Multidimensional Screening and Menu Design in Health Insurance Markets

Hector Chade, Victoria Marone, Amanda Starc, and Jeroen Swinkels*

December 2023

Abstract

We study a general screening model that encompasses the problem facing a price-setting insurer offering vertically differentiated contracts to consumers with multiple dimensions of private information. We show how even with minimal assumptions on consumer valuations and costs, progress can be made to understand the solution in two ways. First, we derive conditions that any optimal menu must satisfy, and show how they can be used to shed light on insurer incentives. Second, we propose a tractable method to approximate the solution, and show how the quality of the approximation can be ex-post evaluated in any practical application. Applying our method empirically in the context of health insurance, we find that the approximation comes within one percent of the true solution. We illustrate the usefulness of the approximation for understanding the solution graphically as well as for numerically evaluating optimal policy interventions in a monopoly market. Our analysis highlights the importance of strategic insurer responses and endogenous contract characteristics in evaluating the effects of policy in these markets.

JEL: I11, C26, I13

*We are grateful to Zack Cooper, Stuart Craig, Michael Dickstein, David Dranove, Eilidh Geddes, Ben Handel, Kate Ho, Amanda Kowalski, Tim Layton, Ilse Lindenlaub, Neale Mahoney, Alessandro Pavan, Rob Porter, Mike Powell, Mark Shepard, Ashley Swanson, Ben Vatter, Mike Whinston and seminar participants at Northwestern, the Utah Winter Business Economics Conference, Stanford, UCLA, Paris School of Economics, Michigan State, Penn, the NBER IO program, University of Nevada-Reno, Arizona State, Karlsruhe, HSBC Business School, and Penn State for their helpful comments.

Chade: hector.chade@asu.edu, Arizona State University; Marone: marone@utexas.edu, University of Texas at Austin and NBER; Starc: amanda.starc@kellogg.northwestern.edu, The Kellogg School of Management, Northwestern University, and NBER; Swinkels: j-swinkels@kellogg.northwestern.edu, The Kellogg School of Management, Northwestern University.

1 Introduction

Asymmetric information is central to our understanding of major sectors 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, health insurers sell insurance contracts to consumers who know both their health status and their taste for insurance. In each of these settings, agents hold multiple important dimensions of private information, and the space of possible contracts across which they 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 (e.g., Rothschild and Stiglitz, 1976; Stiglitz, 1977), or else a multidimensional problem with a highly restricted contract space (e.g., Einav et al., 2010a; Veiga and Weyl, 2016).

One reason for this dearth of evidence is, of course, that screening problems become substantially less tractable with both multidimensional agents and endogenous characteristics across many potential products. Since Wilson (1993), theorists have attempted to characterize optima of multidimensional screening models of this type. But despite being a natural and highly relevant generalization of the single-dimensional agent case, sufficient conditions for optimality have proved elusive. Existing results rely on assumptions imposed on consumer utility as a function of their type and allocation. Unfortunately, these assumptions fail in many applied settings. In health insurance, for example, the utility function is derived from the certainty equivalent of a lottery over health outcomes, which is not likely to have the relevant properties.[1] As a result, once one takes seriously both that consumers can vary along several private dimensions and that insurers may offer new contracts in response to this heterogeneity, we have only a limited understanding of optimal menus in these markets.

This paper asks what can be learned about the solution to multidimensional screening problems that are sufficiently general to accommodate real-world applications. We make progress on two fronts. First, we derive intuitive necessary conditions for a solution in a setting with only minimal assumptions on primitive objects. We show how these conditions can nonetheless be used to derive interesting properties of the solution. Second, we provide a highly tractable method with which the solution can be approximated, and show how the quality of the approximation can be ex-post evaluated in any practical application. We view this latter approach as a tool that may be especially useful for applied researchers in this area. Indeed, menu design in health insurance markets is a topic on which the progress of empirical analysis appears to have outpaced theoretical advancement.[2] While endogenous determination of contract characteristics has been recognized as a centrally important force, applied researchers must often abstract from it in order to make empirical progress. Our approach accommodates this force, permitting tractable analysis in settings that have traditionally been considered prohibitively complex.

[1]The standard simplifying assumption in the theoretical literature is that consumers' valuation of product quality or quantity exhibits a single crossing with respect to consumer type. This assumption is rejected by standard parameterizations and estimates of consumer demand in health insurance markets (Marone and Sabety, 2022).

[2]As noted by Einav and Finkelstein (2011a) (and emphasized by Veiga and Weyl, 2016), "On the theoretical front, we currently lack clear characterizations of the equilibrium in a market in which firms compete over contract dimensions as well as price, and in which consumers may have multiple dimensions of private information."

Our model 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 dictates the extent of coverage for different levels of healthcare utilization. The consumer's outside option is an exogenous base level of coverage provided by the government (which may be no coverage). The insurer's payoff is a weighted average of consumer surplus, profits, and government spending. Our model thus subsumes a range of cases, from a monopolist insurer to a utilitarian social planner. The timing is as follows. The insurer offers a menu of contracts. Consumers observe the menu, learn their type, and choose a contract. Consumers then privately learn their health state and choose their healthcare utilization.

We begin by deriving conditions that any optimal menu must satisfy. The key assumption required is that consumer utility exhibits single crossing with respect to contract quality and *some* dimension of the consumer's type. There are no restriction on the interaction between consumer utility and all remaining dimensions of consumers' private information. We then show how these necessary conditions can be used to establish properties of the solution, related to the insurers' incentive to screen and exclude consumers as well as the existence of positive trade in the market. The flexibility of our model with respect to the insurer's objective function also permits comparative statics along this dimension. We show that an insurer aiming to maximize social surplus (a planner) will generically exclude fewer consumers than an insurer maximizing only producer surplus (a monopolist).[3]

While our necessary conditions provide useful extensions of standard screening results to a multidimensional setting, the problem lacks sufficient structure to allow full characterization. Following Wilson (1993), we make progress through a novel application of the "demand profile" approach. The approach relies on a reinterpretation of the problem. Instead of setting the premium of each contract, suppose the insurer instead sets the *incremental* premium of each *incremental* level of coverage. A full premium schedule can then be reverse engineered, accumulating over all optimal incremental premiums. Recast in this way, the insurer's problem may fully "decouple," in the sense that the optimality condition for each incremental premium is independent of all other increments. Decoupling substantially simplifies analysis of the problem, yielding a potentially powerful tool for exploring its properties.

This approach relies critically on an enumerable set of incremental coverage levels, and thus on a fixed and finite set of possible contracts. Before proceeding, we therefore first establish a novel link between two key versions of the menu design problem: (i) the base case, in which the insurer can offer a continuum of contracts, in principle achieving full separation across types (Stiglitz (1977), Mussa and Rosen (1978)),Maskin and Riley (1984)); and (ii) the case in which the insurer is restricted to offering a fixed and finite set of contracts, on which it simply sets prices. In the

[3]For clarity with respect to the application, we state the model in the health insurance context. More generally, the analysis applies to a setting in which consumers with a multidimensional type derives utility from consuming a good of different qualities (or quantities). A single firm supplies the good, at cost that depends on the good's quality and can as well depend on the consumer's type. Both value and cost are increasing in quality, and satisfy a suitable single-crossing property in the good's quality and in one of the dimensions of type. The firm chooses an optimal price schedule, which specifies a price for each level of quality.

latter case, there may be at a minimum simply one fixed contract plus an outside option, as in Akerlof (1970) and Einav et al. (2010a). We show that as the number of fixed contracts increases, an insurer's optimal payoff—as well as the associated price schedule and allocation—converges to its counterpart in the continuum case. This result provides an important missing link between settings in which product characteristics are "fully endogenous" (the continuum case) and settings in which product characteristics are pre-specified. As long as firms can offer a sufficiently large number of contracts, these two versions of the problem are *not distinct*. The characteristics of traded contracts are endogenous in either case.

Proceeding with a fixed set of contracts, the critical questions then become *when* the insurer's problem will fully decouple, and if it does not, *how close* will the reverse engineered solution be to the true solution? The key to understanding these questions is the relationship between the optimal price schedule and consumer demand. The problem will fully decouple only if all consumers' payoffs are quasiconcave in coverage level under the optimal price schedule. With quasiconcave payoffs, consumers will only consume incremental coverage up to the point at which its incremental value becomes negative. When anticipating consumers' choices, one therefore needs only to check *local* incentive constraints. In contrast, under a price schedule that does not uniformly induce quasiconcavity, checking local incentive constraints would be insufficient. In such case, the demand profile approach—which by its nature checks only local constraints—may incorrectly anticipate consumers' choices, yielding a non-optimal solution.

The degree to which the demand-profile solution is non-optimal will depend on two things: (i) for how many consumers was the optimal choice incorrectly anticipated, and (ii) how much damage did these deviations do the insurer's payoff? As discussed in Deneckere and Severinov (2017), it is difficult to answer these questions analytically, or even to draw general conclusions about them based on assumptions on primitives. Optimal price schedules are equilibrium objects that depend on the distribution of consumer preferences and costs in the market, something that becomes substantially complex when consumers are multidimensional. The observation we make in this paper, however, is that it is straightforward to answer these questions *ex post* in any practical application.

We use this observation propose a method to applied researchers for approximating the solution to multidimensional screening problems with vertically differentiated products. The key choice variable is *how many* products to use to span the relevant contract space. In the health insurance context, this might be a range of coverage levels between no insurance and full insurance. The demand-profile solution will be more likely to approximate the true solution if there are fewer potential products. Indeed, if there are just two products, *any* price schedule will trivially induce quasiconcave payoffs for all consumers, since each consumer's payoff function contains only two points. Using more contracts, on the other hand, will more fully capture the limiting environment in which product characteristics are fully endogenous, but will introduce more possibilities for price schedules to violate quasiconcavity. We provide a procedure that aids applied researchers in approaching this trade-off in a principled fashion.

If the approximation is of acceptable quality in the setting of interest, it provides two valuable contributions. First, the applied researcher can use the approximation to solve an optimal delegation problem, where an insurer's optimal menu must be evaluated within an inner loop of an optimization routine. Evaluating a social planner's optimal tax on a monopolist insurer, for example, would require such a routine.[4] The second contribution of our approximation approach is to shed light on the solution to the problem by way of a graphical framework, in the spirit of Einav et al. (2010a). Whereas Einav et al. restricted the insurer to offering only two contracts, this paper provides the conditions under which it can be applied with *any* number of vertically-ordered contracts, as well as the caveats necessary when these conditions are not met. The demand-profile approach lends itself directly to graphical analysis because the problem can be solved separately on each incremental level of coverage. Intuitively, equilibrium outcomes at each increment depend on the demand curves for incremental coverage, the associated marginal revenue curves, and the marginal cost of providing incremental coverage. 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.

We then study the health insurance menu design problem numerically using a simulated population of consumers calibrated to match demographics of the under-65 US population and parameter estimates from Marone and Sabety (2022). Our analysis allows for a flexible and empirically grounded distribution of consumer types. We implement the model using 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-only contract, and that the government covers the cost associated with base coverage regardless of what coverage level a consumer selects.

We first show that convergence in the density of the contract space is remarkably fast. In our setting, the insurer can capture over 98 percent of the available payoff with as few as five contracts. Using five potential contracts, we then 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 the analytical results derived from the necessary conditions, we find that a monopolist insurer excludes substantially more consumers than a social planner facing no excess cost of funds. Likewise, the monopolist screens more than the social planner, separating consumers across a range of coverage levels, and ultimately offering much less coverage overall. 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 the planner. As the cost of public funds rises, however, losses in the market becomes more costly for the planner, and it begins acting more like the monopolist.

We then evaluate the quasiconcavity conditions under which the demand-profile approach approximates the true problem. We find that for our three focal insurers (planner, planner with excess cost of funds, monopolist), the optimal premium schedules of both the true and simplified versions of the problem are consistent with quasiconcavity for 100 percent, 91 percent,

[4] As a monopolist can strategically respond to the tax, finding the optimal tax requires finding a fixed point. Under any proposed tax, solving for the monopolist's optimal response under the true multidimensional screening problem could take hours of computational time. Solving for an optimal tax vector in an outer loop would therefore be all but impossible in finite time (at least at current computing speeds). Evaluating the monopolist's *approximate* best response, however, would be nearly instantaneous, allowing such an analysis to proceed.

and 96 percent of consumers, respectively. Because quasiconcavity violations occur for only a small number of consumers, the solutions derived using the demand-profile approach are not meaningfully different from the true solutions. For each type of insurer, the payoff from solving the simplified problem agrees with the payoff from solving the true problem within a margin of 1 percent.

Finally, we use our framework to explore how a regulator might best intervene on behalf of consumers in a monopoly market. Here, reasoning in terms of incremental coverage levels is useful for two reasons. First, we show how it is possible to gain intuition about the impacts of pricing regulation in a setting where the characteristics of traded contracts are endogenous. Just as in a setting with two contracts, our graphical analysis can be used to visually evaluate the impacts of regulatory intervention. Second, we show how our approach makes it feasible to compute counterfactual equilibria in which a regulator optimizes over a policy tool to which the insurer strategically responds. By eliminating calls to a nested optimization problem without a closed-form solution, we can analyze strategic reactions to alternative regulations. Though our quantitative results are of course specific to our numerical setting, our analysis illustrates the usefulness of the simplified problem as a pragmatic approach to deriving novel insights in multidimensional screening problems.

Our paper is related to a large empirical literature on health insurance as well as an extensive theoretical literature on screening. With respect to the empirical literature, our model of consumer demand for health insurance builds on a workhorse introduced by Cardon and Hendel (2001), which has been used in several subsequent papers (for example, Einav et al., 2013; Azevedo and Gottlieb, 2017; Ho and Lee, 2021; Marone and Sabety, 2022). We 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 framework in Einav et al. (2010b), who focus on competitive markets and two potential contracts. A central contribution of our paper is to show under what conditions this approach can be extended to an arbitrary number of contracts. 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 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). These papers similarly analyze a principal-agent problem with private information, but consider 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), which has been surveyed by Rochet and Stole (2003).[5] There is also a recent theoretical literature on competitive markets with mul-

[5]Our class of problems belongs to Section 5 of Rochet and Stole (2003), 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

tidimensional private information, such as Azevedo and Gottlieb (2017), who provide a new equilibrium concept in settings with adverse selection, and Farinha Luz et al. (2022), who focus on risk classification. Insurers in these papers are price-takers, while the insurer in our setting is a price-setter. In a recent independent contribution, Gottlieb and Moreira (2023) analyze optimal monopoly insurance with multidimensional types. They too derive an optimal exclusion result under permissive primitives, but only with binary losses. Moreover, they show that competitive firms provide less coverage than monopoly for those who have a higher willingness to pay for coverage. The two papers complement one another nicely with respect to properties of optimal menus with market power.

Our approximation method relies on the pioneering work of Wilson (1993), which focused on nonlinear pricing in a setting without common values (that is, without *selection*). While this approach has been used in a number of theoretical contexts, there are few tests of its applicability in real-world settings.[6] Our finding that this approach provides an excellent approximation to the true multidimensional screening problem in health insurance markets presents a promising avenue for new theoretical and empirical exploration.

The paper has distinct theoretical and numerical analyses, which are targeted to different groups of readers. It is organized as follows. Section 2 describes the model. Sections 3.1 and 3 present theoretical results, including the optimality conditions and convergence. These sections can be skipped by the applied reader without loss of continuity. Section 4 presents the simplified reformulation of the problem and the graphical analysis. In Section 5, we discuss our numerical application, solve for the optimal menus under different insurer objectives, and apply our analysis to evaluate the impact of regulatory intervention. Section 6 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 for moral hazard and selection on moral hazard in the sense of Einav et al. (2013). The government may 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. To help the reader who is more interested in the application, we separate much of our discussion of the technical contributions 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

for a class of problems with two-dimensional private information, and Veiga and Weyl (2016) characterize the solution to a problem very similar to ours, but in which the insurer is limited to offering just one contract.

[6] One recent example is Gaynor et al. (2023), who use the demand profile approach to finding the optimal nonlinear reimbursement contract to offer healthcare providers.

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 $[0, \bar{l}]$.[7] 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.[8] We assume that G has a continuous density function g.[9] 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. It is convenient in what follows to assume that F has a finite-dimensional parametrization $\tilde{F}(\cdot|\mathbf{t})$, where $\mathbf{t} \in [0, 1]^n$, and where for reasons we will discuss later, $\tilde{F}(\cdot|\mathbf{t})$ is strictly *MLRP* increasing in the first coordinate of $\mathbf{t}$. Thus, $\mathcal{F}$ is the collection of all the $\tilde{F}(\cdot|\mathbf{t})$'s as $\mathbf{t}$ varies, and we can as is convenient work with G or with an equivalent measure $\tilde{G}$ on $[0, \bar{\omega} \times [0, \bar{\psi}] \times [0, 1]^n$.[10]

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.[11] That is, an agent is hurt by a worse health outcome, but is 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.[12] 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 example belongs to a class of b functions $b(a, l, \omega) = \hat{b}(a - l, \omega)$, 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 different levels 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 $c_{ax} < 0$.[13] That is, out-of-pocket costs are increasing and concave in the level of healthcare spending, and as x increases, 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.[14]

[7] 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.

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

[9] We will abuse notation by also denoting by G and g several conditional and marginal distributions and densities.

[10] 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 $G(Y) = \tilde{G}\{(\omega, \psi, \mathbf{t}) | (\omega, \psi, \tilde{F}(\cdot|\mathbf{t})) \in Y)\}$.

[11] 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.

[12] 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.

[13] 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.

[14] Both concavity and monotonicity of out-of-pocket cost functions are natural properties of health insurance

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.[15] 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 $v(x,\theta)$ and the premium.

The Government. The government exogenously 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 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.[16] 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 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,

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

contracts (see e.g., Zeckhauser, 1970).

[15] 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.

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

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.

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

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

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, which we will maintain throughout the paper:

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.2 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

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

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. 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
Firm offering insurance to employees	1	1	0

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

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

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).

Note that the fact that consumers privately observe their health state allows for (ex-post) moral hazard in the model. While our theoretical analysis still applies absent moral hazard, we incorporate this complication because it is a first-order concern in real-world health insurance markets (Manning et al., 1987). The presence of moral hazard also lets us explore the standard trade-off between risk protection and over-consumption of healthcare. 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 (Pauly, 1968; Zeckhauser, 1970; Marone and Sabety, 2022).

While we focus on the health insurance application, note that we ultimately 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). Consumer valuations are given by $v(x, \theta)$, and costs are given by $\gamma^I(x, \theta)$, with product quality indexed by x. Our model thus subsumes, for example, 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). We will discuss these relationships to the one-dimensional screening literature in more detail below.

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.5).

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.

3 Optimal Menu Design

We now derive necessary conditions that any optimally designed menu must satisfy. We consider two versions of the insurer's menu design problem: (i) the case in which the insurer is restricted to offering a finite set of contracts with fixed characteristics, and (ii) the case in which the insurer can offer a continuum of contracts, such that it can also control the qualities of all contracts offered. In both cases, we derive necessary conditions for optimality of ρ, which 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 we designed the problem to be sufficiently general that it could be taken directly to data, it does not have enough structure for the insurer's payoff to be quasiconcave in ρ. Despite this limitation, which is a serious stumbling block in most of the literature on multidimensional screening, our conditions shed light on several important properties of optimal menus, including the incentive to exclude and screen consumers and the existence of positive trade in the market.

We then show that these two versions of the problem are closely related. 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.

3.1 Consumer Demand for Insurance

As a building block, we first analyze the consumer's demand for insurance as a function of their type. Recall that the consumer is characterized by a coefficient of absolute risk aversion ψ, a distribution of health states F, and a taste for healthcare utilization ω. It will be 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.[17] 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 can now shed some light on the comparative statics of χ with respect to θ. We do so by analyzing how v_x changes with ω, ψ, and F, respectively, since this will pin down the behavior of χ. Given ω and F, demand increases in the consumer's absolute risk aversion parameter ψ. Knowing that the consumer behaves in a monotone fashion along one dimension of her type will prove useful below. Given ω and ψ, demand increases when F increases in an *MLRP* sense, namely, when the consumer becomes sicker. Finally, while comparative statics with respect to ω are straightforward in the case of linear or convex functions c, they are ambiguous under our concavity assumption on c. The formal statement and proofs of these results are in Online Appendix B.1.

3.2 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, define

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

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, and $(w^G - w^I)\gamma^G$ is 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 is important in what follows that $\mathcal{S}$ does not depend on p.

We can now interpret the marginal gains from trade from insurance in familiar terms. The

[17] The marginal utility of income also in principle depends on $y - p$, but due to CARA, this involves the term $e^{-\psi(y-p)}$, which cancels.

derivative of consumer-specific 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 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.[18]

Technical Remark 3 (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.

3.3 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

[18] 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 $\rho''(\hat{x}) \leq \rho'(x) + d(\rho', \rho'')$, and vice versa. Formally,

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

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.

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 a contract $0 \leq k < K$ 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 contract $\underline{x} = x^{\underline{k}}$ 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.[19] Multiplying r by $g(\hat{\psi})$ captures effects (i)–(iii).

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 a set of contract $\{x^k\}_{k=0}^K$, 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 proof is in Appendix A.1. The role of $\rho(x^k) < \rho(x^{k+1})$ is that on increments where the price schedule is flat (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 optimality condition must therefore

[19] 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 contract 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.

hold with equality only on increments where the price schedule is strictly increasing.[20]

3.4 Optimally Pricing 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 a contract x and raise the price of all strictly higher contracts by ε. Given x and some fixed (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 to think of r as reflecting a limit where $\bar{x} - \underline{x}$ is strictly positive but small. 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.3. 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 4 in Appendix A.3.

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 additional condition would also be necessary in the case in which the insurer could offer a fixed *number* of contracts, and could freely set their prices and qualities. This case is presented in Appendix A.2.

3.5 Some Relationships to the Literature

Our derivation of optimality conditions relies solely on perturbations to the premium schedule, which allows us to generalize several results from the literature on principal-agent problems with private information.

First, our optimality condition $\int \mathcal{V} dG\,(\omega, F) = 0$ 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 4, and so we defer the details.

Second, the conditions subsume as special cases the well-known analogous conditions when private information is one dimensional. To see this, 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. The setting then reduces to a standard *one-dimensional principal-agent problem.* In this case, our main condition equating the integral of (7) to zero reduces to

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

[20]For another example where an elementary perturbation argument on the price schedule leads to economically intuitive optimality conditions, see Saez (2001). He derives an intuitive first-order condition in the canonical Mirrlees optimal taxation problem via a simple perturbation of the optimal tax schedule.

for all x, and so reflects the standard efficiency versus information-rents trade-off. 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). The usual approach in the literature is to solve for the allocation (in this case, the contract assigned to each type) in the relaxed problem that omits the monotonicity constraint, and then "iron" it if the solution fails to be monotone. Online Appendix B.6 shows that the ensuing condition based on our perturbation provides a direct route to the standard "ironing" condition (Fudenberg and Tirole, 1991, Chapter 7),

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

In short, our conditions in the multidimensional case are the natural generalizations of the textbook cases in which private information is one-dimensional.

Third, return to multidimensional types, and assume that we restrict the monopolist to choosing a single contract, which is a special case of our setting with a finite *number* of contracts (discussed in Appendix A.2). 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.[21,22]

Finally, 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.

3.6 Incentives to Exclude, Screen, and Trade

The optimality condition $\int \mathcal{V} dG = 0$ also provides insight into the insurer's incentives to exclude and screen types. In Online Appendix B.7, we show that from the point of view of the social planner, a monopolist excludes too many consumers the market. We also show that if ω was the only source of private information, a social planner would completely pool types, while a monopolist insurer may completely sort types, an extreme example of differential incentives to screen.

Another question of interest is whether a monopolist insurer always trades (that is, makes strictly positive profits).[23] Assume that the government's costs γ^G increase with consumers'

[21] 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."

[22] 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.

[23] This question has received attention both in the empirical insurance literature (see the no-trade result in Hendren,

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. The proof is in Appendix A.4, 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. Thus, giving a little extra insurance to some types has a first-order gain in terms of the insurance motive, but only a second-order cost in terms of wasteful medical spending covered by the monopolist.

3.7 Convergence

We now 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. Definition 1 defines convergence of sets of price schedules. Theorem 2 shows that if $\mathcal{P}^n$ converges to $\mathcal{P}^0$, then the payoff to the insurer does as well, and similarly for the optimal solutions. The proof is in Appendix A.5.

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 $\hat{\rho}$.*

Intuitive as it may be, Theorem 2 has two very useful 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.[24] 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 considered in Appendix A.2. We can therefore work in the case of a (large) fixed set of contracts or in the continuum, whichever is more convenient.

4 Demand-Profile Approach

We now present a reformulation of the insurer's problem and discuss the conditions under which its solution is optimal or approximately optimal in the original problem. Our approach builds

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).

[24]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.

directly on the "demand profile" approach proposed by Wilson (1993) and discussed in Rochet and Stole (2003) and Armstrong (2016), although the existence of selection in our setting introduces additional complications. Formally, we show how—and when—the problem of setting a full price schedule on a fixed set of contracts can be reduced to a set of entirely independent sub-problems of setting the marginal price of incremental coverage. We then demonstrate how this decoupling yields substantially more powerful analytical results and permits visual analysis using a familiar graphical framework.

4.1 The Reformulation

The key to this approach is to reformulate the problem 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.

Fix a set of potential contracts $X = x^0 < x^1 < \cdots < x^K \leq 1$. 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 and $\gamma^k(\theta) = \gamma^I(x^k, \theta) - \gamma^I(x^{k-1}, \theta)$ be the marginal insured cost. For simplicity in this portion of the analysis, suppose that the government's cost of providing base coverage is invariant to the consumer's chosen contract, so that marginal government spending is zero: $\gamma^G(x^0, x^k, \theta) - \gamma^G(x^0, x^{k-1}, \theta) = 0$.[25] Note that since contracts are vertically differentiated and the premium schedule is increasing in coverage level, p^k, v^k, and γ^k are all weakly positive. The insurer's marginal payoff from charging type θ a marginal premium p^k to move from contract $k-1$ to k is then

$$S^k(p^k, \theta) = w^C(v^k(\theta) - p^k) + w^I(p^k - \gamma^k(\theta)).$$

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 $S^0(\theta) = w^C v(x^0, 0) + w^I \gamma^I(x^0, \theta) + w^G \gamma^G(x^0, \theta)$ is the payoff from putting type θ into contract x^0, and where we take the second term to be zero when the consumer's optimal contract is x^0.

As in the original problem, solving for the consumer's optimal contract is complicated because

[25] This corresponds to, for example, the Medicare Advantage market, where the government makes a fixed payment to insurers for enrolling each consumer, regardless of what coverage level the consumer ultimately selects. It is straightforward to proceed with a government cost of base coverage that *is* linked to the consumer's chosen coverage level, but there is limited additional economic insight and substantially more notation.

it is defined by a set of non-local incentive constraints. Under certain conditions, however, it can become substantially simpler. Say that a price schedule ρ is *quasiconcave consistent (QC)* for a consumer of type θ if the consumer's payoff $v(x,\theta)-\rho(x)$ is single-peaked in x. 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. Now, the optimal contract choice $\tilde{k}$ 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, we can rewrite the insurer's payoff on type θ as the sum of the payoffs on all increments that the consumer is—in isolation—willing to pay for:

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

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

$$\tilde{\Pi}(\rho)\equiv\int S^0(\theta)dG(\theta)+\sum_{k=1}^{K}\tilde{\Pi}^k(p^k). \tag{10}$$

To see this, integrate (9) with respect to G and swap the order of integration and summation.[26]

So, consider the problem in which the insurer maximizes $\tilde{\Pi}(\rho)$:

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

Note that 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 therefore be constructed from the set of optimal incremental prices $(\tilde{p}^1,\ldots,\tilde{p}^K)$, where on each increment, $\tilde{p}^k\in\arg\max_{p^k}\tilde{\Pi}^k(p^k)$. Because $\tilde{P}$ can be solved one contract at a time, it is a much simpler problem than P.

4.2 Analyzing the Reformulated Problem

The reformulated problem allows us to think about the insurer's optimal price increments p^k one at a time. We use this simplicity to study the solution to $\tilde{P}$, and recast that solution in familiar

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

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

terms. To begin, rewrite the insurer's payoff 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.[27]

Let

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

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^k be the derivative of C^k. Similarly, 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 $V^k_q(q^k) = P^k(q^k)$. It is now straightforward to verify that

$$\tilde{\Pi}^k(P^k(q^k)) = w^C[V^k(q^k) - P^k(q^k)q^k] + w^I[P^k(q^k)q^k - C^k(q^k)], \tag{11}$$

and thus we can think of the insurer as solving $\max_{q^k} \tilde{\Pi}^k(P^k(q^k))$ at each coverage level increment. We can also now usefully decompose the insurer's incremental payoff into a "benefit" equal to $w^C V^k(q^k) + (w^I - w^C)P^k(q^k)q^k$ and a "cost" equal to $w^I C^k(q^k)$. 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, $C^k(q^k)$.

Denoting the price-elasticity of demand by ϵ, so that $1/\epsilon = P^k_{q^k} q^k / P^k$, we can then write the derivative of the insurer's objective function as

$$\left(\tilde{\Pi}^k(P^k(q^k))\right)_{q^k} = \underbrace{P^k(q^k)\left(w^I + (w^I - w^C)\tfrac{1}{\epsilon}\right)}_{Marginal\ benefit} - \underbrace{w^I MC^k(q^k)}_{Marginal\ cost}. \tag{12}$$

The first term is the insurer's marginal benefit of giving more consumers incremental coverage level k.[28] 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 $w^I MC^k(q^k)$ is the incremental insured cost of the marginal consumer.

At an interior optimum, marginal benefit is equal to marginal cost, yielding a familiar markup equation. When the insurer is a monopolist, the optimality condition reduces to $P^k(1 + (1/\epsilon)) = MC^k$. 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 $\int \mathcal{V} dG = 0$.[29] This makes intuitive sense, as increasing

[27] 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.

[28] 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.

[29] See Lemma 3 in Appendix A.6.

the marginal premium p^k raises the price schedule ρ for all contracts at or above k, and so is effectively the perturbation discussed in Section 3.3.

4.3 Graphical Analysis

The reformulated problem $\tilde{P}$ is composed of a set of independent two-contract problems, one between each pair of adjacent contracts. It can therefore be analyzed graphically in the spirit of Einav et al. (2010b). Indeed, Einav and Finkelstein (2011b) suggest the possibility of generalizing their model to more than two contracts. Geruso et al. (2019) take a first step in this direction by extending the graphical analysis to accommodate three contracts. A central contribution of our analysis is to formalize 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). In what follows, we normalize the insurer's weight on its own profits w^I to 1.

Figure 1 illustrates the insurer's problem for one incremental coverage level. It shows the inverse demand function P^k, the associated marginal revenue function $MR^k = P^k(1 + (1/\epsilon))$, and the insurer's marginal cost curve MC^k.[30] The insurer's marginal benefit of serving more consumers is given by $MB^k(q^k) = P^k(q^k)(1 + (1 - w^C)\frac{1}{\epsilon})$, 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 $w^C \in (0, 1)$.[31] The insurer's optimal quantity $\tilde{q}^k$ obtains where $MB^k = MC^k$.[32]

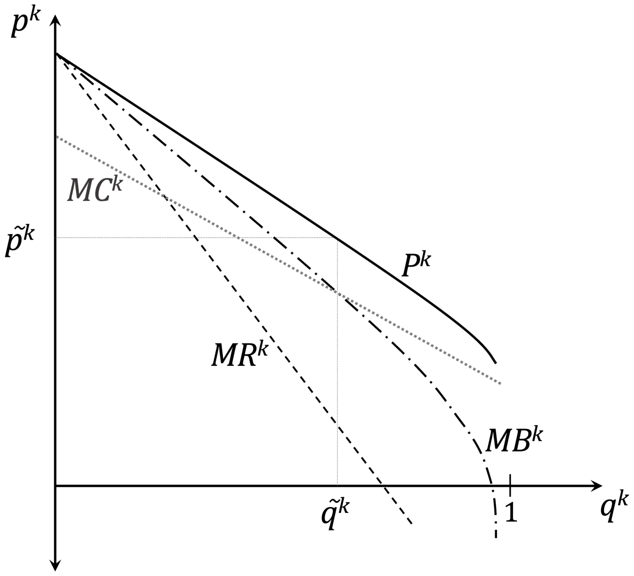
Figure 1 subsumes a number of cases. For a monopolist ($w^C = 0$), MB^k coincides with the marginal revenue curve MR^i, and the optimal quantity solves $MR^k = MC^k$. For a utilitarian social planner with no excess cost of public funds ($w^C = 1$), MB^k coincides with the inverse demand curve, and the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

[30] 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., 2010b).

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

[32] 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 multiple crossings (in which case the solution may not be interior), or of either a single crossing from below or no crossing at all (in which case it certainly will not be).

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 insurer's optimal quantity $\tilde{q}^k$ obtains where the marginal benefit curve intersects the marginal cost curve.

4.4 Comparative Statics, Exclusion, and Screening

The simplified problem yields some simple *monotone comparative statics* of economic interest, many of which can be seen clearly in Figure 1. First, when the consumer is weighted more heavily in the insurer's objective function, the marginal benefit curve rotates upwards 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).[33] 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).

It is also now possible to derive stronger results with respect to exclusion and screen. 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. The reason for this is that as 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 $\tilde{q}^k$ will be strictly less than one.

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

The proof is in Appendix A.7. Note that Proposition 1 applies at every incremental coverage

[33]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.

level, including the first. 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.[34]

Proposition 1 also sheds light on differential incentives to screen. To see this, note that under any price schedule that is QC for all consumers, within the range of traded coverage levels, every contract will be traded. No contracts will be "skipped." In the case of the monopolist, Proposition 1 tells us that the range of traded coverage levels will *always* include base coverage x_0. If the monopolist's optimal price schedule satisfies quasiconcave consistency, then it sells every contract in a range that includes x^0 at the lower end. The social planner, on the other hand, may very well choose a higher coverage level as the lower end of the traded range. In Figure 1, for example, the demand curve everywhere exceeds the marginal cost curve, and so the social planner wishes to allocate all consumers to a contract weakly greater than k. The monopolist therefore uses more contracts than the social planner in its optimal menu. Note that this also implies that a monopolist will overall offer less coverage than a social planner.

4.5 Relationship Between $\tilde{P}$ and P

The reformulation of the insurer's objective under the true problem P into its objective under $\tilde{P}$ relies on a price schedule that is QC for all consumers. The solution to $\tilde{P}$ therefore coincides with the solution to P only if the solution to P is QC for all consumer types. Nothing in our model guarantees that this will be the case (indeed, Deneckere and Severinov (2017) cast serious doubt on whether such primitives generally exist). The optimal price schedule is an equilibrium object that depends on the distribution $G(\theta)$ of consumer types in the economy as well as on the set of potential contracts X under consideration. Deviations from quasiconcavity indicates a desire to "bundle" adjacent levels of coverage, a phenomenon that by its nature cannot be captured by a decoupled analysis. When such deviations arise, what then can we learn from the reformulated problem?

First, $\Pi(\tilde{\rho})$, the payoff under P (which understands consumers globally optimize) evaluated at the solution to $\tilde{P}$ (which assumes consumers only locally optimize), provides a lower bound on the insurer's true optimal payoff.[35] To the extent one is interested in evaluating what is achievable under different insurer objective or sets of potential contracts, this can be a useful starting point. Second, if a given price schedule ρ is QC for all consumer types except a small set, then $\Pi(\rho) \cong \tilde{\Pi}(\rho)$, with the approximation arbitrarily good as the measure of the set of types where ρ is not QC goes to zero. As the measure of the set of consumer types for whom QC fails goes to zero under the solution to the true problem, the solution to the reformulated problem therefore becomes an arbitrarily good approximation. We will show in Section 5 that in applied analysis, this approximation can be extremely valuable.

While we are unaware of assumptions on primitives that justify the QC property, it is still

[34] Optimal exclusion has precedent in the literature, but only without common values (Armstrong, 1996; Deneckere and Severinov, 2017; Barelli et al., 2014).

[35] Because deviations from quasiconcavity may also arise in the solution to $\tilde{P}$, $\tilde{\Pi}(\tilde{\rho})$ provides only an approximate lower bound on the insurer's payoff under P. As the measure of the set of consumers for whom QC fails goes to zero, the lower bound becomes exact.

possible to build intuition. For the consumer's problem to be quasiconcave in coverage level, the marginal price schedule p^k must intersect the consumer's demand curve for coverage v^k (at most) once from below. Downward-sloping consumer demand curves—which is to say, declining marginal willingness to pay for incremental coverage—will therefore help push in the right direction.[36] Naturally, the number of potential contracts considered—and their *placement* in coverage level space—are also centrally related to the ability to universally satisfy quasiconcavity. With only two potential contracts, any price schedule is trivially QC, since the consumer payoff schedule $v(\theta, x) - \rho(x)$ has only two points. Adding more contracts adds more opportunities for quasiconcavity to fail. Finally, in their study of the demand-profile approach, Rochet and Stole (2003) (Section 9.1) present an example with two products plus an outside option (two contracts plus x_0, in our case). They suggest that a key factor that will help ensure the demand-profile solution will closely approximate the true solution is positive dependence in consumers' marginal valuations. In other words, consumers who have a higher marginal value for the first increment of quality *also* have a higher marginal value for the second ("the *order statistics* [of consumer marginal valuations] *are positively correlated*," Rochet and Stole (2003), emphasis in original). We will discuss this point further in our empirical analysis.

While each of these contributing factors provides intuition, none will (to our knowledge) be sufficient to guarantee quasiconcave consistency. In the end, the extent to which quasiconcave consistency holds in a given problem must be checked empirically.

5 Numerical Analysis

We now study the insurer's problem numerically using an empirically calibrated model of a health insurance market. Importantly, the generality of our model allows us to directly apply state-of-the-art empirical estimates of the distribution of multidimensional consumer types in a population. Beyond allowing us to assess the empirical validity of quasiconcave consistency, the numerical analysis also allows us to demonstrate the usefulness of the reformulated problem in gaining intuition about the solution as well as in evaluating the effects of policy interventions in a monopoly market.

5.1 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).[37] 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

[36] Note that in general, the concept of "coverage level" x has no cardinal interpretation. Cardinality must be inherited from the parameterization of contracts, which for generality we avoid in this paper. If contracts are linear, x can be thought of as the fraction of total healthcare spending paid by the insurer. In this case, quasiconcavity of a price schedule could be guaranteed by the combination of concave consumer demand functions ($v_{xx} \leq 0$) and a convex price schedule ($\rho_{xx} \geq 0$).

[37] Details of the simulation procedure are provided in Online Appendix B.8.

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 \$10,684 under a 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 a depiction of the relationship between consumer types and willingness to pay for insurance in this 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 are reported relative thousands of dollars.

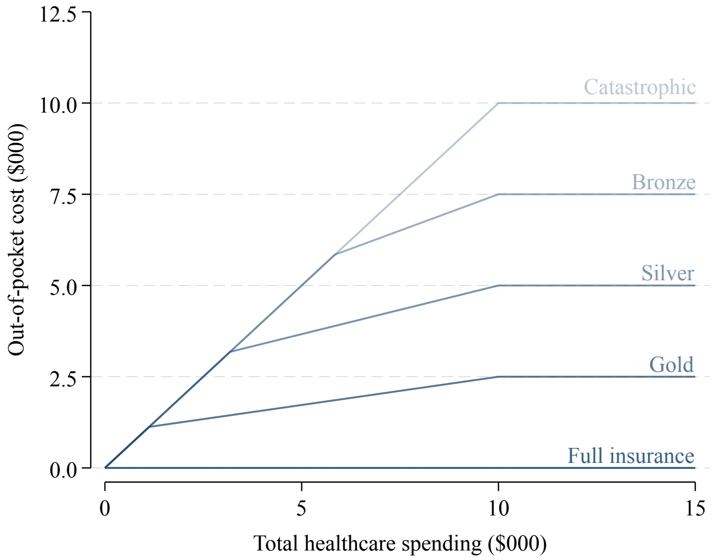
Insurance Contracts. We consider a set of contracts that are piecewise linear, with a deductible, coinsurance region, and out-of-pocket maximum design. We suppose that the base level of coverage x^0 is a "Catastrophic" contract with a deductible and out-of-pocket maximum of \$10,000. Our baseline set of potential 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.[38] As will become clear, the returns to allowing an increasingly "dense" contract space are economically small.

5.2 Convergence

Theorem 2 states 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 will occur. We illustrate and investigate this result by computing optimal menus on

[38] 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. In our population of consumers, the actuarial value of the five contracts 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.

an increasingly dense set of allowable contracts. Figure 2 depicts a set of five allowable contracts, 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 potential 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.[39] For each set of potential contracts, we solve for the optimal menu that would be offered by 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.[40] Figure 3 plots insurer payoffs as a function of the number of contracts in the potential contract space. While insurer payoffs are of course increasing in contract density, in practice the returns to additional density are small. We find that after five contracts, the gains from moving to 65 contracts do not exceed $19 per household per year for any insurer. After nine contracts, gains do not exceed $10.[41]

There are, however, economically meaningful gains from moving 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, 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

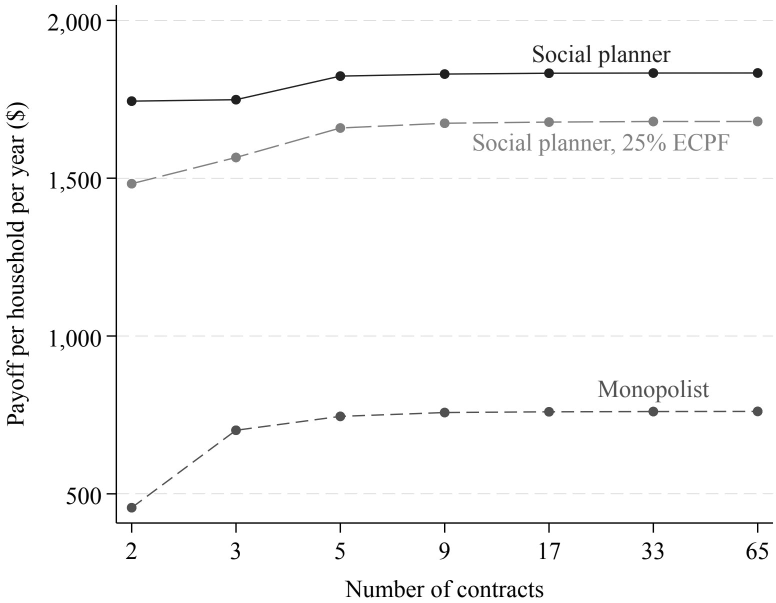
[39]We increase the set of allowable contracts by successively adding a contract between each pair of adjacent contracts. We proceed in this iterative manner so that under successively "dense" contract spaces, all previously allowable contracts remain allowable.

[40]Optimal menus are calculated using a numerical algorithm that relies on the necessary condition derived in Section 3.3. The algorithm is described in detail in Online Appendix B.8.

[41]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.

number of contracts are needed to closely approximate the limiting environment, there are potentially meaningful consequences of over-restricting the contract space. 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 base coverage and full insurance.

Figure 3. Convergence



Notes: The figure shows optimal insurer payoffs as a function of the number of contracts used in the potential 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 their desired allocation under a sparser contract space. All numerical results are thus quite robust to the density of the contract space.

5.3 Performance of the Reformulated Problem

Given the speed of convergence, we proceed with five contracts, which enables a simple presentation of results. We next investigate how well the reformulated version of the problem (presented in Section 4) approximates the true problem (presented in Section 2). For our three focal insurers, we solve the true problem P as well as reformulated problem $\tilde{P}$. Table 3 reports these results. Specifically, it reports the optimal premium schedule, the associated allocations, and the associated insurer payoff.

Recall that the condition under which the two versions of the problem coincide is that consumer payoffs are quasiconcave in coverage level at the optimal menu under the true problem P, and that the solution under $\tilde{P}$ provides a lower bound on the optimal payoff. The final column of Table 3 reports the fraction of consumers for whom the given price schedule is quasiconcave consistent. We find that in every case, it holds for nearly all consumers. With respect to the

Table 3. Performance of the Reformulated Problem

Insurer	Premiums $000s Brnz.	Slvr.	Gold	Full	Allocations Pct. of households Cstr.	Brnz.	Slvr.	Gold	Full	Insurer Payoff $000s Π	Pct. QC
Social planner											
Solution to P	0.16	0.32	0.67	3.21	<0.01	–	<0.01	1.00	–	1.823	1.00
Solution to $\tilde{P}$	0.13	0.30	0.68	3.19	<0.01	–	<0.01	1.00	–	1.823	1.00
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	0.91
Solution to $\tilde{P}$	1.32	2.85	4.73	7.23	0.13	0.03	0.13	0.71	–	1.655	0.99
Monopolist											
Solution to P	2.02	4.09	6.46	9.02	0.39	0.03	0.32	0.26	–	0.745	0.96
Solution to $\tilde{P}$	2.00	4.13	6.50	9.00	0.38	0.06	0.29	0.26	–	0.745	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 true problem P and the reformulated problem $\tilde{P}$. The table also reports the associated allocations and insurer payoffs. The insurer payoff Π is the objective of problem P, meaning consumers globally optimize with respect to prevailing premiums. 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.

reformulated problem, we find that the solutions are QC for over 99 percent of consumers. With respect to the true problem, we find slightly more frequent violations of quasiconcavity, indicating returns to bundling adjacent coverage levels for some consumers. Even so, the gains available from doing so are economically small: at most \$4 per household per year (in the case of the social planner facing an excess cost of funds).

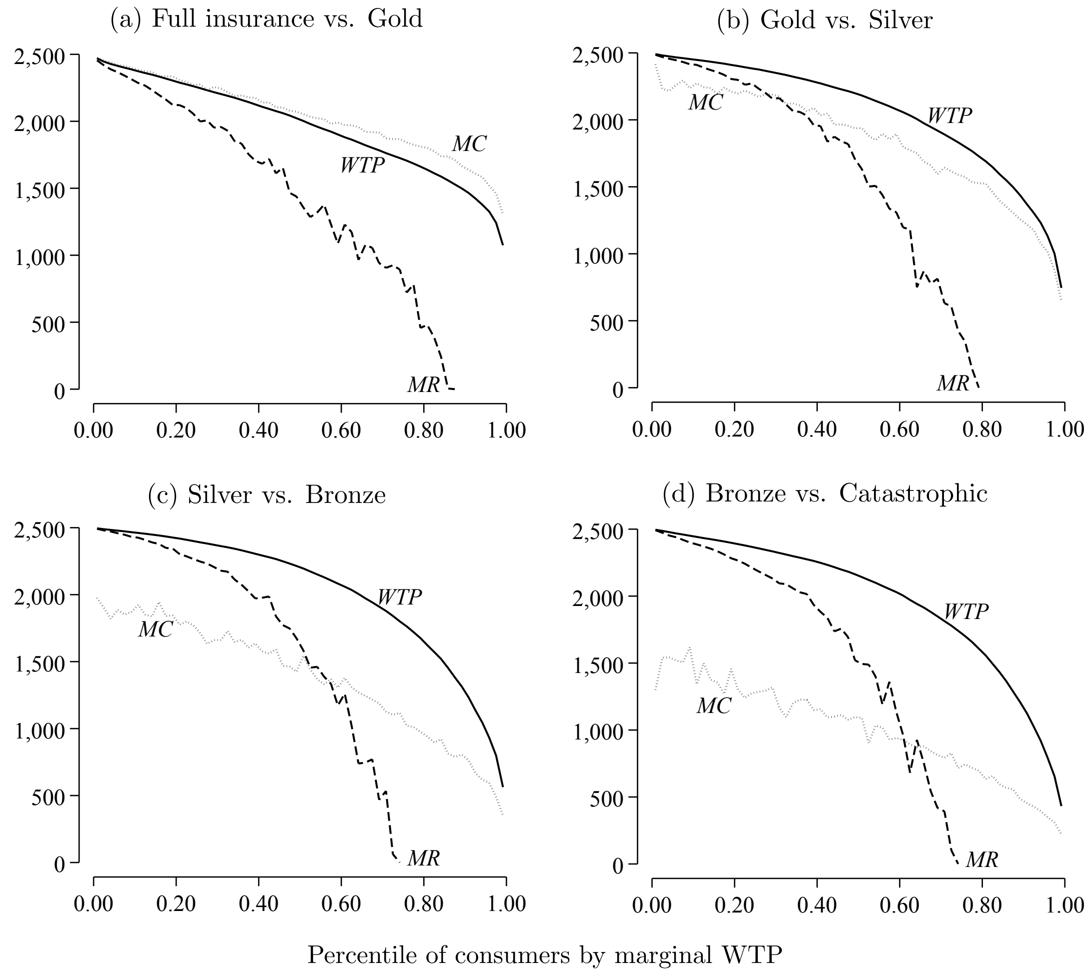
These findings show that in this population and this set of potential contracts, the reformulated problem is an excellent approximation to the true problem. All of the analysis in Section 4 is therefore applicable in thinking about the solution. Moreover, our convergence results in Section 5.2 show that this set of 5 potential contracts is a good approximation of the solution under as many as 65 potential contracts, suggesting that even if quasiconcave consistently were more likely to fail if we added contracts, doing so is not necessary to capture the first-order economic forces at play.

We propose the following procedure to applied researchers aiming to analyze a problem of this type. Begin with a fixed set of contracts, at a density supposed to be sufficient for the purpose at hand. Our results suggest this need not be excessively many contracts, but this likely involves some experimentation, as in Figure 3. Then solve the reformulated problem $\tilde{P}$, and evaluate the percentage of consumers for whom the resulting premium schedule is QC. If the solution is not QC for a substantial number of consumers, begin iteratively dropping contracts and re-solving until a sufficient level of quasiconcave consistency is reached.[42] At this point, solve the true problem using the current set of potential contracts. If the payoffs between the two versions of the problem are within an acceptable tolerance, then the reformulated problem can be confidently used for further analysis. In the following subsections, we show what further analysis might look like.

[42]Our graphical analysis in the next subsection provides some guidance for determining which contracts are creating problems.

Graphical Analysis. Figure 1 described how to solve the insurer's problem graphically on a single incremental coverage level. When considering more than two potential contracts, there are more increments to consider. Figure 4 demonstrates how to carry out the graphical analysis on all increments simultaneously, in order to solve visually for the optimal menu across the full set of potential contracts. The four panels represent 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 coverage level increment, the associated marginal revenue curve MR, and the marginal cost curve MC associated with providing that coverage level increment.[43]

Figure 4. Illustration of Graphical Analysis for Monopolist and Social Planner



Notes: The figure illustrates the graphical analysis of the reformulated 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 incremental level of coverage. Consumers are ordered on the horizontal axes according to their marginal willingness to pay for each coverage level increment. The solid line (WTP) represents consumers' willingness to pay, 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.

[43]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 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. This would have the effect of substantially lowering the insurer's optimal quantity on that increment, likely introducing violations of quasiconcave consistency.

To solve the insurer's problem, one simply needs to find the intersection of the marginal benefit and marginal cost curves in each panel. While all insurers have the same marginal cost curves, marginal benefit curves depend on the insurer's objective. As discussed in Section 4.3, 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, at the increment between Bronze and Catastrophic coverage, marginal revenue exceeds marginal cost for about the first 60 percent of consumers, 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 incremental premium ($2,021) can then be read from the value of the willingness to pay curve at this quantity. At the increment 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. Since the monopolist optimally allocates 61 percent of consumers to coverage above Catastrophic, it excludes the remaining 39 percent from the market for incremental coverage.[44] The same exercise can be repeated for a social planner with zero cost of funds using the intersections of the WTP and MC curves.

The graphical analysis in Figure 4 also provides a visual test of quasiconcave consistency under the solution to the reformulated 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 consumer types that purchase higher coverage level increments will be a subset of those that purchase lower coverage level increments. 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 increments.[45] Violations of this property 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 at which different components of willingness to pay increase 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

[44] Who are these consumers? Understanding the sorting of consumers with multidimensional types to contracts is notoriously difficult, since the distribution of types can in principle exhibit an arbitrary dependence structure. Inspection of Online Appendix Figure B.1, however, provides some insight into these patterns. The figure shows that the lowest willingness-to-pay consumers in the population are substantially healthier and also less risk averse than the average consumer in the population. Since incentive compatibility dictates that these consumers would choose the least coverage, it is these consumers that are excluded under the monopolist's optimal allocation.

[45] It is not necessary for consumers' position on the demand curve to be exactly consistent across coverage level increments 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).

for another.[46] With multidimensional consumer types, this type of reordering is almost 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).

5.4 Welfare

Unsurprisingly, social welfare is lower under monopoly than under the social planner's solution. We now quantify these welfare differences and show how the reformulated problem can be used to evaluate the impacts of various policy interventions.

Table 4 reports outcomes under a set 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 allocations. The welfare outcomes are average per-household per-year social surplus (SS), consumer surplus (CS), and producer surplus (PS), each measured relative to the allocation of all consumers to the Catastrophic contract. The spending outcomes are average per-household per-year government spending (Gov), premiums ($Prem$), and expected out-of-pocket spending (OOP).

Table 4. Welfare Outcomes and Policy Simulations

		Welfare outcomes *\$000 per household*			Spending outcomes *\$000 per household*			Allocations *Pct. of households*				
	Scenario	$SS^{\dagger}$	$CS^{\dagger}$	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	1.02	1.02	–	5.64	1.23	4.27	0.07	0.81	0.12	<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)	Restrict which plans allowed	0.94	0.48	0.46	5.64	4.49	1.70	0.46	–	–	–	0.54
(ii)	Adjust base coverage	1.82	1.82	–	9.86	–	1.80	–	–	–	1.00	–
(iii)	Linear taxes/subsidies	0.74	0.64	0.10	5.15	2.70	3.44	0.58	0.04	–	0.38	<0.01
(iv)	Nonlinear subsidies	1.30	0.48	0.82	5.78	3.55	2.79	0.30	0.01	0.28	0.42	–

Notes: The table shows welfare outcomes, spending outcomes, and allocations under various scenarios. Welfare is evaluated using a zero excess cost of public funds. 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).

Panel A presents the benchmark cases, which serve as useful points of comparison. 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

[46] Recall Rochet and Stole's highlighting of positively correlated order statistics.

outcome. In the first three benchmarks, we assume all surplus accrues to consumers. For the fourth, we calculate the competitive equilibrium proposed by Azevedo and Gottlieb (2017). Allocating all consumers to their socially efficient contract (the "first best") results in social surplus that is $1,860 per household per year higher than allocating all consumers to the Catastrophic contract. Allocating all consumers to full insurance results in social surplus of $1,743. The competitive outcome features substantial unravelling, but as few consumers are fully excluded, still generates a large amount of social surplus.

Panel B then reports outcomes at the optimal menus chosen by our three focal insurers. Social surplus under a planner facing no excess cost of public funds is $1,825 per household per year. Under a monopolist, social surplus falls to $1,081, and consumer surplus falls to $330. Consistent with the theoretical results in Sections 4, the monopolist provides less coverage than the social planner, uses more contracts in its optimal menu, and excludes more consumers from the market. When the insurer is a planner facing an excess cost of public funds, it begins behaving somewhat more like the monopolist, in that it now places more weight on insurer profits than on consumer surplus.[47] As the excess cost of funds increases from 0 to 0.25, the planner begins to both screen and exclude, and the optimal amount of coverage provided decreases. A similar logic can be applied to an employer. If it weighted consumer surplus and health insurance costs equally, its optimal menu would coincide with the planner facing zero excess cost of funds. Yet absent the tax exclusion for employer-sponsored insurance, compensating employees with health insurance would become more expensive, leading it to act more like a planner facing an excess cost of funds.

Interestingly, social surplus is higher under the monopolist relative to a perfectly competitive market. This could not happen in a one-contract setting absent substantial moral hazard (Mahoney and Weyl, 2017). Though the monopolist indeed constrains output relative to a social planner, output is even further constrained by unravelling in the competitive equilibrium. Examining contract-by-contract markups provides some insight (see Online Appendix Table B.1). While competitive insurers must break even on every contract, the monopolist can internalize pricing externalities that one contract has on the profitability of others. At its optimal menu, the monopolist chooses a 457 percent markup on Bronze coverage, diverting many consumers to higher coverage. In a one-contract example, the monopolist has no such tool at its disposal (as there are no other contracts on which to internalize externalities). That a monopolist can increase efficiency in selection markets is consistent with Diamond (1992), who suggests that 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.[48] Of course, the monopolist captures the majority of the surplus it generates. Consumers are better off under competition.

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

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

[48] It is also consistent with Veiga and Weyl (2016), who suggest that market power may increase efficiency in insurance markets relative to perfect competition.

($\omega \approx 0$ for all consumers), the social planner optimally pools all consumers at full insurance. The monopolist again provides less coverage, screens consumers across more contracts, and excludes some consumers from the market entirely.

5.5 Policy Analysis

Panel C of Table 4 reports the impact of strategies a regulator might use to intervene on behalf of consumers in a monopoly market. These strategies are: (i) banning the monopolist from offering certain contracts; (ii) raising the level of base coverage; (iii) linear taxes or subsidies (constant for each consumer that obtains coverage); and (iv) nonlinear subsidies (provided only on *additional* consumers that obtain coverage). In each case, we assume the regulator aims to maximize consumer surplus, taking into account the tax burden associated with government spending (which is available at zero excess cost).

Policies (i) and (ii) are straightforward to implement in our numerical simulations. The set of options the regulator has in each case is enumerable, meaning we can simply solve the monopolist's problem a finite number of times and choose the option under which consumers fare best. Policies (iii) and (iv), on the other hand, are more complex. Supposing that the taxes (or if negative, subsidies) can be continuous, it is not possible to enumerate all possible interventions the regulator could make. Solving for the regulator's optimal set of taxes therefore involves a nested optimization routine in which the regulator chooses a vector of taxes, and the monopolist chooses an optimal menu given those taxes. Unlike in a perfectly competitive environment, there is no closed form way in which to map changes in the vector of taxes into the insurer's choice of optimal menu. The ability to approximate the monopolist's optimal menu therefore becomes extremely valuable in solving for the optimal vector of taxes in finite time.

Under policy (i), the regulator simply chooses which (if any) contracts the monopolist is banned from offering. This requires no additional government spending. Banning a given contract in effect forces the monopolist to "bundle" two adjacent coverage level increments. In the language of the graphical analysis, this would mean vertically summing two adjacent coverage level increments and re-solving for the optimal quantity on that combined increment. This unambiguously hurts the monopolist (given that it has less flexibility), but has the potential to benefit consumers. We find that in our population, the optimal policy is to ban the Bronze, Silver, and Gold contracts, forcing the monopolist to maximally bundle. While this strategy is effective in shifting surplus to consumers, it also increases the amount of exclusion in the market (by 7 percentage points relative to the unregulated monopoly outcome). This results in a redistribution of consumer surplus from lower willingness-to-pay households (who are now excluded) to higher willingness-to-pay households (who now obtain full insurance).[49]

Policy (ii) considers the case in which the regulator can raise the base 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 optimal feasible allocation (since the optimal menu may

[49]This result is consistent with recent analysis by Gottlieb and Moreira (2023), who describe this as a tension between "distortion at the bottom" and "distortion at the top."

involve screening), but it will always be effective at preventing under-insurance in the market. In our setting, the regulator can in fact restore maximal consumer surplus by raising base coverage to the level of the Gold contract. This solution can be read directly off the graphical analysis in Figure 4. Since the demand curve lies everywhere above the marginal cost curve on all increments below full insurance, a planner facing zero excess cost of funds will clearly find it optimal to raise base coverage up through each of those increments. Once consumers receive the Gold contract for free, the monopolist cannot profitably offer the full insurance contract, and therefore effectively exits the market.[50]

Under policy (iii), the regulator has the ability to levy taxes or subsidies that are linear in the number of consumers served. Contracts can still be banned (via an infinite tax), but the regulator can now also mediate prices in a more subtle way. The approximation allows us to solve this problem numerically.[51,52] We find that the regulator's best course of action is to tax the monopolist at every incremental coverage level. The intuition for this result is that supply is fairly inelastic at the monopolist's optimal menu, any subsidization would large be a transfer to the monopolist, while any tax is largely a transfer towards the government (and thus consumers). There is still some effect on quantity, so this strategy does lower the average level of coverage in the market. Taken together, consumer surplus derived from insurance coverage is reduced by \$178, but the consumer's tax burden falls by \$493. On net, consumer surplus increases by \$315 per household per year relative to the unregulated monopoly outcome. While consumer surplus (as measured) does increase, linear taxes/subsidies are an ineffective tool for increasing coverage levels in a monopoly market. Because subsidies must be paid on *all* consumers served, increasing quantity is prohibitively expensive.

Under policy (iv), we therefore consider a policy in which the regulator is able to announce that subsidies are only available to the monopolist on *marginal* consumers. That is, the subsidy policy takes as an input the monopolist's unregulated optimal allocation, and applies subsidies only to additional consumers served. In this way, the regulator avoids paying subsidies on inframarginal consumers. We find that with this policy tool, the regulator optimally *raises* the level of coverage obtained in the market. Consumer gains from higher coverage amount to \$298 per household per year, while government spending increases by only \$143 per household per year. On net, consumer surplus increases by \$155 per household per year relative to the unregulated monopolist outcome.

Our results highlight the challenges associated with regulating a non-competitive insurance market, echoing findings in applied work such as Tebaldi (2022) and Jaffe and Shepard (2017).

[50] The monopolist cannot profitably offer full insurance when Gold coverage is free for the same reason the full insurance contract was not part of the social planner's optimal menu in the first place: the high willingness-to-pay consumers are not willing to pay their own marginal cost for incremental coverage.

[51] In particular, we allow the regulator to optimize over a vector of taxes that lies in $\mathbb{R}^5$. At each candidate vector, we approximate the monopolist's best response (its optimal menu) using the solution to the reformulated problem. The optimization routine converges in approximately 10,000 iterations. Within each iteration, it takes seconds to solve the reformulated problem, but would take hours to solve the monopolist's true problem.

[52] Using the analytical solution to the reformulated problem, we are also able to derive a closed-form expression for the impact of a tax on any *single* incremental coverage level. The logic follows directly from how one would think about the impact of an excise tax within the graphical supply-demand framework, which is useful for providing intuition. We provide and discuss this derivation in Online Appendix B.9.

While linear taxes or subsidies are sufficient for restoring the optimal feasible allocation in a competitive market (Azevedo and Gottlieb, 2017), the monopolist's ability to strategically respond to such an intervention renders it far less useful.

6 Conclusion

We analyze 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 utilitarian social planner, a monopolist, or an insurer with any other objective weighting between consumer surplus, government spending, and profits. Our analysis thus extends the analytical characterization of this class of health insurance problems beyond the perfectly competitive setting (Azevedo and Gottlieb, 2017). In order to capture the many complexities of the health insurance setting, we make few assumptions on primitives, and ask what we can (and cannot) learn generally about the problem. Our approach thus stands in contrast to much of the existing theoretical literature on multidimensional screening, where parameterizations are often highly restrictive. The theoretical tools we develop can be used to describe the optimal menu of contracts and to study positive trade, optimal exclusion, and incentives to screen.

We also further develop the "demand-profile" approach to multidimensional screening problems. We show how this approach can be applied in a selection market, as well as how its application sheds light on when multidimensional screening problems can be approximately interpreted as a series of one-dimensional problems. Our results provide the conditions for how and when the familiar graphical analysis of health insurance markets introduced by Einav et al. (2010a) can be extended to an arbitrary number of vertically-ordered contracts. In the spirit of the original analysis, we view this development as a useful tool with which one can understand the basic theory of multidimensional screening in selection markets with endogenous product quality and the associated implications for welfare and public policy.

Finally, we quantify the magnitudes of theoretically identified effects and use a numerical model to evaluate and illustrate a number of our key results. We find that only a small number of contracts are needed to capture the key economic features of the problem, but that over-restricting the contract space (for example to only two contracts) does have material implications. We also find that in our setting, the reformulated problem provides an excellent approximation of the true problem, and therefore represents a powerful tool for understanding its solution and carrying out policy counterfactuals. We view further exploration of the validity of the demand-profile approach in other empirical settings to be of prime interest.

A number of caveats are in order. First, while our assumption of a single principal is appropriate when thinking about a government or a monopolist, there are many interesting settings in which the market is oligopolistic. Though our results do not speak to such settings, the tools developed here form a basis for a new direction of theoretical exploration. Second, the assumption of CARA preferences crucially suppresses interaction between contract premiums and consumer valuations, and there is good reason to believe that these effects may be important, particularly at lower levels of coverage than we permit in our numerical analysis. As we would no longer be able to work with quasi-linear utility, relaxing this assumption would require a large technical leap. Third, while our insurer objective function is quite flexible, it does not permit differential welfare weights on different types of consumers, something that is likely of central policy interest (as well as directly related to the question of income effects). We view these issues as exciting directions for future work.

References

Akerlof, G. A. (1970). Quality uncertainty and the market mechanism. *Quarterly journal of Economics 84*(3), 488–500.

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.

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

Barelli, P., S. Basov, M. Bugarin, and I. King (2014). On the optimality of exclusion in multi-dimensional screening. *Journal of Mathematical Economics 54*, 74–83.

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.

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. and A. Finkelstein (2011a). Selection in insurance markets: theory and empirics in pictures. *Journal of Economic Perspectives 25*(1), 115–138.

Einav, L. and A. Finkelstein (2011b). Selection in insurance markets: Theory and empirics in pictures. *Journal of Economic Perspectives 25*(1), 115–38.

Einav, L., A. Finkelstein, and M. R. Cullen (2010a). Estimating welfare in insurance markets using variation in prices. *The quarterly journal of economics 125*(3), 877–921.

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 (2010b). Estimating Welfare in Insurance Markets Using Variation in Prices. *The Quarterly Journal of Economics CXV*(3).

Farinha Luz, V., P. Gottardi, and H. Moreira (2022). Risk classification in insurance markets with risk and preference heterogeneity.

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

Gaynor, M., N. Mehta, and S. Richards-Shubik (2023). Optimal contracting with altruistic agents: Medicare payments for dialysis drugs. *American Economic Review*.

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.

Gottlieb, D. and H. Moreira (2023). Market power and insurance coverage.

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.

Jaffe, S. and M. Shepard (2017). Price-Linked Subsidies and Health Insurance Markups. *Working Paper*.

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.

Rothschild, M. and J. Stiglitz (1976). Equilibrium in competitive insurance markets: An essay on the economics of imperfect information. *The Quarterly Journal of Economics 90*(4), 629–649.

Saez, E. (2001). Using elasticities to derive optimal income tax rates. *The review of economic studies 68*(1), 205–229.

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.

Tebaldi, P. (2022, March). Estimating equilibrium in health insurance exchanges: Price competition and subsidy design under the aca. Working Paper 29869, National Bureau of Economic Research.

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.3.

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.*[53]

A.1 Proof of Theorem 1

Recall that by Part (i) of Proposition 3, 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 3 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 $\underline{x}\equiv\underline{x}(\hat{\psi},\rho)$.[54] 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

[53]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.

[54]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 $\hat{\psi}$.

$\underline{x}$ 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

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

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),

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

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 $\rho(x^k) < \rho(x^{k+1})$. □

A.2 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

$$r^l = \frac{\mathcal{S}(x, \psi^l) - \mathcal{S}(\underline{x}^l, \psi^l)}{v_\psi(x, \psi^l) - v_\psi(\underline{x}^l, \psi^l)}, \tag{13}$$

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

$$r^h = \frac{\mathcal{S}(\bar{x}^h, \psi^h) - \mathcal{S}(x, \psi^h)}{v_\psi(\bar{x}^h, \psi^h) - v_\psi(x, \psi^h)}. \tag{14}$$

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

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

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.3 Optimality in the Continuum

In this section, we state the analogs to Theorems 1 and 3 when the insurer can offer a continuum of contracts. 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 4 (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.4 Proof of Positive Trade Result

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, that* $\bar{F}$ *is non-degenerate, and that* $x^0 < 1$*. Then 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.*

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.5 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 3 (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. □

Proof of Theorem 2 Immediate from Lemmas 1 and 2. □

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.6 Equation (12) 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 $\tilde{\psi}^k(p^k,\omega,F)$.[55] 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

(15)
$$\begin{aligned}\tilde{\mathcal{V}}^k(p^k,\omega,F) &= (w^I - w^C)(1 - G(\tilde{\psi}^k(p^k,\omega,F)|\omega,F)) \\ &\quad - \tilde{\psi}^k_{p^k}(p^k,\omega,F)(w^I(p^k - \gamma^{I,k}(\theta)) - (w^G - w^I)\gamma^{G,k}(x^0,\theta))g(\tilde{\psi}^k(p^k,\omega,F)|\omega,F).\end{aligned}$$

[55] 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.

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

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

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,

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

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.7 Proof of Proposition 1

Let $\underline{v}^k = \min_{\omega \in [\underline{\omega},\bar{\omega}]} v^k(\omega,\underline{\psi},\underline{F})$, noting that by Lemma 3, $v^k(\theta) \geq \underline{v}^k$ 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 Demand for Insurance

Proposition 3 (Properties of Insurance Demand) *The consumer's demand for insurance satisfies 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 $\{f(\cdot|t)\}_{t\in[0,1]}$ is a parameterized family of densities ordered by strict monotone-likelihood-ratio property (MLRP), then $v_{xt} > 0$, and thus $\chi(\omega, \psi, \cdot)$ is increasing in t;*[56] *(iii) if $b(a, l, \omega) = \hat{b}(a - l, \omega)$ and c is convex in a (including the case in which c is linear in a), then $v_{x\omega} > 0$, and thus $\chi(\cdot, \psi, F)$ is increasing in ω.*

Proof 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, t). But, for $t_h > t_\ell$ we have

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

since $\{f(\cdot|t)\}_{t\in[0,1]}$ 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

$$\int b_{\omega a} a_x^* m dl + \int b_\omega m_x dl > 0. \tag{16}$$

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 (16) is 0. □

Note that part (iii) includes the case in which c is linear in a (linear insurance contracts).

Technical Remark 5 (Demand for Insurance and ω) One can also show that $\chi(\cdot, \psi, F)$ is increasing when F is dirac at 0.[57] Beyond these cases, the comparative statics in ω are complex.

[56] 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.

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

One can show that $v_{x\omega} > 0$ if and only if

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

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.

B.2 Differentiability of Costs

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 $\bar{\kappa}$, $c(\cdot,x) = \min_{\kappa\in\{1,\dots,\bar{\kappa}\}} \tilde{c}(\cdot,x,\kappa)$ 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,

$$\begin{aligned}
\text{(17)} \qquad \tilde{l}_\omega(x, \omega, \kappa', \kappa'') &= -\frac{\tilde{z}_\omega(\tilde{l}, x, \omega, \kappa'') - \tilde{z}_\omega(\tilde{l}, x, \omega, \kappa')}{\tilde{z}_l(\tilde{l}, x, \omega, \kappa'') - \tilde{z}_l(\tilde{l}, x, \omega, \kappa')} \\
&= -\frac{b_\omega(\tilde{a}(\tilde{l}, x, \omega, \kappa''), \tilde{l}, \omega) - b_\omega(\tilde{a}(\tilde{l}, x, \omega, \kappa'), \tilde{l}, \omega)}{b_l(\tilde{a}(\tilde{l}, x, \omega, \kappa''), \tilde{l}, \omega) - b_l(\tilde{a}(\tilde{l}, x, \omega, \kappa'), \tilde{l}, \omega)} \\
&= -\int_{\tilde{a}(\tilde{l},x,\omega,\kappa')}^{\tilde{a}(\tilde{l},x,\omega,\kappa'')} \frac{b_{\omega a}(a, \tilde{l}, \omega)}{b_{la}(a, \tilde{l}, \omega)} \frac{b_{la}(a, \tilde{l}, \omega)}{\int_{\tilde{a}(\tilde{l},x,\omega,\kappa')}^{\tilde{a}(\tilde{l},x,\omega,\kappa'')} b_{la}(a, \tilde{l}, \omega) da} da,
\end{aligned}$$

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 (17), 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. $\square$

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,* $\left(\int_0^{\bar{l}} c(a^*(l, x, \omega), x) f(l) dl\right)_x$ *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.3 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 $\tilde{x} \in [0, 1]$. Let $\hat{x}^n = \max(x^n - d(\rho^n, \rho), 0)$, 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 18) 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

$$v(x^n,\theta^n) - \rho^n(x^n) \geq v(\grave{x}^n,\theta^n) - \rho^n(\grave{x}^n), \tag{18}$$

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 (18), $\limsup_n(v(x^n, \theta^n) - \rho^n(x^n)) \geq \limsup_n(v(\grave{x}^n, \theta^n) - \rho^n(\grave{x}^n))$, 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.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).[58] 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

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

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

[58] 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.

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

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

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

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.5 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.2 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

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

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

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

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 $\widetilde{\mathcal{W}}(\hat{x}, \omega, F | \tau, k) +$

$O(\hat{x}, \omega, F|\tau, k)$, where

$$(23)\quad \tilde{\mathcal{W}}(\hat{x}, \omega, F|\tau, k) = -v_x(x^k_{j^h(\tau,k)}, \psi^k_{j^h(\tau,k)})r^k_{j^h(\tau,k)}g(\psi^k_{j^h(\tau,k)}) + \int_{\psi^k_{j^l(\tau,k)-1}}^{\psi^k_{j^h(\tau,k)}} S_x(\chi^k(\psi), \psi)dG(\psi)$$
$$+v_x(x^k_{j^l(\tau,k)}, \psi^k_{j^l(\tau,k)-1})r^k_{j^l(\tau,k)-1}g(\psi^k_{j^l(\tau,k)-1})).$$

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,

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

and so, since

$$O(\hat{x}, \omega, F|\tau, k) = \sum_{j^l(\tau,k)}^{j^h(\tau,k)-1} (v_x(x^k_{j+1}, \psi^k_j) - v_x(x^k_j, \psi^k_j))r^k_j g(\psi^k_j).$$

we have

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

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

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,

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

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

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

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 $\max\{\psi | \underline{x}(\psi, \rho^k) \leq x\}$ and $\psi^*(x)$ as $\max\{\psi | \underline{x}(\psi, \rho) \leq x\}$. 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, $\underline{x}^*(x) < x \leq \lim_{k\to\infty} \bar{x}^k(x) \in X(\psi^*(x), \rho)$, 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. $\square$

B.6 Ironing and the One-Contract Case

We asserted in Section 3.5 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 3.6 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 $\int \mathcal{V}^M(x^0, \omega, F) dG(\omega, F) = 0$.[59] The term $\rho(x^1) - (\gamma^I(x^1) - \gamma^G(x^1, x^0))$ represents the profit the insurer was making on consumers 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.[60]

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.[61] 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

---

[59] 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).

[60] 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.

[61] 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.

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, which is echoed in our numerical analysis. We now provide a theoretical example, albeit in a one-dimensional setting, 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.[62] We must limit attention to the one-dimensional problem to gain analytical tractability. 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 $b(a, l, \omega) \equiv \hat{b}(a - l, \omega)$.[63] 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 5.

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 $x\varphi_{(1-x)\omega}(1 - x, \omega)$.[64] One can show that this is *strictly*

[62] 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.

[63] 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.

[64] 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

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.[65] 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.[66] 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."

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.[67] 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.[68] 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.[69]

into the cross-partial derivative of the objective function we obtain $v_{x\omega} - \int a^*_\omega dF - x\int a^*_{x\omega}dF = x\varphi_{(1-x)\omega}$.

[65] 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.

[66] 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 $U(\omega) = U(\underline{\omega}) + \int_{\underline{\omega}}^{\omega} v_\omega(\chi(s),s)ds$.

[67] The algebraic steps are as follows. First, use the incentive compatibility characterization in Footnote 66 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.

[68] 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.

[69] 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

B.8 Computational Details

Simulated Population of Consumers. 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. 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 65. We assume that 90 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 65. We further assume each household has between 1 and 4 children, where each child exists with 15 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.[70] 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 Online Appendix Figure B.1.

Numerical Algorithm for Computing Optimal Menus. We calculate optimal premium schedules numerically given a fixed set of potential contracts $\{x^k\}_{k=0}^K$. Note that we cannot calculate optimal menus in the case of a continuum of contracts because there is not a closed form solutions for key required objects: consumer utility $v(\theta, x)$ and insurer costs $\gamma^I(\theta, x)$. These objects must therefore be pre-calculated for each consumer type θ and each pre-specified contract x. The largest number of contracts on which we calculate optimal menus is 65. A powerful implication of our theoretical convergence result is that this approach approximates the continuum case.

Our numerical algorithm for finding optimal prices mirrors the logic of the perturbation ar-

(33) is equal to $-x - 0 + x((g'/g)((1-G)/g) + 1)$, which is strictly positive if $g' > 0$.

[70] 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.

gument underlying our necessary conditions stated in Theorem 1. The algorithm proceeds as follows. Start from a candidate price schedule $\rho(x)$, and let $p^k = \rho(x^k) - \rho(x^{k-1})$ be the incremental premium between adjacent contracts. Starting from the first increment $k = 1$, consider a small perturbation to the incremental premium p^k, holding all other incremental premiums fixed. According to Theorem 1, at any optimal menu, this perturbation should not have a first order effect on the insurer's payoff. Use an unconstrained optimizer to find a locally optimal p^1, around which the insurer's payoff cannot be improved.[71] Proceed to the second increment $k = 2$ and repeat this process, now optimizing over p^2, holding all other incremental premiums fixed. Proceed through all remaining increments up to K. At this point, restart at $k = 1$, and repeat the entire loop again. Once payoffs are unresponsive (within some tolerance) to small perturbations at every incremental premium k, a price schedule that fulfills the necessary conditions for local optimality has been found.

Given the complexity of this nonlinear optimization problem, the standard caveat applies that it is impossible to guarantee that a local optimum is a global optimum. In principle, this is exactly the same roadblock that prevents us from analytically deriving sufficient conditions for optimality in the true problem. We calculate local optima starting from the solution to the simplified version of the problem, as well as starting from 100 random starting values. The random starting values in general do not do better than starting from the solution to the simplified problem.

B.9 Derivation of BFIB

Using the simplified version of the problem, we can derive an analytical expression for the local impact of taxes or subsidies in a monopoly market. To begin, fix a single incremental coverage level of interest and suppress k. Define a subsidy scheme $\sigma(q|s)$, where $s \in \mathbb{R}$ is a generosity parameter and σ represents the total dollar amount of subsidies paid to the monopolist when it serves q consumers. Assume that $s = 0$ corresponds to no subsidies ($\sigma(q|0) \equiv 0$) and that higher s corresponds to higher subsidies both as an absolute and at the margin ($\sigma_s(q|s) \geq 0$ and $\sigma_{qs}(q|s) \geq 0$). Positive s therefore corresponds to a subsidy, while negative s corresponds to a tax (we will use the generic term *subsidy* throughout). A linear subsidy scheme (that is, linear in the number of consumers served) would be given by $\sigma(q|s) = sq$.

Given a subsidy function $\sigma(q|s)$, the goal is to find a regulator's optimal subsidy level s. To this end, denote the optimal quantity of consumers served by the monopolist facing subsidy level s is given by

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

We assume the subsidy function σ has enough regularity that $q(s)$ is well-defined and continuously differentiable. Suppose the regulator has an objective function that takes the same form as an insurer's—as given by equation (3)—just with different weights. Suppose the regulators objective weights are $\tilde{w} = (\tilde{w}^C, \tilde{w}^I, \tilde{w}^G)$. The payoff to the regulator of implementing subsidy level s in a

[71] We use the commercial optimization packages available through MATLAB.

monopoly insurance market is then given by

$$(34)\quad \underbrace{\tilde{w}^C \int_0^{q(s)} P(q')dq' + (\tilde{w}^I - \tilde{w}^C)P(q(s))q(s) - \tilde{w}^I C(q(s)) + \tilde{w}^I \sigma(q(s)|s)}_{\text{Benefit of subsidy, } \beta(s)} - \underbrace{\tilde{w}^G \sigma\left(q(s)|s\right)}_{\text{Cost of subsidy}},$$

Let *bang for incremental buck* (BFIB) be defined as $\beta_s(s)/(\sigma(q(s)|s))_s$, where $(\cdot)_s$ denotes the total derivative with respect to s. It is the marginal benefit the regulator realizes on an extra dollar spent on subsidies starting from level s. Note that the marginal cost of subsidies is linear in the weight placed on government spending, $\tilde{w}^G$. Assuming that the regulator's problem of choosing s is characterized by the first-order condition, comparing BFIB to $\tilde{w}^G$ therefore tells us whether the regulator wishes to increase or decrease the subsidy. If BFIB $> \tilde{w}^G$, then the marginal benefit of an increase in subsidy level exceeds its marginal cost, and the regulator will optimally increase s. If BFIB $< \tilde{w}^G$, then the opposite is true, and the regulator optimally reduces the subsidy.

Differentiate the regulator's benefit function $\beta(s)$, we obtain

$$\beta_s(s) = q_s(s)\big(\tilde{w}^C P(q(s)) - \tilde{w}^I MC(q(s)) + (\tilde{w}^I - \tilde{w}^C)MR(q(s))\big) + \tilde{w}^I\big(\sigma_q(q(s)|s)q_s(s) + \sigma_s(q(s)|s)\big).$$

We can simplify by applying the identity $MR(q(s)) = MC(q(s)) - \sigma_q(q(s)|s)$, which holds by the optimality of $q(s)$. After simplification, we can rewrite $\beta_s(s)$ as follows:

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

Hence, BFIB is given by the following expression

$$\text{BFIB} = \frac{q_s(s)\tilde{w}^C\big(P(q(s)) - MC(q(s)) + \sigma_q(q(s)|s)\big) + \tilde{w}^I \sigma_s(q(s)|s)}{\sigma_q(q(s)|s)q_s(s) + \sigma_s(q(s)|s)}.$$

We can simplify further by differentiating the identity $MR(q(s)) - MC(q(s)) + \sigma_q(q(s)|s)) = 0$. This yields an expression for the marginal impact of the subsidy policy on the monopolist's chosen quantity:

$$q_s(s) = \frac{\sigma_{qs}(q(s)|s)}{MC_q(q(s)) + \sigma_{qq}(q(s)|s) - MR_q(q(s))},$$

which is positive (since the numerator is positive by assumption and the denominator is positive since it is equal to the negative of the second-order necessary condition). The monopolist therefore always increases quantity in response to increased subsidies.

Inserting $q_s(s)$ into the expression for BFIB yields

$$\text{BFIB} = \frac{\sigma_{qs}(q(s)|s)\tilde{w}^C \frac{P(q(s)) - MC(q(s)) + \sigma_q(q(s)|s)}{MC_q(q(s)|s) + \sigma_{qq}(q(s)|s) - MR_q(q(s)|s)} + \tilde{w}^I \sigma_s(q(s)|s)}{\sigma_q(q(s)|s)\frac{\sigma_{qs}(q(s)|s)}{MC_q(q(s)) + \sigma_{qq}(q(s)|s) - MR_q(q(s)|s)} + \sigma_s(q(s)|s)}.$$

Given a linear subsidy $\sigma(q|s) = qs$ and regulatory objective weights $\tilde{w} = (1, 0, 1)$, BFIB reduces to

$$\text{BFIB} = \frac{\frac{P(q(s))-MC(q(s))+s}{MC_q(q(s))-MR_q(q(s))}}{\frac{s}{MC_q(q(s))-MR_q(q(s))} + q(s)}.$$

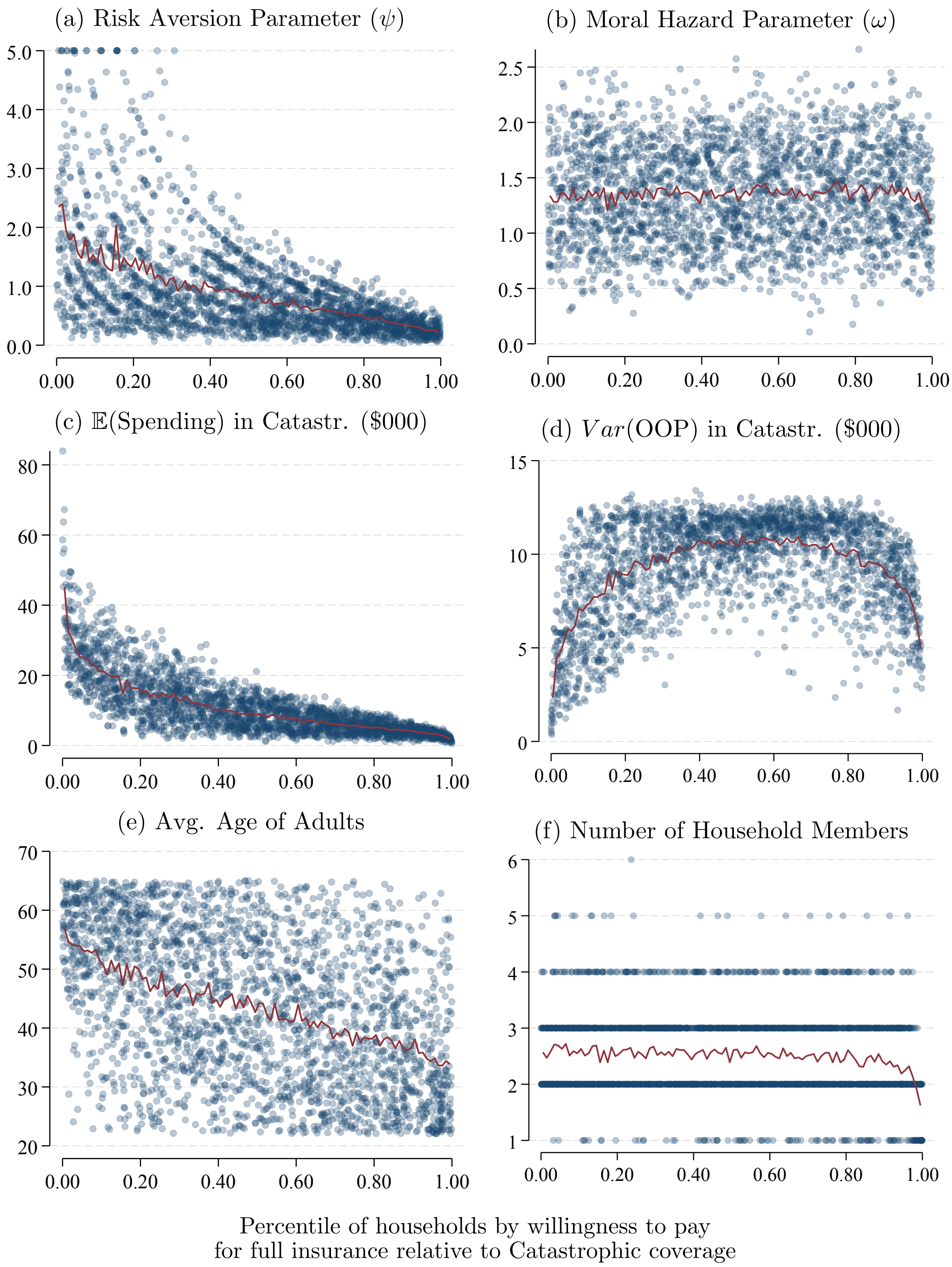
We can use this expression to analyze the special case of *whether* to subsidize (or tax) at all. Consider the first dollar of subsidy by setting $s = 0$, so that $q(0) = q^m$ (the monopolist's optimal quantity). In this case, the expression simplifies to:

$$\text{BFIB} = \frac{1}{q^m} \frac{P(q^m) - MC(q^m)}{MC_q(q^m) - MR_q(q^m)},$$

and $BFIF$ can be easily interpreted. First, $1/q^m$ is the amount by which the first dollar spent on subsidies lowers the monopolist's effective marginal cost curve. Next, $1/(MC_q - MR_q)$ is the amount by which a decrease in marginal cost leads the monopolist to increase quantity. Finally, $P - MC$ reflects the effect of a change in quantity on consumer surplus. We can evaluate BFIB numerically on each incremental coverage level at the monopolist's optimal allocation. In each case, we find that it falls below one, meaning that starting from zero subsidies, the regulator would—at the margin—be better off *taxing* incremental coverage, consistent with our numerical results.[72]

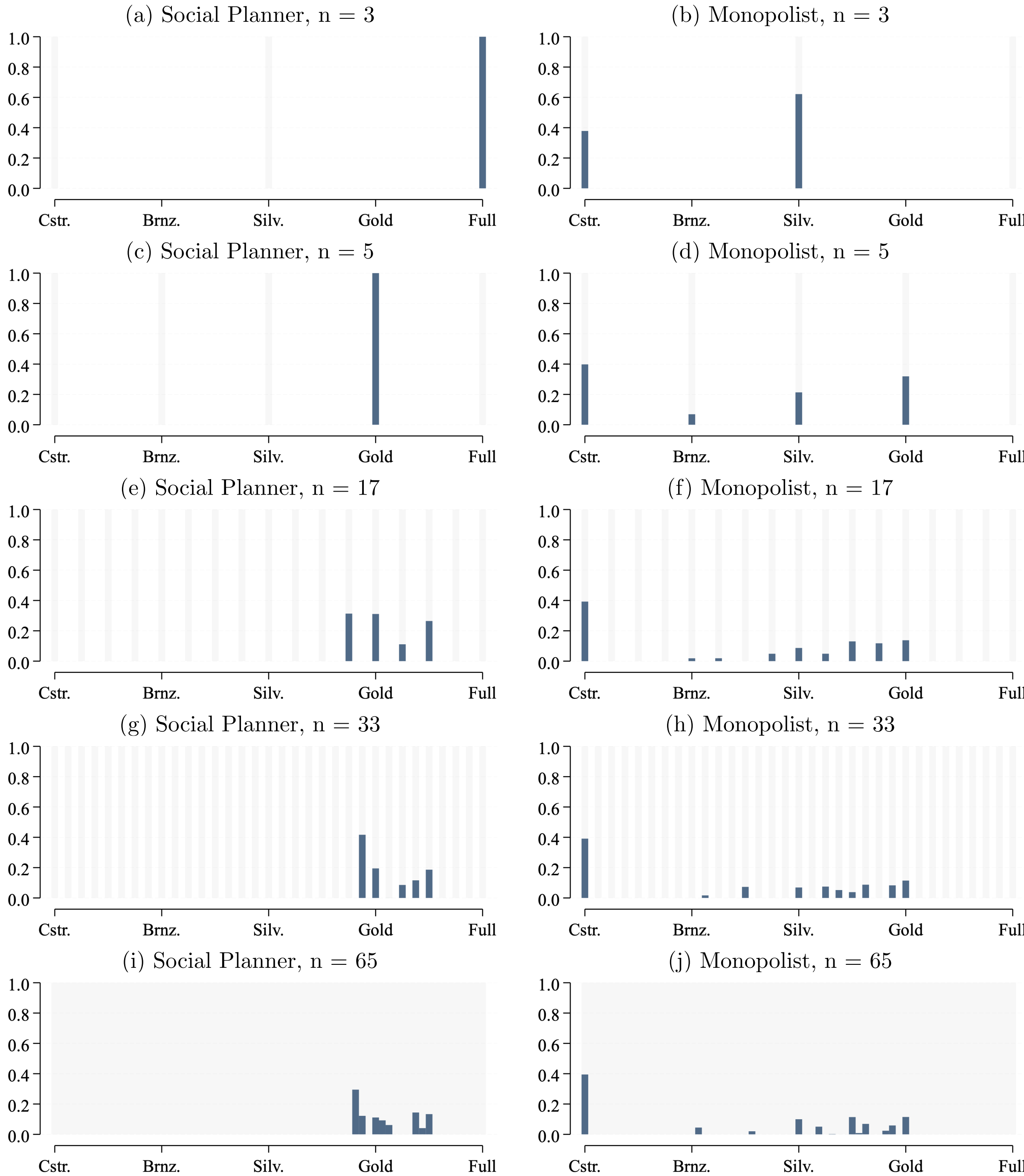
[72]On the margin between full insurance and Gold, BFIB does not exist because q^m is zero. On the margins between Gold and Silver, Silver to Bronze, and Bronze to Catastrophic, BFIB is 0.33, 0.34, and 0.30, respectively.

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.

Table B.1. Contract-by-Contract Markups

	Average cost *\$000s*				Premium *\$000s*				Premium / Average cost			
Scenario	Brnz.	Slvr.	Gold	Full	Brnz.	Slvr.	Gold	Full	Brnz.	Slvr.	Gold	Full
Social planner		0.16	4.22		0.16	0.34	0.77	3.27		2.17	0.18	
Social planner, 25% ECPF	0.13	1.02	4.86	9.78	1.49	2.80	4.64	7.10	11.52	2.75	0.95	0.73
Monopolist	0.44	2.34	5.91		2.01	4.13	6.49	9.01	4.57	1.76	1.10	
Competitive equilibrium	0.97	3.42	5.92	8.41	0.97	3.42	5.92	8.41	1.00	1.00	1.00	1.00

Notes: The table shows average costs, premiums, and markups (the ratio of premium to average cost) at the optimal menu of our three focal insurers as well as at the competitive equilibrium. Costs are incremental relative to the Catastrophic contract, and note that the premium of the Catastrophic contract is always zero.