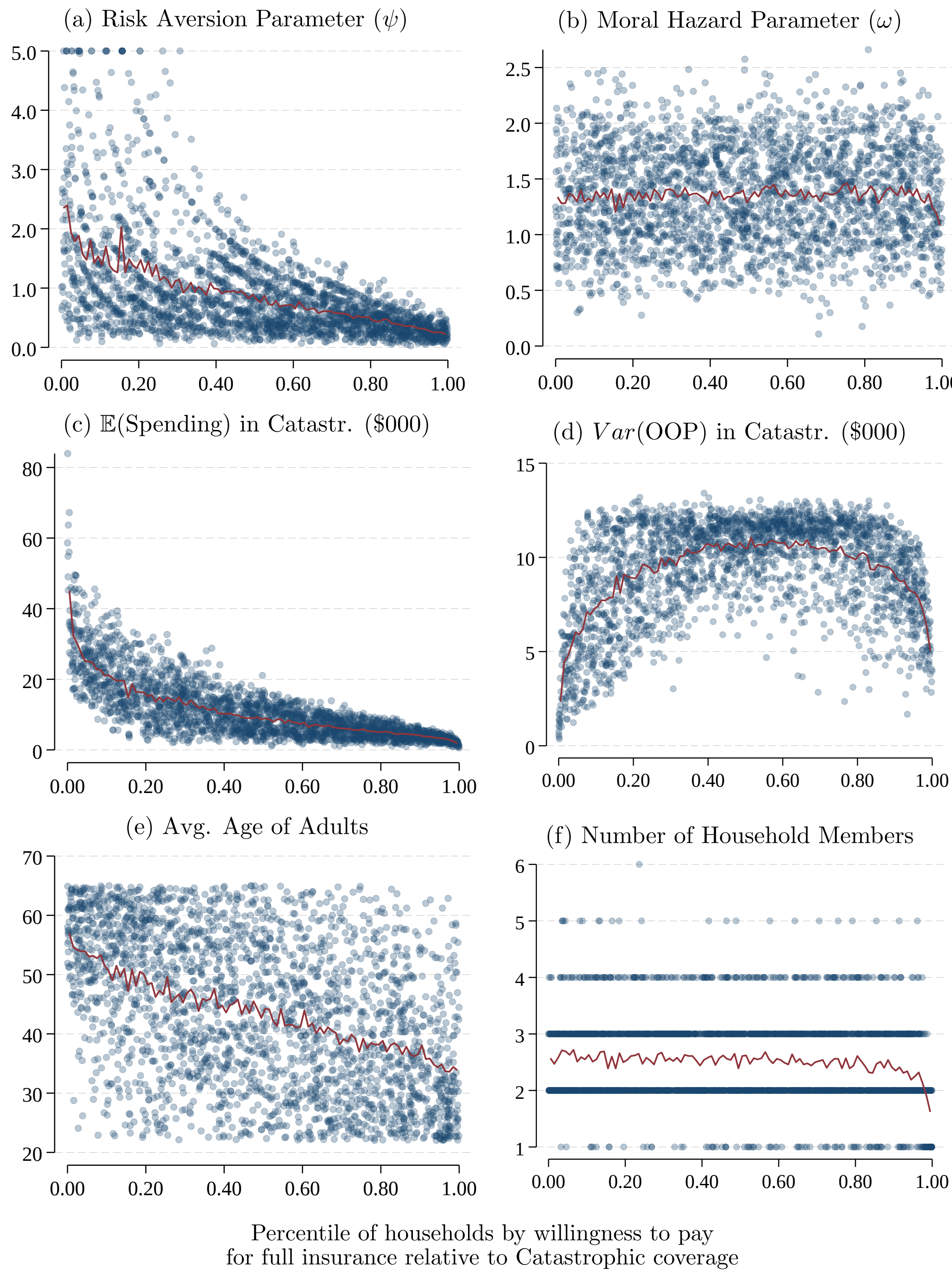
We simulate a population of consumers using the parameter estimates reported in Column 3 of Table 3 and Appendix Table A.8 of Marone and Sabety (2022). We first construct a population of households in terms of simple demographic characteristics (such as age and gender), and then construct each household's type θ using the reported parameters. As in Marone and Sabety (2022), we model a household as a group of individuals, each of whom is characterized by an age, a gender, and a health risk score.

We construct a population of households to match characteristics of the U.S. population, as reported in Section 6.1. We start the construction of each household with a "head of household." This person is female with 50 percent probability and has a uniform distribution of age between 22 and 60. We assume that 50 percent of households have a spouse present, and when present, that the spouse is of the opposite gender to the head of household. Spouses draw an age from a normal distribution with mean equal to the age of the head of household and a standard deviation of 4, subject to bounds between 22 and 60. We further assume each household has between 1 and 4 children, where each child exists with 25 percent probability, independently of one another and of the presence of a spouse. Conditional on existing, each child is female with 50 percent probability and draws their age from a uniform distribution between 0 and 18. Finally, we assume that all individuals draw a risk score from a log-normal distribution with mean positively related to age, such that for individual i: $\log(riskscore_i) \sim N(\frac{age_i}{20}, 1)$. We censor the right tail of the risk score distribution such that no individual can have a risk score that is more than five standard deviations above the uncensored mean. Our baseline population contains 10,000 households. Increasing the number of households does not change our results.

With this simulated population in hand, we then apply the parameter estimates to construct $\theta = (\psi, \omega, F)$ for each household. We make one adjustment, which is to cap the risk aversion parameter at a value of 5.[43] Summary statistics on the population distribution of demographics and resulting household types are reported in Table 2. In addition, the joint distributions of various household characteristics and households' willingness to pay (for full insurance relative to Catastrophic coverage) are shown in Figure B.1.

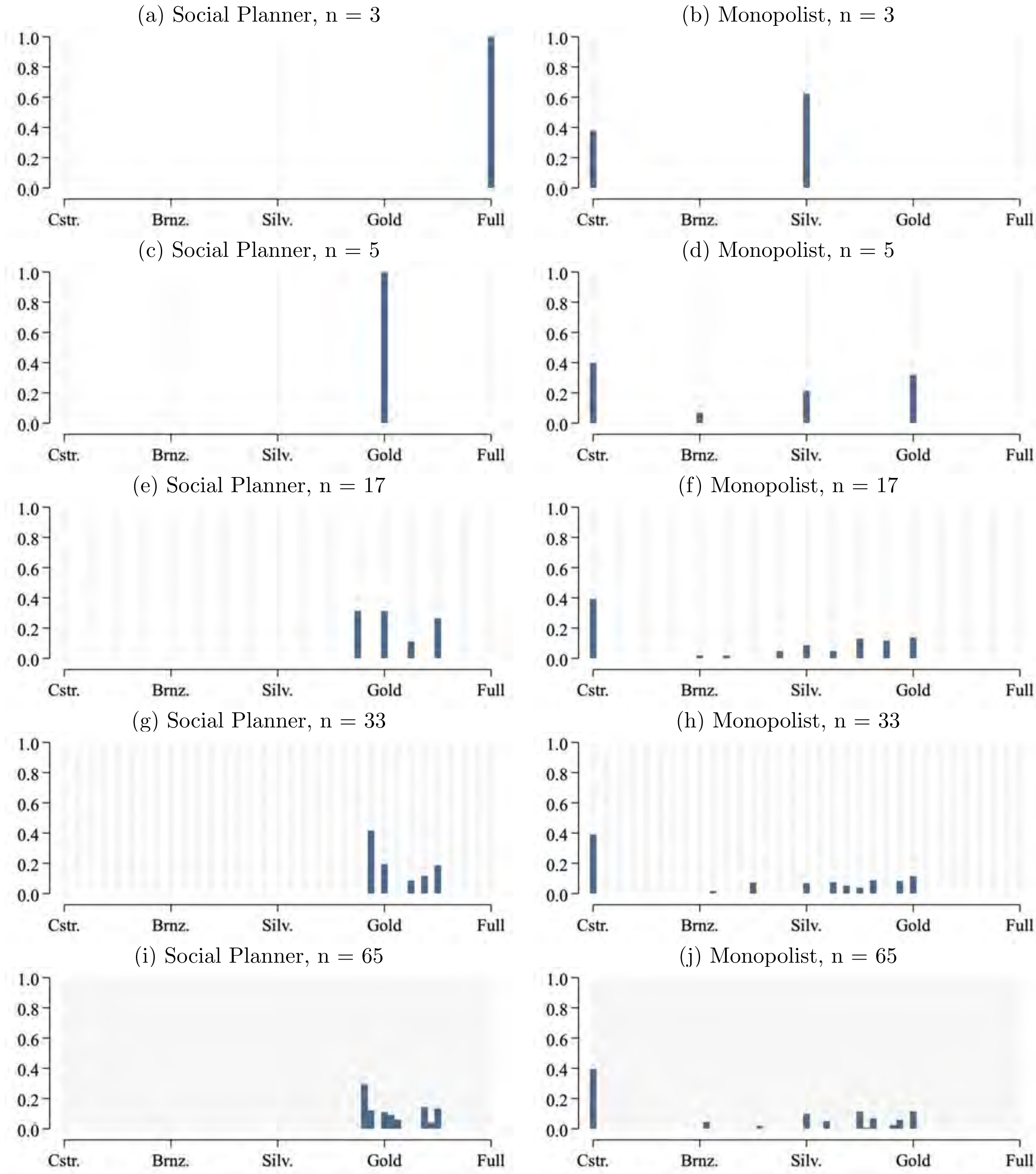
[43] We express monetary amounts in thousands of dollars, so dividing our coefficients of absolute risk aversion by 1,000 makes them comparable to other settings where monetary amounts are measured in dollars.

Figure B.1. Distribution of Household Types in Simulated Population



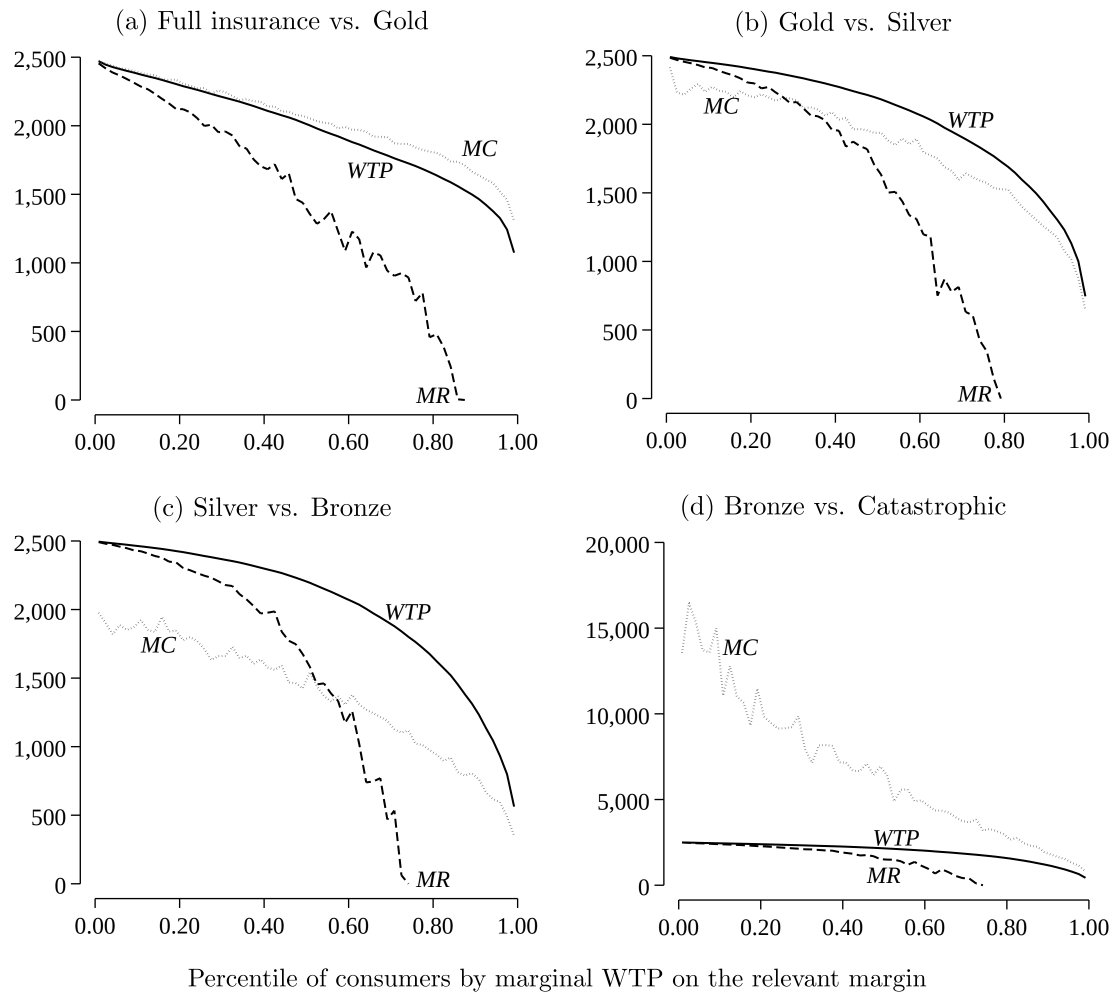
Notes: The figure shows the distribution across households of (a) the risk aversion parameter, (b) the moral hazard parameter, (c) households' expected total healthcare spending under the Catastrophic contract, (d) households' variance of out-of-pocket spending under the Catastrophic contract, (e) the average age of adults in the household, and (f) the number of individuals in the household. An adult is anyone 18 and older. Households are arranged on the horizontal axis in order of their willingness to pay for full insurance relative to the Catastrophic contract. Each dot represents a household, for a 25 percent random sample of households. The line in each panel is a connected binned scatter plot, representing the mean value of the vertical axis variable at each percentile of willingness to pay.

Figure B.2. Optimal Allocations as Density of Contract Space Increases



Notes: The figure shows the percentage of consumers allocated to each contract under the optimal menus chosen by a social planner and a monopolist as the density of the contract space increases. The gray bars identify the set of *potential* contracts available to the menu designer, while the blue bars show the actual allocations. The left-hand side panels show the allocations chosen by the social planner, while the right-hand side panels show the allocations chosen by the monopolist. The rows correspond to 3, 5, 17, 33, and 65 potential contracts, respectively.

Figure B.3. Illustration of Graphical Analysis Under "Total Pricing"



Notes: The figure demonstrates the graphical analysis of the simplified problem when the government implements "total pricing," meaning that it only covers the cost of Catastrophic coverage in the event a consumer enrolls in the Catastrophic contract. Each panel represents the "market for incremental coverage" between each pair of adjacent contracts. The vertical axes are measured in dollars. The horizontal axes report the percentage of consumers choosing a given marginal level of coverage. Consumers are ordered on the horizontal axes according to their marginal willingness to pay for the additional coverage offered on each margin. The solid line (WTP) represents consumers' willingness to pay on each margin, the dotted line (MC) represents the marginal cost curve, and the dashed line (MR) represents a monopolist's marginal revenue curve. The MC and MR curves are constructed as connected binned scatter plots using 100 points.

B.2 Proof of Lemma 1

We first establish that V is continuous. Let $(\theta^n, \rho^n) \to (\theta, \rho)$. Let us show first that $V(\rho, \theta) \geq \limsup_n V(\rho^n, \theta^n)$. For each n, choose $x^n \in X(\rho^n, \theta^n)$. Without loss of generality, x^n converges to some $\tilde{x} \in [0, 1]$. Letting $\hat{x}^n = \max(x^n - d(\rho^n, \rho), 0)$, and note that since $\hat{x}^n$ is a feasible choice,

$$\begin{aligned} V(\rho, \theta) &\geq v(\hat{x}^n, \theta^n) - \rho(\hat{x}^n) \\ &= v(x^n, \theta^n) - \rho^n(x^n) + v(\hat{x}^n, \theta^n) - v(x^n, \theta^n) + \rho^n(x^n) - \rho(\hat{x}^n) \\ &\geq V(\rho^n, \theta^n) + v(\hat{x}^n, \theta^n) - v(x^n, \theta^n) - d(\rho^n, \rho). \end{aligned}$$

where the third inequality uses that $V(\rho^n, \theta^n) = v(x^n, \theta^n) - \rho^n(x^n)$ and that $\rho(\hat{x}^n) \leq \rho^n(x^n) + d(\rho^n, \rho)$ by definition of d (see Footnote 13) and by construction of $\hat{x}^n$. But then, since v is continuous with $\lim \hat{x}^n = \lim x^n = \tilde{x}$, and since $d(\rho^n, \rho) \to 0$, we can apply $\limsup_n$ on each side to arrive at $V(\rho, \theta) \geq \limsup_n V(\rho^n, \theta^n)$ as desired. Showing that $V(\rho, \theta) \leq \liminf_n V(\rho^n, \theta^n)$ is similar. In particular, choose $\check{x} \in X(\rho, \theta)$, let $\check{x}^n = \max(\check{x} - d(\rho^n, \rho), 0)$, and observe that for all n,

$$\begin{aligned} V(\rho^n, \theta^n) &\geq v(\check{x}^n, \theta^n) - \rho^n(\check{x}^n) \\ &= v(\check{x}, \theta) - \rho(\check{x}) + v(\check{x}^n, \theta^n) - v(\check{x}, \theta) + \rho(\check{x}) - \rho^n(\check{x}^n) \\ &\geq V(\rho, \theta) + v(\check{x}^n, \theta^n) - v(\check{x}, \theta) - d(\rho^n, \rho). \end{aligned}$$

Thus, since v is continuous, and since $\rho^n \to \rho$, we have $\liminf_n V(\rho^n, \theta^n) \geq V(\rho, \theta)$. Hence, V is continuous.

Now, let us show that X is upper hemicontinuous. To do so, let $(x^n, \rho^n, \theta^n) \to (x, \rho, \theta)$ where for each n, $x^n \in X(\rho^n, \theta^n)$. We desire to show $x \in X(\rho, \theta)$. So, choose any $\grave{x}$, and for each n, let $\grave{x}^n = \max(\grave{x} - d(\rho^n, \rho), 0)$. Since $x^n \in X(\rho^n, \theta^n)$, we have

$$v(x^n, \theta^n) - \rho^n(x^n) \geq v(\grave{x}^n, \theta^n) - \rho^n(\grave{x}^n), \tag{17}$$

for all n. We will show that this implies that $v(x, \theta) - \rho(x) \geq v(\grave{x}, \theta) - \rho(\grave{x})$. Since $\grave{x}$ is arbitrary, this would establish that $x \in X(\rho, \theta)$.

Consider the *lhs*. Let us argue first that $\limsup_n(-\rho^n(x^n)) \leq -\rho(x)$. To see this, let $\tilde{x}^n = \max(x^n - d(\rho^n, \rho), 0)$, and note that $-\rho^n(x^n) = -\rho(\tilde{x}^n) + \rho(\tilde{x}^n) - \rho^n(x^n) \leq -\rho(\tilde{x}^n) + d(\rho^n, \rho)$.

But, $d(\rho^n, \rho) \to 0$ by construction, and so, since $-\rho$ is upper semicontinuous, and since $\tilde{x}^n \to x$, $\limsup_n(-\rho^n(x^n)) \leq -\rho(x)$, establishing the claim. Using this, it follows that $v(x, \theta) - \rho(x) \geq \limsup_n(v(x^n, \theta^n) - \rho^n(x^n))$, and so, since from (17), $\limsup_n(v(x^n, \theta^n) - \rho^n(x^n)) \geq \limsup_n(v(\grave{x}^n, \theta^n) - \rho^n(\grave{x}^n))$, we would be done if $\limsup_n(v(\grave{x}^n, \theta^n) - \rho^n(\grave{x}^n)) \geq v(\grave{x}, \theta) - \rho(\grave{x})$ or, since v is continuous and $\grave{x}^n \to \grave{x}$, if $\limsup_n(-\rho^n(\grave{x}^n)) \geq -\rho(\grave{x})$. But, $-\rho^n(\grave{x}^n) = -\rho(\grave{x}) + \rho(\grave{x}) - \rho^n(\grave{x}^n) \geq -\rho(\grave{x}) - d(\rho^n, \rho)$, and the result follows immediately since $d(\rho^n, \rho) \to 0$. $\square$

B.3 Differentiability of γ^I and γ^G

In this section, we provide primitives for γ^I and γ^G to be almost everywhere differentiable with bounded derivatives. We do so by restricting c to a class where we can tame the way in which the consumer jumps from one a to another as l changes.

Assumption 2 *For some finite $\bar{\kappa}$, $c(\cdot, x) = \min_{\kappa \in \{1,\ldots,\bar{\kappa}\}} \tilde{c}(\cdot, x, \kappa)$ where for each κ, $\tilde{c}(\cdot, x, \kappa)$ is twice continuously differentiable, with $b_{aa}(\cdot, l, \omega) - \tilde{c}_{aa}(\cdot, x, \kappa)$ strictly bounded away from zero, and $\tilde{c}_a(a, x, \cdot)$ strictly decreasing.*

That is, while c can have kinks, on each segment where it is differentiable, $b - c$ is concave. An example is when $c(\cdot, x)$ is piecewise linear for each x.

We also need a condition on how the marginal value of healthcare changes with ω and l.

Assumption 3 *The ratio $b_{\omega a}(\cdot, l, \omega)/b_{la}(\cdot, l, \omega)$ is strictly monotone (of either sign).*

In the canonical example, $b_{\omega a}(\cdot, l, \omega)/b_{la}(\cdot, l, \omega) = (a - l)/\omega$ which is strictly increasing in a. In general, Assumption 3 asks that $b_\omega(\cdot, l, \omega)$ is strictly either more or less concave than $b_l(\cdot, l, \omega)$.

Let $A(l, x, \omega)$ be the optimal correspondence of the consumer's choice of a when the health state is l, the contract is x, and the consumer's taste for healthcare is ω. Assume that A has 2BRP: for all x, there is a finite subset of $[\underline{\omega}, \bar{\omega}]$ such that except on this set, $A(\cdot, x, \omega)$ has at most two elements. Let us first provide primitives for 2BRP.

Lemma 4 *Let Assumptions 2 and 3 be true. Then, A satisfies 2BRP.*

Proof Let $\tilde{a}(l, x, \omega, \kappa) = \arg\max_a (b(a, l, \omega) - \tilde{c}(a, x, \kappa))$ be the consumer's best action facing $\tilde{c}(\cdot, x, \kappa)$ and let $\tilde{z}(l, x, \omega, \kappa)$ be the associated value function. Since $b_{aa}(\cdot, l, \omega) - \tilde{c}_{aa}(\cdot, x, \kappa)$ is bounded below zero, $\tilde{a}(\cdot, \cdot, \cdot, \kappa)$ is uniquely defined by $b_a(\tilde{a}, l, \omega) = c_a(\tilde{a}, x)$ and is continuously differentiable. For example, the Envelope Theorem and the Implicit Function Theorem gives us

$$\tilde{a}_x(l, x, \omega, \kappa) = \frac{c_{ax}(\tilde{a}, x)}{b_{aa}(\tilde{a}, l, \omega) - c_{aa}(\tilde{a}, x)},$$

which is uniformly bounded. Note that since $\tilde{c}_a(a, x, \cdot)$ is strictly decreasing, $\tilde{a}(l, x, \omega, \cdot)$is strictly increasing. The consumer's optimal choice is then given by maximizing $\tilde{z}(l, x, \omega, \kappa)$ over κ, and then taking the associated $\tilde{a}(l, x, \omega, \kappa)$. The consumer has more than one best response if and only if $\tilde{z}(l, x, \omega, \cdot)$ has more than one maximizer.

For $\kappa'' > \kappa'$, let $\tilde{l}(x, \omega, \kappa', \kappa'')$ be the l that solves $\tilde{z}(l, x, \omega, \kappa') = \tilde{z}(l, x, \omega, \kappa'')$. The Envelope Theorem, $b_{al} > 0$, and $\tilde{a}$ strictly increasing in κ yield

$$\tilde{z}_l(l, x, \omega, \kappa') = b_l(\tilde{a}(l, x, \omega, \kappa'), l, \omega) < b_l(\tilde{a}(l, x, \omega, \kappa''), l, \omega) = \tilde{z}_l(l, x, \omega, \kappa''),$$

and so $\tilde{l}(x, \omega, \kappa', \kappa'')$ is unique. A consumer with proclivity to spend on healthcare ω will be indifferent between $\tilde{a}(l, x, \omega, \kappa')$ and $\tilde{a}(l, x, \omega, \kappa'')$ facing insurance quality x only if their health realization is $\tilde{l}(x, \omega, \kappa', \kappa'')$. By the Envelope Theorem and the Implicit Function Theorem,

$$
\begin{aligned}
\tilde{l}_\omega(x, \omega, \kappa', \kappa'') &= -\frac{\tilde{z}_\omega(\tilde{l}, x, \omega, \kappa'') - \tilde{z}_\omega(\tilde{l}, x, \omega, \kappa')}{\tilde{z}_l(\tilde{l}, x, \omega, \kappa'') - \tilde{z}_l(\tilde{l}, x, \omega, \kappa')} \\
&= -\frac{b_\omega(\tilde{a}(\tilde{l}, x, \omega, \kappa''), \tilde{l}, \omega) - b_\omega(\tilde{a}(\tilde{l}, x, \omega, \kappa'), \tilde{l}, \omega)}{b_l(\tilde{a}(\tilde{l}, x, \omega, \kappa''), \tilde{l}, \omega) - b_l(\tilde{a}(\tilde{l}, x, \omega, \kappa'), \tilde{l}, \omega)} \\
&= -\int_{\tilde{a}(\tilde{l}, x, \omega, \kappa')}^{\tilde{a}(\tilde{l}, x, \omega, \kappa'')} \frac{b_{\omega a}(a, \tilde{l}, \omega)}{b_{la}(a, \tilde{l}, \omega)} \frac{b_{la}(a, \tilde{l}, \omega)}{\int_{\tilde{a}(\tilde{l}, x, \omega, \kappa')}^{\tilde{a}(\tilde{l}, x, \omega, \kappa'')} b_{la}(a, \tilde{l}, \omega) da} da,
\end{aligned} \tag{18}
$$

where the last equality follows from the Fundamental Theorem of Calculus applied to numerator and denominator, and by multiplying and dividing the integrand in the numerator by $b_{al} > 0$. That is, $\tilde{l}_\omega(x, \omega, \kappa', \kappa'')$ is an expectation of $b_{\omega a}/b_{la}$ over the interval $(\tilde{a}(\tilde{l}, x, \omega, \kappa'), \tilde{a}(\tilde{l}, x, \omega, \kappa''))$.

Consider the case $b_{\omega a}/b_{la}$ strictly increasing. Then, by (18), if $\kappa' < \kappa'' < \kappa'''$ then, since $\tilde{a}(\tilde{l}, x, \omega, \cdot)$ is strictly increasing, $\tilde{l}_\omega(x, \omega, \kappa'', \kappa''') - \tilde{l}_\omega(x, \omega, \kappa', \kappa'') > 0$ (the it has sign opposite to that of the strict monotonicity of $b_{\omega a}/b_{la}$). Thus, for each x, there is at most one ω such that $\tilde{l}(x, \omega, \kappa', \kappa'') = \tilde{l}(x, \omega, \kappa'', \kappa''')$. But, $\tilde{l}(x, \omega, \kappa', \kappa'') = \tilde{l}(x, \omega, \kappa'', \kappa''')$ is a necessary condition for $\tilde{a}(l, x, \omega, \kappa')$, $\tilde{a}(l, x, \omega, \kappa')$, and $\tilde{a}(l, x, \omega, \kappa'')$ to all be elements of $A(l, x, \omega)$. Since $\bar{\kappa}$ is finite, there are a finite set of triples $\kappa', \kappa'', \kappa'''$ to check, and so there are, for each x, at most a finite set of ω where there are more than two best responses at any l. □

Lemma 5 *Let Assumption 2 hold. Let A have 2BRP. Then, for each x, and for any θ with ω not in the exceptional set, $\left(\int_0^{\bar{l}} c\left(a^*(l, x, \omega), x\right) f(l) dl\right)_x$ exists and is uniformly bounded.*

Proof Fix x, and fix ω such that 2BRP holds. Then, we claim, there are $1 \leq \bar{j} \leq \bar{\kappa}$ points $0 = l^0 < l^1 < l^2 < \ldots < l^{\bar{j}} = \bar{l}$, such that on (l^{j-1}, l^j) there is a unique best κ^j. The case $\bar{j} < \bar{\kappa}$ occurs when the consumer chooses not to use some segments of c. By 2BRP, at l^j, the two best κ's are κ^j and κ^{j+1}, with all other κ's strictly worse. That is,

$$\tilde{z}(l^j, x, \omega, \kappa^j) = \tilde{z}(l^j, x, \omega, \kappa^{j+1}) > \max_{\kappa \neq \kappa^j, \kappa^{j+1}} \tilde{z}(l^j, x, \omega, \kappa).$$

Let $K = \left\{\kappa^1, \ldots, \kappa^{\bar{j}}\right\} \subseteq \{1, \ldots, \bar{\kappa}\}$ be the set of indexes that ωuses given x.

We claim that for all x' in a neighborhood of x, ω chooses exactly the elements of K when facing x'. To see this, note that any $\kappa' \notin K$ is never an optimal choice for ω facing x. That is,

$$\max_{\kappa \in K} \tilde{z}(l, x, \omega, \kappa) - \tilde{z}(l, x, \omega, \kappa') > 0.$$

This follows because if $l \in (l^{j-1}, l^j)$ then the only optimal κ is κ^j, while if $l = l^j$ then the only best responses are κ^j and κ^{j+1}. But, then, since both sides are continuous in l and x over the bounded set of l and x, the same is true for all x' in some neighborhood of x. Also, for any $\kappa^j \in K$, choose any $\hat{l} \in (l^{j-1}, l^j)$. Then, since

$$\tilde{z}(\hat{l}, x, \omega, \kappa^j) > \max_{\kappa \neq \kappa^j} \tilde{z}(\hat{l}, x, \omega, \kappa),$$

the same is true on a neighborhood of x, and κ^j is sometimes chosen.

It follows that there are $0 = l^0 < l^1(x') < l^2(x') < \ldots < l^{\bar{j}} = \bar{l}$ such that κ^j is chosen by ω facing x' on the interval $(l^{j-1}(x), l^j(x))$, where $l^j(x')$ is defined by

$$\tilde{z}(l^j(x'), x', \omega, \kappa^j) = \tilde{z}(l^j(x'), x', \omega, \kappa^{j+1}).$$

But, as in the proof of Lemma 4, by the Envelope Theorem and the Implicit Function Theorem, $l^j_x(x')$ is differentiable on a neighborhood of x with

$$l^j_x(x') = \frac{b_{\omega a}(a, l^j(x'), \omega)}{b_{la}(a, l^j(x'), \omega)}$$

for some $a \in (\tilde{a}(l, x', \omega, \kappa^j), \tilde{a}(l, x', \omega, \kappa^j)$. By assumption, this is bounded. We can then write

$$\int_0^{\bar{l}} c(a^*(l, x', \omega), x') f(l) dl = \sum_{j=1}^{\bar{j}} \int_{l^{j-1}(x')}^{l^j(x')} c(\tilde{a}(l, x', \omega, \kappa^j) f(l) dl,$$

and so the derivative of the *rhs* with respect to x evaluated at $x = x'$ is

$$\left(\int_0^{\bar{l}} c\left(a^*(l, x, \omega), x\right) f(l) dl \right)_x = \sum_{j=1}^{\bar{j}} \left(\begin{array}{c} l^j_x(x) c(\tilde{a}(l^j, x', \omega, \kappa^j), x', \kappa^j) f(l^j) \\ -l^{j-1}_x(x) c(\tilde{a}(l^{j-1}, x', \omega, \kappa^j), x', \kappa^j) f(l^{j-1}) \\ + \int_{l^{j-1}(x')}^{l^j(x')} c_a(\tilde{a}(l, x', \omega, \kappa^j), x', \kappa^j) \tilde{a}_x(l, x', \omega, \kappa^j) f(l) dl \end{array} \right),$$

each part of which is uniformly bounded. □

We can now show that $\gamma^I(x, \theta)$ and $\gamma^G(x, \theta)$ have the requisite differentiability properties.

Proposition 4 *Let Assumptions 2 and 3 hold. Then, $\gamma^I(x, \theta)$ and $\gamma^G(x, \theta)$ are differentiable in x for almost all θ, with $\gamma^I_x(x, \theta)$ and $\gamma^G_x(x, \theta)$ uniformly bounded.*

Proof Consider any θ for which for which 2BRP holds for the relevant ω. Then, from above, $\int c(a^*(l, x, \omega), x) dF(l)$ is differentiable with a uniformly bounded derivative. Taking the case where $c(\cdot, x)$ is the identity shows that $\int a^*(l, x, \omega) dF(l)$ has the same property. But then, γ^G is differentiable with a uniformly bounded derivative. Taking $x = x^0$ then covers γ^I. □

B.4 Proof of Theorem 3

Let $x = x^k$ be the quality level being modified, and let ρ^ε be the premium schedule in which the step in ρ at x has been replaced by a step at $x+\varepsilon$. Formally, let $\rho^\varepsilon(x') = \rho(x')$ for $x' \notin (\min\{x, x+\varepsilon\}, \max\{x, x+\varepsilon\})$, while $\rho^\varepsilon(x') = \rho(x)$ for $x' \in (\min\{x, x+\varepsilon\}, \max\{x, x+\varepsilon\})$.

Fix (ω, F) such that *2BRP* holds and, similar to the construction involving Δ in the previous proof, such that neither $\underline{\psi}$ nor $\bar{\psi}$ is indifferent between x and their next best choice. To lessen the notational load, suppress (x, ω, F) in what follows. If x is not a best response for any ψ, then for small ε, x remains unattractive for all ψ and so the perturbation has no effect. Hence, since $\mathcal{W}(x) = 0$ by definition in this case, we have that $\mathcal{W}(x) = (\Pi(\rho^\varepsilon))_\varepsilon = 0$.

So assume that x is a best response for some ψ. Then, by *2BRP*, $X(\rho, \psi) = x$ on a positive-lengthed interval (ψ^l, ψ^h).[44] In a minor abuse of notation, let $(\psi^l(\varepsilon), \psi^h(\varepsilon))$ be the nonempty interval on which $X(\rho^\varepsilon, \psi) = x + \varepsilon$, noting that $\psi^l = \psi^l(0)$ and $\psi^h = \psi^h(0)$. Arguing as in the previous proof, if we let $\bar{x}^h \equiv \bar{x}(\psi^h, \rho) \geq x^{k+1} > x$, then types just to the right of $\psi^h(\varepsilon)$ choose $\bar{x}^h$, and similarly, types just to the left of $\psi^l(\varepsilon)$ choose $\underline{x}^l \equiv \underline{x}(\psi^l, \rho) \leq x^{k-1} < x$. If $\psi^h = \bar{\psi}$,then $\psi^h(\varepsilon) = \bar{\psi}$ for small ε, and so $\psi^h_\varepsilon(0) = 0$. Otherwise, the defining condition for $\psi^h(\varepsilon)$ is

$$v(x+\varepsilon, \psi^h(\varepsilon)) - \rho(x) = v(\bar{x}^h, \psi^h(\varepsilon)) - \rho(\bar{x}^h),$$

where the *lhs* is the payoff to $\psi^h(\varepsilon)$ of choosing $x+\varepsilon$ and the *rhs* the payoff of switching to $\bar{x}^h$, and so for small ε,

$$\psi^h_\varepsilon(\varepsilon) = \frac{v_x(x+\varepsilon, \psi^h(\varepsilon))}{v_\psi(\bar{x}^h, \psi^h(\varepsilon)) - v_\psi(x+\varepsilon, \psi^h(\varepsilon))} > 0,$$

using $v_{\psi x} > 0$ and $v_x > 0$. Similarly, if $\psi^l = \underline{\psi}$, then for small ε, $\psi^l_\varepsilon(\varepsilon) = 0$, while where ψ^l is interior,

$$\psi^l_\varepsilon(\varepsilon) = \frac{-v_x(x+\varepsilon, \psi^l(\varepsilon))}{v_\psi(x+\varepsilon, \psi^l(\varepsilon)) - v_\psi(\underline{x}^l, \psi^l(\varepsilon))} < 0.$$

For ε small, we then have

$$\begin{aligned}
\Pi(\rho^\varepsilon) - \Pi(\rho) \quad = \quad & \int_{\psi^h}^{\psi^h(\varepsilon)} (S(\rho(x), x+\varepsilon, \psi) - S(\rho(\bar{x}^h), \bar{x}^h, \psi)) dG(\psi) \\
& + \int_{\psi^l}^{\psi^h} (S(\rho(x), x+\varepsilon, \psi) - S(\rho(x), x, \psi)) dG(\psi) \\
& + \int_{\psi^l(\varepsilon)}^{\psi^l} (S(\rho(x), x+\varepsilon, \psi) - S(\rho(\underline{x}^l), \underline{x}^l, \psi)) dG(\psi),
\end{aligned} \tag{19}$$

where the first integral reflects that types in $(\psi^h, \psi^h(\varepsilon))$ switch their quality choice from $\bar{x}^h$ to $x+\varepsilon$, the second integral reflects that those in (ψ^l, ψ^h) "switch" from x to $x+\varepsilon$, and the third integral

[44] If x was chosen by a single type ψ, then there would be three best responses at ψ, with one representing the action taken by types just below ψ, and one the action of types just above ψ.

reflects that types in $(\psi^l(\varepsilon), \psi^l)$ switch their quality choice from $\underline{x}^l$ to $x+\varepsilon$.

Thus,

$$\begin{aligned}(\Pi(\rho^\varepsilon))_\varepsilon &= \psi^h_\varepsilon(\varepsilon)(S(\rho(x), x+\varepsilon, \psi^h(\varepsilon)) - S(\rho(\bar{x}^h), \bar{x}^h, \psi^h(\varepsilon)))g(\psi^h(\varepsilon)) \\ &\quad + \int_{\psi^l(\varepsilon)}^{\psi^h(\varepsilon)} S_x(\rho(x), x+\varepsilon, \psi)dG(\psi) \\ &\quad - \psi^l_\varepsilon(\varepsilon)(S(\rho(x), x+\varepsilon, \psi^l(\varepsilon)) - S(\rho(\underline{x}^l), \underline{x}^l, \psi^l(\varepsilon)))g(\psi^l(\varepsilon)),\end{aligned}$$

where the passing of the derivative through the integral is valid by $LDCT$, noting that

$$S_x(p, x, \theta) = w^C v_x(x, \theta) - w^I \gamma^I(x, \theta) + (w^I - w^G)\gamma^G_x(x, x^0, \theta), \tag{20}$$

where $v_x = \int(-c_x)zdl$ is defined everywhere and bounded, and where by assumption, γ^I and γ^G are differentiable in x for almost every θ, with uniformly bounded derivatives (recall that we provide primitives backing this assumption).

Thus,

$$\begin{aligned}(\Pi(\rho^\varepsilon))_\varepsilon|_{\varepsilon=0} &= \psi^h_\varepsilon(0)(S(\rho(x), x, \psi^h) - S(\rho(\bar{x}^h), \bar{x}^h, \psi^h))g(\psi^h) \\ &\quad + \int_{\psi^l}^{\psi^h} S_x(\rho(x), x, \psi)dG(\psi) \\ &\quad - \psi^l_\varepsilon(0)(S(\rho(x), x, \psi^l) - S(\rho(\underline{x}^l), \underline{x}^l, \psi^l))g(\psi^l).\end{aligned} \tag{21}$$

Now, if ψ^l is interior, then as in the previous proof,

$$S(\rho(x), x, \psi^l) - S(\rho(\underline{x}^l), \underline{x}^l, \psi^l) = \mathcal{S}(x, \psi^l) - \mathcal{S}(\underline{x}^l, \psi^l)$$

and so

$$-\psi^l_\varepsilon(0)(S(\rho(x), x, \psi^l) - S(\rho(\underline{x}^l), \underline{x}^l, \psi^l)) = v_x(x, \psi^l)\frac{\mathcal{S}(x, \psi^l) - \mathcal{S}(\underline{x}^l, \psi^l)}{v_\psi(x, \psi^l) - v_\psi(\underline{x}^l, \psi^l)} = v_x(x, \psi^l)r^l,$$

while if $\psi^l = \underline{\psi}$, then

$$\psi^l_\varepsilon(0)(S(\rho(x), x, \psi^l) - S(\rho(\underline{x}^l), \underline{x}^l, \psi^l)) = 0 = v_x(x, \psi^l)r^l,$$

and similarly,

$$\psi^h_\varepsilon(0)(S(\rho(x), x, \psi^h) - S(\rho(\bar{x}^h), \bar{x}^h, \psi^h)) = -v_x(x, \psi^h)r^h.$$

Also, $S_x(\rho(x), x, \psi) = \mathcal{S}_x(x, \psi) - (w^I - w^C)v_x(x, \psi)$, and so, making the relevant substitutions,

$$(\Pi(\rho^\varepsilon))_\varepsilon|_{\varepsilon=0} = -v_x(x, \psi^h)r^h g(\psi^h) + \int_{\psi^l}^{\psi^h}(\mathcal{S}_x(x, \psi) - (w^I - w^C)v_x(x, \psi))dG(\psi) + v_x(x, \psi^l)r^l g(\psi^l).$$

Reinstating (x, ω, F), we have $(\Pi(\rho^\varepsilon, \omega, F))_\varepsilon|_{\varepsilon=0} = \mathcal{W}(x, \omega, F)$ as asserted.

But then, as above, we can apply *LDCT* to see that

$$0 = \left(\int \Pi(\rho^\varepsilon, \omega, F) dG(\omega, F)\right)_\varepsilon\Big|_{\varepsilon=0} = \int (\Pi(\rho^\varepsilon, \omega, F))_\varepsilon|_{\varepsilon=0}\, dG(\omega, F) = \int \mathcal{W}(x, \omega, F) dG(\omega, F),$$

where the first equality reflects that ρ^ε is a feasible perturbation and ρ is optimal. □

B.5 Proof of Theorem 4

Since it is more intricate, we begin by showing that $\int \mathcal{W} dG(\omega, F) = 0$. We will then use the machinery developed to analyze $\int \mathcal{V} dG(\omega, F)$. We proceed in a sequence of steps.

Step 1 Let (ρ, χ) be optimal in the continuum. Let $\mathcal{P}^k$ be the subset of $\mathcal{P}$ that are step functions with at most k steps. Consider the problem of the insurer restricted to $\mathcal{P}^k$ and has payoff function $\tilde{\Pi}(\rho') = \Pi(\rho') - d^2(\rho, \rho')$. That is, the insurer is penalized for choosing ρ' different than ρ according to the square of the Levy distance from ρ' to ρ. For each k, let $\rho^k \in \arg\max_{\rho' \in \mathcal{P}^k} \tilde{\Pi}(\rho')$ be an optimum of this problem. We claim that $d(\rho^k, \rho) \to 0$. Thus, with the penalty function, the solution to the continuum problem is well-approximated by nearby solutions of the discrete problem.

Proof Fix $\delta > 0$. By Lemma 2 (in Section A.6 below), for all k large enough, there is ρ^* with at most k steps with $d^2(\rho^*, \rho) \leq \delta/2$ and $\Pi(\rho^*) \geq \Pi(\rho) - \delta/2$. But then, since ρ^* is feasible while ρ^k is optimal,

$$\tilde{\Pi}(\rho^k) \geq \tilde{\Pi}(\rho^*) = \Pi(\rho^*) - d^2(\rho^*, \rho) \geq \Pi(\rho) - \delta.$$

Now, since ρ is optimal in the original problem, it follows that $\Pi(\rho) - d^2(\rho^k, \rho) \geq \Pi(\rho^k) - d^2(\rho^k, \rho) = \tilde{\Pi}(\rho^k)$, and so we must have $d^2(\rho^k, \rho) \leq \delta$. Since δ was arbitrary, it follows that $d(\rho^k, \rho) \to 0$. Let χ^k be the associated allocations. That is, χ^k is a selection from $X(\cdot, \rho^k)$, recalling that this is unique G-almost everywhere.

Step 2 For any given $\tau > 0$ for any given k, and for any given $\hat{x}$ offered by ρ, consider the perturbation in which each x that is offered under ρ^k and is contained in $(\hat{x} - \tau, \hat{x} + \tau)$ is increased by ε. We will first calculate the value of this perturbation by breaking it up into a set of perturbations of the type analyzed in Section A.3 and then summing as appropriate. Then, we will consider the form of the limiting expression as one first takes $k \to \infty$ and then takes $\tau \to 0$.

Step 3 Let us begin with some definitions. For this and the next several steps, we will work with a fixed (ω, F) which we will suppress, and reintroduce only later when it is needed. So, for example, we will write $\chi(\psi)$ when we mean properly $\chi(\omega, \psi, F)$. We will asssume (ω, F) satisfies *2BRP* relative to ρ.

Let $\psi^l(\tau) = \inf\{\psi | \chi(\psi) \geq \hat{x} - \tau\}$ and $\psi^h(\tau) = \sup\{\psi | \chi(\psi) \leq \hat{x} + \tau\}$. So, $[\psi^l(0), \psi^h(0)]$ is the (possibly empty) interval over which the consumer chooses $\hat{x}$ under χ. Let $\psi^{l,k}(\tau) = \inf\{\psi | \chi^k(\psi) \geq$

$\hat{x}-\tau\}$ and $\psi^{h,k}(\tau)=\sup\{\psi|\chi^k(\psi)\leq\hat{x}+\tau\}$ be the analogous objects when χ is replaced by χ^k.

Let $\{x_j^k\}_j^{J^k}$ list, in order from smallest to largest, the contracts actually chosen by (ω, F) facing ρ^k. That is, $\{x_j^k\}$ is the range of χ^k. Let ψ_j^k be the jump point from x_j^k to x_{j+1}^k, where we take $\psi_0^k=\underline{\psi}$, and $\psi_{J^k}^k=\bar{\psi}$. Let

$$r_j^k(\tau,k)=\frac{\mathcal{S}(x_{j+1}^k,\psi_j^k)-\mathcal{S}(x_j^k,\psi_j^k)}{v_\psi(x_{j+1}^k,\psi_j^k)-v_\psi(x_j^k,\psi_j^k)}.$$

Finally, let $j^l(\tau,k)=\min\{j|x_j^k>\hat{x}-\tau\}$, and let $j^h(\tau,k)=\max\{j|x_j^k<\hat{x}+\tau\}$. Note that this implies that types between $\psi_{j^l(\tau,k)-1}^k$ and $\psi_{j^h(\tau,k)}^k$ choose some $x\in(\hat{x}-\tau,\hat{x}+\tau)$ while other types do not. Note also that by $CMVT$,

$$r_j^k(\tau,k)=\frac{\mathcal{S}_x(x,\psi_j^k)}{v_{\psi x}(x,\psi_j^k)} \tag{22}$$

for some $x\in[x_j^k,x_{j+1}^k]$.

Step 4 Fix some k and some j with $j^l(\tau,k)\leq j\leq j^h(\tau,k)$. Consider first the perturbation of raising x_j^k (and only x_j^k) by ε. From (21) the derivative of payoffs with respect to this perturbation, ignoring the impact of the perturbation on $d(\rho^k,\rho)$ and evaluated at $\varepsilon=0$ can be written as

$$\pi_\varepsilon(0,j)=-v_x(x_j^k,\psi_j^k)r_j^kg(\psi_j^k)+\int_{\psi_{j-1}^k}^{\psi_j^k}S_x(x_j^k,\psi)dG(\psi)+v_x(x_j^k,\psi_{j-1}^k)r_{j-1}^kg(\psi_{j-1}^k),$$

where, as in the proof of Theorem 3, $S_x=\mathcal{S}_x-(w^I-w^C)v_x$ does not depend on p, and so we suppress that argument.

Step 5 Let us sum this expression over the appropriate set of indexes. For notational convenience, abbreviate $j^l(\tau,k)$ to j^l, and $j^h(\tau,k)$ to j^h. We have

$$\begin{aligned}\sum_{j^l}^{j^h}\pi_\varepsilon(0,j) &= \sum_{j^l}^{j^h}\left(-v_x(x_j^k,\psi_j^k)r_j^kg(\psi_j^k)+\int_{\psi_{j-1}^k}^{\psi_j^k}S_x(x_j^k,\psi)dG(\psi)+v_x(x_j^k,\psi_{j-1}^k)r_{j-1}^kg(\psi_{j-1}^k)\right)\\ &= -v_x(x_{j^h}^k,\psi_{j^h}^k)r_{j^h}^kg(\psi_{j^h}^k)-\sum_{j^l}^{j^h-1}v_x(x_j^k,\psi_j^k)r_j^kg(\psi_j^k)+\sum_{j^l}^{j^h}\int_{\psi_{j-1}^k}^{\psi_j^k}S_x(x_j^k,\psi)dG(\psi)\\ &\quad+\left(\sum_{j^l+1}^{j^h}\left(v_x(x_j^k,\psi_{j-1}^k)r_{j-1}^kg(\psi_{j-1}^k)\right)\right)+v_x(x_{j^l}^k,\psi_{j^l-1}^k)r_{j^l-1}^kg(\psi_{j^l-1}^k).\end{aligned}$$

Now, reindex the sum in the large brackets in the last line to sum from j^l to j^{h-1}, and combine it with the sum in the second term to arrive at

$$O(\hat{x}, \omega, F|\tau, k) = \sum_{j^l}^{j^h-1} (v_x(x_{j+1}^k, \psi_j^k) - v_x(x_j^k, \psi_j^k))r_j^k g(\psi_j^k).$$

Note for interpretation that O captures all of the "internal" spillovers as the consumer switches between the set of x's in $(\hat{x}-\tau, \hat{x}+\tau)$. Also, recognize that by construction, $x_j^k = \chi^k(\psi)$ for $\psi \in (\psi_{j-1}^k, \psi_j^k)$, and so the summation of integrals can be rewritten as $\int_{\psi_{j^l-1}^k}^{\psi_{j^h}^k} S_x(\chi^k(\psi), \psi) dG(\psi)$. We thus have that the profit of the perturbation facing (ω, F) and given τ and k is $\tilde{\mathcal{W}}(\hat{x}, \omega, F|\tau, k) + O(\hat{x}, \omega, F|\tau, k)$, where

$$\begin{aligned}(23)\tilde{\mathcal{W}}(\hat{x}, \omega, F|\tau, k) \quad = \quad & -v_x(x_{j^h(\tau,k)}^k, \psi_{j^h(\tau,k)}^k) r_{j^h(\tau,k)}^k g(\psi_{j^h(\tau,k)}^k) + \int_{\psi_{j^l(\tau,k)-1}^k}^{\psi_{j^h(\tau,k)}^k} S_x(\chi^k(\psi), \psi) dG(\psi) \\ & + v_x(x_{j^l(\tau,k)}^k, \psi_{j^l(\tau,k)-1}^k) r_{j^l(\tau,k)-1}^k g(\psi_{j^l(\tau,k)-1}^k)).\end{aligned}$$

Note for what follows that all terms of this are uniformly bounded. In particular, as in the discussion immediately following (20), S_x is uniformly bounded, and using (22) the r terms are bounded as well. The density g is continuous on a compact set, and so is bounded. Finally, since $v_x = \mathbb{E}_z[-c_x]$, where c_x is bounded, v_x is bounded as well.

Step 6 Let

$$\mu \equiv \max_{x,\psi} |v_{xx}(x, \psi)| \max_{x,\psi} \left(\frac{\mathcal{S}_x(x, \psi)}{v_{\psi x}(x, \psi)} g(\psi) \right) < \infty,$$

noting that μ is finite since all of the relevant objects are continuous on the compact set $[0, 1] \times [\underline{\psi}, \bar{\psi}]$, and since $v_{\psi x}(x, \psi)$ is strictly positive. Then, for all τ and k, $|O(\hat{x}, \omega, F|\tau, k)| \leq 2\tau\mu$.

Proof Using the claim at the end of Step 3,

$$\left| r_j^k g(\psi_j^k) \right| \leq \max_{x,\psi} \left(\frac{\mathcal{S}_x(x, \psi)}{v_{\psi x}(x, \psi)} g(\psi) \right),$$

and so, since

$$O(\hat{x}, \omega, F|\tau, k) = \sum_{j^l(\tau,k)}^{j^h(\tau,k)-1} (v_x(x_{j+1}^k, \psi_j^k) - v_x(x_j^k, \psi_j^k)) r_j^k g(\psi_j^k).$$

we have

$$|O(\hat{x}, \omega, F|\tau, k)| \leq \left(\sum_{j^l(\tau,k)}^{j^h(\tau,k)-1} (x_{j+1}^k - x_j^k) \right) \max_{x,\psi} |v_{xx}(x, \psi)| \max_{x,\psi} \left(\frac{\mathcal{S}_x(x, \psi)}{v_{\psi x}(x, \psi)} g(\psi) \right),$$

hence noting that $x_{j^h}^k < \hat{x} + \tau$ and $x_{j^l}^k > \hat{x} - \tau$.

Step 7 For any given τ and k, let $\rho^k(\varepsilon)$ be the perturbation of ρ^k in which contracts in $(\hat{x}-\tau, \hat{x}+\tau)$ are increased by ε. Then,

$$\left| \int \tilde{\mathcal{W}}(\hat{x}, \omega, F|\tau, k) dG(\omega, F) \right| \leq 2d(\rho^k(\varepsilon), \rho) + 2\tau\mu.$$

Proof We have that $\tilde{\Pi}(\rho^k(\varepsilon)) = \Pi(\rho^k(\varepsilon)) - d^2(\rho^k(\varepsilon), \rho)$, and so since ρ^k is optimal,

$$0 = (\tilde{\Pi}(\rho^k(\varepsilon)))_\varepsilon|_{\varepsilon=0} = \left[(\Pi(\rho^k(\varepsilon)))_\varepsilon - 2d(\rho^k(\varepsilon), \rho) d(\rho^k(\varepsilon), \rho)_\varepsilon\right]\Big|_{\varepsilon=0},$$

where by Step 5,

$$(\Pi(\rho^k(\varepsilon)))_\varepsilon\Big|_{\varepsilon=0} = \int \left(\tilde{\mathcal{W}}(\hat{x}, \omega, F|\tau, k) + O(\hat{x}, \omega, F|\tau, k)\right) dG(\omega, F),$$

where we used $LDCT$ to exchange the integral and the derivative, which is valid by the discussion immediately following (23). But, $(d(\rho^k(\varepsilon), \rho)_\varepsilon\big|_{\varepsilon=0}$ can take on values only in $\{-1, 0, 1\}$, since the effect of increasing the relevant set of x's is to either increase d at rate one, decrease d at rate one, or leave d unchanged. Hence, by Step 6, $\left|\int \tilde{\mathcal{W}}(\hat{x}, \omega, F|\tau, k) dG(\omega, F)\right| \leq 2d(\rho^k(\varepsilon), \rho) + 2\tau\mu$.

Step 8 We have $\lim_{k\to\infty} \psi^k_{j^l(\tau,k)-1} = \psi^l(\tau)$ and $\lim_{k\to\infty} \psi^k_{j^h(\tau,k)} = \psi^h(\tau)$.

Proof By construction, $\psi^k_{j^l(\tau,k)-1} = \inf\{\psi|\chi^k(\psi) \geq \hat{x} - \tau\}$. But $\psi^l(\tau) = \inf\{\psi|\chi(\psi) \geq \hat{x} - \tau\}$, and the first claim follows since almost everywhere convergence of χ^k to χ implies that, considered as a function of ψ alone, the sequence of increasing functions χ^k converges to χ in the Levy Metric. The other case is the same.

Step 9 Consider any (ω, F) such that $\bar{x}(\psi^l(\tau), \rho) = \underline{x}(\psi^l(\tau), \rho)$ (and so both equal $\hat{x} - \tau$). Then, $\lim_{k\to\infty} x^k_{j^l(\tau,k)-1} = \lim_{k\to\infty} x^k_{j^l(\tau,k)} = \hat{x} - \tau$, and

$$\lim_{k\to\infty} r^k_{j^l(\tau,k)-1} = r^l(\tau) \equiv \frac{\mathcal{S}_x(\hat{x} - \tau, \psi^l(\tau))}{v_{\psi x}(\hat{x} - \tau, \psi^l(\tau))}.$$

Similarly, if $\bar{x}(\psi^h(\tau), \rho) = \underline{x}(\psi^h(\tau), \rho)$ (and so both equal $\hat{x}+\tau$), then $\lim_{k\to\infty} x^k_{j^h(\tau,k)} = \lim_{k\to\infty} x^k_{j^h(\tau,k)+1} = \hat{x} + \tau$, and

$$\lim_{k\to\infty} r^k_{j^h(\tau,k)} = r^h(\tau) \equiv \frac{\mathcal{S}_x(\hat{x} + \tau, \psi^h(\tau))}{v_{\psi x}(\hat{x} + \tau, \psi^h(\tau))}.$$

Proof Following Step 8, and from the best response correspondence being upper hemicontinous, we obtain that $\lim_{k\to\infty} x^k_{j^l(\tau,k)-1} = \lim_{k\to\infty} x^k_{j^l(\tau,k)} = \hat{x} - \tau$. But then,

$$\lim_{k\to\infty} r^k_{j^l(\tau,k)-1} = \lim_{k\to\infty} \frac{\mathcal{S}(x^k_{j^l(\tau,\kappa)}, \psi^k_j) - \mathcal{S}(x^k_{j^l(\tau,k)-1}, \psi^k_j)}{v_\psi(x^k_{j^l(\tau,k)}, \psi^k_j) - v_\psi(x^k_{j^l(\tau,k)-1}, \psi^k_j)} = \frac{\mathcal{S}_x(\hat{x} - \tau, \psi^l(\tau))}{v_{\psi x}(\hat{x} - \tau, \psi^l(\tau))},$$

using $CMVT$. The case at $\psi^h(\tau)$ is the same.

Step 10 Consider any (ω, F) such that $\underline{x}(\psi^l(\tau), \rho) < \hat{x} - \tau < \bar{x}(\psi^l(\tau), \rho)$. Then, $\lim_{k\to\infty} x^k_{j^l(\tau,k)-1} = \underline{x}(\psi^l(\tau), \rho)$, $\lim_{k\to\infty} x^k_{j^l(\tau,k)} = \bar{x}(\psi^l(\tau), \rho)$, and

$$\lim r^k_{j^l(\tau,k)-1} = r^l(\tau) \equiv \frac{\mathcal{S}(\bar{x}(\psi^l(\tau), \rho), \psi^l(\tau)) - \mathcal{S}(\underline{x}(\psi^l(\tau), \rho), \psi^l(\tau))}{v_\psi(\bar{x}(\psi^l(\tau), \rho), \psi^l(\tau)) - v_\psi(\underline{x}(\psi^l(\tau), \rho), \psi^l(\tau))}.$$

Similarly, if $\underline{x}(\psi^h(\tau), \rho) < \hat{x} + \tau < \bar{x}(\psi^h(\tau), \rho)$, then, $\lim_{k\to\infty} x^k_{j^h(\tau,k)-1} = \underline{x}(\psi^h(\tau), \rho)$, $\lim_{k\to\infty} x^k_{j^h(\tau,k)+1} = \bar{x}(\psi^h(\tau), \rho)$, and

$$\lim r^k_{j^h(\tau,k)} = r^h(\tau) \equiv \frac{\mathcal{S}(\bar{x}(\psi^h(\tau), \rho), \psi^h(\tau)) - \mathcal{S}(\underline{x}(\psi^h(\tau), \rho), \psi^h(\tau))}{v_\psi(\bar{x}(\psi^h(\tau), \rho), \psi^h(\tau)) - v_\psi(\underline{x}(\psi^h(\tau), \rho), \psi^h(\tau))}.$$

Proof Note that by upper hemicontinuity and Step 8, any cluster point of $x^k_{j^l(\tau,k)-1}$ is a best response to ρ for $\psi^l(\tau)$ which is, by construction, at or below $\hat{x} - \tau$. But then by *2BRP*, it must be that this cluster point is $\underline{x}(\psi^l(\tau), \rho)$, and so $\lim_{k\to\infty} x^k_{j^l(\tau,k)-1} = \underline{x}(\psi^l(\tau), \rho)$. Similarly, $\lim_{k\to\infty} x^k_{j^l(\tau,k)} = \bar{x}(\psi^l(\tau), \rho)$. The claimed form for $\lim r^k_{j^h(\tau,k)}$ then follows immediately.

Step 11 Let $T(\tau, \omega, F) = 1$ if $\chi(\omega, \cdot, F)$ has either a jump ending at $\hat{x} - \tau$ or a jump beginning at $\hat{x} + \tau$, and zero otherwise. Let $Q(\tau) = \{(\omega, F) | T(\tau, \omega, F) = 0\}$. We claim that if $(\omega, F) \in Q(\tau)$ then $\lim_{k\to\infty} \tilde{\mathcal{W}}(\hat{x}, \omega, F|\tau, k)$ exists, is uniformly bounded, and (in an abuse of notation) is equal to

$$\begin{aligned} \tilde{\mathcal{W}}(\hat{x}, \omega, F|\tau) \quad \equiv \quad & -v_x(\bar{x}(\psi^h(\tau), \rho), \psi^h(\tau)) r^h(\tau) g(\psi^h(\tau)) \\ & + \int_{\psi^l(\tau)}^{\psi^h(\tau)} S_x(\chi(\psi), \psi) dG(\psi) + v_x(\underline{x}(\psi^l(\tau), \rho), \psi^l(\tau)) r^l(\tau) g(\psi^l(\tau)), \end{aligned}$$

where we remind the reader that all of the objects on the *rhs* depend on (ω, F).

Proof For given τ, $Q(\tau)$ is the set of (ω, F) such that either Step 9 or Step 10 applies, so that the various limiting objects are well-behaved. The result is then immediate from (23) and from Steps 8-10, with *LDCT* telling us that the limit can be passed through the integral.

Step 12 We claim that for almost all τ, the set $Q(\tau)$ has full measure. That is, $G_{\omega,F}(Q(\tau)) = 1$.

Proof For each (ω, F), $\chi(\omega, \cdot, F)$ jumps at most a countable number of times, and so there is at most a countable set of τ such that $T(\tau, \omega, F) = 1$. Hence, $\int_0^1 T(\tau, \omega, F) d\tau = 0$. But then, $\int \left(\int_0^1 T(\tau, \omega, F) d\tau \right) dG(\omega, F) = 0$, and so $\int_0^1 \left(\int T(\tau, \omega, F) dG(\omega, F) \right) d\tau = 0$. But then, for almost all $\tau \in [0, 1]$, $\int T(\tau, \omega, F) dG(\omega, F) = 0$, or equivalently, $G_{\omega,F}(Q(\tau)) = 1$.

Step 13 Let τ be such that $G_{\omega,F}(Q(\tau)) = 1$. Then, $\left| \int \tilde{\mathcal{W}}(\hat{x}, \omega, F; \tau) dG(\omega, F) \right| \leq 2\tau\mu$.

Proof By Step 1, $d(\rho^k, \rho) \to 0$. The result follows from Steps 7 and 11, once again invoking *LDCT*.

Step 14 Using Step 12, choose a sequence $\tau^n \to 0$ where for each n, $G_{\omega,F}(Q(\tau)) = 1$. Then, for all $(\omega, F) \in \cap_n Q(\tau^n)$ we have $\lim_{n\to\infty} \tilde{\mathcal{W}}(\hat{x}, \omega, F|\tau^n) = \mathcal{W}(\hat{x}, \omega, F)$.

Proof Fix $(\omega, F) \in \cap_n Q(\tau^n)$. Assume first that $\underline{x}(\omega, \psi^l, F, \rho) < \bar{x}(\omega, \psi^l, F, \rho) = \hat{x}$. Then, for all $\tau^n < \bar{x}(\omega, \psi^l, F, \rho) - \underline{x}(\omega, \psi^l, F, \rho)$, $\psi^l(\tau^n, \omega, F) = \psi^l(\omega, F)$, and so $\lim_{n\to\infty} r^l(\tau^n) = r^l$. If instead $\underline{x}(\omega, \psi^l, F, \rho) = \bar{x}(\omega, \psi^l, F, \rho) = \hat{x}$, then by upper hemicontinuity of the best response correspondence, we must have that $\lim_{n\to\infty} \underline{x}(\omega, \psi^l(\tau^n), F, \rho) = \lim_{n\to\infty} \bar{x}(\omega, \psi^l(\tau^n), F, \rho) = \hat{x}$, and so again $\lim_{n\to\infty} r^l(\tau^n) = r^l$. Similarly, in both relevant cases, $\lim_{n\to\infty} r^h(\tau^n) = r^h$.

Finally, consider $\underline{x}(\omega, \psi^l, F, \rho) < \hat{x} < \bar{x}(\omega, \psi^l, F, \rho)$. That is, at ψ^l, the consumer jumps from strictly below $\hat{x}$ to strictly above $\hat{x}$. Then, by *2BRP*, $\hat{x}$ is not a best response to ψ^l, and so changing $\hat{x}$ a small amount has no effect on ψ^l, and so for n large, no effect on $\psi^l(\tau^n, \omega, F) = \psi^l(\omega, F)$.

Step 15 The result that $\int \mathcal{W} dG(\omega, F) = 0$ then follows from Steps 13 and 14 and *LDCT*.

So, let us turn to $\int \mathcal{V} dG(\omega, F)$. Define ρ^k as in Step 1 above. Much as in Step 12, there is a set $Y \subseteq [0, 1]$ whose complement is countable, such that for each $x \in Y$ and for G-almost all (ω, F), $\bar{x}(\omega, \cdot, F, \rho)$ (or equivalently $\underline{x}(\omega, \cdot, F, \rho)$) does not have a jump beginning or ending at x.

Choose any $x \in Y$. Fix (ω, F) such that $\underline{x}(\omega, \cdot, F, \rho)$ does not have a jump beginning or ending at x. As in the proof of Theorem 1, also choose (ω, F) such that neither neither $\Delta(\omega, \underline{\psi}, F, \rho) = 0$ nor $\Delta(\omega, \bar{\psi}, F, \rho) = 0$, where we make explicit the dependence of Δ on the premium schedule. Note that by continuity, it follows that for k large enough, if $\Delta(\omega, \underline{\psi}, F, \rho) > 0$ then also $\Delta(\omega, \underline{\psi}, F, \rho^k) > 0$ and hence facing ρ^k, (ω, F) chooses above x regardless of ψ, and has a strict preference for doing so. Hence, the appropriate $r(x, \omega, F)$ is zero both for ρ and for ρ^k. The situation is similar if $\Delta(\omega, \bar{\psi}, F, \rho) < 0$.

So, in what follows, let us concentrate on the interesting case where $\Delta(\omega, \underline{\psi}, F, \rho) < 0 < \Delta(\omega, \bar{\psi}, F, \rho)$ so that there is for large enough k an interior risk aversion parameter where the optimal action shifts from x or below to above x. Suppressing (ω, F), define $\psi^k(x)$ as $\max\{\psi|\underline{x}(\psi, \rho^k) \le x\}$ and $\psi^*(x)$ as $\max\{\psi|\underline{x}(\psi, \rho) \le x\}$. Let $\underline{x}^k(x) = \underline{x}(\psi^k(x), \rho^k)$ and $\bar{x}^k(x) = \bar{x}(\psi^k(x), \rho^k)$, noting that because ρ^k is a step function, $\bar{x}^k(x) > \underline{x}^k(x)$. Similarly, let $\underline{x}^*(x) = \underline{x}(\psi^*(x), \rho)$ and $\bar{x}^*(x) = \bar{x}(\psi^*(x), \rho)$. Then, it follows from the upper hemicontinuity of X that $\psi^k(x) \to \psi^*(x)$, $\underline{x}^k(x) \to \underline{x}^*(x)$, and $\bar{x}^k(x) \to \bar{x}^*(x)$. To see this, assume first that $\underline{x}^*(x) < \bar{x}^*(x)$ so that there is a jump in $\underline{x}(\cdot, \rho)$ at $\psi^*(x)$. Then, $x \in (\underline{x}^*(x), \bar{x}^*(x))$ by choice of (ω, F). Thus, along any convergent subsequence, $\underline{x}^*(x) < x \le \lim_{k\to\infty} \bar{x}^k(x) \in X(\psi^*(x), \rho)$, and so by *2BRP*, $\lim_{k\to\infty} \bar{x}^k(x) = \bar{x}^*(x)$. Similarly, $\lim_{k\to\infty} \underline{x}^k(x) = \underline{x}^*(x)$. If instead $\underline{x}^*(x) = \bar{x}^*(x) = x$ then since along any convergent subsequence, $\lim_{k\to\infty} \underline{x}^k(x) \in X(\psi^*(x), \rho) = \{x\}$, we have that $\lim_{k\to\infty} \underline{x}^k(x) = x$, and similarly $\lim_{k\to\infty} \bar{x}^k(x) = x$.

It follows that if we define

$$\tilde{r}^k(x) \equiv \frac{\mathcal{S}(\bar{x}^k(x), \psi^k(x)) - \mathcal{S}(\underline{x}^k(x), \psi^k(x))}{v_\psi(\bar{x}^k(x), \psi^k(x)) - v_\psi(\underline{x}^k(x), \psi^k(x))}$$

then $\tilde{r}^k(x) \to_k r^*(x)$, where

$$r^*(x) \equiv \frac{\mathcal{S}(\bar{x}^*(x), \psi^*(x)) - \mathcal{S}(\underline{x}^*(x), \psi^*(x))}{v_\psi(\bar{x}^*(x), \psi^*(x)) - v_\psi(\underline{x}^*(x), \psi^*(x))} \text{ or } \frac{\mathcal{S}_x(x, \psi^*(x))}{v_{\psi x}(x, \psi^*(x))} \tag{24}$$

as appropriate, and use *CMVT* when $\bar{x}^k(x) - \underline{x}^k(x) \to 0$.

Say that x' is offered by ρ^k if $\rho^k(x'') > \rho^k(x')$ for all $x'' > x'$. Note that a non-offered contract is never a best response for the consumer, since they can have more coverage at the same price. Let $\tilde{x}^k(x)$ be the largest quality offered by ρ^k that is at or below x. Let us show that $\psi^k(\tilde{x}^k(x)) = \psi^k(x)$. To see this, recall that $\psi^k(x) = \max\{\psi | \underline{x}(\psi, \rho^k) \leq x\}$ and $\psi^k(\tilde{x}^k(x)) = \max\{\psi | \underline{x}(\psi, \rho^k) \leq \tilde{x}^k(x)\}$. Thus, $\psi^k(\tilde{x}^k(x)) \leq \psi^k(x)$. Assume $\psi^k(\tilde{x}^k(x)) < \psi^k(x)$. Then, for $\psi \in (\psi^k(\tilde{x}^k(x)), \psi^k(x))$ we have that $\underline{x}(\psi, \rho^k) > x$, since $\tilde{x}^k(x)$ is the largest offered quality at or below x and by definition of $\psi^k(\tilde{x}^k(x))$, any $\psi > \psi^k(\tilde{x}^k(x))$ has a lowest best response strictly above $\tilde{x}^k(x)$. Hence, since there are no contracts offered between $\tilde{x}^k(x)$ and x, it is strictly above x. But $\underline{x}(\psi, \rho^k) \leq x$ by definition of $\psi^k(x)$, which is a contradiction. Thus, $\psi^k(\tilde{x}^k(x)) = \psi^k(x)$, and so

$$-\tilde{r}^k(x) g(\psi^k(x)) + 1 - G(\psi^k(x)) = -\tilde{r}^k(\tilde{x}^k(x)) g(\psi^k(\tilde{x}^k(x))) + 1 - G(\psi^k(\tilde{x}^k(x))).$$

Then, reinstating (ω, F), we have by the result for the finite case that

$$\int [-\tilde{r}^k(\tilde{x}^k(x), \omega, F) g(\psi^k(\tilde{x}^k(x), \omega, F)) + 1 - G(\psi^k(\tilde{x}^k(x), \omega, F))] dG(\omega, F) = 0,$$

and so

$$\int [-\tilde{r}^k(x, \omega, F) g(\psi^k(x, \omega, F)) + 1 - G(\psi^k(x, \omega, F))] dG(\omega, F) = 0.$$

But, the integrand in this expression is uniformly bounded, and so, by *LDCT*, we have that

$$\int \mathcal{V}(x, \omega, F) dG(\omega, F) = \int [-r^*(x, \omega, F) g(\psi^*(x, \omega, F)) + 1 - G(\psi^*(x, \omega, F))] dG(\omega, F) = 0,$$

as claimed. □

B.6 Ironing and the One-Contract Case

We asserted in Section 4.4 that the optimality condition that obtains from the perturbation of one contract x reduces, in the one-dimensional case where only ψ is stochastic, to the standard ironing condition. In this case the optimality condition is simply $\mathcal{W} = 0$, and so for each offered $x > x^0$,

$$- v_x(x, \psi^h) r^h g(\psi^h) + \int_{\psi^l}^{\psi^h} (\mathcal{S}_x(x, \psi) - v_x(x, \psi)) g(\psi) d\psi + v_x(x, \psi^l) r^l g(\psi^l) = 0. \tag{25}$$

From the perturbation of the price schedule, we obtain in this case that $\mathcal{V} = 0$, or $r^h g(\psi^h) = 1 - G(\psi^h)$ and $r^l g(\psi^l) = 1 - G(\psi^l)$, and thus (25) becomes

$$-v_x(1 - G(\psi^h)) + \int_{\psi^l}^{\psi^h} (\mathcal{S}_x(x,\psi) - v_x(x,\psi))g(\psi)d\psi + v_x(x,\psi^l)(1 - G(\psi^l)) = 0. \tag{26}$$

Integrating by parts $-\int_{\psi^l}^{\psi^h} v_x g d\psi$ and then multiplying and dividing the integrand by g yields

$$\begin{aligned}
-\int_{\psi^l}^{\psi^h} v_x(x,\psi)g(\psi)d\psi &= (1-G)v_x(x,\psi)|_{\psi^l}^{\psi^h} - \int_{\psi^l}^{\psi^h} v_{x\psi}\frac{1-G}{g} g d\psi \\
&= (1 - G(\psi^h))v_x(x,\psi^h) - (1 - G(\psi^l))v_x(x,\psi^l) - \int_{\psi^l}^{\psi^h} v_{x\psi}\frac{1-G(\psi)}{g(\psi)}g(\psi)d\psi.
\end{aligned}$$

Inserting this expression into (26) and rearranging yields

$$\int_{\psi^l}^{\psi^h}\left(\mathcal{S}_x(x,\theta) - v_{x\psi}\frac{1-G(\psi)}{g(\psi)}\right)g(\psi)d\psi = 0,$$

which is the standard optimality condition in the ironing case.

Consider now the multidimensional case with just one contract (p, x), as in Veiga and Weyl (2016). In this case, simple algebra reveals that $\mathcal{W}$ reduces to

$$\mathcal{W}(x,\omega,F) = -\int_{\psi^l}^{\bar{\psi}} \gamma_x^I(x,\psi)dG(\psi) + v_x(x,\psi^l)\frac{p - \gamma^I(x,\psi^l)}{v_\psi(x,\psi^l) - v_\psi(x^0,\psi^l)}g(\psi^l), \tag{27}$$

where $p = v(x,\psi^l) - v(x^0,\psi^l)$ since type (w,ψ^l,F) is indifferent between choosing x and x^0.

In turn, $\mathcal{V}$ reduces to

$$\mathcal{V}(x,\omega,F) = 1 - G(\psi^l) - \frac{p - \gamma^I(x,\psi^l)}{v_\psi(x,\psi^l) - v_\psi(x^0,\psi^l)}g(\psi^l). \tag{28}$$

It is easy to show that $\int \mathcal{V}(x,\omega,F)dG(\omega,F) = 0$ is the same as the first-order condition with respect to p in Veiga and Weyl (2016), and is given by

$$p - \int \gamma^I(x,\psi^l)dR(\omega,F,x,x^0) = \frac{N}{s}, \tag{29}$$

where $N = \int(1 - G(\psi^l|\omega,F))dG(\omega,F)$ is the total mass of types served by the firm, s is the mass of types that switch from x to outside option x^0, given by

$$s = \int \frac{g(\psi^l(\omega,F)|\omega,F)}{v_\psi(x,\psi^l(\omega,F),\omega,F) - v_\psi(x^0,\psi^l(\omega,F),\omega,F)}dG(\omega,F),$$

and R is the cdf of types that switch, and can be obtained by integrating its density r given by

$$r(\omega, F, x, x^0) = \frac{\frac{g(\psi^l(\omega,F)|\omega,F)}{v_\psi(x,\psi^l(\omega,F),\omega,F)-v_\psi(x^0,\psi^l(\omega,F),\omega,F)}}{s}.$$

From (29), $p = \int \gamma^I dR + (N/s)$. Inserting this expression for p into (27), integrating the resulting expression with respect to ω, F, and manipulating yields that $\int \mathcal{W} dG = 0$ can be written as follows:

$$0 = -\int \gamma_x^I(x, \psi^l) \frac{1 - G(\psi^l|\omega, F)}{N} dG(\omega, F) + \int v_x(x, \psi^l) dQ(\omega, F, x, x^0) - \frac{cov_r(v_x, \gamma^I)}{\frac{N}{s}},$$

where $cov_r(v_x, \gamma^I)$ is the covariance between v_x and γ^I calculated using the density r of switching types, and is the same as the first-order condition with respect to x in Veiga and Weyl (2016).

B.7 Incentives to Exclude and Screen

We mentioned in Section 4.6 that the optimality condition of our main perturbation can be used to shed light on the insurer's incentives to exclude types from any insurance above x^0, and also to screen types. Here we present the analytical support for that comment.

Incentives to Exclude. By varying the weights w, our optimality conditions highlight the differential incentives of insurers with different objectives. We now show that the monopolist has a greater incentive to exclude consumers than the social planner. Fix a level x^0 of government-provided insurance. We start with the incentives of a monopolist insurer. To simplify notation, assume that any consumer who is taking the outside option has the lowest offered level of incremental coverage as their second best choice, and denote this contract by x^1. Fix and suppress x^0, ω, and F, and let the marginal type who is excluded by the monopolist be ψ^* (since $v_{x\psi} > 0$, the set of types excluded is an interval beginning at $\psi = 0$). Then, since for the monopolist, $w^I = 1$ while $w^C = w^G = 0$, it is easy to show that $\mathcal{V}$, the effect on payoffs of an increase in the premium of all contracts x^1 and above (and hence of moving some people from an inside option to the outside option x^0) is given by

$$\mathcal{V}^M = -\frac{\rho(x^1) - (\gamma^I(x^1) - \gamma^G(x^1, x^0))}{v_\psi(x^1, \psi^*) - v_\psi(x^0, \psi^*)} g(\psi^*) + 1 - G(\psi^*),$$

where the superscript M stands for monopolist. Optimal exclusion requires that $\int \mathcal{V}^M(x^0, \omega, F) dG(\omega, F) = 0$.[45] The term $\rho(x^1) - (\gamma^I(x^1) - \gamma^G(x^1, x^0))$ represents the profit the insurer was making on con-

[45]For a monopolist, $\mathcal{S}(x, \theta) \equiv v(x, \theta) - \gamma^I(x, \theta) + \gamma^G(x, x^0, \theta)$, and so

$$\mathcal{S}(x^1, \psi^*) - \mathcal{S}(x^0, \psi^*) = v(x^1, \psi^*) - \gamma^I(x^1) + \gamma^G(x^1, x^0) - \left(v(x^0, \psi^*) - \gamma^I(x^0) + \gamma^G(x^0, x^0)\right).$$

But, $v(x^1, \psi^*) - v(x^0, \psi^*) = \rho(x^1)$ and $\gamma^G(x^0, x^0) = \gamma^I(x^0)$, and so $\mathcal{S}(x^1, \psi^*) - \mathcal{S}(x^0, \psi^*) = \rho(x^1) - \gamma^I(x^1) + \gamma^G(x^1, x^0)$, and the expression follows by substituting into (7).

sumers it now excludes. The other parts of the first term reflect the speed at which types are excluded as premiums are raised. The last part of the expression $1 - G$ is the impact on revenue from inframarginal consumers.[46]

From a regulator's perspective, is the monopolist excluding too little or too much? To answer this question, consider the setting where $w^I = w^C = 1 \leq w^G$, so that the regulator equally weights consumer surplus and monopolist profits, and respects any excess cost of public funds. Under these weights, the effect of increasing premiums on all contracts x^1 and above is

$$\mathcal{V}^G = -\frac{\rho(x^1) - (\gamma^I(x^1) - \gamma^I(x^0)) - (w^G - 1)(\gamma^G(x^1, x^0) - \gamma^I(x^0))}{v_\psi(x^1, \psi^*) - v_\psi(x^0, \psi^*)} g(\psi^*),$$

where the superscript G stands for "government", and $\rho(x^1) - (\gamma^I(x^1) - \gamma^I(x^0))$ measures the change in consumers' willingness to pay less the cost of serving them, while $(w^G - 1)(\gamma^G(x^1, x^0) - \gamma^I(x^0))$ measures the cost of increased government spending.[47] Note that $\gamma^G(x^1, x^0) \geq \gamma^G(x^0, x^0) = \gamma^I(x^0)$.

Comparing the impact of incremental exclusion from the perspective of the monopolist versus the utilitarian regulator yields,

$$\mathcal{V}^M - \mathcal{V}^G \equiv -w^G \frac{\gamma^G(x^1, x^0) - \gamma^I(x^0)}{v_\psi(x^1, \psi^*) - v_\psi(x^0, \psi^*)} g(\psi^*) + 1 - G(\psi^*). \tag{30}$$

The first term is negative and reflects the social cost of increased government spending that arises when consumers receive higher coverage. The second term is positive and reflects that the monopolist values transfers from the consumer while the regulator is indifferent. Overall the comparison is ambiguous. Because γ^G depends on consumers' behavior in their chosen contract (in this case x^1), the monopolist in effect does not bear the full cost of additional healthcare spending due to higher coverage. This subsidy encourages the monopolist to serve more consumers than it otherwise would. However, under the alternative rule where government spending depends only on consumers' behavior had they chosen x^0—i.e., when γ^G is fixed at $\gamma^I(x^0)$—then the first term cancels out, and the regulator unambiguously wants the monopolist to exclude fewer consumers.

Incentives to Screen. Marone and Sabety (2022) provide an empirical illustration where the social planner chooses to pool all consumers in a single contract, and our numerical analysis also provides a strong incentive for the planner to pool types. We now provide a theoretical example to illustrate how nonresponsiveness in the planner's problem can drive this outcome, as well as how it contrasts with the outcome that would be chosen by a monopolist.[48] To gain traction analytically,

[46] In the case where ω and F are not stochastic (the one-dimensional case), optimal exclusion requires that $\mathcal{V}^M = 0$, which rearranges to the classic "virtual profit" condition.

[47] In this case, $\mathcal{S} = v - \gamma^I - (w^G - 1)\gamma^G$, and so since $v(x^1, \psi^*) - v(x^0, \psi^*) = \rho(x^1)$,

$$\mathcal{S}(x^1, \psi^*) - \mathcal{S}(x^0, \psi^*) = \rho(x^1) - (\gamma^I(x^1) - \gamma^I(x^0)) - (w^G - 1)(\gamma^G(x^1, x^0) - \gamma^I(x^0))$$

and the expression for $\mathcal{V}^{2,G}$ follows.

[48] Nonresponsiveness holds when, as a function of the consumer's type, the allocation of contracts to types that is incentive compatible has the opposite monotonicity property than the efficient allocation.

we limit attention to the one-dimensional problem. Specifically, we assume that the consumer's only private information is ω, and that the distribution of ψ and F is degenerate. We also restrict attention to linear out-of-pocket cost functions of the form $c(a,x) = (1-x)a$, and assume that $b(a,l,\omega) \equiv \hat{b}(a-l,\omega)$.[49] Finally, for simplicity, we assume that $\gamma^G = 0$ and that the social planner assigns the same weight to the insurer and to the consumer.

The assumptions on c and b yield a very convenient expression for $v(x,\omega)$ (we omit ψ and F from θ since they are fixed in this section). To see this, note first that from $\hat{b}_a(a-l,\omega) = c_a(a,x) = 1-x$, we obtain $a^*(l,x,\omega) = l + \varphi(1-x,\omega)$, where φ is the inverse of $\hat{b}_a$ with respect to its first argument. Inserting the optimal choice of a into v we obtain

$$v(x,\omega) = \hat{b}(\varphi(1-x,\omega),\omega) - (1-x)\varphi(1-x,\omega) - \frac{1}{\psi}\log\int e^{\psi(1-x)l}dF(l),$$

and

$$v_x(x,\omega) = \varphi(1-x,\omega) - \frac{1}{\psi}\log\int e^{-\psi l}dF(l), \tag{31}$$

which yields $v_{x\omega} = \varphi_\omega(1-x,\omega) > 0$, as discussed in Technical Remark 4.

Consider first the social planner's problem without adverse selection (the 'first-best' case). Since $a - c(a,x) = xa$ in this case, the planner solves, for each ω,

$$\max_{x\in[0,1]}\left(v(x,\omega) - x\int a^*(l,x,\omega)dF(l)\right).$$

Using $a^*(l,x,\omega) = l + \varphi(1-x,\omega)$ and (31), we obtain that the cross-partial derivative of the objective function with respect to (x,ω) is $x\varphi_{(1-x)\omega}(1-x,\omega)$.[50] One can show that this is *strictly negative* for all $x > 0$ if $\hat{b}_{a\omega}/\hat{b}_{aa}$ is strictly decreasing in a, a condition that is satisfied by the canonical example.[51] By a standard monotone comparative statics argument, this implies that the efficient allocation of contracts to types in the first best is *decreasing* in ω.

But in this case a necessary condition for $\chi(\cdot)$ to be incentive compatible is that it be *increasing* in ω.[52] It follows from this conflicting monotonicity that when ω is the only source of private information, the social planner's optimal allocation of contracts to types is "flat."

[49] Under this out-of-pocket cost functions, we know that $v_{x\omega} > 0$, and in the parametrization used in the numerical simulations we obtain closed-form solutions for a^* and z.

[50] To see this, note that $v_{x\omega} = \varphi_\omega(1-x,\omega)$, $\int a^*_\omega dF = \varphi_w$, and $\int a^*_{x\omega}dF = -\varphi_{(1-x)\omega}$. Inserting these expressions into the cross-partial derivative of the objective function we obtain $v_{x\omega} - \int a^*_\omega dF - x\int a^*_{x\omega}dF = x\varphi_{(1-x)\omega}$.

[51] To see this, from $\hat{b}_a(\varphi(1-x,\omega),\omega) = 1-x$, differentiate twice and use the derivative of the inverse function φ to obtain $\varphi_{(1-x)\omega} = -(\hat{b}^3_{aa})^{-1}(\hat{b}_{aa\omega}\hat{b}_{aa} - \hat{b}_{aaa}\hat{b}_{a\omega})$, and this is strictly negative if the term in parenthesis is, that is, when $\hat{b}_{a\omega}/\hat{b}_{aa}$ is strictly decreasing in a. If $b(a,l,\omega) = a - l - (1/(2\omega))(a-l)^2$, then $b_{a\omega}/b_{aa} = -(a-l)/\omega$, which is clearly strictly decreasing in ω.

[52] The standard incentive compatibility characterization states that χ is incentive compatible if and only if it is increasing and the consumer's indirect utility when her type is ω, $U(\omega) = v(\chi(\omega),\omega) - \rho(\chi(\omega))$ can be written as $U(\omega) = U(\underline{\omega}) + \int_{\underline{\omega}}^{\omega} v_\omega(\chi(s),s)ds$.

Proposition 5 (Social Planner and Pooling) *Assume that only ω is private information, that $b(a,l,\omega) = \hat{b}(a-l,\omega)$, that $b_{a\omega}/b_{aa}$ is strictly decreasing in a for each (l,ω), and that $c(a,x) = (1-x)a$. Then the optimal χ for the social planner entails complete pooling of types.*

Consider now the profit-maximizing monopolist's problem. After some algebra that is standard in screening with one-dimensional private information, the monopolist's problem becomes

$$\max_{\chi(\cdot)} \left(\int \left(v(\chi(\omega),\omega) - x \int a^*(l,\chi(\omega),\omega) dF(l) - v_\omega(\chi(\omega),\omega) \frac{1-G(\omega)}{g(\omega)} \right) dG(\omega) - v(x^0, \underline{\omega}) \right)$$

s.t. χ increasing.[53] If we ignored the monotonicity constraint, we could maximize, for each ω,

$$v(x,\omega) - x \int a^*(l,x,\omega) dF(l) - v_\omega(x,\omega) \frac{1-G(\omega)}{g(\omega)} \tag{32}$$

with respect to x. If this expression had a strictly negative cross-partial derivative with respect to (x,ω), then once again we would have complete pooling.[54] But, unlike the planner's objective function, whose cross-partial is strictly negative when $\hat{b}_{a\omega}/\hat{b}_{aa}$ is strictly decreasing in a, we have an extra term, $-v_\omega(x,\omega)((1-G(\omega))/g(\omega))$. As a result, the cross-partial derivative of (32) is

$$x\varphi_{(1-x)\omega} - v_{x\omega\omega} \frac{1-G}{g} - v_{x\omega} \left(\frac{1-G}{g} \right)_\omega = x\varphi_{(1-x)\omega} - \varphi_{\omega\omega} \left(\frac{1-G}{g} \right) - \varphi_\omega \left(\frac{1-G}{g} \right)_\omega . \tag{33}$$

To see that this expression need not be strictly negative, assume $\hat{b}(a-l,\omega) = a-l-(1/(2\omega))(a-l)^2$, and that g is a strictly increasing density with $g' > 0$. Then one can show that (33) is actually *strictly positive*, which implies that the monopolist *completely sorts* types at the optimal menu, providing a drastic contrast with the social planner's solution.[55]

B.8 Omitted Algebra from Section 7

The payoff to the government with subsidy scheme s is

$$\underbrace{w^C \int_0^{q(s)} P(q')dq' - w^I C(q(s)) + (w^I - w^C) P(q(s))q(s) + w^I \sigma(q(s)|s)}_{\beta(s)} - w^G \sigma(q(s)|s), \tag{34}$$

[53] The algebraic steps are as follows. First, use the incentive compatibility characterization in Footnote 52 to write the monopolist's problem as $\max_{\rho(\chi(\cdot)),\chi(\cdot)} \left(\int \left(\rho(\chi(\omega)) - x \int a^*(l,\chi(\omega),\omega) dF(l) \right) dG(\omega) \right)$ subject to χ increasing and $\rho(\chi(\omega)) = v(\chi(\omega),\omega) - v(x^0,\underline{\omega}) - \int_{\underline{\omega}}^{\omega} v_\omega(\chi(s),s)ds$, where we have set $U(\underline{\omega}) = v(x^0,\underline{\omega})$, which is optimal for the monopolist. Second, insert the expression for ρ into the objective function. Finally, integrate by parts a double integral that appears after the replacement and rearrange.

[54] The solution to the relaxed problem (which ignores the monotonicity constraint) would be decreasing, and thus there would be a need for ironing, which would yield a flat allocation of contracts to types.

[55] To see that (33) is strictly positive, note that from $\hat{b}_a = 1-x$ we obtain $\varphi(1-x,\omega) = \omega(1-(1-x))$, and thus (33) is equal to $-x - 0 + x((g'/g)((1-G)/g) + 1)$, which is strictly positive if $g' > 0$.

where $\beta(s)$ is the benefit of the subsidy and $w^G\sigma(q(s)|s)$ is the cost. Differentiating the government's payoff function (34) we obtain

$$\left(w^C P(q(s)) - w^I MC(q(s)) + (w^I - w^C)MR(q(s))\right) q_s(s) + (w^I - w^G)(\sigma_q(q(s)|s)q_s(s) + \sigma_s(q(s)|s)),$$

which has the same sign as

$$\frac{\left(w^C P(q(s)) - w^I MC(q(s)) + (w^I - w^C)MR(q(s))\right) q_s(s)}{\sigma_q(q(s)|s)q_s(s) + \sigma_s(q(s)|s)} + w^I - w^G = \frac{\beta_s(s)}{\sigma_q(q(s)|s)q_s(s) + \sigma_s(q(s)|s)} - w^G$$

Evaluating the last expression at $s = 0$, and recalling that $q(0) = q^m$ and $\sigma_q(q|0) = 0$ we obtain equation (12). And since at q^m we have $MR = MC$, we obtain

$$\text{BFFB} = w^C \frac{P(q^m) - MC(q^m)}{\sigma_s(q^m|0)} q_s(0) + w^I.$$

But, differentiating the identity $MR(q(s)) - (MC(q(s)) - \sigma_q(q(s)|s))) = 0$ yields

$$q_s(0) = \frac{\sigma_{qs}(q^m|0)}{MC_q(q^m) - MR_q(q^m)},$$

and so

$$\text{BFFB} = w^C(P(q^m) - MC(q^m)) \frac{1}{MC_q(q^m) - MR_q(q^m)} \frac{\sigma_{qs}(q^m|0)}{\sigma_s(q^m|0)} + w^I$$

To see the intuition for this equation, start from the right and work left, and note first that w^I reflects the direct benefit to the government of the fact that the monopolist receives a dollar of subsidies. Since the monopolist has chosen price optimally, their payoff is otherwise unchanged. Next, σ_{qs}/σ_s is the amount by which the first dollar spent on subsidies lowers the effective marginal cost of the monopolist. The term $1/(MC_q - MR_q)$ reflects the amount by which a decrease in the marginal cost of the monopolist leads the monopolist to increase output. Finally, the effect of a change in output by the monopolist on the surplus of the consumer is $P - MR$, which, since we are at the monopolist's optimum is the same as $P - MC$.

Now, $P - MC = -P_q q^m$ from the standard monopoly optimum formula, while $MR_q = 2P_q + q^m P_{qq}$, and so, substituting and manipulating,

$$\text{BFFB} = w^C q^m \frac{\sigma_{qs}}{\sigma_s} \frac{1}{2 + \left(\frac{MC_q}{-P_q} + q^m \frac{P_{qq}}{P_q}\right)} + w^I,$$

which is the expression in the text. □