Price Impact of Insurance*

Andrea Di Giovan Paolo[†] Jose Higueras[‡]

Job Market Paper

September 4, 2026

Abstract

This paper analyzes optimal insurance design when the insurer internalizes the effect of coverage on third-party service prices. A monopolistic insurer contracts with risk-averse agents who have sequential two-dimensional private information and preferences represented by Yaari's dual utility. Insurance contracts shape service demand and, through a market-clearing condition, determine equilibrium third-party prices. We show that contracts take simple forms—either full coverage after a deductible is paid or limited coverage with an out-of-pocket maximum—closely mirroring real-world insurance plans. Technically, we formulate the problem as a sequential screening model and solve it using tools from optimal transport theory.

Keywords: insurance, mechanism design, market design

JEL Codes: I11, I13, D47, D81, D82, D86

1 Introduction

Insurance plays a crucial role in mitigating risks by covering a portion of the monetary losses customers incur following adverse events. These monetary losses typically depend on prices set by third-party service markets. For example, in healthcare insurance, the monetary loss is directly tied to the price of hospital services. In car insurance, the loss is determined by the price of auto repair services, whereas in home insurance, it depends on the price for property restoration services. While the mechanism design literature has extensively studied the optimal design of insur-

*We are deeply grateful to Piotr Dworczak, Jeff Ely and Alessandro Pavan for their guidance. All errors remain our own.

[†]Department of Economics, Northwestern University. Email: andreadigiovanpaolo@u.northwestern.edu

[‡]Department of Economics, Northwestern University. Email: josehiguerascorona2025@u.northwestern.edu

ance contracts under adverse selection and moral hazard (Stiglitz, 1977; Grossman and Hart, 1983; Chade and Schlee, 2012; Gershkov et al., 2023), less attention has been paid to the equilibrium effects of insurance. The type and level of insurance coverage can influence the prices charged by third-party providers. In health insurance, for instance, more generous coverage encourages greater utilization of hospital services, raising overall demand and potentially driving medical prices upward (Feldstein, 1970). These higher prices, in turn, feed back into the insurer's decision about both the extent and structure of coverage: they increase the cost of providing insurance and, more subtly, affect the value that policyholders attach to being insured. Similar feedback loops can arise in auto and home insurance, where contract design influences the demand for repair and restoration services, thereby affecting market prices. Given that large[1] insurance companies are likely to internalize these feedback effects when designing contracts, a key question arises: how does the interaction between insurance and third-party service markets shape both the *types* of contracts offered and the *price* for these services?

This paper addresses these questions by incorporating equilibrium effects into a mechanism design framework. We study the problem faced by a risk-neutral monopolistic insurer offering contracts to a population of risk-averse agents, and building on Gershkov, Moldovanu, Strack and Zhang (2023) we model their preferences using Yaari's dual utility (Yaari, 1987). In contrast to the standard expected utility framework—where risk-averse agents evaluate lotteries by applying a concave utility function to outcomes—dual utility models assume linear utility over outcomes and capture risk aversion through probability distortions.

Agents are sequentially and privately informed about two characteristics: their ex-ante risk type—such as the probability of developing an illness or suffering a car accident—and their ex-post loss-related valuation, which reflects the monetary loss they associate with not treating or repairing the damages resulting from an adverse event. This valuation determines their willingness to pay for third-party services.

In the first period, when the agent knows only her risk type and her loss-related valuation has not yet been realized, she can purchase an insurance contract from the monopolistic provider. An insurance contract has two parts. First, a premium that the agent must pay to obtain coverage.

[1] According to the latest report on competition in health insurance (American Medical Association, 2024), the American Medical Association (AMA) finds that, based on the DOJ/FTC Merger Guidelines, 95% of metropolitan commercial markets were highly concentrated in 2023, with a Herfindahl–Hirschman Index (HHI) above 1800. The average commercial market exhibited an HHI of 3458. Moreover, in 89% of metropolitan markets at least one insurer held a market share of 30% or more, and in 47% of markets a single insurer's share reached at least 50%. These findings indicate that many health insurance markets display monopolistic characteristics.

Second, a set of options, where each option specifies a quantity or fraction of services covered by insurance together with the out-of-pocket payment the agent would face if she later uses that option.

In the second period, once the adverse event occurs and the agent learns her loss-related valuation, she can repair damages or seek treatment from the third-party service provider. Services are supplied competitively at a market price, and the agent may demand any quantity. If she purchased an insurance contract in the first period, she may exercise one of the coverage options: doing so entitles her to obtain the specified quantity of services by paying only the associated out-of-pocket cost.

The insurance provider chooses a menu of contracts to maximize profits. Importantly, the insurer internalizes that each contract affects an agent's decision of whether to seek third-party services and how much to consume. As a result, the menu of contracts shapes aggregate demand in the competitive service market. Through the market-clearing condition—requiring demand to equal available supply—this demand determines the equilibrium service price.

We adopt a two-stage approach to determine the structure of the contracts offered by the insurer and the resulting equilibrium service price: 1) In the first-stage problem, we fix a price for services and identify, among the menus that sustain that price in equilibrium, the one that maximizes the insurer's profits. 2) In the second-stage problem we associate each achievable price in equilibrium with the corresponding profits obtained by the insurer, and then solve for the price that maximizes these profits.

Our first main result, Theorem 1, shows that—under a regularity condition—the optimal first-stage contracts take a simple and intuitive form that closely resembles real-world insurance plans. Crucially, one of these contract types arises *because of the insurer's price impact*. Specifically, the contract offered to each risk type takes one of the following two forms:

1. **Simple deductible contract.** This contract is analogous to standard deductible-based insurance: once the agent pays a fixed amount out-of-pocket, insurance covers the remaining cost of the service in full.

2. **Limited coverage deductible contract.** Under this contract, after the agent pays a fixed amount out-of-pocket, insurance covers a fraction $\alpha \in (0, 1)$ of the service. If the agent wants the remaining fraction $1 - \alpha$, she must pay the full service price. Alternatively, by paying a higher out-of-pocket amount, the agent can receive full coverage without limits.

In our model, the insurer faces a novel trade-off that goes beyond the classic screening frictions studied in traditional insurance models (see, for instance, Stiglitz (1977)). On the one hand, offering more generous coverage increases agents' demand for services, which in turn raises the equilibrium service price. On the other hand, offering less coverage reduces aggregate demand and lowers the equilibrium price. In other words, every contract the insurer offers exerts a *price impact*. This price effect provides an economic rationale for the *limited-coverage deductible contract*.

Intuitively, if the insurer ignored this price impact—treating service prices as fixed—he would prefer to offer all agents a *simple deductible plan*. However, once equilibrium price effects are taken into account, such contracts can generate either excessive or insufficient utilization. Under a simple deductible plan, the effective price of services faced by policyholders is entirely determined by the deductible. If the deductible is too low (respectively, too high), utilization becomes too high (respectively, too low).

The insurer's optimal response is therefore to introduce a limited-coverage deductible contract. When utilization is excessive, the insurer can limit coverage beyond the deductible, while still allowing policyholders to obtain full coverage by paying more out of pocket. Conversely, when utilization is too low, the insurer can reduce the deductible but cap coverage, while still allowing policyholders to achieve full coverage by paying the previous higher deductible—that is, by making the previous deductible the out-of-pocket maximum.

This theoretical prediction echoes real-world practice. Coinsurance or cost-sharing arrangements typically cover only a fraction α of treatment costs, with policyholders responsible for the remainder. Many plans also include an out-of-pocket maximum, beyond which the insurer covers all costs. More generally, the parameter α can capture common service restrictions. Medicare Part A covers only up to 90 days of inpatient hospital care per benefit period; additional days must be paid entirely out-of-pocket. Private plans often cap the number of physical therapy or mental health sessions, drug plans restrict the quantity of medication dispensed or the frequency of refills, and advanced diagnostic procedures such as MRI, CT, or PET scans may not be fully covered.

From a technical standpoint, our contribution is to extend sequential screening to environments in which the mechanism designer influences market outcomes—a combination we analyze using tools from optimal transport theory. In our setting, the insurer engages in sequential screening: in the first period, the agent knows only her risk type, and in the second, she learns her loss-related valuation. This two-stage information structure introduces additional complexity relative to standard one-dimensional mechanism design, where the allocation rule maps a single type directly to an outcome.

To address this complexity, we formulate a relaxed version of the insurer's problem that retains all constraints a menu of contracts must satisfy, except for period 1 incentive compatibility (i.e., truthful reporting of risk types). In particular, we impose local but not global incentive compatibility. We then express this relaxed problem in the *quantity space*—an adaptation of Myerson (1981)'s ironing technique, recently applied in related mechanism-design settings (see, i.e., Dworczak et al. (2021); Loertscher and Muir (2022)). Under this formulation, the insurer selects, for each risk type, a lottery over quantities, where quantities represent the measure of agents of that type who obtain coverage, and the corresponding probabilities represent coverage levels. This representation recasts the insurer's problem as an *optimal-transport–like* problem in which each risk type is effectively matched to the mass of agents who receive coverage under the mechanism. Leveraging duality results from optimal transport theory, we characterize key properties of the optimal contracts—most notably, Theorem 1—and use the same dual framework to derive conditions under which the relaxed solution is implementable, and thus optimal.

Our second main result, Theorem 3, shows how the insurer's profits change with the equilibrium service price. This result plays a central role in the second-stage analysis, where the insurer endogenously shapes market prices through the menu of contracts it offers to agents. More specifically, the result reveals two opposing economic forces that drive the insurer's coverage incentives. On the one hand, one force encourages the insurer to *reduce coverage* across the population of risk-averse agents in order to limit utilization, restrain demand, and lower service prices. On the other hand, the opposing force motivates the insurer to *expand coverage*, which raises utilization, increases demand, and drives prices upward.

The first force is mechanical and reflects the *cost of increasing the price*. A unit increase in the price leads to a proportional increase in the insurer's payouts for a given level of coverage.

The second force is more subtle and concerns *surplus extraction*. As the service price increases, insured agents place greater value on their contracts. A higher service price implies greater risk, since larger monetary losses can now occur with positive probability. In other words, the distribution of losses shifts upward in the sense of first-order stochastic dominance. Covered agents are protected from these higher prices, and under dual utility they distort probabilities, assigning a non-linear value to the likelihood of such losses. The insurer can capture this additional value—net of information rents—through premiums.

After identifying these economic forces, we introduce the concept of a *price-taking benchmark*. This benchmark is the price that clears the service market and solves the insurer's first-stage problem in the absence of equilibrium effects. In other words, it is the market-clearing price that would arise if the insurer treated prices as exogenous and ignored its influence on service allocation. We then ask: under what conditions will an insurer with price impact endogenously push the equilibrium price—through the expansion or reduction of coverage—above or below this benchmark?

In Proposition 6, we provide sufficient conditions on the model primitives under which the more subtle *surplus extraction force* dominates the *cost effect* at every price below the benchmark. When this occurs, the insurer has incentives to expand coverage across the population of risk-averse agents relative to the price-taking benchmark, thereby inducing an equilibrium service price above it.

Intuitively, this condition implies that if the surplus extraction force—that is, the value of insurance net of information rents—increases with the service price, or even if it decreases but at a rate bounded by the arrival rate of losses, then the resulting equilibrium price will lie above the benchmark. In this case, the insurer finds it optimal to increase coverage, since the additional surplus it can extract from agents outweighs the higher payouts associated with a higher service price.

Nonetheless, unlike in standard monopoly pricing models—where the monopolist typically sets a price above the price-taking benchmark—in our setting the optimal service price may lie either above or below it. The key distinction lies in the nature of the trade-offs faced by the insurer. In standard models, the monopolist benefits unambiguously from raising the price at the benchmark, as doing so increases profits per unit sold. In contrast, the insurer in our framework balances a potentially positive gain from surplus extraction against a negative effect from higher payouts. As a result, the net benefit of increasing the price at the benchmark can be either positive or negative, leading to qualitatively different pricing behavior.

Finally, we present three extensions of our baseline model that preserve its core economic insights and main results. In the first extension, we show how our framework applies to settings in which the insurance provider acts as a social planner with redistributive objectives, in the spirit of Akbarpour et al. (2024). That is, the insurer places positive weight not only on revenue, but also on agents' utility. This setting captures environments where the state is the main provider of insurance or where insurance markets are highly regulated. In the second extension, we reinterpret our baseline model in a setting where the service price is not determined by a competitive market but instead negotiated through Nash bargaining between the insurer and a capacity-constrained third-party provider, with the insurer holding all bargaining power. We also discuss how the main results extend to settings in which the provider has some bargaining power. The third extension considers a setting in which the provider offers multiple tiered services each with its own price.

Related literature. The relationship between insurance and third-party service prices has received considerable empirical attention since the seminal work of Feldstein (1970). More recent studies estimate hospital–insurer bargaining models, allowing researchers to address key questions such as the division of surplus under hospital capacity constraints (Ho, 2009), the consequences of hospital mergers (Gowrisankaran et al., 2015), and the effects of reduced insurer competition on welfare, hospital prices, and premiums (Ho and Lee, 2017), among others. However, this line of research typically analyzes bargaining dynamics while taking the form of insurance contracts as given and does not consider information frictions such as moral hazard or adverse selection. By contrast, the mechanism design literature—although centered on optimal insurance design under information frictions—has paid relatively little attention to these market interactions. One of this paper's main contributions is therefore to develop a mechanism-design framework that makes it possible to analyze, in a tractable way, optimal insurance design when insurers account for the feedback loop between coverage and service prices.

More precisely, this paper contributes to two main strands of literature. The first is the extensive theoretical body of work on insurance. A large literature following Borch (1960) and Arrow (1963) studies welfare-maximizing insurance policies in the absence of information frictions. For example, Raviv (1979) shows that coinsurance arrangements are welfare maximizing when the insurer faces an exogenous cost of provision that is increasing and convex in coverage.

A pioneering literature starting with Zeckhauser (1970) considers ex post moral hazard, where insurance contracts influence agents' expenditure and treatment decisions, and in richer formulations, physicians' effort (as in Ma and McGuire (1997)). In this papers, the price or cost of care is fixed, so there is no feedback loop between insurance coverage and market prices. Moreover, the insurer's objective is usually to maximize the expected utility of the insured subject to an actuarially fair premium.

The line of work described above does not incorporate adverse selection. Models with adverse selection follow the seminal contributions of Rothschild and Stiglitz (1978), who study competition among insurers, and Stiglitz (1977), who analyzes the monopoly case. Chade and Schlee (2012) provide a comprehensive and up-to-date analysis of monopolistic profit maximization in insurance markets with adverse selection and expected-utility agents. Of particular relevance to our analysis is the recent work of Gershkov et al. (2023), which revisits the monopolistic insurance problem and develops tools that are especially useful for our framework. They depart from the canonical model in two key respects:

First, they allow for general loss distributions that may be correlated with agents' private information. Second, they assume agents are risk-averse with preferences represented by a dual utility function (Yaari, 1987). Unlike the expected utility framework, dual utility captures risk aversion through a non-linear transformation of probabilities rather than payoffs. As shown by Yaari (1987), this representation arises from a single deviation from expected utility theory: rather than assuming independence concerning probabilistic mixtures of lotteries, it assumes independence concerning mixtures of payments associated with those lotteries.

Gershkov et al. (2023) argue that dual utility is a compelling framework for several reasons. First, empirical studies (Sydnor, 2010; Barseghyan et al., 2013, 2016) document insurance choices that are difficult to reconcile with expected utility but align well with models incorporating probability weighting—a feature inherent to dual utility. Second, dual utility disentangles risk aversion from wealth effects, allowing the framework to capture risk aversion while preserving constant marginal utility of wealth, a property that together with not distorting payoffs enhances tractability relative to traditional frameworks.

Motivated by these insights, we also endow agents with Yaari preferences. However, relative to Gershkov et al. (2023) and the broader theoretical literature, our focus is distinct: we study how the structure of optimal insurance contracts changes once we account for the feedback loop between coverage and third-party service prices. To capture this feedback, we analyze a fundamentally different environment. Monetary losses are not exogenous but arise from adverse events that can only be treated through third-party services. Insurance is therefore effective only insofar as it stimulates utilization—implying that coverage affects the demand for services, which must equal supply through market clearing. This interaction determines the equilibrium service price, closing the feedback loop between coverage and market prices. In addition, agents possess two dimensions of private information revealed sequentially, capturing richer selection dynamics and expanding the set of feasible contract forms. Recognizing the insurer's price impact allows us to provide an economic rationale for limited-coverage deductible contracts. Together, these features show that equilibrium price effects fundamentally reshape optimal insurance design and help explain contract structures observed in practice.

The second strand of literature we contribute to is the growing work on partial mechanism design, where designers lack full control over the primary market and must consider the effects of their mechanisms on secondary trading venues. This framework has been applied to financial markets (Philippon and Skreta, 2012; Tirole, 2012), monopoly pricing with resale (Calzolari and Pavan, 2006a; Loertscher and Muir, 2022), information transmission to aftermarkets (Calzolari and Pavan, 2006b; Dworczak, 2020), optimal redistribution (Kang, 2023; Kang and Watt, 2024), contracting between firms (Calzolari and Denicolo, 2015; Kang and Muir, 2022), and policy interventions based on market outcomes (Valenzuela-Stookey and Poggi, 2020). We extend this literature by applying partial mechanism design to insurance markets with risk-averse agents holding non-standard preferences, where third-party service prices are endogenously determined by the insurer's mechanism and the available supply. Technically, our framework differs from prior studies in one key respect: we address the insurer's price impact within a sequential screening framework, similar to Courty and Hao (2000), where mechanisms resemble "menus of menus." To tackle this complexity, we reformulate the problem using an optimal transport approach.

The remainder of the paper is organized as follows. Section 2 introduces the model. Section 3 defines the first-stage problem and presents our first main result, Theorem 1. Section 4 turns to the second-stage problem and presents our second main result, Theorem 3. Section 5 discusses extensions of the baseline model, and Section 6 concludes.

2 Model

Sequential Private Information. A unit mass of agents faces the risk of experiencing an adverse event—such as developing an illness or being involved in a car accident—that may require third-party services tomorrow. Each agent is characterized by a privately observed ex-ante *risk type* $\theta \in \Theta = [\underline{\theta}, \bar{\theta}]$, which parametrizes the distribution of her also privately observed ex-post loss-related *valuation* $b \in B = [0, \bar{b})$, where $\bar{b}$ may be finite or infinite. Specifically, an agent with risk type θ draws her valuation from the cumulative distribution function $H_\theta : B \to [0, 1]$.

An agent who draws valuation b experiences a ex-post payoff of $-b$ (i.e., a *loss* of b). We interpret b as the monetary loss the agent associates with the negative consequences of the adverse event in the absence of treatment or restoration services.

We assume that H_θ increases in θ in the sense of first-order stochastic dominance: higher risk types are more likely to draw higher valuations. The distribution of risk types is denoted by $F : \Theta \to [0, 1]$, with continuous density function $f : \Theta \to (0, \infty)$.

We impose the following regularity conditions. The function H_θ is continuously differentiable in θ, and for each fixed θ, $H_\theta(\cdot)$ is continuously differentiable on B, with a strictly positive derivative $h_\theta : B \to (0, \infty)$. Moreover, for each (θ, b), the cross partial derivative $\frac{\partial^2 H_\theta(b)}{\partial b, \partial \theta}$ exists and is continuous.

Example 1 (Stiglitz family of loss distributions) *Suppose that risk type θ represents the probability of developing an illness. Conditional on becoming ill, the agent's loss-related valuation is distributed on $[0, \bar{b})$ according to the distribution $Q(b)$. In this case, the distribution function H_θ is given by:*

$$H_\theta(b) = 1 - \theta + \theta Q(b).$$

Third-Party Services and the Agent's Ex-Post Payoff. After drawing a valuation b, the agent can visit the service provider to receive treatment or repair damages at a price $p \in \mathbb{R}_+$. In that case, her loss becomes p instead of b. Therefore, the agent seeks the service whenever $b > p$. That is, whenever the monetary loss she associates with the adverse event in the absence of of the service exceeds the service price.

We also allow for the possibility of *partial treatment*, modeled as the agent's ability to purchase any fraction $x \in [0, 1]$ of the service for a total cost of xp. In other words, the service provider uses linear pricing.

An agent with valuation b who chooses to consume x units of service incurs a loss of $xp+(1-x)b$. This loss is additive in two components. The first is the monetary payment of xp made to the provider for the x units of service. The second is the residual loss $(1 - x)b$, which reflects the untreated portion of the adverse event. Since the agent forgoes $1 - x$ units of service and values the full untreated event at b, we assume that these effects also interact linearly.

Fractional service becomes relevant only with insurance; without it, it is inconsequential. If

$b \leq p$, any $x > 0$ yields a loss of $(1-x)b + xp \geq b$, so the agent prefers $x = 0$. If $b > p$, any $x < 1$ yields a loss strictly above p, so the agent chooses $x = 1$. Thus, the agent's loss is $b_p = \min\{b, p\}$.

Finally, we assume that the market for services is competitive and characterized by a strictly increasing supply function, denoted by $S : \mathbb{R}_+ \to \mathbb{R}_+$, where $S(p)$ represents the quantity of services supplied at price p. Additionally, we impose the following technical assumptions: S is continuously differentiable, unbounded and satisfies $S(0) = 0$.

The Agent's Ex-Ante Payoff. As in Gershkov et al. (2023), agents are endowed with a **Yaari (dual) utility** determined by a probability distortion function $g : [0,1] \to [0,1]$, where g is increasing, continuously differentiable, and satisfies $g(q) \leq q$ with $g(0) = 0$ and $g(1) = 1$.

The *certainty equivalent* of suffering a random loss distributed according to H_θ is given by:

$$CE(H_\theta(b)) = -\int_0^{\bar{b}} [1 - g(H_\theta(b))]\, db.$$

In particular, if services are provided at price p, so that the agent's loss is given by $b_p = \min\{b, p\}$, then the certainty equivalent is:

$$-\int_0^{p} [1 - g(H_\theta(b))]\, db,$$

where we use the fact that $\mathbb{P}(b_p \leq b) = H_\theta(b)$ for all $b < p$, $\mathbb{P}(b_p \leq b) = 1$ for all $b \geq p$, and $g(1) = 1$.

As discussed in Gershkov et al. (2023), the agent is risk-averse in the weak sense: the certainty equivalent of any lottery is less than the lottery's expected value. That is:

$$CE(H_\theta(b)) = -\int_0^{\bar{b}} [1 - g(H_\theta(b))]\, db \leq -\int_0^{\bar{b}} [1 - H_\theta(b)]\, db = -\mathbb{E}_{H_\theta}[b],$$

where the inequality follows from $g(q) \leq q$, and $\mathbb{E}_{H_\theta}[b]$ represents the expected value of b under H_θ.

While in the expected utility framework, this form of risk aversion is equivalent to aversion to mean-preserving spreads, in the dual utility framework, aversion to mean-preserving spreads is stronger and is equivalent to g being convex. As in Gershkov et al. (2023), the weak form of risk aversion, which requires only $g(q) \leq q$, is sufficient for our analysis.

To better understand dual utility, we can use integration by parts to express the certainty equivalent of the lottery H_θ as:

$$CE(H_\theta(b)) = -\int_0^{\bar{b}} [1 - g(H_\theta(b))]\, db = -\int_0^{\bar{b}} bg'(H_\theta(b))h_\theta(b)\, db$$

In contrast to the expected utility framework—where the standard expectation operator is modified by distorting payoffs through the von Neumann–Morgenstern utility function—dual utility modifies the expected value of the lottery by distorting probabilities through g'. Here, every loss b is weighted by $g'(H_\theta(b))$. The term $g(H_\theta(b))$ represents the cumulative weight assigned to losses below b, while $1 - g(H_\theta(b))$ represents the cumulative weight assigned to losses above b. The assumption $g(q) \leq q$ implies that the agent overweights the cumulative probability of larger losses and underweights the cumulative probability of smaller losses.

Figure 1 illustrates the certainty equivalents of a risk-neutral agent and a risk-averse agent with Yaari utility given by $g(q) = q^2$, when both face the distribution H_θ described in Example 1, and p is the price of services. The blue curve represents the distribution function $H_\theta(b)$ truncated at p, while the red curve shows $g(H_\theta(b))$, also truncated at p. The area above the blue curve corresponds to the certainty equivalent of the risk-neutral agent. The area above the red curve and below the horizontal dashed line corresponds to the certainty equivalent of the risk-averse agent. The area between the red and blue curves reflects the additional amount the risk-averse agent is willing to pay to eliminate the lottery risk.

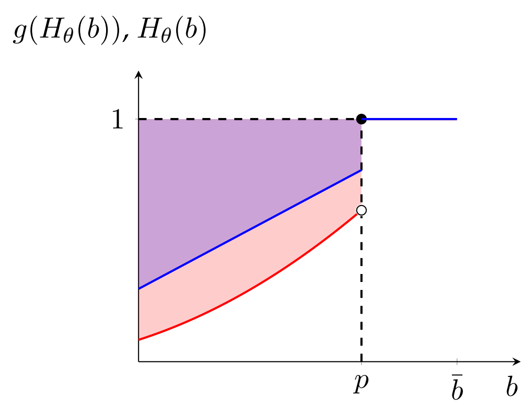

Figure 1: $H_\theta(b) = 1 - \theta + \theta b,\ \theta = 0.7,\ g(q) = q^2$

Finally, another key property of Yaari utility is its additivity with respect to constant random variables t:

$$CE(H_\theta + t) = CE(H_\theta) + t.$$

As noted by Yaari (1987), in the expected utility framework, the agent's attitude toward risk and their attitude toward wealth are inherently linked. Specifically, a risk-averse agent will always exhibit decreasing marginal utility with respect to wealth. In contrast, under dual utility, an agent can maintain constant marginal utility over wealth while still displaying risk aversion toward uncertain outcomes.

Insurance. A risk-neutral monopolistic insurance provider (he) offers a mechanism to the agents. The mechanism is offered when an agent knows only her ex-ante risk type θ before her ex-post valuation b has been realized. We assume that the mechanism specifies the following for each agent: 1) the *quantity of services*, x, covered by insurance; 2) the *per-unit price*, or *per-unit out-of-pocket cost* (OPC), denoted by D, that the agent must pay for the fraction of services covered by insurance; and 3) an upfront payment or *premium*, t, that an agent must pay to access her insurance.

The insurer can set terms only for the insured units; it cannot dictate the agent's consumption of additional units at the prevailing market price. Once insured, agents are free to decide how much treatment to consume—potentially purchasing extra units at the market price. As models of ex-post moral hazard, this discretion implies that coverage influences utilization. The insurer cannot restrict or condition access to services (for instance, by paying the agent not to seek treatment).

Additionally, we restrict our attention to: 1) direct mechanisms and 2) non-randomized mechanisms. Focusing on direct mechanisms is without loss of generality, as the revelation principle applies in our setting. While Gershkov et al. (2023) also study non-randomized mechanisms, they note that this restriction is not without loss: lotteries over insurance contracts could help the provider better screen risk-averse agents. However, focusing on non-randomized mechanisms is appropriate for insurance applications, as lotteries over contracts are rarely observed in practice.

More formally, a mechanism (or *menu*) is represented as:

$$\left\{ \left(\{(x(\theta, b), D(\theta, b))\}_{b \in B}, t(\theta) \right) \right\}_{\theta \in \Theta},$$

where each *menu option* includes a *contract* $\{(x(\theta, b), D(\theta, b))\}_{b \in B}$ and a *premium* $t(\theta) \in \mathbb{R}$. Upon visiting the service provider, an agent with the contract $\{(x(\theta, b), D(\theta, b))\}_{b \in B}$ can choose any of its items. If the agent selects the *contract item* $(x(\theta,b), D(\theta,b))$, insurance covers $x(\theta,b) \in [0,1]$ units of service for a per-unit price or OPC of $D(\theta,b) \in \mathbb{R}_+$. Equivalently, a contract item specifies $x(\theta,b)$ and the total out-of-pocket payment the agent would incur for those units, given by $\hat{D}(\theta,b) = x(\theta,b)D(\theta,b)$. For reasons that will become clear later in the analysis, however, it is more convenient to work with the per-unit out-of-pocket cost.

We present two examples of insurance contracts within our framework that share key features with real-world plans and will become relevant in our later analysis:

- *Simple deductible contract* $\{(1,D)\}$: This contract is analogous to standard deductible-based insurance. If the agent pays the OPC D, she receives the full unit of services, and the insurer covers the remaining $p - D$.
- *Limited coverage deductible contract* $\{(\alpha,D),(1,M)\}$: Under this contract, insurance covers up to a fraction α of the service at an OPC of D. If the agent chooses to consume the remaining $1-\alpha$ units, she must pay the per-unit service price p. Alternatively, by paying a higher OPC $M > D$, the agent can bypass the coverage cap α, receive the full unit of service, and have the insurer cover the remaining amount, $p - M$.

Timing. We now describe in more detail the timing of the interaction between the insurer and the agents.

1. The insurer commits to and offers the agents a menu:

$$\left\{\left(\{(x(\theta,b), D(\theta,b))\}_{b\in B}, t(\theta)\right)\right\}_{\theta\in\Theta}.$$

2. The agents privately observe their ex-ante risk type θ, decide whether to participate in the mechanism and, if so, they select an option from the menu, and pay the corresponding premium.
3. The ex-post loss-related valuation b is realized and privately observed by the agent. An agent holding the contract $\{(x(\theta,b), D(\theta,b))\}_{b\in B}$ decides whether to visit the service provider or remain at home. If she visits the provider, she may demand any quantity of service. Payment can be made either by paying the per-unit service price p or by selecting a contract item. If

the agent selects $(x(\theta, b), D(\theta, b))$, she is entitled to $x(\theta, b)$ units at a per-unit price of $D(\theta, b)$. For any additional units beyond $x(\theta, b)$, the full per-unit price p must be paid.

4. The contracts chosen by the agents, along with the items selected from the contracts, generate a demand for services at a given price. The market for services clears at the price where demand equals supply, and payoffs are accrued.

The objective of the insurer is to find a menu that maximizes his profits. However, since a menu influences the equilibrium service price, which in turn affects both the agents' behavior and the insurer's profits, we adopt a two-stage approach to determine the optimal menu: 1) In the first stage, we fix a price for services, p, and identify, among the menus that sustain that price in equilibrium, the one that maximizes the insurer's profits. 2) Once the first stage is complete, we associate each achievable price in equilibrium with the corresponding profits obtained by the insurer, and then solve for the price that maximizes these profits.

2.1 Discussion of the Model

Ex-Post Private Information. In our model, b represents the agent's willingness to pay for third-party services. Since this valuation is subjective and shaped by many personal and contextual factors, we assume that b is private information.

One important determinant of b is the *severity* of the agent's condition: a more severe illness or greater property damage typically generates higher perceived losses and, consequently, higher valuations for treatment or restoration services. If higher values of b correspond to more serious events, it is plausible that service costs are also higher.

In Section 5.3, we present an extension of our baseline model that captures this idea while preserving our main results. The basic setup in this extension is that the third-party provider offers tiered services: a basic one (i.e., a routine checkup) at price p_1, and a more advanced one (i.e., a specialist visit) at price p_2. Only agents with sufficiently high valuations choose the advanced service in addition to the basic one and therefore face higher treatment costs. Since the service received is verifiable by the insurer, he can condition contracts on the type of service provided.

Ex-Post and Ex-Ante Payoff. An important assumption that brings tractability to the model is the linearity of the agent's ex-post payoff in the quantity of services x. As we will show later, this assumption plays a crucial role in ensuring that the insurer's objective is linear in x, which allows us to represent the insurer's problem as an optimal transport problem and to use duality results to characterize the optimal insurance contracts.

The linearity of the agent's ex-post payoff arises from our assumption that the service provider uses linear pricing. As previously mentioned, in Section 5.3 we present an extension in which the provider offers multiple services, each with potentially different base prices. However, we maintain the assumption that, for any given service, partial treatment is possible and the provider charges linearly in the amount of service received.

Another key assumption that ensures linearity of the insurer's objective in x is that the agents are endowed with a Yaari (dual) utility. This implies that the agent's ex-ante utility is also linear in x, since dual utility distorts probabilities rather than payoffs. Moreover, since dual utility is additive over constant random variables, we will show that premiums can be expressed as a linear function in x.

Beyond its analytical convenience, we believe that dual utility is compelling for several substantive reasons. A number of empirical studies (i.e., Sydnor, 2010; Barseghyan et al., 2013, 2016) document patterns in insurance choice that are difficult to reconcile with expected utility theory but are well explained by models involving probability weighting.

Market Structure. The assumption that the insurance provider is a monopolist with price impact is motivated by many settings where insurers hold significant market power. According to the latest report on competition in health insurance (American Medical Association, 2024), the American Medical Association (AMA) finds that, based on the DOJ/FTC Merger Guidelines, 95% of metropolitan commercial markets were highly concentrated in 2023, with a Herfindahl–Hirschman Index (HHI) above 1800. The average commercial market exhibited an HHI of 3458. Moreover, in 89% of metropolitan markets at least one insurer held a market share of 30% or more, and in 47% of markets a single insurer's share reached at least 50%.

Nonetheless, in Section 5.1 we also analyze an environment in which the insurer maximizes the sum of agents' utilities with a weakly positive weight on revenue. This formulation captures settings where the state is the primary insurer, and more informally can also represent a competitive insurance market in which the insurer must leave surplus to agents.

Regarding the assumption of a competitive service market, in Section 5.2 we show that our baseline model can also be interpreted as a setting where prices are determined not by competition, but through Nash bargaining between the insurer and the third-party provider, with the insurer holding all the bargaining power. We also discuss how our main results extend to cases where the provider has bargaining power.

3 First Stage: Optimal Insurance Menus

In this section, we fix the price for hospital services, denoted by $p \in \mathbb{R}_+$. Our goal is to characterize the set of implementable menus that sustain p in equilibrium and maximize the insurer's profits. We proceed in three steps.

Step 1: Characterization of implementable menus. In Section 3.1 we first characterize contracts for which it is ex-post incentive compatible for an agent with loss-related valuation b to select the contract item corresponding to her true valuation. Building on this, we characterize menus for which it is ex-ante individually rational for type θ to select an option, and ex-ante incentive compatible to choose the option corresponding to her true type. We rely on standard techniques for this analysis, though the ex-post stage requires additional care: the agent must also decide whether to seek services at all and whether to pay using the contract or at the full service price. Finally, we define the equilibrium constraint that any menu sustaining p must satisfy.

Step 2: Relaxed insurer's problem and first-stage main result. With Step 1 complete, we now formally define a relaxed version of the insurer's problem (Section 3.2). This version includes all constraints that a menu must satisfy except for ex-ante incentive compatibility (i.e., that agents truthfully report θ). In particular, it incorporates a necessary and sufficient condition for *local* ex-ante incentive compatibility, though not global. In Section 3.3, we then formally state our first-stage main result (Theorem 1), which shows that for any given price p, a menu in which each risk type receives either a *simple deductible contract* or a *limited-coverage contract* is optimal for the insurer's relaxed problem.

Step 3: Proof of Theorem 1: Optimal transport formulation. To prove Theorem 1 we express the insurer's relaxed problem in the *quantity space*—an adaptation of Myerson (1981)'s ironing technique. Under this formulation, the insurer chooses, for each risk type, a lottery over quantities, where quantities represent the measure of agents of that type who receive coverage, and the associated probabilities represent the level of coverage. Then, we illustrate this formulation with an example involving two risk types (Section 3.5), which provides intuition for the main result in Theorem 1 using a *concavification* approach. Finally, in Section 3.6 we address the technical challenges that arise from the infinite-dimensional type space. To handle these, we reformulate the problem so that it closely resembles an optimal transport problem, in which each risk type is effectively matched to the mass of agents who obtain coverage under the mechanism. Leveraging duality results from optimal transport theory, we show Theorem 1 and identify regularity conditions under which this solution is fully implementable—that is, it satisfies ex-ante incentive compatibility globally.

3.1 Implementable Menus

Period 2 Incentive Compatibility. After an agent with risk type θ' selects an option from the menu and her valuation b is realized, she makes a series of decisions. First, she chooses whether to visit the service provider or remain at home. If she decides to seek the service, she has two payment options: covering the service bill entirely out of pocket or utilizing an item from her contract $\{(x(\theta, b), D(\theta, b))\}_{b\in B}$. The agent's *loss function* when engaging with the item $(x(\theta, b'), D(\theta, b'))$ is defined as follows:

$$L(b, b'; \theta) = x(\theta, b') \min\{D(\theta, b'), b_p\} + (1 - x(\theta, b'))b_p,$$

where $b_p = \min\{b, p\}$.

We now explain why this expression represents the agent's loss. The agent will seek the service and utilize the fraction of treatment covered by the contract item, $x(\theta, b)$, if the OPC $D(\theta, b')$ is less than b_p. In this case, insurance covers a fraction $x(\theta, b)$ of the service and the agent pays $x(\theta, b)D(\theta, b')$.

If b is higher than p, the agent will also demand the remaining fraction of the service, $1 - x(\theta, b)$, at the full per-unit price p. The additional service cost is $(1 - x(\theta, b))p$, resulting in a total loss of $x(\theta, b)D(\theta, b') + (1 - x(\theta, b))p$.

However, if b is lower than p, the agent will not demand the remaining portion, and her loss will instead be $x(\theta, b)D(\theta, b') + (1 - x(\theta, b))b$, where $x(\theta, b)D(\theta, b')$ is the monetary payment for the $x(\theta, b)$ units of service and $(1 - x(\theta, b))b$ is the residual loss, which reflects the untreated portion of the adverse event.

Conversely, if $D(\theta, b')$ is higher than b_p, two scenarios may arise: 1) If the service price p is lower than her valuation b, the agent prefers to get the full unit of the service and pay the entire service cost out of pocket. In this situation, her loss is p. 2) If her valuation b is lower than p, the agent prefers to remain at home, and her loss is b.

Note that whenever the agent decides whether to seek services or to pay using a contract item, she compares the per-unit price under insurance, $D(\theta, b')$, with b_p. For this reason, it is more convenient to work with $D(\theta, b')$ rather than the total out-of-pocket payment $\hat{D}(\theta, b) = x(\theta, b)D(\theta, b)$.

Finally, it is also important to note that the agent's risk type θ' is irrelevant in determining her loss function. At this stage, the only relevant variables are her actual valuation b, the contract chosen from the menu (indexed by θ), and the item selected from the contract (indexed by b'). Figure 2 illustrates the agent's behavior and her resulting loss when she selects the item from the contract corresponding to her true valuation.

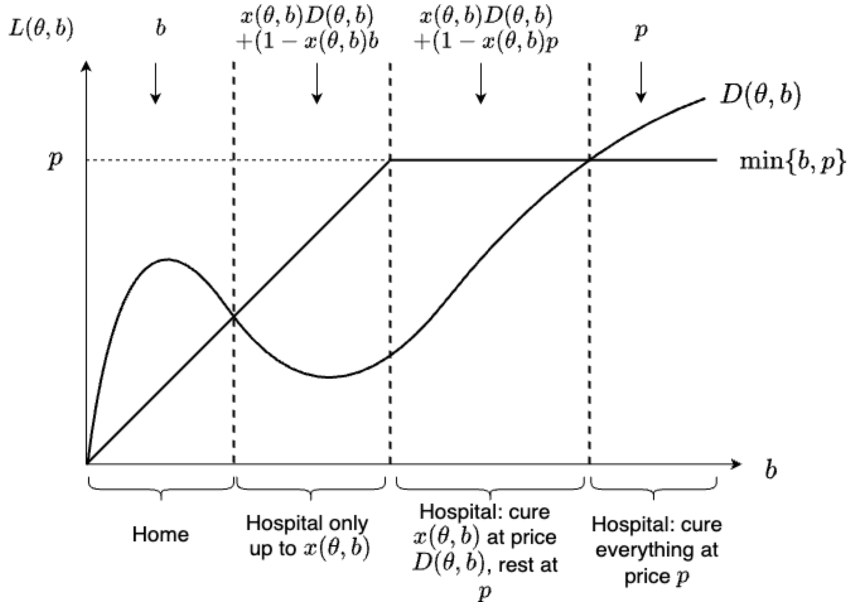

Figure 2: Loss and behavior of an agent that truthfully reports her valuation in the context of health insurance.

Having introduced the agent's period 2 loss function, we can now define the usual incentive compatibility constraint that a contract must satisfy. Specifically, it suffices to focus on contracts where the agent truthfully reports her valuation:

$$L(b;\theta) \equiv L(b,b;\theta) \leq L(b,b';\theta) \text{ for all } b,b' \in B. \tag{IC_2}$$

We now show some properties of contracts that satisfy IC_2:

Lemma 1 *Without loss of generality we can restrict attention to contracts that satisfy $D(\theta,b) \leq b_p$.*

Lemma 2 *A contract satisfies IC_2 only if $L(p;\theta) = L(b;\theta)$ for all $b > p$. Furthermore, without loss of generality, we can restrict attention to contracts that satisfy $x(\theta,b) = 1$ for all $b > p$.*

The formal proofs of Lemmas 1 and 2 can be found in Appendix A.

Lemmas 1 and 2 are intuitive. For Lemma 1, if $D(\theta,b) > b_p$, the agent will not use her insurance. The insurer can replicate this outcome by setting $D'(\theta,b) \leq b_p$ but restricting the agent's access to the service by setting $x'(\theta,b) = 0$.

For Lemma 2, note that all agents with $b \geq p$ incur the same loss from any allocation. These are the agents who would always seek the service in the absence of insurance, effectively behaving as if they have the same valuation. To satisfy IC_2, all such agents must experience the same loss.

Moreover, for any agent with a valuation $b > p$, and for any contract that results in a loss of $L(p;\theta)$ for such an agent, the insurer can replicate the same outcome by setting $x'(\theta,b) = 1$ and $D'(\theta,b) = L(p;\theta)$. This choice preserves the agent's demand for services, their loss (or ex-post payoff), and the insurer's cost.

Lemmas 1 and 2 allow us to express the loss of an agent with $b \in [0,p]$ when truthfully reporting as:

$$L(b;\theta) = x(\theta,b)D(\theta,b) + (1 - x(\theta,b))b.$$

This loss function is linear in b and separable in the OPC. Using standard arguments from mechanism design, we can characterize contracts that satisfy IC_2 by applying the envelope formula and the monotonicity of $x(\theta,\cdot)$:

Lemma 3 *A contract satisfies IC_2 if and only if:*

(1) $L(b;\theta) = \int_0^b (1 - x(\theta,l))\, dl$ for all $b \in [0,p]$, and $L(b;\theta) = L(p;\theta)$ for all $b > p$.

(2) $x(\theta,\cdot) : B \to [0,1]$ is a non-decreasing function, with $x(\theta,b) = 1$ for all $b > p$.

The formal proof of Lemma 3 can be found in Appendix A.

Period 1 Incentive Compatibility and Individual Rationality. By Lemma 3, we know that if an agent with risk type θ selects the menu option:

$$\left(\left\{\left(x(\theta', b), D(\theta', b)\right)\right\}_{b \in B}, t(\theta')\right),$$

the agent's loss function is given by:

$$L(b; \theta') = \int_0^b (1 - x(\theta', l))\, dl.$$

Observe that $L(\cdot; \theta')$ is non-decreasing. Consequently, the lottery faced by risk type θ in period 1 is represented by $H_\theta(L^{-1}(l; \theta'))$, where $L^{-1}(l; \theta) = \sup\{z : L(z; \theta) \le l\}$.

The certainty equivalent that risk type θ gets from her menu option is given by:

$$\begin{aligned}
U(\theta, \theta') &= -t(\theta') - \int_0^{L(\bar{b}; \theta')} \left[1 - g\left(H_\theta\left(L^{-1}(l; \theta')\right)\right)\right]\, dl \\
&= -t(\theta') - \int_0^{\bar{b}} \frac{\partial L(b; \theta')}{\partial b} [1 - g(H_\theta(b))]\, db \\
&= -t(\theta') - \int_0^{p} (1 - x(\theta', b))[1 - g(H_\theta(b))] db.
\end{aligned}$$

The additivity of the premium arises from the fact that dual utility is additive in constant random variables. The second equality follows from the change of variable $b = L^{-1}(l; \theta)$, while the third equality results from $\frac{\partial L(b;\theta')}{\partial b} = 1 - x(\theta', b)$ (Lemma 3, Condition (1)), along with the condition $x(\theta', b) = 1$ for $b > p$ (Lemma 3, Condition (2)).

The certainty equivalent is linear in x, a property that provides crucial tractability for the analysis. This linearity arises from the fact the agent's ex-post payoff is linear in x and the fact that dual utility distorts probabilities rather than payoffs.

If the agent were instead a risk-averse expected utility maximizer, two important aspects would change. First, because the agent's loss $L(b; \theta)$ is evaluated through a concave von Neumann–Morgenstern utility function, the certainty equivalent would no longer be linear in x. Second, since this distortion also applies to the premium, t would no longer be separable from x.

The period 1 incentive compatibility constraint is:

$$U(\theta) \equiv U(\theta,\theta) \geq U(\theta,\theta') \text{ for all } \theta,\theta' \in \Theta, \tag{IC_1}$$

Lemma 4 provides a partial characterization of menus that satisfy IC_1:

Lemma 4

(1) A menu satisfies IC_1 only if

$$U(\theta) = U(\underline{\theta}) + \int_{\underline{\theta}}^{\theta} \int_0^p (1 - x(s,b)) g'(H_s(b)) \frac{\partial H_s(b)}{\partial s} \, db \, ds.$$

(2) If a menu satisfies the above condition and $x(\cdot, b) : \Theta \to [0,1]$ is a non-decreasing function for all $b \in B$, then the menu satisfies IC_1.

The formal proof of Lemma 4 is in Appendix A.

Condition (1) of Lemma 4 is the standard envelope formula. Importantly, the additivity of dual utility in constant random variables, combined with Condition (1), enables us to determine premiums (up to a constant) as a function of x. Additionally, the monotonicity of $x(\cdot, b)$ for all b, together with Condition (1), ensures that a menu satisfies IC_1.

However, a menu that satisfies Condition (1) does not necessarily require $x(\cdot, b)$ to be monotonic for all b. This result is common in screening environments where the allocation chosen by the designer for a fixed agent type is infinite—or multi-dimensional. Nevertheless, monotonicity remains a robust condition, as it does not depend on the functional forms of g or H_θ.

It remains to define the period 1 participation constraint. We begin by specifying the agent's outside option (non-participation payoff), denoted $U_{NP}(\theta)$ for risk type θ. We assume that an uninsured agent faces a service price of $\beta + p$ with $\beta \geq 0$; that is, the uninsured pays a weakly positive (possibly zero) premium β over the baseline price p. This reflects empirical evidence that uninsured individuals typically face higher prices: for example, Anderson (2007) document that uninsured hospital patients are charged, on average, 2.5 times more than insured patients for the same services. It also captures a less visible but important benefit of insurance—access to lower prices (Cook, 2020).

$$U_{NP}(\theta) = -\int_0^{\beta+p} [1 - g(H_\theta(b))]\, db.$$

Note that the insurer can always offer a contract with zero premium and no coverage, ensuring $U(\theta) \geq U_{NP}(\theta)$. Therefore, it is without loss of generality to restrict attention to mechanisms in which all agents participate in period 1. This leads to the following period 1 individual rationality constraint:

$$U(\theta) \geq U_{NP}(\theta) \quad \text{for all } \theta \in \Theta. \tag{IR}$$

Then, we can show that an incentive compatible mechanism satisfies IR if and only if it binds for the lowest type:

Lemma 5 *Suppose a menu satisfies IC_1. Then, the IR condition is satisfied for all types if and only if it is satisfied for the lowest type, $\underline{\theta}$.*

The formal proof of Lemma 5 is in Appendix A.

Lemma 5 follows from the fact that a higher-risk type faces a riskier lottery and, therefore, has a lower certainty equivalent. This implies that both the agent's outside option, U_{NP}, and their utility from participation, U, decrease in θ. However, U decreases more slowly, since the agent receives some coverage (i.e., $x(\theta, b) \in [0, 1]$). Therefore, $U'(\theta) - U'_{NP}(\theta) \geq 0$, and

$$U(\theta) - U_{NP}(\theta) = U(\underline{\theta}) - U_{NP}(\underline{\theta}) + \int_0^\theta [U'(s) - U'_{NP}(s)]\, ds \geq 0,$$

whenever IR holds for type $\underline{\theta}$.

Period 3 Equilibrium Constraint. If each contract in a menu satisfies IC_2 and the menu satisfies IC_1 and IR, the demand for services by an agent with risk type θ and valuation b is $x(\theta, b)$. This follows because the menu and its contracts are both incentive compatible and individually rational, ensuring that the agent selects the contract item $(x(\theta, b), D(\theta, b))$.

Moreover, if $b \leq p$, the agent will not demand any additional units of the service, and her demand is fully determined by $x(\theta, b)$. Conversely, if $b > p$, she would, in principle, demand the remaining portion, $1 - x(\theta, b)$. However, by Lemma 2, $x(\theta, b) = 1$, meaning the remaining portion is zero. Thus, her demand is again fully determined by $x(\theta, b)$.

The demand for services at price p is given by:

$$\begin{aligned}\mathcal{D}(x;p) &= \int_{\underline{\theta}}^{\bar{\theta}} \int_0^{\bar{b}} x(\theta,b)h_\theta(b)\,db f(\theta)\,d\theta \\ &= \int_{\underline{\theta}}^{\bar{\theta}} \int_0^{p} x(\theta,b)h_\theta(b)\,db f(\theta)\,d\theta + \int_{\underline{\theta}}^{\bar{\theta}} [1-H_\theta(p)]f(\theta)\,d\theta,\end{aligned}$$

where the second equality follows from the fact that $x(\theta,b)=1$ for all $b>p$.

It is convenient to define the residual supply of services faced by the insurer at price p as:

$$RS(p) = S(p) - \int_{\underline{\theta}}^{\bar{\theta}} [1-H_\theta(p)]f(\theta)\,d\theta,$$

where the second term reflects the portion of demand that the insurer cannot influence, as no matter what menu the insurer offers, he cannot affect the demand arising from valuations higher than p.

The equilibrium constraint can then be expressed as:

$$\int_{\underline{\theta}}^{\bar{\theta}} \int_0^{p} x(\theta,b)h_\theta(b)\,db f(\theta)\,d\theta = RS(p). \tag{EQ}$$

The lowest demand that the insurer can generate occurs with a menu that sets $\underline{x}(\theta,b)=0$ for all θ and $b\in[0,p]$. In this case, the insurer provides no coverage to any agents, and demand is given by:

$$\mathcal{D}(\underline{x};p) = \int_{\underline{\theta}}^{\bar{\theta}} [1-H_\theta(p)]f(\theta)\,d\theta.$$

Moreover, by our technical assumptions on the supply function S and the distributions of valuations H_θ, there exists a unique strictly positive price p^N that satisfies $\mathcal{D}\left(\underline{x};p^N\right) = S\left(p^N\right)$. That is, p^N is the price that clears the service market in the absence of coverage.

Similarly, the highest demand that the insurer can generate occurs with a menu that sets $\overline{x}(\theta,b)=1$ for all θ and $b\in[0,p]$. In this case, the insurer provides full insurance—or full access to services at zero OPC. Demand in this case is $\mathcal{D}\left(\overline{x};p\right) = \int_{\underline{\theta}}^{\bar{\theta}}[1-H_\theta(0)]f(\theta)\,d\theta$, where $1-H_\theta(0)$ is the probability that a risk type θ experiences an adverse event. Once again, by our technical assumptions, there exists a unique positive price p^F such that $\mathcal{D}\left(\overline{x};p^F\right) = S\left(p^F\right)$, and $p^F > p^N$.

We then have the following result:

Lemma 6 *There exists a function x that satisfies Condition (2) of Lemma 3 and EQ if and only if $p \in [p^N, p^F]$.*

The formal proof of Lemma 6 is provided in Appendix A.

From now on, we assume that p belongs to $[p^N, p^F]$, ensuring that there exists a feasible menu that induces it as an equilibrium price in the service market.

3.2 Insurer's Problem

We are now ready to formulate the first-stage insurer's problem. To do so, the following Lemma provides an expression for the insurer's objective:

Lemma 7 *Suppose that each contract in a menu satisfies IC_2, and the menu satisfies IC_1 and IR. Then, the insurer's profits from the menu are:*

$$\int_{\underline{\theta}}^{\bar{\theta}} \int_0^p x(\theta, b) J(\theta, b) h_\theta(b)\, db f(\theta)\, d\theta - U(\underline{\theta}) - \int_0^p [1 - g(H_{\underline{\theta}}(b))]\, db,$$

where

$$J(\theta, b) = \underbrace{\left[1 - g(H_\theta(b)) + \frac{1 - F(\theta)}{f(\theta)} g'(H_\theta(b)) \frac{\partial H_\theta(b)}{\partial \theta}\right]}_{\text{Period 1 Virtual Value}} \frac{1}{h_\theta(b)} + \underbrace{b - p - \frac{1 - H_\theta(b)}{h_\theta(b)}}_{\text{Period 2 Virtual Value}}.$$

The formal proof of Lemma 7 is in Appendix B.

The insurer's cost of providing any contract is derived from the fact that the portion of the service bill paid by the insurer, when facing an agent with risk type θ and valuation b, is $x(\theta, b)(p - D(\theta, b))$. Using the identity $x(\theta, b)D(\theta, b) = L(b; \theta) - (1 - x(\theta, b))b$ and the expression for $L(b; \theta)$ obtained in Lemma 3, Condition (1), along with integration by parts, allows us to determine the insurer's cost.

Once we have an expression for the insurer's cost, the only remaining step to determine profits is to find an expression for the revenue the insurer receives from premiums. To obtain this, we can apply standard techniques. Specifically, as mentioned earlier, since Yaari utility is additive in constant random variables, t can be expressed as a function of x, up to a constant $U(\underline{\theta})$.

To interpret the insurer's profits, recall that in screening environments with quasi-linear preferences, where a designer allocates an object or good, the classical agent's virtual value is a key concept in determining whether it is profitable to assign the good to the agent. It is defined as the marginal value from an increase in the allocation minus the information rents. Information rents, in turn, are captured by the inverse hazard rate multiplied by the derivative of the agent's marginal value with respect to her type.

In period 1, the agent's private information, or risk type, is denoted by θ, and her utility from an allocation, x, is given by $U(\theta)$. Based on this utility, we see that the agent values an additional unit of hospital services (if her future valuation is b) by $1 - g(H_\theta(b))$. This valuation forms the first term in $J(\theta, b)$, referred to as the *period 1 virtual value*. The derivative of this marginal valuation with respect to the agent's risk type is given by:

$$-g'(H_\theta(b))\frac{\partial H_\theta(b)}{\partial b}.$$

This term is then multiplied by the inverse hazard rate of the distribution F.

In period 2, the agent's private information shifts to her valuation, denoted by b, and her payoff from an allocation is determined by the loss function $L(b;\theta)$. Here, the second term in $J(\theta, b)$, denoted as the *period 2 virtual value*, is also analogous to the classical virtual value in screening problems but accounts for provision costs. Specifically, the agent's marginal valuation for an additional unit of services corresponds to her valuation b. On the other hand, the insurer's marginal cost for providing this additional unit is p.

The *(relaxed) first-stage insurer's problem* is given by:

$$\Pi(p) \equiv \max_x \quad \int_{\underline{\theta}}^{\bar{\theta}} \int_0^p x(\theta, b)J(\theta, b)h_\theta(b)\, db f(\theta)\, d\theta + K(p) \tag{$P1$}$$

$$\text{s.t.} \quad \int_{\underline{\theta}}^{\bar{\theta}} \int_0^p x(\theta, b)h_\theta(b)\, db f(\theta)\, d\theta = RS(p), \tag{EQ}$$

$$x(\theta, \cdot) : B \to [0, 1] \text{ is non-decreasing for all } \theta, \tag{MON}$$

where:

$$K(p) = -U_{NP}(\underline{\theta}) - \int_0^p [1 - g(H_{\underline{\theta}}(b))]\, db.$$

In the formulation of $P1$, we have already used the fact that, in an optimal menu, the insurer will always set the utility of the lowest risk type equal to her outside option. We refer to $P1$ as a relaxed version of the first-stage insurer's problem because it incorporates all necessary conditions that an implementable menu must satisfy. However, a solution to $P1$ is not necessarily implementable. Specifically, it is possible that x does not satisfy IC_1.

The approach, therefore, is to solve $P1$ and identify conditions under which the solution is implementable. For instance, by Lemma 4, Condition (2), we know that if the solution satisfies the property that $x(\cdot, b)$ is non-decreasing for all b, then x is implementable.

The function $\Pi : [p^N, p^F] \to \mathbb{R}$ captures the profits obtained by the insurer from the optimal menu that sustains p in equilibrium.

3.3 First-Stage Analysis and Results

We begin by stating the main result of the first-stage analysis. This result characterizes the structure of the optimal contracts induced by the x^* that solves $P1$:

Theorem 1 *There exists an x^* that solves the first-stage problem $P1$ such that the contract offered to risk type θ takes one of the following forms $\{(1, D(\theta))\}$ with $D(\theta) \in [0, p]$, or $\{(\alpha, D(\theta)), (1, M(\theta))\}$ where $0 \leq D(\theta) < M(\theta) \leq p$ and $\alpha \in (0, 1)$.*

As formalized in Theorem 1, the optimal contract for each risk type takes one of the two forms described in Section 2. The first is a *simple deductible contract*, in which a risk type θ receives the full unit of the service after paying the OPC $D(\theta)$. All agents with a valuation higher than $D(\theta)$ will pay the OPC and receive full treatment.

The second is a *limited coverage deductible contract*. In this case, insurance covers a fraction α of the service at an OPC $D(\theta)$. If the agent chooses to consume the remaining $1 - \alpha$ units, she must pay for them the full service price p. However, if she pays the higher OPC $M(\theta)$, she receives the full service.

Whenever the contract for risk type θ is a limited coverage deductible contract, the allocation function $x^*(\theta, \cdot)$ exhibits two jumps, as shown in Figure 3. Specifically:

- It equals 0 for valuations $b \leq D(\theta)$—these agents forgo treatment and stay at home.
- It equals α for valuations $b \in \left(D(\theta), \frac{M(\theta)-\alpha D(\theta)}{1-\alpha}\right]$—these agents pay the OPC and receive partial coverage.
- It equals 1 for valuations $b > \frac{M(\theta)-\alpha D(\theta)}{1-\alpha}$—these agents pay the higher OPC and receive full coverage.

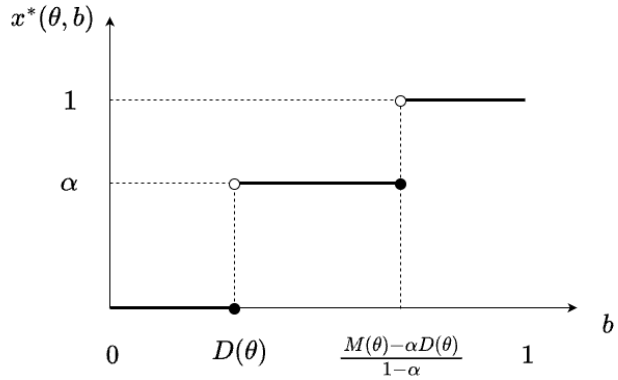

Figure 3: Optimal allocation x^* corresponding to the menu described in Theorem 1.

Finally, note that although the low and high OPCs in the optimal contract may depend on the agent's risk type θ, the service cap does not. In other words, all limited-coverage deductible contracts impose the same cap α, regardless of the agent's risk type. This result follows from our optimal transport formulation, as shown in the proof of Lemma 8.

3.4 Representing the Insurer's Problem in the Quantity Space

In this section we reformulate the insurer's problem in the quantity (or quantile) space. For this purpose, let $H_\theta(\cdot|p) : [0, p] \to [0, 1]$ denote the distribution of valuations for risk type θ, conditional on being strictly greater than 0 and less than or equal to p. Specifically,

$$H_\theta(b \mid p) = \frac{H_\theta(b) - H_\theta(0)}{H_\theta(p) - H_\theta(0)},$$

By our assumption on H_θ, it follows that $H_\theta(\cdot \mid p)$ is invertible. Any non-decreasing, left-continuous function $x(\theta, \cdot) : [0, p] \to [0, 1]$ can then be represented as:

$$x(\theta, b) = \int_0^1 \mathbf{1}_{[b > b_\theta(q)]} \, dG_\theta(q),$$

where $b_\theta(q) = H_\theta^{-1}(1 - q \mid p)$ and G_θ is a distribution over $[0, 1]$.

Economically, this representation allows us to express a mechanism in the quantity space. In the absence of insurance, only the agents with a valuation exceeding p seek the service. As a result, the total demand for services, conditional on risk type θ, is given by $1 - H_\theta(p)$. To induce additional demand of $[H_\theta(p) - H_\theta(0)]q$ units, the insurer can offer a *simple deductible contract* of the form $\{(1, b_\theta(q))\}$. Under this contract, in addition to agents with valuations greater than p, those with valuations in the range $(b_\theta(q), p]$ will also seek the service. For risk type θ, the mass of agents within the interval $(b_\theta(q), p]$ corresponds precisely to $[H_\theta(p) - H_\theta(0)]q$. In other words, the insurer can only influence the demand of agents with valuations within $(0, p]$. Among these agents, the insurer selects a fraction q who will receive full coverage. Moreover, by appropriately randomizing over q (or equivalently, over simple deductible contracts), the insurer can replicate any feasible demand schedule $x(\theta, \cdot)$. Thus, without loss of generality, the optimization can be performed over G_θ instead of directly over $x(\theta, \cdot)$.

Then, the insurer's problem becomes:

$$\max_G \quad \int_{\underline{\theta}}^{\bar{\theta}} \int_0^1 \Phi(\theta, q)\, dG_\theta(q) f(\theta)\, d\theta, \tag{$P1'$}$$

$$\text{s.t.} \quad \int_{\underline{\theta}}^{\bar{\theta}} \int_0^1 [H_\theta(p) - H_\theta(0)]q\, dG_\theta(q) f(\theta)\, d\theta = RS(p), \tag{EQ}$$

where:

$$\Phi(\theta, q) = [H_\theta(p) - H_\theta(0)] \int_0^q \phi(\theta, s) ds,$$

and:

$$\phi(\theta, q) = J(\theta, b_\theta(q)).$$

Note that in formulation $P1'$ we no longer include the constant $K(p)$. In the first stage, this is without loss of generality, as any function G that solves $P1'$ will also be a maximizer if the constant $K(p)$ is included. Therefore, to simplify notation, we omit $K(p)$ for the remainder of this section.

3.5 Two-type Example

In this section, we illustrate the main intuition behind Theorem 1 with an example. Specifically, we consider a setting with only two risk types, $\theta \in \{\theta_L, \theta_H\}$, where $\theta_H > \theta_L$. This simplified case already reveals the economic forces that drive the insurer to offer a *limited coverage deductible contract*.

Using the representation of the insurer's problem derived in Section 3.4, we obtain the following expression for the insurer's problem in the quantity space:

$$\begin{aligned}
\max_{G} \quad & \int_0^1 \Phi(\theta_L, q)\, dG_{\theta_L}(q) \hat{f}(\theta_L) + \int_0^1 \Phi(\theta_H, q)\, dG_{\theta_H}(q) \hat{f}(\theta_H) \\
\text{s.t.} \quad & \int_0^1 q\, dG_{\theta_L}(q) \hat{f}(\theta_L) + \int_0^1 q\, dG_{\theta_H}(q) \hat{f}(\theta_H) = RS(p),
\end{aligned}$$

where

$$\Phi(\theta, q) = \int_0^q \phi(\theta, s) ds,$$

$$\phi(\theta_L, q) = b_{\theta_L}(q) - p + \frac{H_{\theta_L}(b_{\theta_L}(q)) - g(H_{\theta_L}(b_{\theta_L}(q)))}{h_{\theta_L}(b_{\theta_L}(q))} + \frac{1 - f(\theta_L)}{f(\theta_L) h_{\theta_L}(b_{\theta_L}(q))} g(H_{\theta_H}(b_{\theta_H}(q))) - g(H_{\theta_L}(b_{\theta_L}(q))),$$

$$\phi(\theta_H, q) = b_{\theta_H}(q) - p + \frac{H_{\theta_H}(b_{\theta_H}(q)) - g(H_{\theta_H}(b_{\theta_H}(q)))}{h_{\theta_H}(b_{\theta_H}(q))},$$

$$RS(p) = S(p) - f(\theta_L)(1 - H_{\theta_L}(p)) - f(\theta_H)(1 - H_{\theta_H}(p)),$$

and

$$\hat{f}(\theta) = [H_\theta(p) - H_\theta(0)]\, f(\theta).$$

Observe that $\phi(\theta, q)$ accounts for the period 1 and period 2 virtual values adjusted to a discrete risk type space. Specifically, we account for the discrete version of information rents. These rents no longer arise from the envelope formula but instead follow from the standard result that, with two types, the IC_1 constraint for the high type and the IR constraint for the low type always bind.

We now analyze the insurer's problem. For any distribution G_θ that results in an average q given by q_θ, there exists a distribution G'_θ that preserves the same average q_θ while attaining the concave envelope of $\Phi(\theta, \cdot)$ evaluated at q_θ. In other words, let $\overline{\Phi}(\theta, \cdot)$ denote the concave envelope of $\Phi(\theta, \cdot)$. Then, we have:

$$\int_0^1 q\, dG'_\theta(q) = q_\theta = \int_0^1 q\, dG_\theta(q),$$

and

$$\int_0^1 \Phi(\theta, q)\, dG'_\theta(q) = \overline{\Phi}(\theta, q_\theta) \geq \int_0^1 \Phi(\theta, q)\, dG_\theta(q).$$

G' weakly increases the insurer's payoff while still satisfying the EQ condition. Thus, we can also express the insurer's problem as:

$$\max_{q} \quad \overline{\Phi}(\theta_L, q_{\theta_L})\hat{f}(\theta_L) + \overline{\Phi}(\theta_H, q_{\theta_H})\hat{f}(\theta_H)$$
$$\text{s.t.} \quad q_{\theta_L}\hat{f}(\theta_L) + q_{\theta_H}\hat{f}(\theta_H) = RS(p),$$

In the above formulation, for each risk type, the insurer selects the fraction of agents with a valuation in $(0, p]$ who get coverage, denoted by q_θ. This formulation already demonstrates why a solution exists in which the contract offered to each risk type is either a simple deductible contract or a limited coverage deductible contract. Indeed, any point in $\overline{\Phi}(\theta, \cdot)$ can be attained by a distribution G_θ with at most two elements in its support. This, in turn, results in an allocation function $x(\theta, \cdot)$ with at most two jumps, as illustrated in Figure 3.

Moreover, if $\Phi(\theta, \cdot)$ is concave, or equivalently, $\frac{\partial \phi(\theta,q)}{\partial q} \leq 0$ for all q, then the insurer can always achieve $\overline{\Phi}(\theta, q_\theta)$ with a distribution that puts probability one on q_θ. That is, the insurer can always offer risk type θ the simple deductible contract $\{(1, b_\theta(q_\theta))\}$. This observation is formalized in Proposition 4.

To illustrate the economic forces that lead the insurer to include a limited coverage deductible contract, we assume $\theta_L = 0.25$, $\theta_H = 0.35$, $f(\theta_H) = \frac{1}{5}$, and $H_\theta(b) = 1 - \theta + \theta Q(b)$, where $Q(b) = b^2 \cdot \mathbf{1}_{[0 \leq b \leq 1]}$. In this case, we have $q\Phi(\theta, 1) > \Phi(\theta, q)$ for all $q \in (0, 1)$. This implies that the concave envelope of $\Phi(\theta, \cdot)$ is the line segment connecting $\Phi(\theta, 0)$ and $\Phi(\theta, 1)$, i.e., $\overline{\Phi}(\theta, q) = q\Phi(\theta, 1)$, as shown in Figure 4.

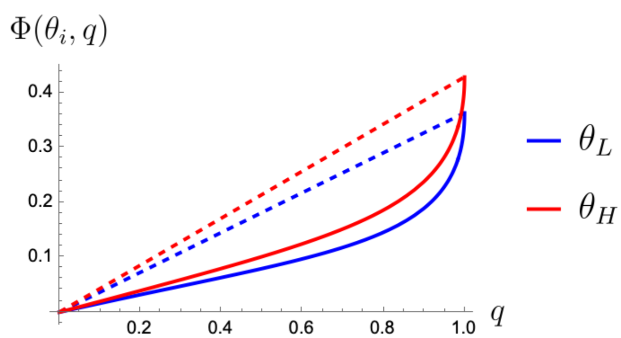

Figure 4: The continuous lines illustrate $\Phi(\theta_L, q)$ (in blue) and $\Phi(\theta_H, q)$ (in red). The dashed lines are the respective concave envelopes. The figure is drawn assuming that $p = 0.8$.

Under these parametric assumptions, the insurer solves:

$$\begin{aligned} \max_{q} \quad & q_{\theta_L}\Phi(\theta_L,1)\hat{f}(\theta_L) + q_{\theta_H}\Phi(\theta_H,1)\hat{f}(\theta_H) \\ \text{s.t.} \quad & q_{\theta_L}\hat{f}(\theta_L) + q_{\theta_H}\hat{f}(\theta_H) = RS(p). \end{aligned}$$

The solution thus depends solely on the relationship between $\Phi(\theta_L, 1)$ and $\Phi(\theta_H, 1)$. In particular, we can verify that $\Phi(\theta_H, 1) > \Phi(\theta_L, 1)$. Therefore, the optimal allocation is:

$$q^*_{\theta_H} = \min\left\{1, \frac{RS(p)}{\hat{f}(\theta_H)}\right\}, \quad q^*_{\theta_L} = \frac{RS(p)}{\hat{f}(\theta_L)} - \min\left\{1, \frac{RS(p)}{\hat{f}(\theta_H)}\right\} \cdot \frac{\hat{f}(\theta_H)}{\hat{f}(\theta_L)}.$$

In other words, the insurer favors the riskier type, allocating as many coverage as possible to θ_H first. If the maximum insurance-driven demand, $\hat{f}(\theta_H)$, is sufficient to meet the residual supply, then $q^*_{\theta_H} = RS(p)/\hat{f}(\theta_H)$ and $q^*_{\theta_L} = 0$. Otherwise, we set $q^*_{\theta_H} = 1$ and further increase demand by including some low-risk types until the market clears.

Furthermore, if $q^*_\theta \in (0, 1)$, the contract for type θ is a limited-coverage deductible contract. In this case, the insurer implements q^*_θ by mixing between the two extremes $q = 0$ and $q = 1$. For example, when $p = 0.8$, we obtain $q^*_{\theta_L} = 0.67$ and $q^*_{\theta_H} = 1$. Thus, the high-risk type receives full coverage (a simple deductible contract with zero deductible), while the low-risk type receives a limited-coverage deductible contract: she may access up to 67% of the service at zero deductible or obtain full coverage by paying an out-of-pocket cost of 0.26.

It is important to note that if we ignored the equilibrium constraint EQ, the insurer would optimally choose $q^*_\theta = 1$ for both types when $p = 0.8$; that is, both types would receive simple deductible contracts with zero deductible. However, this would lead to excessive utilization, since aggregate demand would exceed supply. The insurer therefore responds by limiting coverage for the low-risk type after the zero deductible, while still allowing her to purchase full coverage at a higher out-of-pocket cost.

Finally, because the insurer always favors the higher-risk type, the optimal menu satisfies IC_1. In particular, whenever the high-risk type receives a limited-coverage deductible contract, the low-risk type receives no insurance; and whenever the low-risk type receives some insurance, the high-risk type receives full coverage with zero deductible. Hence, the allocation of coverage is increasing in risk type: $x^*(\theta_H, b) \geq x^*(\theta_L, b)$ for all b. By Lemma 4 and Condition (2), it follows that x^* satisfies IC_1.

The economic forces that lead the insurer to favor the riskier type follow from the fact that the profits of providing any units of hospital services are higher for the high-risk type than for the low-risk type, as illustrated in Figure 4. The latter occurs because $\phi(\theta_H, q) > \phi(\theta_L, q)$, i.e., the virtual value function is higher for the high-risk type than for the low-risk type for any q. This observation is formalized in Proposition 1.

3.6 Optimal Transport and Duality

We now return to the analysis of the insurer's problem in the quantity space, focusing on the infinite-dimensional risk-type space. Consider a collection of distributions over $[0, 1]$, denoted by $\{G_\theta\}_{\theta\in\Theta}$, which the insurer selects in problem $P1'$. Additionally, recall that F denotes the prior distribution over risk types. Together, these define a probability measure π on $\Theta \times [0, 1]$, where the marginal of π on Θ is F. We can then reformulate problem $P1'$ as follows:

$$\max_{\pi} \quad \int_{\underline{\theta}}^{\bar{\theta}} \int_0^1 \Phi(\theta, q)\, d\pi(\theta, q) \tag{OT}$$

$$\text{s.t.} \quad \int_0^{\theta} \int_0^1 d\pi(s, q) = F(\theta), \quad \forall \theta \in \Theta, \tag{BP}$$

$$\int_{\underline{\theta}}^{\bar{\theta}} \int_0^1 [H_\theta(p) - H_\theta(0)]q\, d\pi(\theta, q) = RS(p). \tag{EQ}$$

In the *primal problem* OT, the insurer is effectively matching risk types with the mass of agents who obtain coverage and therefore seek the service. If risk type θ is matched with q, then among the agents that the insurer can influence to seek the service—those with a valuation in the range $(0, p]$—a fraction q will get coverage. However, this allocation is subject to two constraints. The first constraint, BP, ensures that the insurer does not assign more individuals of risk type θ than the total available, as determined by the prior distribution F. In other words, the marginal of π on Θ must coincide with F, mirroring a Bayes-plausibility constraint in Bayesian persuasion models. The second constraint, EQ, is the equilibrium condition described earlier.

The similarity between OT and an optimal transport problem is advantageous because it allows us to leverage known duality results to analyze the insurer's optimization problem more effectively. To this end, we introduce the corresponding *dual problem*:

$$\min_{v,\lambda} \quad \int_{\underline{\theta}}^{\bar{\theta}} v(\theta) f(\theta) d\theta \tag{D}$$

$$\text{s.t.} \quad v(\theta) \geq \max_{q \in [0,1]} \ \Phi(\theta, q) + \lambda \left([H_\theta(p) - H_\theta(0)] q - RS(p) \right), \quad \text{for } f\text{-almost all } \theta. \tag{SP}$$

In this dual formulation, $v(\theta)$ represents the shadow price of matching risk type θ. The multiplier λ corresponds to the value of relaxing the equilibrium constraint EQ. Furthermore, constraint SP ensures that $v(\theta)$ is at least as large as the insurer's valuation of matching risk type θ with any q, where this valuation consists of the insurer's profit, $\Phi(\theta, q)$, plus the term multiplied by λ, which captures the excess demand or supply generated by risk type θ when assigned to q.

Then, we have the following duality result:

Theorem 2 (Monge-Kantorovich Duality) *A solution to both the primal problem OT and the dual problem D exists. Furthermore, there is no duality gap: the optimal value of the primal problem OT equals the optimal value of the dual problem D.*

The formal proof of Theorem 2 can be found in the technical Appendix E.

The absence of a duality gap is a direct consequence of the Monge-Kantorovich duality in optimal transport problems (see, for example, Villani et al., 2009). The existence of a solution to the primal problem OT follows from our technical assumptions that guarantee the continuity of the function Φ and the fact that the maximization is performed over a compact space. Finally, leveraging the absence of a duality gap and the existence of a solution to the primal problem, we explicitly construct a solution to the dual problem D.

Implications of Zero Duality Gap. An immediate consequence of the absence of a duality gap is that, given a solution to the dual problem D, we obtain necessary and sufficient conditions that any joint distribution π must satisfy to be a solution of the primal problem OT. To establish these conditions, consider any π that satisfies BP. Let $\pi(\cdot \mid \theta)$ denote the distribution over $[0, 1]$ conditional on risk type θ, defined as:

$$\pi(q \mid \theta) = \frac{\pi(q, \theta)}{f(\theta)}.$$

With this, we obtain the following complementary slackness result:

Corollary 1 (Complementary Slackness) *Suppose that (v^*, λ^*) is a solution to the dual problem D. Then, any π that satisfies conditions BP and EQ is a solution to the primal problem OT if and only if f-almost every θ has $\pi(\cdot \mid \theta)$ supported on:*

$$Q(\theta) \equiv \arg\max_{q\in[0,1]} \quad \Phi(\theta, q) + \lambda^* \left([H_\theta(p) - H_\theta(0)]q - RS(p)\right). \tag{Supp}$$

The formal proof of Corollary 1 is in the technical Appendix E.

The proof of Corollary 1 follows directly from the absence of a duality gap and the fact that if (v^*, λ^*) is a solution to D, then we have

$$v^*(\theta) = \max_{q\in[0,1]} \quad \Phi(\theta, q) + \lambda^* \left([H_\theta(p) - H_\theta(0)]q - RS(p)\right),$$

where the maximum is well-defined due to the continuity of $\Phi(\theta, \cdot)$ and the compactness of the set $[0, 1]$.

Corollary 1 is particularly useful for proving Theorem 1. To this end, we introduce the following lemma, from which Theorem 1 follows immediately:

Lemma 8 (Solution to OT) *There exists a solution π^* to the primal problem OT such that, for all θ and q, $\pi^*(q \mid \theta) \in \{0, \alpha, 1\}$ for some $\alpha \in (0, 1)$. Moreover, this π^* implies that each risk type θ is assigned either a **simple deductible contract** $\{(1, D(\theta))\}$ with $D(\theta) \in [0, p]$, or a **limited-coverage deductible contract** $\{(\alpha, D(\theta)), (1, M(\theta))\}$ where $0 \leq D(\theta) < M(\theta) \leq p$ and $\alpha \in (0, 1)$.*

The proof of Lemma 8 in Appendix C.

Theorem 1 follows directly from Lemma 8. Once we know that π^* is a solution to the primal problem OT, it immediately implies that the following function x^* solves the insurer's problem $P1$:

$$x^*(\theta, b) = \int_0^1 \mathbf{1}_{[b > b_\theta(q)]} \, d\pi^*(q \mid \theta).$$

Since $\pi^*(\cdot \mid \theta)$ consists of at most two jumps, and the size of these jumps is independent of the risk type, the allocation function $x^*(\theta, \cdot)$ can take one of two forms for any risk type θ. It may be a step function that jumps from 0 to 1 at some value of b given by $D(\theta)$, which corresponds to the *simple deductible contract*. Alternatively, $x^*(\theta, \cdot)$ may exhibit two jumps, as shown in Figure 3.

The intuition behind the proof of Lemma 8 is as follows: By Theorem 2 and Corollary 1, we know that there exists a π' that solves OT, a (v^*, λ^*) that solves D, and that $\pi'(\cdot \mid \theta)$ is supported on $Supp$.

By Carathéodory's theorem, we can construct a new distribution π^* such that $\operatorname{supp} \pi^*(\cdot \mid \theta)$ contains at most two elements, is a subset of $\operatorname{supp} \pi'(\cdot \mid \theta)$, and results in the same average quantity of services for type θ. This implies that π^* still satisfies the EQ constraint. Furthermore, since $\operatorname{supp} \pi^*(\cdot \mid \theta) \subset \operatorname{supp} \pi'(\cdot \mid \theta)$, we can again apply Corollary 1 to conclude that π' is optimal.

Next, we examine the conditions under which the solution to problem OT is fully implementable. In particular, we must verify that the allocation x^* induced by the optimal plan π^* described in Lemma 8 satisfies condition IC_1. By Lemma 4, it is sufficient to check whether x^* is non-decreasing in θ for all b. Using a monotone comparative-statics result, this will hold if the following condition is satisfied:

Proposition 1 *Suppose that:*

$$J(\theta, b)h_\theta(b) + \lambda^* h_\theta(b)$$

is non-decreasing in θ for all b. Let π^* denote the solution to the primal problem described in Lemma 8, and let x^* be the allocation function induced by π^*. Then x^* is non-decreasing in θ for all b, and hence x^* satisfies condition IC_1.

The proof of Proposition 1 in Appendix C. The condition above is a standard regularity requirement ensuring that the virtual value function is non-increasing in θ, with the only difference that here equilibrium effects enter through the multiplier λ^*. Economically, an increasing virtual value in θ, net of equilibrium effects, implies stronger incentives for the insurer to offer greater coverage to higher-risk types. The mathematical intuition for the result is as follows. If the virtual value function remains non-increasing in θ once equilibrium effects are accounted for, then, by Topkis (1978)'s monotonicity theorem, we obtain that whenever $\theta_1 > \theta_2$,

$$\min_{q \in Q(\theta_1)} b_{\theta_1}(q) \leq \min_{q \in Q(\theta_2)} b_{\theta_2}(q) \quad \text{and} \quad \max_{q \in Q(\theta_1)} b_{\theta_1}(q) \leq \max_{q \in Q(\theta_2)} b_{\theta_2}(q).$$

Hence, for the constructed solution π^* to the primal problem described in Lemma 8, the corresponding allocation x^* exhibits its first (and, when applicable, second) jump earlier for θ_1 than for θ_2. Moreover, because the magnitude of these jumps is type-independent, x^* is increasing in θ for all b.

A simple way to make this condition robust to the value of the multiplier is to use the fact that λ^* is bounded. More precisely,

$$|\lambda^*| \leq \sup_{\theta \in \Theta,\, b \in [0,\bar{p}]} |J(\theta, b)| \;\equiv\; U.$$

Then, we obtain the following result:

Corollary 2 *Suppose that:*

$$\frac{\partial\big(J(\theta,b)h_\theta(b)\big)}{\partial\theta} \;-\; U\left|\frac{\partial h_\theta(b)}{\partial\theta}\right| \geq 0,$$

for all θ and b. Let π^* denote the solution to the primal problem described in Lemma 8, and let x^* be the allocation function induced by π^*. Then x^* is non-decreasing in θ for all b, and hence x^* satisfies condition IC_1.

The proof of Corollary 2 in Appendix C.

A weaker sufficient condition for obtaining the desired monotone comparative statics result is to use the *interval dominance order* introduced by Quah and Strulovici (2012). In particular, we can establish the result if, for all $\theta_1 > \theta_2$, there exists a positive, non-decreasing function β (which may depend on θ_1 and θ_2) such that:

$$\beta(b)J(\theta_1,b)h_{\theta_1}(b) - J(\theta_2,b)h_{\theta_2}(b) - U\big|\beta(b)\,h_{\theta_1}(b) - h_{\theta_2}(b)\big| \geq 0,$$

for all b. Note that when $\beta(b) = 1$, this condition is equivalent to that of Corollary 2.

More generally, the interval dominance order of Quah and Strulovici (2012) allows us to obtain monotone comparative statics results that are independent of the sign or magnitude of the multiplier λ^*. To formalize this, and with a slight abuse of notation, denote by V the following function:

$$V(\theta,b) = J(\theta,b)h_\theta(b) = H_\theta(b) - g(H_\theta(b)) + \frac{1-F(\theta)}{f(\theta)}\,g'(H_\theta(b))\,\frac{\partial H_\theta(b)}{\partial b} + [b-p]\,h_\theta(b).$$

Proposition 2 *Suppose that $V(\cdot,b)$ is non-decreasing in θ for all b, that $h_\theta(b)$ is non-increasing (respectively, non-decreasing) in θ for all b, and that for any $\theta_1 > \theta_2$ there exists a positive, non-decreasing function*

β (which may depend on θ_1 and θ_2) such that:

$$\beta(b)V(\theta_1, b) \geq V(\theta_2, b) \quad \textit{and} \quad \beta(b)h_{\theta_1}(b) \geq (\leq)\, h_{\theta_2}(b),$$

for all b. Let π^* denote the solution to the primal problem described in Lemma 8, and let x^* be the allocation function induced by π^*. Then x^* is non-decreasing in θ for all b, and therefore x^* satisfies condition IC_1.

The proof of Proposition 2 is provided in Appendix C. The key challenge in establishing monotone comparative statics arises from the fact that the multiplier λ^* may take either positive or negative values. Proposition 2 addresses this by employing different monotonicity theorems depending on the sign of λ^*. Specifically, when $h_\theta(b)$ is non-increasing in θ and $\lambda^* \leq 0$, we apply Topkis (1978)'s monotonicity theorem. Conversely, when $\lambda^* > 0$, we rely on the monotonicity result of Quah and Strulovici (2012).

An immediate corollary of Proposition 2 is that when the ratio $\frac{h_{\theta_2}(b)}{h_{\theta_1}(b)}$ is non-decreasing in b, one can verify whether $\alpha(b) = \frac{h_{\theta_2}(b)}{h_{\theta_1}(b)}$ satisfies the conditions of Proposition 2. Moreover, if h_θ is non-decreasing in θ, verifying these conditions reduces to checking whether the virtual value function $J(\cdot, b)$ is non-decreasing in θ. This is precisely the case for the Stiglitz family of loss distributions introduced in Example 1, which yields the following result.

Corollary 3 *Suppose that H_θ is given by:*

$$H_\theta(b) = 1 - \theta + \theta Q(b).$$

Furthermore, suppose that $J(\cdot, \theta)$ is non-decreasing in θ for all b. Let π^* denote the solution to the primal problem described in Lemma 8, and let x^* be the allocation function induced by π^*. Then x^* is non-decreasing in θ for all b, and hence x^* satisfies condition IC_1.

We omit the proof of Corollary 3, since as stated before follows from Proposition 2.

Example 2 (An implementable Stiglitz distribution over loss-related valuations) *Suppose that H_θ belongs to the Stiglitz family of loss distributions (Example 1). Let $g(q) = q^2$ and assume $\theta \sim \mathrm{Unif}[0, \bar{\theta}]$ with $\bar{\theta} \leq \frac{2}{3}$, so that the inverse hazard rate is $c(\theta) = \frac{1-F(\theta)}{f(\theta)} = \bar{\theta} - \theta$ and its derivative is $c'(\theta) = -1$. A short calculation shows that the virtual value is increasing in θ whenever:*

$$2c(\theta) - 2\theta c'(\theta) + \theta^2(1 - Q(b))\big(2c'(\theta) - 1\big) = 2\bar{\theta} - 3\theta^2\big(1 - Q(b)\big) \geq 0,$$

which holds for all b because $\theta \leq \frac{2}{3}$. Therefore, by Corollary 3, for any π^* that solves the primal problem, the induced allocation x^* is non-decreasing in θ for all b, and hence x^* satisfies condition IC_1.

Moreover, we can also conduct monotone comparative statics directly in the quantity space and then verify whether the resulting solution satisfies condition IC_1.

Proposition 3 *Suppose that $\frac{\partial b_\theta(q)}{\partial \theta} \leq 0$ and $\frac{\partial \phi(\theta,q)}{\partial \theta} \geq 0$ for all θ and q. Let π^* denote the solution to the primal problem described in Lemma 8, and let x^* be the allocation function induced by π^*. Then x^* is non-decreasing in θ for all b, and hence x^* satisfies condition IC_1.*

The proof of Proposition 3 in Appendix C. The intuition is that when $\frac{\partial b_\theta(q)}{\partial \theta} \leq 0$, it is sufficient for the marginal revenue from expanding coverage—represented by ϕ—to be increasing in θ in order to apply Topkis (1978)'s monotonicity theorem in the quantity space and then use the fact that b_θ is non-increasing in θ to conclude that whenever $\theta_1 > \theta_2$:

$$\min_{q \in Q(\theta_1)} b_{\theta_1}(q) \leq \min_{q \in Q(\theta_2)} b_{\theta_2}(q) \quad \text{and} \quad \max_{q \in Q(\theta_1)} b_{\theta_1}(q) \leq \max_{q \in Q(\theta_2)} b_{\theta_2}(q).$$

Observe that Corollary 3 also follows from Proposition 3, since for the Stiglitz family of loss distributions $b_\theta(q)$ is independent of θ, that is, $\frac{\partial b_\theta(q)}{\partial \theta} = 0$.

Example 3 (A complicated distribution over loss-related valuations) *Suppose that*

$$H_\theta(b) = 1 - \theta^2 + \theta^2 Q(b)^{1+\varepsilon(1-\theta)},$$

where Q is an arbitrary distribution over b, and $\varepsilon \in (0,1)$. Under this specification, θ captures both the probability that an adverse event occurs and the distribution of loss-related valuations conditional on the event.

It can be verified that $\frac{\partial b_\theta(q)}{\partial \theta} \leq 0$. Hence, one of the conditions required by Proposition 3 is satisfied, and the regularity condition to be checked reduces to $\frac{\partial \phi(\theta,q)}{\partial \theta} \geq 0$.

The figure below illustrates a specific convex distortion function g, inverse hazard rate for the distribution over risk types F, and distribution Q such that $\frac{\partial \phi(\theta,q)}{\partial q} \geq 0$. Under these specifications, the solution π^* to the primal problem described in Lemma 8 induces an allocation x^* that satisfies condition IC_1.

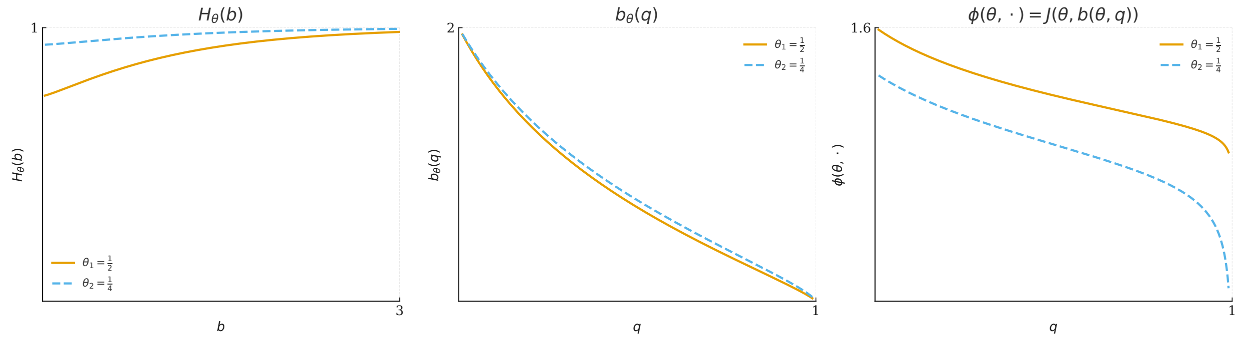

Figure 5: Example with $g(q) = q^{1.2}$, $p = 2$, $Q(b) = 1 - e^{-b}$ and inverse hazard rate $\frac{1-F(\theta)}{f(\theta)} = \frac{0.2}{1+\theta}$.

Finally, we provide conditions on the primitives under which each risk type is offered a simple deductible contract.

Proposition 4 (Simple Deductible Contracts) *Fix any risk type θ and suppose that $\frac{\partial \phi(\theta,q)}{\partial q} \leq 0$ for all q. Let π^* denote the solution to the primal problem described in Lemma 8, then $\pi^*(q \mid \theta) \in \{0,1\}$ for all q. Under this solution, all agents are assigned a* ***simple deductible contract****.*

The proof of Proposition 4 is provided in Appendix C. Proposition 4 once again relies on our complementary slackness result. To illustrate, suppose that $\frac{\partial \phi(\theta,q)}{\partial q} > 0$ holds with strict inequality. Then, the objective in $Supp$ is strictly concave in q, implying that the set $Supp$ contains only one element. Moreover, by our complementary slackness result, we know that $\pi^*(\cdot \mid \theta)$ is supported on $Supp$, leading to the desired conclusion.

4 Second-Stage: Price of Services

Throughout section 3, we took the price of services, p, as given and characterized the shape of the optimal menu that sustains p in equilibrium. This allowed us to associate each achievable p with the corresponding profits obtained by the insurer, denoted by $\Pi(p)$.

We now proceed to the second-stage analysis, which consists of solving for the price that maximizes Π. Formally, the second-stage problem is:

$$\max_{p \in [p^N, p^F]} \Pi(p). \tag{$P2$}$$

To analyze problem $P2$, we introduce the following result, which characterizes the derivative of Π:

Theorem 3 *Let x^* be a solution to the first-stage problem $P1$ and λ^* be a solution to the dual problem D. Then, the function $\Pi : [p^N, p^F] \to \mathbb{R}_+$ is differentiable in p, and its derivative is given by:*

$$\Pi'(p) = \int_{\underline{\theta}}^{\overline{\theta}} x^*(\theta, p) J(\theta, p) h_\theta(p) f(\theta)\, d\theta - \int_{\underline{\theta}}^{\overline{\theta}} \int_0^p x^*(\theta, b) h_\theta(b)\, db f(\theta)\, d\theta$$
$$- \lambda^* \left(S'(p) + \int_{\underline{\theta}}^{\overline{\theta}} (1 - x^*(\theta, p)) h_\theta(p) f(\theta) d\theta \right) - [g(H_{\underline{\theta}}(p+\beta)) - g(H_{\underline{\theta}}(p))]$$

The formal proof of Theorem 3 is in Appendix D. We invoke Theorem 4 of Milgrom and Segal (2002), which provides an envelope formula for saddle points. In particular, our duality result (Theorem 2) implies that a solution π^* to the primal problem OT, together with the multiplier λ^* that solves the dual problem D, form a saddle point. Under our technical assumptions, this allows us to apply Theorem 4 of Milgrom and Segal (2002) to obtain a formula for the derivative of the insurer's profit in terms of π^* and λ^*. Finally, using the relationship between π^* and x^*, we derive the desired result. We postpone the economic interpretation of Π' in the case where the insurer offers a simple deductible contract to all risk types, as this setting allows for a clearer explanation.

Since Π is differentiable, it is also continuous. Moreover, because problem $P2$ involves maximizing Π over a compact domain, the *Extreme Value Theorem* ensures that a maximizer exists. Whenever the maximizer lies in the interior of the domain, it must satisfy the first-order condition:

Corollary 4 *A solution p^* to second-stage problem $P2$ exists. Moreover, if $p^* \in (p^N, p^F)$, then it satisfies $\Pi'(p^*) = 0$.*

Before turning to the case where the insurer offers only simple deductible contracts, we show that as the premium β decreases, the equilibrium service price is weakly higher.

Corollary 5 *Suppose that $\beta > \beta'$. Then, there does not exist a price p that solves the second-stage problem $P2$ under β and is strictly greater than all prices that solve the second-stage problem $P2$ under β'.*

The formal proof of Corollary 5 is provided in Appendix D. The result is stated to accommodate the possibility of multiple maximizers. When the maximizer of $P2$ is unique under both β and β', the corollary simply states that the price solving $P2$ under β is lower than the one under β'.

The mathematical intuition is straightforward: the derivative of the insurer's profit under β is always weakly lower than under β'. The economic intuition is similarly simple: when the premium β is very high, an increase in p has little effect on the agent's outside option. Consider the extreme case in which β is high enough that $\beta + p \geq \bar{b}$; in this case the agent will never seek the service in the absence of insurance, so further increases in p do not change this fact, and her outside option—and thus her willingness to pay for insurance—remains the same. Then, any incentive to increase p in order to worsen the agent's outside option, raise her willingness to pay for insurance, and extract more surplus is shut down.

Simple Deductible Contracts. Throughout this section, we assume that for all p and θ, the function $\phi(\theta, \cdot)$ is strictly decreasing. Recall that $\phi(\theta, q) = J(\theta, b_\theta(q))$, where $b_\theta(q) = H_\theta^{-1}(1-q \mid p)$, and $H_\theta(\cdot \mid p)$ denotes the distribution of valuations for risk type θ, conditional on being in $(0, p]$. It is easy to verify that $b_\theta(\cdot)$ is strictly decreasing. Therefore, the assumption is equivalent to requiring that the virtual value function $J(\theta, \cdot)$ is strictly increasing for all θ.

Then, Proposition 4 implies that for all p, it is optimal for the insurer to offer each risk type a simple deductible contract, denoted by $\{(1, D(\theta; p))\}$, where $D(\theta; p) \in [0, p]$. With a slight abuse of notation, we now make explicit the dependence of the contract on the price p. Similarly, we use the notation $\lambda^*(p)$ to reflect the dependence of the multiplier solving the dual problem D on p. Applying Theorem 3, we obtain the following result:

Corollary 6 *Suppose that $J(\theta, \cdot)$ is strictly increasing for all θ. Then, the function $\Pi : [p^N, p^F] \to \mathbb{R}+$ is differentiable in p, and its derivative is given by:*

$$\Pi'(p) = \int_{\underline{\theta}}^{\bar{\theta}} \mathbf{1}_{[D(\theta;p)<p]} \left[\underbrace{1 - g(H_\theta(p)) + \frac{1-F(\theta)}{f(\theta)} g'(H_\theta(p)) \frac{\partial H_\theta(p)}{\partial \theta}}_{\textit{Surplus Extraction}} - \underbrace{(1 - H_\theta(D(\theta;p)))}_{\textit{Cost Effect}} \right] f(\theta) d\theta$$
$$\underbrace{- \lambda^*(p) \left[S'(p) + \int_{\underline{\theta}}^{\bar{\theta}} \mathbf{1}_{[D(\theta;p)=p]} h_\theta(p) f(\theta) d\theta \right]}_{\textit{Equilibrium effects}} - \underbrace{[g(H_{\underline{\theta}}(p+\beta)) - g(H_{\underline{\theta}}(p))]}_{\textit{Outside Option}}.$$

We omit the proof of Corollary 6 as it is an immediate consequence of Theorem 3 using the fact that $x^*(\theta, b)$ is equal to zero for values of b lower or equal than $D(\theta; p)$, it is equal to one for values of b strictly higher than $D(\theta; p)$ and $D(\theta; p) \in [0, p]$.

Corollary 6 allows us to provide a clear interpretation of the economic forces that determine the value of inducing a marginal increase in the service price. For each θ, the term $1 - g(H_\theta(p))$ captures the ex-ante increase in the value of insurance. Agents who receive coverage—i.e., those for whom $D(\theta; p) < p$—now face potentially higher losses as the service price increases. Moreover, these losses are protected by insurance. Since $1 - g(H_\theta(p))$ reflects the cumulative weight assigned to losses above p, this term measures the agent's marginal increase in willingness to pay for insurance.

If θ were observable, the insurer could fully appropriate this increase in willingness to pay by raising premiums accordingly, thereby increasing profits. However, when θ is private information, the insurer can only appropriate the agent's first-period virtual value—that is, her willingness to pay net of information rents. Recall that the term $\frac{1-F(\theta)}{f(\theta)} g'(H_\theta(b)) \frac{\partial H_\theta(b)}{\partial \theta}$ is negative, reflecting these rents.

The second term, $1 - H_\theta(D(\theta; p))$, captures the insurer's cost from increasing the price. If a risk type θ receives a simple deductible contract $\{(1, D(\theta; p))\}$ and is covered, then all agents with valuations in the range $(D(\theta; p), \bar{b})$ will choose to seek the service and pay the OPC. This implies a total demand equal to $1 - H_\theta(D(\theta; p))$. Therefore, a unit increase in price leads to a proportional increase in payouts.

In addition, we must account for equilibrium effects. The sign of this component is governed by the multiplier $\lambda^*(p)$, since the bracketed term is always positive and represents the increase in residual supply. Specifically, a higher price increases the supply of services, captured by $S'(p)$, and reduces demand among those not covered under the optimal contract (i.e., those for whom $D(\theta; p) = p$).

Whenever $J(\theta, \cdot)$ is strictly increasing, we can derive a comparative statics result describing how the multiplier $\lambda^*(\cdot)$ and the deductible $D(\theta; \cdot)$ change with the price.

Proposition 5 *Suppose that $J(\theta, \cdot)$ is strictly increasing for all θ. Then, $\lambda^*(\cdot)$ is continuous and strictly increasing, and for all θ, $D(\theta; \cdot)$ is continuous and strictly decreasing whenever $D(\theta; p) \in (0, p)$ and is non-increasing when $D(\theta; p) = 0$.*

The proof of Proposition 5 is in Appendix D. The intuition is straightforward: when the price increases, supply also increases. To clear the market, demand must rise as well. To incentivize higher demand, the multiplier must increase. In turn, agents require greater coverage, which the insurer provides by lowering the deductible.

Using our comparative statics result, we now address the question of whether the insurer chooses a "high" or "low" price in the second-stage problem $P2$. To do so, we must first define what constitutes a high or low price. For this purpose, we introduce the concept of a *price-taking benchmark*, which is the price that clears the services market and solves the first-stage problem $P1$, ignoring the equilibrium constraint EQ. We denote this benchmark price by p^U.

Note that the value of relaxing the equilibrium constraint EQ at p^U is zero, i.e., $\lambda^*(p^U) = 0$. Moreover, p^U has a clear economic interpretation: it is the market price that would arise if the insurer took prices as given and did not internalize its impact on the market for services. The following lemma establishes the existence and uniqueness of the price-taking benchmark.

Lemma 9 *Suppose that $J(\theta, \cdot)$ is strictly increasing for all θ. Then, there exists a unique price p^U such that $\lambda^*\left(p^U\right) = 0$.*

The existence and uniqueness of p^U follow essentially from the fact that $\lambda^*(\cdot)$ is strictly increasing and continuous. The proof of Lemma 9 is provided in Appendix D.

We now establish a sufficient condition on the model primitives under which the price that maximizes the second-stage problem $P2$ is higher than p^U. This implies that the insurer provides more coverage across the population of agents relative to the benchmark. Indeed, achieving a higher price than p^U requires an increase in the aggregate demand for services. To sustain this higher demand, the insurer must expand coverage across agents, thereby stimulating utilization of third-party services and raising the equilibrium service price.

For this purpose, let $SE_\theta(b)$ denote the *surplus extraction force* at any loss-related valuation b:

$$SE_\theta(b) = 1 - g(H_\theta(b)) + \frac{1 - F(\theta)}{f(\theta)} g'(H_\theta(b)) \frac{\partial H_\theta(b)}{\partial \theta}.$$

We say that SE_θ satisfies *H_θ-regularity* if

$$SE_\theta(b_1) - SE_\theta(b_2) \geq -[b_1 - b_2] h_\theta(b_2) \quad \text{for all } b_1 > b_2. \qquad (H_\theta\text{-regularity})$$

Condition H_θ-regularity states that as the loss-related valuation increases from b_2 to b_1, the decline (if any) in the surplus extraction force cannot be so steep that it exceeds the additional cost the insurer incurs by extending coverage to all realizations in $[b_2, b_1]$.

We then have the following result.

Proposition 6 *Suppose that $J(\theta, \cdot)$ is strictly increasing, $\beta = 0$ and that SE_θ satisfies condition H_θ-regularity for all θ. Then, for every type θ that receives coverage (i.e., $D(\theta; p) < p$), the surplus extraction force strictly dominates the cost effect force at every service price below the benchmark ($p < p^U$). Moreover, any optimal price p^* that solves the second-stage problem $P2$ satisfies $p^* \geq p^U$.*

The proof is provided in Appendix D. Intuitively, when SE_θ satisfies H_θ-regularity, an increase in the service price does not reduce the surplus extraction force faster than it raises the cost of providing additional coverage. Furthermore, by Proposition 5, we have that $-\lambda^*(p) > -\lambda^*(p^U) = 0$ for all $p < p^U$. Hence, equilibrium effects are positive at all prices below the benchmark. At such prices, relaxing the equilibrium constraint is valuable: the insurer would like to expand coverage further but is constrained by limited supply. Together, these two observations imply the desired result.

A simple sufficient condition for H_θ-regularity is that SE_θ be non-decreasing. Another sufficient condition arises when H_θ is concave. In that case, a convenient differential condition is

$$\frac{\partial SE_\theta(b)}{\partial b} + h_\theta(b) \geq 0 \quad \text{for all } b,$$

that is, the sum $SE_\theta(b) + H_\theta(b)$ is non-decreasing in b. Since H_θ is a distribution function and thus non-decreasing, this condition requires that the rate of decline of SE_θ be bounded by the loss density h_θ, ensuring that the extraction force does not decay too quickly relative to the arrival rate of losses.

Example 4 *Suppose there are two risk types, $\theta \in \{\theta_L, \theta_H\}$, with $\theta_H > \theta_L$. In this discrete case, adjusting for the discrete version of information rents, the surplus extraction forces and virtual value functions are given by:*

$$\begin{aligned}
SE_{\theta_H}(b) &= 1 - g(H_{\theta_H}(b)), \\
SE_{\theta_L}(b) &= 1 - g(H_{\theta_L}(b)) + \frac{1 - f(\theta_L)}{f(\theta_L)}\left[g(H_{\theta_H}(b)) - g(H_{\theta_L}(b))\right], \\
J(\theta, b) &= \frac{SE_\theta(b)}{h_\theta(b)} + b - p - \frac{1 - H_\theta(b)}{h_\theta(b)}, \quad \textit{for } \theta \in \{\theta_L, \theta_H\}.
\end{aligned}$$

Suppose that $H_\theta(b) = 1 - \theta + \theta\beta b$. That is, conditional on an adverse event, the loss-related valuation

b is uniformly distributed on $\left[0, \frac{1}{\beta}\right]$. Let $g(q) = q^\gamma$ with $\gamma > 1$, so that g is strictly convex and agents are risk-averse in the strong sense. Finally, let the supply function be $S(p) = s \cdot p$, where s is the slope.

Consider the parameter values $\theta_H = 0.95$, $\theta_L = 0.85$, $f(\theta_H) = 0.72$, $\gamma = 1.75$, and $s = 0.3$ (other combinations yield similar results). For these values, the set of implementable prices is $[2.35, 3.1]$, and we obtain the following:

- *SE_θ satisfies condition H_θ-regularity;*
- *$J(\theta, \cdot)$ is non-decreasing in b (so, by Proposition 4, each θ receives a simple deductible contract);*
- *$J(\cdot, b)$ is non-decreasing in θ (so, by Corollary 3, $D(\cdot; p)$ is non-increasing in θ, and the solution to the first-stage problem $P1$ satisfies IC_1).*

Hence, by Proposition 6, the surplus extraction force dominates the cost effect force at every price below the benchmark, and the optimal solution p^* to problem $P2$ satisfies $p^* \geq p^U$.

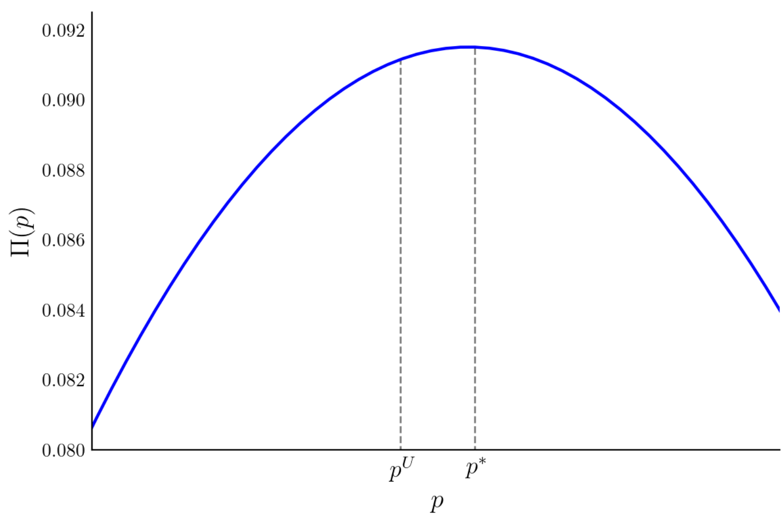

Figure 6: In this case, the optimal price is $p^* = 2.61$ and the price-taking benchmark is $p^U = 2.64$.

5 Extensions

5.1 Social Planner

In this section, we show how our model can also be used to analyze the price impact of a public provider of insurance. We are particularly motivated by countries where the state is the main provider of insurance, while the majority of service providers are privately owned, or countries where insurance markets are highly regulated.[2]

To analyze this setting, we consider a social planner with an objective similar to that proposed by Akbarpour et al. (2024). Specifically, in the first-stage problem, given a price p, the planner designs an insurance menu that maximizes a weighted average of insurer profits and social surplus:

$$\underbrace{\delta\left[\int_{\underline{\theta}}^{\bar{\theta}}\int_0^p x(\theta,b)J(\theta,b)h_\theta(b)\,db f(\theta)\,d\theta - U(\underline{\theta}) - \int_0^p [1-g(H_{\underline{\theta}}(b))]\,db\right]}_{\text{Profits}} + \underbrace{\int_{\underline{\theta}}^{\bar{\theta}} \omega(\theta)U(\theta)f(\theta)\,d\theta}_{\text{Social Surplus}},$$

where the planner assigns a weight δ to profits and a weight $\omega(\theta)$ to an agent with ex-ante risk type θ. Profits are given by Lemma 7, and correspond to those generated by an implementable insurance menu—that is, a menu that satisfies conditions IC_2, IC_1, and IR. Similarly, $U(\theta)$ denotes the ex-ante payoff—or certainty equivalent—that an agent of type θ obtains from an implementable menu (see Lemma 4, Condition (1)).

Using the expression for $U(\theta)$ provided in Lemma 4, Condition (1), we can simplify the objective as follows:

$$\int_{\underline{\theta}}^{\bar{\theta}}\int_0^p x(\theta,b)V(\theta,b)h_\theta(b)\,db f(\theta)\,d\theta + \int_{\underline{\theta}}^{\bar{\theta}}\int_0^p \Omega(\theta)I(\theta,b)\,db f(\theta)\,d\theta + (\bar{\omega}-\delta)U(\underline{\theta}) - \delta\int_0^p [1-g(H_{\underline{\theta}}(b))]\,db,$$

where $I(\theta,b)$ represents the period 1 information rents:

$$I(\theta,b) = \frac{1-F(\theta)}{f(\theta)} g'(H_\theta(b))\frac{\partial H_\theta(b)}{\partial\theta}.$$

$\Omega(\theta)$ represents the average social weight assigned to all agents with a risk type greater than θ:

$$\Omega(\theta) = \mathbb{E}_F[\omega(s)\mid s\geq\theta] = \frac{1}{1-F(\theta)}\int_\theta^{\bar{\theta}} \omega(s)f(s)\,ds.$$

$V(\theta,b)$ represents the social value of allocating a unit of the service to an agent with risk type θ

[2] For instance, in Germany, the government mandates insurance coverage through a statutory health insurance system (SHI), while roughly 50% of hospitals are privately owned. In Japan, there is universal public insurance through government-managed or employer-based plans, with state regulation, while the majority of hospitals are privately owned.

and valuation b:

$$V(\theta,b)=\delta J(\theta,b)-\Omega(\theta)\frac{I(\theta,b)}{h_\theta(b)}.$$

This social value function has a similar economic interpretation to the one in Akbarpour et al. (2024), but in the context of insurance. In particular, it consists of a weighted sum of the period 1 and period 2 virtual values, captured by $J(\theta,b)$ (corresponding to revenue), and the period 1 information rents, $I(\theta,b)$, weighted by $\Omega(\theta)$. The weight $\Omega(\theta)$ on the period 1 information rents arises from the envelope formula (Lemma 2, Condition (1)), which implies that, for the menu to satisfy condition IC_1, any increase in the ex-ante utility of risk type θ must also be reflected in the utility of all higher risk types.

Finally, $\bar\omega$ represents the average social weight across all risk types:

$$\bar\omega=\mathbb{E}_F[\omega(\theta)]=\int_{\underline{\theta}}^{\bar\theta}\omega(\theta)f(\theta)\,d\theta,$$

where, whenever $\bar\omega$ exceeds the weight on profits, the planner assigns the highest possible ex-ante utility to the lowest type and, through premiums, implements the largest possible transfer to all risk types.

Then, the first-stage social planner's problem is given by:

$$W(p)\equiv\max_{x,U(\underline{\theta})}\quad\int_{\underline{\theta}}^{\bar\theta}\int_0^p x(\theta,b)V(\theta,b)h_\theta(b)\,db f(\theta)\,d\theta+(\bar\omega-\delta)U(\underline{\theta})+K(p)\qquad (P1-SP)$$

$$\text{s.t.}\quad\int_{\underline{\theta}}^{\bar\theta}\int_0^p x(\theta,b)h_\theta(b)\,db f(\theta)\,d\theta=RS(p),\qquad (EQ)$$

$$x(\theta,\cdot):B\to[0,1]\text{ is non-decreasing for all }\theta,\qquad (MON)$$

$$U(\underline{\theta})\in\left[U_{NP}(\underline{\theta}),\bar U\right],\qquad (IR)$$

where:

$$K(p)=\int_{\underline{\theta}}^{\bar\theta}\int_0^p\Omega(\theta)I(\theta,b)\,db f(\theta)\,d\theta)-\delta\int_0^p[1-g(H_{\underline{\theta}}(b))]\,db.$$

The function $W:\left[p^N,p^F\right]\to\mathbb{R}$ represents the welfare attained by the planner from the optimal menu that sustains price p in equilibrium. Note that, even in the context of a social planner, the implementable prices remain constrained to the interval $\left[p^N,p^F\right]$ (see Lemma 6).

Nonetheless, unlike the first-stage insurer's problem $P1$, it is not immediate that the planner will set the ex-ante utility of type $\underline{\theta}$ equal to her no-participation utility, $U_{NP}(\underline{\theta})$. As argued above, if $\bar{\omega} > \delta$, the planner prefers to set $U(\underline{\theta})$ as high as possible. Therefore, to ensure the maximization problem is well-defined, we assume that the maximum ex-ante utility the planner can assign to type $\underline{\theta}$ is bounded above by some $\bar{U}$ satisfying $U_{NP}(\underline{\theta}) \leq \bar{U} < \infty$.

Once we define problem $P1 - SP$, it is straightforward to see that we can apply the techniques described in Sections 3.4 and 3.6 to represent $P1 - SP$ in the quantity space and rewrite it as an optimal transport problem, along with its corresponding dual formulation. Moreover, Lemma 8 allows us to characterize the solutions to this optimal transport problem. As in the case of a monopolist insurer, we conclude that the social planner will offer to each risk type either a simple deductible contract or a limited coverage deductible contract:

Corollary 7 (Optimal Contracts Social Planner) *There exists an x^* that solves the problem $P1 - SP$ such that the contract offered to risk type θ takes one of the following forms: a **simple deductible contract** $\{(1, D(\theta))\}$ with $D(\theta) \in [0, p]$, or a **limited coverage deductible contract** $\{(\alpha, D(\theta)), (1, M(\theta))\}$ where $0 \leq D(\theta) < M(\theta) \leq p$ and $\alpha \in (0, 1)$.*

Next, we turn to the second-stage planner's problem, which consists of choosing the price that maximizes the welfare function W:

$$\max_{[p^N, p^F]} W(p) \tag{$P2 - SP$}$$

Using an argument analogous to the one employed in the proof of Theorem 3, we can compute the derivative of the welfare function with respect to p. For simplicity and clarity of exposition, assume that $\bar{\omega} > \delta$ and that $V(\theta, \cdot)$ is strictly increasing for all θ. The first assumption implies that $U(\underline{\theta}) = \bar{U}$, and the second implies, via Proposition 4, that the planner will offer each risk type a simple deductible contract of the form $\{(1, D(\theta; p))\}$, where we again make explicit the dependence of the contract on the price p. Then, the derivative of W is given by:

Corollary 8 (Welfare Derivative) *The function W is differentiable in p and its derivative is given by:*

$$W'(p) = \int_{\underline{\theta}}^{\bar{\theta}} \mathbf{1}_{[D(\theta;p)<p]} \delta \left[\underbrace{1 - g(H_\theta(p)) + I(\theta,p)}_{\text{Virtual value}} - \underbrace{(1 - H_\theta(D(\theta;p)))}_{\text{Payouts}} \right] f(\theta) d\theta$$
$$\underbrace{- \lambda^*(p) \left[S'(p) + \int_{\underline{\theta}}^{\bar{\theta}} \mathbf{1}[D(\theta;p) = p] h_\theta(p) f(\theta) d\theta \right]}_{\text{Equilibrium effects}} + \int_{\underline{\theta}}^{\bar{\theta}} \underbrace{\mathbf{1}_{[D(\theta;p)=p]} \Omega(\theta) I(\theta,p) f(\theta)\, d\theta}_{\text{Information rents}}$$
$$- \underbrace{\delta[1 - g(H_{\underline{\theta}}(p))]}_{\text{Outside Option}}.$$

Compared to the derivative of profits, $\Pi'(p)$, the derivative of welfare assigns an additional negative weight to the period 1 information rents of agents who do not receive coverage (i.e., those for whom $D(\theta;p) = p$). This is captured by the term $\Omega(\theta)I(\theta,p)$, where, recall, $I(\theta,p) \leq 0$. Intuitively, this term reflects that a social planner dislikes high prices because uninsured agents are unprotected against them.

As in the case of a monopolist provider, when $V(\theta,\cdot)$ is strictly increasing for all θ, we can apply an argument analogous to the one in Proposition 5 to show that $\lambda^*(\cdot)$ is strictly increasing and continuous, and that $D(\theta;\cdot)$ is continuous and strictly increasing whenever $D(\theta;p) \in (0,p)$, while it is non-increasing when $D(\theta;p) = 0$. Moreover, we define the price-taking benchmark p^U as the price that clears the services market and solves problem $P1-SP$ without the EQ constraint (i.e., such that $\lambda^*(p^U) = 0$). Existence and uniqueness of p^U then follow from the argument in Lemma 9. Finally, using the derivative of the welfare function, we can show that:

Proposition 7 *Suppose that $V(\theta,\cdot)$ is strictly increasing for all θ, $\delta = 0$, and $p^U \in (p^N, p^F)$. Then, the price chosen by the planner in the second-stage problem $P2-SP$ is strictly lower than p^U.*

We omit the proof of Proposition 7, as it follows from the fact that $W'(p) < 0$ for all $p \geq p^U$. Indeed, since $\delta = 0$, the planner assigns zero weight to revenue. As a result, the derivative of W depends only on the term capturing forgone information rents, which is always negative, and on the term capturing equilibrium effects, which is also negative above the price-taking benchmark. This is because the bracketed term is always positive, and $\lambda^*(\cdot)$ is strictly increasing. Hence, $-\lambda^*(p) < -\lambda^*(p^U) = 0$ for all $p > p^U$.

5.2 Nash Bargaining

In this section, we show how our baseline model can be extended to a setting in which, in the second stage, the insurer does not necessarily have full control over the service price, as the service provider may also possess bargaining power. Under this formulation, our baseline model corresponds to the special case in which the insurer holds all the bargaining power.

More precisely, the timing and interaction between the service provider and insurer is as follows:

1. First, the insurer and the service provider bargain over the service price p and a capacity cap S, which specifies the maximum number of patients the provider is willing to accept from the insurer. The pair (p, S) is determined through simultaneous Nash bargaining, with the outcome maximizing the Nash product:

   $$\Pi_i(p, S)^{\beta} \Pi_s(p, S)^{1-\beta},$$

   where Π_i and Π_s denote the profits of the insurer and the service provider, respectively, and $\beta \in [0, 1]$ is the insurer's bargaining weight.

   The fact that the insurer and the provider bargain over S is motivated by empirical evidence showing that providers often face real or strategic capacity constraints. For example, Ho (2009) analyzes a setting in which hospitals encounter such constraints during bargaining, either due to physical limitations—such as limited numbers of beds or nurses per bed—or because a hospital is highly attractive to consumers and thus less reliant on insurers to secure patient volume. Ho (2009) finds that hospitals expected to fill their beds can leverage this position to obtain approximately \$6,900 more per patient than average. Similarly, Town and Vistnes (2001) argue that hospitals with little excess capacity tend to have higher reservation prices and, as a result, stronger bargaining positions.

   Moreover, in many countries such constraints are explicitly written into contracts. For instance, in European health systems, contracts often include explicit volume caps, sometimes called "activity caps" or "contracted volumes" (Appleby et al., 2012).

   Finally, this formulation also accommodates a simpler version in which only the price p is negotiated. In that case, p indirectly determines the provider's capacity constraint through a function $S : \mathbb{R}_+ \to \mathbb{R}_+$, where, for example, $S(p) = S$ for all p, implying that the hospital's capacity constraint is invariant to price.

2. Once p and S are agreed upon, the insurer offers an insurance menu to the agents:

$$\left\{\left(\{(x(\theta,b),D(\theta,b))\}_{b\in B}\,,\,t(\theta)\right)\right\}_{\theta\in\Theta}.$$

   The menu is offered at a stage when each agent observes only her ex-ante risk type θ, before her loss-related valuation b is realized. The agent also observes the agreed price p, and decides whether to select a contract from the menu. If she does, she pays the corresponding premium.

3. The agent's private loss-related valuation b is then realized. Based on this information, she chooses whether to the seek the service or remain at home. If she seeks the service, she can either use the benefit specified in her contract to cover the service costs, or pay entirely out of pocket.

4. The collection of contracts selected by the agents, along with the specific items used within those contracts, generates total demand for services at the agreed price p. This demand must satisfy the capacity constraint S—that is, the total volume of customers referred by the insurer must not exceed S.

Note that for any fixed price p, we can still apply Lemmas 3 and 4 to characterize an implementable insurance menu—that is, a menu that satisfies conditions IC_2, IC_1, and IR. Moreover, this implies that the insurer's profits from any implementable menu are still determined by Lemma 7.

Then, the insurer's problem is given by:

$$\Pi_i(p) \equiv \max_x \quad \int_{\underline{\theta}}^{\bar{\theta}} \int_0^p x(\theta,b)J(\theta,b)h_\theta(b)\,db f(\theta)\,d\theta + K(p) \qquad (P1-NB)$$

$$\text{s.t.} \quad \int_{\underline{\theta}}^{\bar{\theta}} \int_0^p x(\theta,b)h_\theta(b)\,db f(\theta)\,d\theta + \int_{\underline{\theta}}^{\bar{\theta}} [1-H_\theta(p)]f(\theta)\,d\theta \le S, \qquad (CAP)$$

$$x(\theta,\cdot): B \to [0,1] \text{ is non-decreasing for all } \theta, \qquad (MON)$$

where:

$$K(p) = -U_{NP}(\underline{\theta}) - \int_0^p [1-g(H_{\underline{\theta}}(b))]\,db.$$

The main difference between problem $P1$ and problem $P1-NB$ is that we replace the equilibrium constraint EQ with the capacity constraint CAP. This capacity constraint requires that the volume of agents referred by the insurer to the service provider not exceed S.

It is straightforward to apply the techniques described in Sections 3.4 and 3.6 to represent problem $P1 - NB$ in quantity space and rewrite it as an optimal transport problem, along with its corresponding dual formulation. Moreover, Lemma 8 allows us to characterize the solutions to this optimal transport problem. As in the case of a competitive market for services, we conclude that the insurer offers each risk type θ either a simple deductible contract or a limited coverage contract:

Corollary 9 (Optimal Contracts under Nash Bargaining) *There exists an x^* that solves the problem $P1-NB$ such that the contract offered to risk type θ takes one of the following forms: a* ***simple deductible contract*** $\{(1, D(\theta))\}$ *with* $D(\theta) \in [0, p]$*, or a* ***limited coverage deductible contract*** $\{(\alpha, D(\theta)), (1, M(\theta))\}$ *where* $0 \leq D(\theta) < M(\theta) \leq p$ *and* $\alpha \in (0, 1)$*.*

Using the solution to problem $P1 - NB$, we can associate each price p and capacity cap S with the corresponding insurer profits, denoted $\Pi_i(p, S)$. We then define the *Nash bargaining price and capacity cap* as the solution to the problem:

$$\max_{p,S} \; \Pi_i(p, S)^{\beta} \Pi_s(p, S)^{1-\beta}, \tag{$P2$-SP}$$

Note that when $\beta = 1$—that is, when the insurer holds all the bargaining power—and when the capacity cap S is indirectly determined by the agreed price p, problem $P2$-SP reduces to the second-stage problem $P2$ of the baseline model.

Moreover, we can use the argument made in Theorem 3 to express the derivative of Π_i with respect to p and S:

Corollary 10 (Insurer's Profit Derivative under Nash Bargaining) *The function Π_i is differentiable* *in p and S; its partial derivatives are:*

$$\begin{aligned}\frac{\partial \Pi_i(p,S)}{\partial p} =& \int_{\underline{\theta}}^{\bar{\theta}} x^*(\theta,p)J(\theta,p)h_\theta(p)f(\theta)\,d\theta - \int_{\underline{\theta}}^{\bar{\theta}}\int_0^p x^*(\theta,b)h_\theta(b)\,db f(\theta)\,d\theta \\ & - \lambda^*\left(\int_{\underline{\theta}}^{\bar{\theta}}(1-x^*(\theta,p))h_\theta(p)f(\theta)d\theta\right) - [g(H_{\underline{\theta}}(p+\beta)) - g(H_{\underline{\theta}}(p))], \\ \frac{\partial \Pi_i(p,S)}{\partial S} =& - \lambda^*.\end{aligned}$$

Corollary 5.2 is key for analyzing the Nash-bargaining problem $P2\text{-}SP$. For this problem, we have not yet imposed any assumptions on the service provider's profit function Π_s. In practice, depending on the setting, these profits may even depend on the insurance menu chosen by the insurer in problem $P1-NB$. For example, one could assume:

$$\Pi_s(p) = \int_{\underline{\theta}}^{\bar{\theta}}\int_0^p x^*(\theta,b)(p-mc)h_\theta(b)\,db\,f(\theta)\,d\theta + \int_{\underline{\theta}}^{\bar{\theta}}[1-H_\theta(p)](p-mc)f(\theta)\,d\theta,$$

where the service provider faces constant marginal cost mc, and $x^*(\theta,b)$ denotes the units of the service provided to an agent of type θ with loss-related valuation $b \leq p$ in the optimal menu solving problem $P1-NB$. Recall from Lemma 3 that $x^*(\theta,b) = 1$ for all $b > p$.

5.3 Pricing Tiers

Service providers typically offer *tiered services*, differentiated by the level of care, specialization, and quality provided. For example, a hospital may offer primary care or routine checkups at a base price p_1, while secondary care—such as treatment from a specialist—is available for an additional cost p_2. Similarly, basic lab panels may be priced at p_1, whereas more advanced diagnostics like imaging require an extra payment of p_2. In terms of amenities, a standard room might be available at price p_1, while private rooms or more luxurious accommodations come at an added cost p_2.

In our initial model, it is natural to assume that agents with higher valuations are more likely to demand additional services, and thus incur higher treatment costs. To formalize this idea, we continue to assume that each agent is characterized by her risk type $\theta \in \Theta$ and valuation $b \in B$. The service provider now offers *tiered pricing*: a basic service is available for a price $p_1 \in \mathbb{R}_+$ and additional (or higher-quality) services are offered for an extra price $p_2 \in \mathbb{R}_+$.

An agent's loss when her valuation is b and she receives only basic services is given by $p_1 +$

$\max\{b-b^*,0\}$, for some cap $b^* \in B$. Conversely, her loss when she only receives additional services is $\min\{b,b^*\}+p_2$. This implies that the agent's valuation for basic services is $\min\{b,b^*\}$, reflecting a cap or threshold b^* on the loss that basic care can prevent. The valuation for additional services is then $\max\{b-b^*,0\}$—only agents with $b>b^*$ value these services.

Recall that b represents the agent's willingness to pay for third-party services. Since this valuation is subjective and shaped by many personal and contextual factors, we model b as private information.

One key factor influencing b is the *severity* of the agent's condition: more severe illnesses typically generate higher perceived losses and, consequently, higher valuations for treatment. Thus, a higher value of b corresponds to a more serious health condition. This link between severity and valuation is explicitly captured in this extension of our baseline model.

In particular, when an agent's valuation b exceeds a threshold b^*, we can interpret this as the agent suffering a more severe adverse event than one with $b<b^*$. In such cases, a basic service (i.e., a routine checkup, denoted as service 1) is insufficient to fully treat or repair damages. The agent also requires a more specialized intervention (i.e., seeing a specialist, denoted as service 2). As a result, treating the more severe condition is costlier: the agent must pay the combined cost p_1+p_2 rather than just p_1.

We assume that $p_1 \in [0,b^*]$ such that this structure yields three types of behavior:

- Agents with $b \le p_1$ do not seek the service and incur a loss of b.
- Agents with $b \in (p_1, b^*+p_2]$ receive only basic services and incur a loss of $p_1+\max\{b-b^*,0\}$.
- Agents with $b > b^*+p_2$ receive both basic and additional services and incur a loss of p_1+p_2.

It is easy to verify that the certainty equivalent for a risk type θ agent when a service provider offers tiered services is given by:

$$-\sum_{j=1}^{2}\int_{b_j}^{b_j+p_j}[1-g(H_{\underline{\theta}}(b))]\,db,$$

where $b_1=0$ and $b_2=b^*$.

Moreover, an insurance menu is now represented as:

$$\left\{\Big(\{(x_j(\theta,b),D_j(\theta,b))\}_{b\in B,\,j\in\{1,2\}},\ t(\theta)\Big)\right\}_{\theta\in\Theta},$$

where a contract $\{(x_j(\theta,b), D_j(\theta,b))\}_{b\in B,\, j\in\{1,2\}}$ specifies, for each service j, a fixed per unit price $D_j(\theta,b)$ that must be paid for the $x_j(\theta,b)$ units of the service covered by the insurer.

The agent's loss function when her true valuation is b and she picks contract item $\{(x_j(\theta,b'), D_j(\theta,b'))\}_{j\in\{1,2\}}$ is defined as follows:

$$L(b,b';\theta) = L_1(b,b';\theta) + L_2(b,b';\theta),$$

where

$$L_j(b,b';\theta) = x_j(\theta,b')\min\{D_j(\theta,b'), b_{p_j}\} + (1 - x_j(\theta,b'))b_{p_j},$$

and $b_{p_j} = \min\{b - b_j, p_j\}$.

We can then use an argument similar to the one used in the proof of Lemma 3 to characterize the contracts that satisfy the period 2 incentive compatibility constraint IC_2. In particular, a contract satisfies IC_2 if and only if:

1. The agent's loss function is given by:

$$L(b;\theta) \equiv L(b,b;\theta) = \int_0^{\min\{b,p_1\}} (1 - x_1(\theta,l)\, dl + \int_0^{\min\{\max\{b^*,b\},b^*+p_2\}} (1 - x_2(\theta,l)\, dl,$$

2. $x_j(\theta,\cdot) : B \to [0,1]$ is a non-decreasing function, with $x_j(\theta,b) = 1$ for all $b > b_j + p_j$ and $x_2(\theta,b) = 0$ for all $b \le b^*$.

Then, if risk type θ selects the menu option:

$$\left(\left\{\left(x_j(\theta',b), D_j(\theta',b)\right)\right\}_{b\in B,\, j\in\{1,2\}},\ t(\theta')\right),$$

the agent's certainty equivalent is given by:

$$U(\theta,\theta') = -t(\theta') - \sum_{j=1}^{2} \int_{b_j}^{b_j+p_j} (1 - x_j(\theta,b))[1 - g(H_\theta(b))]\, db,$$

Applying an argument similar to the one used in Lemma 4, we can show that any menu satisfying the period 1 incentive compatibility constraint IC_1 must also satisfy the envelope formula:

$$U(\theta) \equiv U(\theta,\theta) = U(\underline{\theta}) + \sum_{j=1}^{2} \int_{\underline{\theta}}^{\theta} \int_{b_j}^{b_j+p_j} (1 - x_j(s,b)) g'(H_s(b)) \frac{\partial H_s(b)}{\partial s}\, db\, ds.$$

For simplicity, we also assume that $\beta_j = 0$, so that if the insurer implements price p_j, then p_j is the price agents face if they choose not to participate in the mechanism. Using the argument from Lemma 5, we can then show that a menu satisfies the period 1 individual rationality constraint IR if and only if:

$$U(\underline{\theta}) \geq -\sum_{j=1}^{2} \int_{b_j}^{b_j+p_j} [1 - g(H_{\underline{\theta}}(b))] \, db$$

Additionally, if the supply of service j is given by a strictly increasing function $S_j : \mathbb{R}+ \to \mathbb{R}+$, then we can define the residual supply of service j faced by the insurer at price p_j as:

$$RS_j(p) = S_j(p) - \int_{\underline{\theta}}^{\bar{\theta}} [1 - H_\theta(b_j + p_j)] f(\theta) \, d\theta.$$

This residual supply accounts for the fact that, regardless of the menu offered by the insurer, all agents with a valuation $b > b_j + p_j$ will demand the full unit of service j.

Then, the equilibrium constraint faced by the insurer for service j is given by:

$$\int_{\underline{\theta}}^{\bar{\theta}} \int_{b_j}^{b_j+p_j} x_j(\theta, b) h_\theta(b) \, db f(\theta) \, d\theta = RS_j(p).$$

We can then apply an argument similar to the one used in Lemma 6 to show that the set of implementable prices for service j lies in the compact interval $\left[p_j^N, p_j^F\right]$, where, recall, p_j^N is the price that clears the market for service j when $x_j(\theta, b) = 0$ for all $b \in [b_j, b_j + p_j]$, and p_j^F is the price that clears the market when $x_j(\theta, b) = 1$ for all $b \in [b_j, b_j + p_j]$.

Additionally, to remain consistent with our assumption that $p_1 \in [0, b^*]$, we assume that $p_1^F \leq b_2 = b^*$.

Next, applying the same logic as in Lemma 7, we can show that the insurer's profits are given by:

$$\sum_{j=1}^{2} \left[\int_{\underline{\theta}}^{\bar{\theta}} x_j(\theta, b) J_j(\theta, b) h_\theta(b) \, db f(\theta) \, d\theta - \int_{b_j}^{b_j+p_j} [1 - g(H_{\underline{\theta}}(b))] \, db \right] - U(\underline{\theta}),$$

where:

$$J_j(\theta, b) = \left[1 - g(H_\theta(b)) + \frac{1 - F(\theta)}{f(\theta)} g'(H_\theta(b)) \frac{\partial H_\theta(b)}{\partial \theta} \right] \frac{1}{h_\theta(b)} + b - b_j - p_j - \frac{1 - H_\theta(b)}{h_\theta(b)}.$$

From here, it is straightforward to see that the first-stage problem can be separated by service j, with the corresponding problem given by:

$$\Pi_j(p) \equiv \max_{x_j} \quad \int_{\underline{\theta}}^{\bar{\theta}} \int_{b_j}^{b_j+p_j} x_j(\theta,b) J_j(\theta,b) h_\theta(b)\, db f(\theta)\, d\theta \tag{$P1_j$}$$

$$\text{s.t.} \quad \int_{\underline{\theta}}^{\bar{\theta}} \int_{b_j}^{b_j+p_j} x_j(\theta,b) h_\theta(b)\, db f(\theta)\, d\theta = RS_j(p), \tag{EQ_j}$$

$$x_j(\theta,\cdot) : B \to [0,1] \text{ is non-decreasing for all } \theta, \tag{MON_j}$$

where we have already used the fact that the insurer sets the utility of the lowest type $\underline{\theta}$ equal to her outside option.

It is straightforward to apply the techniques described in Sections 3.4 and 3.6 to represent problem $P1_j$ in quantity space and to rewrite it as an optimal transport problem, along with its corresponding dual formulation. Moreover, Lemma 8 allows us to characterize the solutions to this optimal transport problem. We conclude that the insurer offers each risk type θ either a simple deductible contract or a limited coverage deductible contract:

Corollary 11 (Optimal Contracts Tiered Services) *There exists an x^* that solves the problem $P1_j$ such that the contract offered to risk type θ for service j takes one of the following forms: a **simple deductible contract** $\{(1, D_j(\theta))\}$ with $D_j(\theta) \in [0, p_j]$, or a **limited coverage deductible contract** $\{(\alpha_j, D_j(\theta)), (1, M_j(\theta))\}$ where $0 \le D_j(\theta) < M_j(\theta) \le p_j$ and $\alpha_j \in (0, 1)$.*

Finally, we can also decompose the second-stage problem by service j. In particular, the problem is given by:

$$\max_{p_j \in [p_j^N, p_j^F]} \Pi_j(p),, \tag{$P2_j$}$$

where using Theorem 3 we can characterize the derivative of Π_j:

Corollary 12 (Insurer's Profit Derivative Tiered Services) *The function Π_j is differentiable in p and its derivative is given by:*

$$\Pi_j'(p) = \int_{\underline{\theta}}^{\overline{\theta}} x_j^*(\theta, b_j + p_j) J_j(\theta, b_j + p_j) h_\theta(p) f(\theta)\, d\theta - \int_{\underline{\theta}}^{\bar{\theta}} \int_{b_j}^{b_j+p_j} x_j^*(\theta,b) h_\theta(b)\, db f(\theta)\, d\theta$$
$$- \lambda^* \left(S_j'(p) + \int_{\underline{\theta}}^{\overline{\theta}} (1 - x_j^*(\theta, b_j + p_j)) h_\theta(p) f(\theta) d\theta \right).$$

6 Conclusion

This paper develops a mechanism design framework to study how insurance contracts interact with third-party service markets through their impact on service prices. We analyze a setting in which a monopolistic insurer offers contracts to risk-averse agents with dual utility preferences and sequential two-dimensional private information. Crucially, the insurer internalizes the effect of coverage on the demand for services, which in turn determines equilibrium prices. We approach this problem in two stages: first, we characterize optimal insurance menus for any given service price; second, we identify the equilibrium price that maximizes the insurer's profits.

Our first main result shows that optimal contracts take a simple and realistic form: either full coverage after a deductible or limited coverage with an out-of-pocket maximum. The limited coverage design emerges as a tool not only to manage risk but also to control service utilization and price levels. Our second main result identifies three forces that shape the insurer's pricing decision—surplus extraction, higher payouts, and equilibrium effects—offering sufficient conditions under which the insurer sets prices above or below a price-taking benchmark.

We also extend the model to settings with redistributive objectives, negotiated prices, and pricing tiers. These extensions preserve the core insights of the baseline model while broadening its applicability to both policy-oriented and institutional contexts. Overall, our results contribute to a better understanding of how insurers may strategically design contracts in environments where prices and coverage are jointly determined through market interactions.

References

Akbarpour, Mohammad, Piotr Dworczak, and Scott Duke Kominers, "Redistributive allocation mechanisms," *Journal of Political Economy*, 2024, *132* (6), 1831–1875.

American Medical Association, "Competition in Health Insurance: A Comprehensive Study of U.S. Markets — 2024 Update," Technical Report 23rd edition, American Medical Association, Division of Economic and Health Policy Research, Chicago, IL 2024.

Anderson, Gerard F, "From 'soak the rich'to 'soak the poor': recent trends in hospital pricing," *Health Affairs*, 2007, *26* (3), 780–789.

Appleby, John, Anthony Harrison, Loraine Hawkins, and Anna Dixon, *Payment by results: how can payment systems help to deliver better care?.*, King's Fund London, 2012.

Arrow, Kenneth, "Uncertainty and the Welfare Economics of Medical Care," *American Economic Review*, 1963, pp. 941–961.

Barseghyan, Levon, Francesca Molinari, and Joshua C Teitelbaum, "Inference under stability of risk preferences," *Quantitative Economics*, 2016, *7* (2), 367–409.

_ , _ , Ted O'Donoghue, and Joshua C Teitelbaum, "The nature of risk preferences: Evidence from insurance choices," *American economic review*, 2013, *103* (6), 2499–2529.

Borch, Karl, "The safety loading of reinsurance premiums," *Scandinavian Actuarial Journal*, 1960, *1960* (3-4), 163–184.

Calzolari, Giacomo and Alessandro Pavan, "Monopoly with resale," *The RAND Journal of Economics*, 2006, *37* (2), 362–375.

_ and _ , "On the optimality of privacy in sequential contracting," *Journal of Economic theory*, 2006, *130* (1), 168–204.

_ and Vincenzo Denicolo, "Exclusive contracts and market dominance," *American Economic Review*, 2015, *105* (11), 3321–3351.

Chade, Hector and Edward Schlee, "Optimal insurance with adverse selection," *Theoretical Economics*, 2012, *7* (3), 571–607.

Cook, Amanda, "Do the uninsured demand less care? Evidence from Maryland's hospitals," *International journal of health economics and management*, 2020, *20* (3), 251–276.

Courty, Pascal and Li Hao, "Sequential screening," *The Review of Economic Studies*, 2000, *67* (4), 697–717.

Dworczak, Piotr, "Mechanism design with aftermarkets: Cutoff mechanisms," *Econometrica*, 2020, *88* (6), 2629–2661.

_, Scott Duke Kominers, and Mohammad Akbarpour, "Redistribution through markets," *Econometrica*, 2021, *89* (4), 1665–1698.

Feldstein, Martin S, "The rising price of physician's services," *The Review of Economics and Statistics*, 1970, pp. 121–133.

Gershkov, Alex, Benny Moldovanu, Philipp Strack, and Mengxi Zhang, "Optimal Insurance: Dual Utility, Random Losses, and Adverse Selection," *American Economic Review*, 2023, *113* (10), 2581–2614.

Gowrisankaran, Gautam, Aviv Nevo, and Robert Town, "Mergers when prices are negotiated: Evidence from the hospital industry," *American Economic Review*, 2015, *105* (1), 172–203.

Grossman, Sanford J. and Oliver D. Hart, "An Analysis of the Principal-Agent Problem," *Econometrica*, 1983, pp. 7–45.

Ho, Kate and Robin S Lee, "Insurer competition in health care markets," *Econometrica*, 2017, *85* (2), 379–417.

Ho, Katherine, "Insurer-provider networks in the medical care market," *American Economic Review*, 2009, *99* (1), 393–430.

Kang, Zi Yang, "The public option and optimal redistribution," Technical Report, Working paper 2023.

_ and Ellen V Muir, "Contracting and Vertical Control by a Dominant Platform.," in "EC" 2022, pp. 694–695.

_ and Mitchell Watt, "Optimal Redistribution Through Subsidies," Technical Report, Working paper 2024.

Loertscher, Simon and Ellen V Muir, "Monopoly pricing, optimal randomization, and resale," *Journal of Political Economy*, 2022, *130* (3), 566–635.

Ma, Ching-To Albert and Thomas G McGuire, "Optimal health insurance and provider payment," *The American Economic Review*, 1997, pp. 685–704.

Milgrom, Paul and Ilya Segal, "Envelope theorems for arbitrary choice sets," *Econometrica*, 2002, *70* (2), 583–601.

Myerson, Roger B, "Optimal auction design," *Mathematics of operations research*, 1981, *6* (1), 58–73.

Philippon, Thomas and Vasiliki Skreta, "Optimal interventions in markets with adverse selection," *American Economic Review*, 2012, *102* (1), 1–28.

Quah, John K-H and Bruno Strulovici, "Aggregating the single crossing property," *Econometrica*, 2012, *80* (5), 2333–2348.

Raviv, Artur, "The design of an optimal insurance policy," in "Foundations of Insurance Economics: Readings in Economics and Finance," Springer, 1979, pp. 251–263.

Rothschild, Michael and Joseph Stiglitz, "Equilibrium in competitive insurance markets: An essay on the economics of imperfect information," in "Uncertainty in economics," Elsevier, 1978, pp. 257–280.

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

Sydnor, Justin, "(Over) insuring modest risks," *American Economic Journal: Applied Economics*, 2010, *2* (4), 177–199.

Tirole, Jean, "Overcoming adverse selection: How public intervention can restore market functioning," *American economic review*, 2012, *102* (1), 29–59.

Topkis, Donald M, "Minimizing a submodular function on a lattice," *Operations research*, 1978, *26* (2), 305–321.

Town, Robert and Gregory Vistnes, "Hospital competition in HMO networks," *Journal of health economics*, 2001, *20* (5), 733–753.

Valenzuela-Stookey, Quitzé and Francisco Poggi, "Market-Based Policies," Technical Report, Working Paper 2020.

Villani, Cédric et al., *Optimal transport: old and new*, Vol. 338, Springer, 2009.

Yaari, Menahem E, "The dual theory of choice under risk," *Econometrica*, 1987, pp. 95–115.

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

A Proofs of Results in Section 3.1

Proof of Lemma 1. Suppose that a contract satisfies IC_2 and that $D(\theta, b) > b_p$ for some b. In this case, the agent will not use her insurance and her loss is given by $L(b; \theta) = b_p$.

Now, suppose the insurer modifies the contract by setting $D'(\theta, b) = b_p$ and $x'(\theta, b) = 0$. Under this modified contract, the loss remains unchanged at $L'(b; \theta) = b_p$. The treatment cost for the insurer remains equal to 0, the demand for hospital services remains 1 if $b > p$ and 0 otherwise. Finally, for every $b' \neq b$, $L'(b', b; \theta) = b_p \geq L(b'; \theta)$. Therefore, IC_2 is still satisfied. ■

Proof of Lemma 2. Suppose that a contract satisfies IC_2. Towards a contradiction, assume that $L(b; \theta) > L(p; \theta)$ for some $b > p$. Then:

$$L(b, p; \theta) = x(\theta, p) \min\{D(\theta, p), p\} + (1 - x(\theta, p))p = L(p; \theta) < L(b; \theta),$$

which contradicts the assumption that IC_2 is satisfied. Therefore, we must have $L(b; \theta) \leq L(p; \theta)$. A similar argument shows that $L(b; \theta) \geq L(p; \theta)$, hence $L(p; \theta) = L(b; \theta)$ for all $b > p$.

To see why we can, with out loss of generality, set $x(\theta, b) = 1$ for all $b > p$, suppose instead that $x(\theta, b) < 1$ for some $b > p$. Consider a modified contract where $x'(\theta, b) = 1$ and $D'(\theta, b) = L(p; \theta)$ for all $b > p$. Then, $L'(b; \theta) = \min\{L(p; \theta), p\} = L(p; \theta) = L(b; \theta)$. Thus, the loss remains unchanged. Moreover, $p - D'(\theta, b) = p - L(p; \theta) = p - L(b; \theta) = x(\theta, b)(p - D(\theta, b))$, meaning the treatment cost to the insurer also remains the same. The demand for hospital services is still 1. Finally, for any $b \leq p$, we have:

$$L(b; \theta) \leq L(b, p; \theta) \leq x(\theta, p) \min\{D(\theta, p), p\} + (1 - x(\theta, p))p = L(p; \theta).$$

Additionally, we know that $L(b; \theta) \leq b$. Hence, $L(b; \theta) \leq \min\{L(p; \theta), b\}$. Since $L'(b, p; \theta) = \min\{L(p; \theta), b\}$, IC_2 is still satisfied. ■

Proof of the Necessity part of Lemma 3. Suppose that a contract satisfies IC_2. Then, for any $b, b' \in [0, p]$, we have:

$$L(b; \theta) \leq L(b, b'; \theta) = (1 - x(\theta, b'))b + x(\theta, b') \min\{D(\theta, b'), b\} \leq (1 - x(\theta, b'))b + x(\theta, b')D(\theta, b').$$

Thus:

$$L(b;\theta)=\min_{b'\in[0,p]}\quad x(\theta,b')D(\theta,b')+(1-x(\theta,b'))b.$$

Using the envelope theorem (see Milgrom and Segal, 2002 Theorem 2), it follows that:

$$L(b;\theta)=L(0;\theta)+\int_0^b(1-x(\theta,l))\,dl.$$

Furthermore, since $0\le L(b;\theta)\le b$, we have $L(0;\theta)=0$. Thus:

$$L(b;\theta)=\int_0^b(1-x(\theta,l))dl.$$

We now show Condition (2) of Lemma 3. From IC_2, for $p\ge b>b'$, we have:

$$\begin{aligned}(1-x(\theta,b))b+x(\theta,b)D(\theta,b)&\le(1-x(\theta,b'))b+x(\theta,b')\min\{D(\theta,b'),b\}\\&=(1-x(\theta,b'))b+x(\theta,b')D(b',\theta),\\(1-x(\theta,b'))b'+x(\theta,b')D(\theta,b')&\le(1-x(\theta,b))b'+x(\theta,b)\min\{D(\theta,b),b'\}\\&\le(1-x(\theta,b))b'+x(\theta,b)D(b,\theta).\end{aligned}$$

If we combine both inequalities we get:

$$(1-x(\theta,b))(b-b')\le(1-x(\theta,b'))(b-b')$$
$$x(\theta,b')\le x(\theta,b).$$

■

Proof of the Sufficiency part of Lemma 3. Suppose a contract satisfies Conditions (1) and (2) of

Lemma 3. For $b, b' \in [0, p]$ with $b' \leq b$, we have:

$$\begin{aligned}
L(b, b'; \theta) &= x(b', \theta)D(b', \theta) + (1 - x(b', \theta))b \\
&= L(b'; \theta) + (b - b')(1 - x(\theta, b')) \\
&= \int_0^{b'} (1 - x(\theta, l))\, dl + \int_{b'}^{b} (1 - x(\theta, b'))\, dl \\
&\geq \int_0^{b'} (1 - x(\theta, l))\, dl + \int_{b'}^{b} (1 - x(\theta, l))\, dl \\
&= L(b; \theta),
\end{aligned}$$

where the inequality follows from Condition (2), and the last equality from Condition (1). This shows IC_2 is satisfied for $b, b' \in [0, p]$ with $b' \leq b$.

Similarly, for $b, b' \in [0, p]$ with $b \leq b'$, $L(b, b'; \theta)$ can take two possible values. If $D(\theta, b') > b$, $L(b, b'; \theta) = b \geq L(b; \theta)$ and IC_2 is trivially satisfied. If $C(\theta, b') \leq b$, we have:

$$\begin{aligned}
L(b, b'; \theta) &= L(b'; \theta) - (b' - b)(1 - x(\theta, b')) \\
&= \int_0^{b'} (1 - x(\theta, l))\, dl - \int_{b}^{b'} (1 - x(\theta, b'))\, dl \\
&\geq \int_0^{b'} (1 - x(\theta, l))\, dl - \int_{b}^{b'} (1 - x(\theta, l))\, dl \\
&= L(b; \theta),
\end{aligned}$$

This shows IC_2 is satisfied for $b, b' \in [0, p]$ with $b \leq b'$.

Next, for $b > p$ and $b' \in [0, p]$, we have that $L(b, b'; \theta) = L(p, b'; \theta)$ and $L(p; \theta) = L(b; \theta)$. From earlier arguments we know that $L(p; \theta) \leq L(p, b'; \theta)$. Then, $L(b; \theta) \leq L(b, b'; \theta)$ and IC_2 is satisfied for $b > p$ and $b' \in [0, p]$.

Finally, for $b \in [0, p]$ and $b' > p$, we have that $x(\theta, b') = 1$ and $L(b'; \theta) = L(p, \theta)$. Thus, $L(b, b'; \theta) = \min\{L(p; \theta), b\}$. Furthermore, we know that $L(b; \theta) \leq b$ and that:

$$\begin{aligned}
L(p; \theta) &= x(\theta, p)D(\theta, p) + (1 - x(\theta, p))p \\
&\geq x(\theta, p)D(\theta, p) + (1 - x(\theta, p))b \\
&= L(b, p; \theta) \\
&\geq L(b, \theta),
\end{aligned}$$

where the last inequality follows from earlier arguments. Therefore, $L(b,;\theta) \leq \min\{L(p;\theta), b\} = L(b,b';\theta)$ and IC_2 is satisfied for $b \in [0,p]$ and $b' > p$. ■

Proof of Condition (1) of Lemma 4. Suppose that a menu satisfies IC_1. Using the envelope theorem (see Milgrom and Segal (2002) Theorem 2), the agent's certainty equivalent is given by:

$$U(\theta) = U(\underline{\theta}) + \int_{\underline{\theta}}^{\theta} \int_0^p (1 - x(s,b)) g'(H_s(b)) \frac{\partial H_s(b)}{\partial s} \, db \, ds.$$

■

Proof of Condition (2) of Lemma 4. Suppose that a menu satisfies Condition (1) of Lemma 4 and $x(\cdot,b) : B \to [0,1]$ is a non-decreasing function for every θ. Then, for any $\theta > \theta'$, we have that:

$$\begin{aligned}
U(\theta,\theta') &= -t(\theta') - \int_0^p (1 - x(\theta',b))[1 - g(H_\theta(b))] \, db \\
&= U(\theta') + \int_0^p (1 - x(\theta',b))[1 - g(H_{\theta'}(b))] \, db - \int_0^{\bar{b}} (1 - x(\theta',b))[1 - g(H_\theta(b))] \, db \\
&= U(\theta') + \int_{\theta'}^{\theta} \int_0^p (1 - x(\theta',b)) g'(H_s(b)) \frac{\partial H_s(b)}{\partial s} \, db \, ds \\
&= U(\theta) + \int_{\theta'}^{\theta} \int_0^p [x(s,b) - x(\theta',b)] g'(H_s(b)) \frac{\partial H_s(b)}{\partial s} \, db \, ds \\
&\leq U(\theta),
\end{aligned}$$

where the second equality follows from the definition of $U(\theta')$, while the fourth equality is a consequence of Condition (1). The inequality stems from the assumptions that $x(\cdot,b)$ is a non-decreasing function, g is increasing, and $\frac{\partial H_\theta(b)}{\partial \theta} \leq 0$ for all b. This last fact arises because H_θ increases in the sense of first-order stochastic dominance. An equivalent argument shows that same inequality holds when $\theta < \theta'$. ■

Proof of Lemma 5. The necessity part follows trivially. In order to show the sufficiency part, suppose that a menu satisfies IC_1 and that IR holds for type $\underline{\theta}$. Additionally, we assume that we are in the *endogenous case* (an equivalent proof applies to the *exogenous case*) such that the certainty equivalent that type θ gets from not participating is:

$$\begin{aligned}
U_{NP}(\theta) &= -\int_0^{\beta+p} [1 - g(H_\theta(b))] \, db \\
&= U_{NP}(\underline{\theta}) + \int_{\underline{\theta}}^{\theta} \int_0^{\beta+p} g'(H_s(b)) \frac{\partial H_s(b)}{\partial s} \, db \, ds.
\end{aligned}$$

Then, by Lemma 4 Condition (1) we get that:

$$U(\theta)-U_{NP}(\theta)=U(\underline{\theta})-U_{NP}(\underline{\theta})-\int_{\underline{\theta}}^{\theta}\int_{0}^{p}x(s,b)g'(H_s(b))\frac{\partial H_s(b)}{\partial s}\,db\,ds$$
$$-\int_{\underline{\theta}}^{\theta}\int_{p}^{\beta+p}g'(H_s(b))\frac{\partial H_s(b)}{\partial s}\,db\,ds.$$

By assumption $U(\underline{\theta})-U_{NP}(\underline{\theta})\geq 0$. Furthermore, $x(\theta,b)\geq 0$, g is increasing and $\frac{\partial H_\theta(b)}{\partial\theta}\leq 0$. Hence: $U(\theta)-U_{NP}(\theta)\geq 0$. ■

Proof of the Necessity part of Lemma 6. Suppose that for some x satisfying Lemma 3, Condition (2), and some $p\in\mathbb{R}_+$, we have $\mathcal{D}(x;p)=S(p)$.

Since $\mathcal{D}(x;p)\leq\int_{\underline{\theta}}^{\bar{\theta}}[1-H_\theta(0)]f(\theta)\,d\theta$, $S\left(p^F\right)=\int_{\underline{\theta}}^{\bar{\theta}}[1-H_\theta(0)]f(\theta)\,d\theta$, and $S(p)$ is strictly increasing, it immediately follows that $p\leq p^F$.

Recall that $\mathcal{D}(x;p)=S(p)$ can also be expressed as

$$\int_{\underline{\theta}}^{\bar{\theta}}\int_{0}^{p}x(\theta,b)h_\theta(b)\,db\,f(\theta)\,d\theta=RS(p).$$

The smallest achievable value on the left-hand side of the equation is zero. For the right-hand side, we know that $RS\left(p^N\right)=0$, and our technical assumptions imply that $RS(p)$ is strictly increasing. Therefore, the equality cannot hold for any $p<p^N$. Thus, $p\geq p^N$. ■

Proof of the Sufficiency part of Lemma 6. Suppose that $p\in\left(p^N,p^F\right)$, and for any $\alpha\in[0,1]$, define the following function:

$$ED(\alpha;p)=\int_{\underline{\theta}}^{\bar{\theta}}\int_{0}^{p}\alpha h_\theta(b)\,db\,f(\theta)\,d\theta-RS(p).$$

Note that $ED(\alpha;p)$ is continuous in α and satisfies $ED(0;p)<0$ and $ED(1;p)>0$. By the intermediate value theorem, there exists an $\alpha^*\in(0,1)$ such that $ED(\alpha^*;p)=0$.

Next, define $x(\theta,b)=\alpha^*$ for all $b\in[0,p]$. It follows that x satisfies IC_2, Condition (2), and EQ.

■

B Proofs of Results in Section 3.2

Proof of Lemma 7. If an agent has risk type θ and disease level b, the insurance provider pays $x(\theta,b)(p-D(\theta,b))$ to the hospital. Furthermore, we know that

$$x(\theta,b)D(\theta,b)=L(b;\theta)-(1-x(\theta,b))b,$$

and by Lemma 3, we have:

$$L(b;\theta)=\int_0^b(1-x(\theta,l))\,dl \text{ for all } b\in[0,p],\ L(p;\theta)=L(b;\theta) \text{ for all } b>p, \text{ and } x(\theta,b)=1 \text{ for all } b>p.$$

Thus, the insurer's cost of providing the contract for risk type θ simplifies as follows:

$$\begin{aligned}
\int_0^{\bar b}x(\theta,b)(D(\theta,b)-p)h_\theta(b)\,db &= \int_0^{\bar b}x(\theta,b)(b-p)h_\theta(b)\,db+\int_0^{\bar b}(L(b;\theta)-b)h_\theta(b)\,db\\
&=\int_0^{p}x(\theta,b)(b-p)h_\theta(b)\,db+\int_p^{\bar b}(b-p)h_\theta(b)\,db\\
&\quad-\int_0^p\left[\int_0^b x(\theta,l)\,dl\right]h_\theta(b)\,db+\int_p^{\bar b}(L(p;\theta)-b)h_\theta(b)\,db\\
&=\int_0^{p}x(\theta,b)(b-p)h_\theta(b)\,db-p(1-H_\theta(p))\\
&\quad-\int_0^p x(\theta,b)\,dbH_\theta(p)+\int_0^p x(\theta,b)H_\theta(b)\,db+\int_p^{\bar b}L(p;\theta)h_\theta(b)\,db\\
&=\int_0^{p}x(\theta,b)(b-p)h_\theta(b)\,db-p(1-H_\theta(p))\\
&\quad+(L(p;\theta)-p)H_\theta(p)+\int_0^p x(\theta,b)H_\theta(b)\,db+L(p;\theta)(1-H_\theta(p))\\
&=\int_0^{p}x(\theta,b)(b-p)h_\theta(b)\,db+\int_0^p x(\theta,b)H_\theta(b)\,db+L(p;\theta)-p\\
&=\int_0^{p}x(\theta,b)(b-p)h_\theta(b)\,db+\int_0^p x(\theta,b)H_\theta(b)db-\int_0^p x(\theta,b)\,db\\
&=\int_0^{p}x(\theta,b)\left[b-p-\frac{1-H_\theta(b)}{h_\theta(b)}\right]h_\theta(b)\,db,
\end{aligned}$$

where in the third equality we used integration by parts to express the term $\int_0^p\left[\int_0^b x(\theta,l)\,dl\right]h_\theta(b)\,db$ as:

$$\int_0^p x(\theta,b)\,dbH_\theta(p)-\int_0^p x(\theta,b)H_\theta(b)\,db.$$

Using Lemma 4, $t(\theta)$ can be expressed as:

$$
\begin{aligned}
t(\theta) =& -U(\underline{\theta}) - \int_0^p (1-x(\theta,b))[1-g(H_\theta(b))]\,db - \int_{\underline{\theta}}^{\theta}\int_0^p (1-x(s,b))g'(H_s(b))\frac{\partial H_s(b)}{\partial s}\,db\,ds \\
=& -U(\underline{\theta}) - \int_0^p [1-g(H_{\underline{\theta}}(b))]\,db + \int_0^p x(\theta,b)[1-g(H_\theta(b))]\,db \\
&+ \int_{\underline{\theta}}^{\theta}\int_0^p x(s,b)g'(H_s(b))\frac{\partial H_s(b)}{\partial s}\,db\,ds.
\end{aligned}
$$

Integration by parts shows:

$$
\int_{\underline{\theta}}^{\bar{\theta}}\int_0^p\int_{\underline{\theta}}^{\theta} x(s,b)g'(H_s(b))\frac{\partial H_s(b)}{\partial s}\,ds\,db f(\theta)\,d\theta = \int_{\underline{\theta}}^{\bar{\theta}}\int_0^p x(\theta,b)\frac{1-F(\theta)}{f(\theta)}g'(H_\theta(b))\frac{\partial H_\theta(b)}{\partial s}\,db f(\theta)\,d\theta.
$$

Thus, we have that:

$$
\begin{aligned}
\int_{\underline{\theta}}^{\bar{\theta}} t(\theta)f(\theta)\,d\theta =& -U(\underline{\theta}) - \int_0^p [1-g(H_{\underline{\theta}}(b))]\,db \\
&+ \int_{\underline{\theta}}^{\bar{\theta}}\int_0^p x(\theta,b)\left[1-g(H_\theta(b)) + \frac{1-F(\theta)}{f(\theta)}g'(H_\theta(b))\frac{\partial H_\theta(b)}{\partial b}\right]\,db f(\theta)\,d\theta.
\end{aligned}
$$

Then, the insurer's profits from offering a menu:

$$
\int_{\underline{\theta}}^{\bar{\theta}}\left[t(\theta) - \int_0^{\bar{b}} x(\theta,b)(p-D(\theta,b))\,dH_\theta(b)\right]f(\theta)\,d\theta,
$$

simplify to:

$$
\int_{\underline{\theta}}^{\bar{\theta}}\int_0^p x(\theta,b)J(\theta,b)h_\theta(b)\,db f(\theta)\,d\theta - U(\underline{\theta}) - \int_0^p [1-g(H_{\underline{\theta}}(b))]\,db,
$$

where:

$$
J(\theta,b) = \left[1-g(H_\theta(b)) + \frac{1-F(\theta)}{f(\theta)}g'(H_\theta(b))\frac{\partial H_\theta(b)}{\partial \theta}\right]\frac{1}{h_\theta(b)} + b - p - \frac{1-H_\theta(b)}{h_\theta(b)}.
$$

■

C Proofs of Results in Section 3.6

Proof of Lemma 8. Let π' be any solution to the primal problem OT, whose existence is guaranteed by Theorem 2. Define q_θ as the expected value of q under the conditional distribution $\pi'(\cdot \mid \theta)$, i.e., $q_\theta = \int_0^1 q \, d\pi^*(q \mid \theta)$.

Since q_θ lies in the convex hull of $\operatorname{supp} \pi'(\cdot \mid \theta) \subset \mathbb{R}$, Carathéodory's theorem ensures the existence of a distribution $\pi''(\cdot \mid \theta)$ such that $\operatorname{supp} \pi''(\cdot \mid \theta)$ contains at most two elements, satisfies $\operatorname{supp} \pi''(\cdot \mid \theta) \subseteq \operatorname{supp} \pi'(\cdot \mid \theta)$, and $q_\theta = \int_0^1 q \, d\pi''(q \mid \theta)$.

Define V as the set of risk types for which $\operatorname{supp} \pi''(\cdot \mid \theta)$ contains exactly two elements. That is, for any $\theta \in V$, we have $\operatorname{supp} \pi''(\cdot \mid \theta) = \{q_{1,\theta}, q_{2,\theta}\}$, where, without loss of generality, $q_{1,\theta} < q_{2,\theta}$, and:

$$q_\theta = q_{1,\theta}\pi''(q_{1,\theta} \mid \theta) + q_{2,\theta}(1 - \pi''(q_{1,\theta} \mid \theta)).$$

Next, define:

$$q_V = \int_V [H_\theta(p) - H_\theta(0)] q_\theta f(\theta) \, d\theta.$$

Since:

$$q_V \in \left(\int_V [H_\theta(p) - H_\theta(0)] q_{1,\theta} f(\theta) \, d\theta, \int_V [H_\theta(p) - H_\theta(0)] q_{2,\theta} f(\theta) \, d\theta \right),$$

there exists some $\alpha \in (0,1)$ such that:

$$q_V = \alpha \int_V [H_\theta(p) - H_\theta(0)] q_{1,\theta} f(\theta) \, d\theta + (1-\alpha) \int_V [H_\theta(p) - H_\theta(0)] q_{2,\theta} f(\theta) \, d\theta.$$

Define a new distribution π^* as follows: For all $\theta \in V$, set:

$$\pi^*(\theta, q_{1,\theta}) = \alpha f(\theta), \quad \pi^*(\theta, q_{2,\theta}) = (1-\alpha) f(\theta),$$

and $\pi^*(\theta, q) = 0$ for all other q. For all $\theta \in V^c$, set

$$\pi^*(\theta, q_\theta) = f(\theta),$$

and $\pi^*(\theta, q) = 0$ for all other q.

By construction, π^* satisfies condition BP. Moreover,

$$\begin{aligned}\int_{\underline{\theta}}^{\bar{\theta}} \int_0^1 [H_\theta(p) - H_\theta(0)] q \, d\pi^*(\theta, q) &= q_V + \int_{V^c} [H_\theta(p) - H_\theta(0)] q_\theta f(\theta) \, d\theta \\ &= \int_{\underline{\theta}}^{\bar{\theta}} \int_0^1 [H_\theta(p) - H_\theta(0)] q_\theta f(\theta) \, d\theta \\ &= \int_{\underline{\theta}}^{\bar{\theta}} \int_0^1 [H_\theta(p) - H_\theta(0)] q \, d\pi'(\theta, q) \\ &= RS(p).\end{aligned}$$

Thus, π^* also satisfies condition EQ.

Finally, by Theorem 2, there exists (v^*, λ^*) solving the dual problem D. By Corollary 1, for f-almost every θ, the distribution $\pi'(\cdot \mid \theta)$ is supported on $Supp$. Since $\operatorname{supp} \pi^*(\cdot \mid \theta) \subseteq \operatorname{supp} \pi'(\cdot \mid \theta)$, it follows that f-almost every θ also has $\pi^*(\cdot \mid \theta)$ supported $Supp$. Using Corollary 1 again, we conclude that π^* is a solution to the primal problem. ■

Proof of Proposition 1. Suppose that:

$$J(\theta, b) h_\theta(b) + \lambda^* h_\theta(b),$$

is non-decreasing in θ for all b, and let π^* denote the solution to the primal problem OT.

By Corollary 1, the measure $\pi^*(\cdot \mid \theta)$ is supported on $Q(\theta)$ ($Supp$). Define:

$$G(\theta, q) = [H_\theta(p) - H_\theta(0)] \int_0^q J\big(\theta, b_\theta(s)\big) ds + \lambda^* [H_\theta(p) - H_\theta(0)] q,$$

and see that $Q(\theta) = \arg\max_{q \in [0,1]} G(\theta, q)$.

Now perform the change of variable $b = b_\theta(q)$, we can equivalently write

$$G(\theta, q) = F(\theta, b) = \int_b^p J(\theta, l) h_\theta(l) \, dl + \lambda^* [H_\theta(p) - H_\theta(b)].$$

Define $B(\theta) = \arg\max_{b \in [0,p]} F(\theta, b)$. Observe that if $s \in Q(\theta)$, then $b_\theta(s) \in B(\theta)$; conversely, if $l \in B(\theta)$, then $b_\theta^{-1}(l) \in Q(\theta)$.

If $B(\theta)$ decreases in θ in the strong set order, then for any $\theta_1 > \theta_2$, we have that:

$$\min_{q \in Q(\theta_1)} b_{\theta_1}(q) \le \min_{q \in Q(\theta_2)} b_{\theta_2}(q) \quad \text{and} \quad \max_{q \in Q(\theta_1)} b_{\theta_1}(q) \le \max_{q \in Q(\theta_2)} b_{\theta_2}(q).$$

Then, for the π^* constructed in the proof of Lemma 8 we can always guarantee that:

$$\min_{q \in \text{supp } \pi(\cdot|\theta_1)} b_{\theta_1}(q) \leq \min_{q \in \text{supp } \pi(\cdot|\theta_2)} b_{\theta_2}(q) \quad \text{and} \quad \max_{q \in \text{supp } \pi(\cdot|\theta_1)} b_{\theta_1}(q) \leq \max_{q \in \text{supp } \pi(\cdot|\theta_2)} b_{\theta_2}(q).$$

This in turn implies that the allocation x^* induced by π^* is non-decreasing in θ for all b. Indeed, x^* exhibits its first (and, when applicable, second) jump earlier for θ_1 than for θ_2. Moreover, because the magnitude of these jumps is type-independent, x^* is increasing in θ for all b.

By Topkis (1978)'s monotonicity theorem, $B(\theta)$ decreases in θ in the strong set order whenever $F(\theta, b)$ is submodular in (θ, b). Submodularity of F is equivalent to the cross-partial $\frac{\partial F(\theta,b)}{\partial b}$ being non-increasing in θ, or equivalently, to $-\frac{\partial F(\theta,b)}{\partial b}$ being non-decreasing in θ. Note that

$$-\frac{\partial F(\theta, b)}{\partial b} = J(\theta, b)h_\theta(b) + \lambda^* h_\theta(b),$$

which is non-decreasing in θ by assumption. Hence F is submodular, $B(\theta)$ decreases in θ in the strong set order, and the resulting allocation x^* is non-decreasing in θ for all b, completing the proof. ■

Proof of Corollary 2. We first show that:

$$|\lambda^*| \leq \sup_{\theta \in \Theta,\, b \in [0, \bar{p}]} |J(\theta, b)| \;\equiv\; U.$$

If there exists a type θ such that $q(\theta) \in (0, 1)$, the first-order condition implies:

$$J\big(\theta, b_\theta(q(\theta))\big) = \lambda^*,$$

which immediately yields $|\lambda^*| \leq U$.

If, instead, there exist types θ_R and θ_L such that $q_R = 1$ and $q_L = 0$, respectively, then the first-order conditions imply:

$$\lambda^* \geq -J\big(\theta_R, b_{\theta_R}(1)\big) \geq -U, \qquad \lambda^* \leq -J\big(\theta_L, b_{\theta_L}(0)\big) \leq U.$$

Hence $-U \leq \lambda^* \leq U$, and again $|\lambda^*| \leq U$.

If $q(\theta) = 1$ for all θ, then setting $\lambda^* = -U$ satisfies the first-order conditions, and the same argument applies symmetrically if $q(\theta) = 0$ for all θ.

Finally, given $|\lambda^*| \leq U$, we have

$$\lambda^* \frac{\partial h_\theta(b)}{\partial \theta} \;\geq\; -\left| \lambda^* \frac{\partial h_\theta(b)}{\partial \theta} \right| \;\geq\; -U \left| \frac{\partial h_\theta(b)}{\partial \theta} \right|,$$

which leads directly to the desired inequality:

$$\frac{\partial \big(J(\theta, b) h_\theta(b)\big)}{\partial \theta} - U \left| \frac{\partial h_\theta(b)}{\partial \theta} \right| \geq 0,$$

completing the proof. ■

Proof of Proposition 2. By the proof of Proposition 1, the desired result follows if the correspondence

$$B(\theta) = \arg\max_{b \in [0,p]} F(\theta, b)$$

is decreasing in θ in the strong set order. Recall that:

$$F(\theta, b) = \int_b^p V(\theta, l)\, dl + \lambda^* \left[H_\theta(p) - H_\theta(b)\right],$$

so that:

$$\frac{\partial F(\theta, b)}{\partial \theta} = -V(\theta, b) - \lambda^* h_\theta(b).$$

By assumption, $V(\cdot, b)$ is non-decreasing in θ for all b, $h_\theta(b)$ is non-increasing in θ for all b, and for any $\theta_1 > \theta_2$ there exists a positive, non-decreasing function β such that:

$$\beta(b) V(\theta_1, b) \geq V(\theta_2, b) \quad \text{and} \quad \beta(b) h_{\theta_1}(b) \geq h_{\theta_2}(b) \quad \text{for all } b.$$

First, consider the case $\lambda^* \leq 0$. In this case, $\frac{\partial F(\theta,b)}{\partial \theta}$ is non-increasing in θ, implying that $F(\theta, b)$ is submodular. By Topkis (1978)'s monotonicity theorem, $B(\theta)$ is then decreasing in θ in the strong set order.

Next, consider the case $\lambda^* > 0$. For any $\theta_1 > \theta_2$, we have

$$-\beta(b) V(\theta_1, b) \leq -V(\theta_2, b) \quad \text{and} \quad -\lambda^* \beta(b) h_{\theta_1}(b) \leq -\lambda^* h_{\theta_2}(b) \quad \text{for all } b,$$

which together imply that

$$\beta(b)\frac{\partial F(\theta_1, b)}{\partial b} \leq \frac{\partial F(\theta_2, b)}{\partial b} \quad \text{for all } b.$$

By Proposition 2 of Quah and Strulovici (2012), $F(\theta_2, \cdot)$ *interval-order dominates* $F(\theta_1, \cdot)$, and by Theorem 1 therein, $B(\theta_2) \geq B(\theta_1)$ in the strong set order.

An analogous argument applies when $h_\theta(b)$ is non-decreasing in θ for all b and $\beta(b)h_{\theta_1}(b) \leq h_{\theta_2}(b)$. ■

Proof of Proposition 3. Suppose that $\frac{\partial \phi(\theta,q)}{\partial \theta} \geq 0$ for all θ and q.

By Corollary 1, the measure $\pi^*(\cdot \mid \theta)$ is supported on $Q(\theta)$ (*Supp*). Multiplying the objective on *Supp* by the positive constant $\frac{1}{H_\theta(p)-H_\theta(0)}$—which is independent of q—does not alter the set of maximizers. Hence, $\pi^*(\cdot \mid \theta)$ is also supported on

$$Q(\theta) \;=\; \arg\max_{q\in[0,1]} \left\{ \frac{\Phi(\theta, q)}{H_\theta(p) - H_\theta(0)} + \lambda^* \left(q - \frac{RS(p)}{H_\theta(p) - H_\theta(0)} \right) \right\}.$$

The derivative of the objective with respect to q is

$$\phi(\theta, q) + \lambda^*,$$

and the derivative of $\phi(\theta, q)$ with respect to θ is

$$\frac{\partial \phi(\theta, q)}{\partial \theta} \geq 0.$$

Hence, the objective function is supermodular in (θ, q). By Topkis (1978)'s monotonicity theorem, $Q(\theta)$ increases in θ in the strong set order.

Now let $\theta_1 > \theta_2$, using the fact that Q increases θ in the strong set order, $b_\theta(q)$ is strictly decreasing in q and, by assumption, non-increasing in θ, it follows that:

$$\min_{q\in Q(\theta_1)} b_{\theta_1}(q) \leq \min_{q\in Q(\theta_2)} b_{\theta_2}(q) \quad \text{and} \quad \max_{q\in Q(\theta_1)} b_{\theta_1}(q) \leq \max_{q\in Q(\theta_2)} b_{\theta_2}(q).$$

Then, for the π^* constructed in the proof of Lemma 8 we can always guarantee that:

$$\min_{q\in \text{supp } \pi(\cdot|\theta_1)} b_{\theta_1}(q) \leq \min_{q\in \text{supp } \pi(\cdot|\theta_2)} b_{\theta_2}(q) \quad \text{and} \quad \max_{q\in \text{supp } \pi(\cdot|\theta_1)} b_{\theta_1}(q) \leq \max_{q\in \text{supp } \pi(\cdot|\theta_2)} b_{\theta_2}(q).$$

This in turn implies that the allocation x^* induced by π^* is non-decreasing in θ for all b. Indeed, x^* exhibits its first (and, when applicable, second) jump earlier for θ_1 than for θ_2. Moreover, because the magnitude of these jumps is type-independent, x^* is increasing in θ for all b.

■

Proof of Proposition 4. Fix a risk type θ and suppose that $\frac{\partial \phi(\theta,q)}{\partial q} \leq 0$ for all q. This assumption implies that the objective in *Supp* is concave in q.

Let π be any solution to the primal problem *OT*, whose existence is guaranteed by Theorem 2. By Corollary 1, we know that $\pi(\cdot \mid \theta)$ is supported on *Supp*. Moreover, by the concavity of the objective in *Supp*, it follows that the average q induced by $\pi(\cdot \mid \theta)$ also belongs to *Supp*. Consequently, for the measure π^* constructed in the proof of Lemma 8, the support of $\pi^*(\cdot \mid \theta)$ contains only a single element. ■

D Proof of Results in Section 4

Proof of Theorem 3. To prove the result, we will show that our setting satisfies all the assumptions needed to apply Theorem 4 of Milgrom and Segal (2002). For this purpose, let $\mathcal{M}^+(\Theta \times [0,1])$ denote the set of positive measures over $\Theta \times [0,1]$, and let $L^1(f)$ denote the set of all measurable functions $v : \Theta \to \mathbb{R}$ that satisfy:

$$\int_{\underline{\theta}}^{\bar{\theta}} |v(\theta)| f(\theta)\, d\theta < \infty.$$

Then, define the function $\mathcal{L} : \mathcal{M}^+(\Theta \times [0,1]) \times (L^1(f) \times \mathbb{R}) \times [p^N, p^F] \to \mathbb{R}$ by:

$$\mathcal{L}(\pi, (v, \lambda), p) = \int_{\underline{\theta}}^{\bar{\theta}} \int_0^1 \Phi(\theta, q)\, d\pi(\theta, q) + \int_{\underline{\theta}}^{\bar{\theta}} v(\theta) \Big[f(\theta)\, d\theta - \int_0^1 d\pi(\theta, q) \Big] + \lambda\, ED(\pi, p),$$

where:

$$ED(\pi, p) = \int_{\underline{\theta}}^{\bar{\theta}} \int_0^1 [H_\theta(p) - H_\theta(0)]\, q\, d\pi(\theta, q) - RS(p).$$

Fix any $p \in [p^N, p^F]$ and let π^* be a solution to the primal problem *OT* and (v^*, λ^*) be a solution to the dual problem *D*. By Lemma 10, $(\pi^*, (v^*, \lambda^*))$ is a saddle point of $\mathcal{L}$:

$$\mathcal{L}(\pi, (v^*, \lambda^*), p) \leq \mathcal{L}(\pi^*, (v^*, \lambda^*), p) = \Pi(p) \leq \mathcal{L}(\pi^*, (v, \lambda), p),$$

for all $\pi \in \mathcal{M}^+(\Theta \times [0,1])$ and $(v, \lambda) \in L^1(f) \times \mathbb{R}$.

It is easy to show that there exist finite positive measures $\overline{\pi}$ and $\underline{\pi}$ such that $ED(\overline{\pi}, p) > 0$ and $ED(\underline{\pi}, p) < 0$ for all $p \in [p^N, p^F]$. Since $(\pi^*, (v^*, \lambda^*))$ is a saddle point, it follows that:

$$\Pi(p) \geq \mathcal{L}(\overline{\pi}, (v^*, \lambda^*), p) = \int_{\underline{\theta}}^{\bar{\theta}} \int_0^1 \Phi(\theta, q)\, d\overline{\pi}(\theta, q) + \int_{\underline{\theta}}^{\bar{\theta}} v(\theta) \Big[f(\theta)\, d\theta - \int_0^1 d\overline{\pi}(\theta, q) \Big] + \lambda^*\, ED(\overline{\pi}, p).$$

Hence:

$$\lambda^* \leq \overline{\lambda} = \sup_{p \in [p^N, p^F]} \frac{\Pi(p) - \int_{\underline{\theta}}^{\bar{\theta}} \int_0^1 \Phi(\theta, q)\, d\overline{\pi}(\theta, q) - \int_{\underline{\theta}}^{\bar{\theta}} v^*(\theta) \Big[f(\theta)\, d\theta - \int_0^1 d\overline{\pi}(\theta, q) \Big]}{ED(\overline{\pi}, p)}.$$

Since the numerator is bounded and the denominator is bounded away from zero by the definition of $\bar{\pi}$ and continuity of $ED(\overline{\pi}, \cdot)$, we conclude that λ^* is bounded above by $\overline{\lambda} < \infty$. A similar argument with $\underline{\pi}$ shows that λ^* is bounded below by some $\underline{\lambda} > -\infty$. Thus, we can assume $\lambda \in [\underline{\lambda}, \overline{\lambda}]$ and $|\lambda| \leq M$ for some $M < \infty$.

Additionally, since $\pi^* \in \Delta(\Theta \times [0,1])$, $(\pi^*, (v^*, \lambda^*))$ is a saddle point even when we restrict to probability measures. Therefore, from now on we restrict to $\pi \in \Delta(\Theta \times [0,1])$.

Next, by our technical assumptions, $\mathcal{L}(\pi, (v, \lambda), \cdot)$ is differentiable with derivative:

$$\frac{\partial \mathcal{L}(\pi, (v, \lambda), p)}{\partial p} = \mathcal{L}_p(\pi, (v, \lambda), p) = \int_{\underline{\theta}}^{\bar{\theta}} \int_0^1 \frac{\partial \Phi(\theta, q)}{\partial p}\, d\pi(\theta, q) - \lambda \Big(S'(p) + \int_{\underline{\theta}}^{\bar{\theta}} \int_0^1 h_\theta(p)(1-q)\, d\pi(\theta, q) \Big).$$

Recall:

$$\Phi(\theta, q) = [H_\theta(p) - H_\theta(0)] \int_0^q \phi(\theta, s)\, ds,$$

where $\phi(\theta, q) = J(\theta, b_\theta(q))$ and:

$$b_\theta(q) = H_\theta^{-1}\Big((1-q)[H_\theta(p) - H_\theta(0)] + H_\theta(0) \Big).$$

A careful calculation yields:

$$
\begin{aligned}
\frac{\partial \Phi(\theta,q)}{\partial p} &= h_\theta(p)\int_0^q J(\theta,b_\theta(s))\,ds + [H_\theta(p)-H_\theta(0)]\int_0^q \frac{\partial J(\theta,b_\theta(s))}{\partial b}\frac{\partial b_\theta(s)}{\partial p}\,ds - [H_\theta(p)-H_\theta(0)]q \\
&= h_\theta(p)\int_0^q J(\theta,b_\theta(s))\,ds - h_\theta(p)\int_0^q (1-s)\frac{\partial J(\theta,b_\theta(s))}{\partial b}\frac{\partial b_\theta(s)}{\partial s}\,ds - [H_\theta(p)-H_\theta(0)]q \\
&= h_\theta(p)[J(\theta,p)-(1-q)J(\theta,q)] - [H_\theta(p)-H_\theta(0)]q.
\end{aligned}
$$

Then:

$$
\mathcal{L}_p(\pi,(v,\lambda),p) = \int_{\underline{\theta}}^{\bar{\theta}}\int_0^1 T(\theta,q,p)\,d\pi(\theta,q) - \lambda\int_{\underline{\theta}}^{\bar{\theta}}\int_0^1 Y(\theta,q,p)\,d\pi(\theta,q),
$$

where:

$$
T(\theta,q,p) = h_\theta(p)\Big[J(\theta,p)-(1-q)J(\theta,b_\theta(q))\Big] - [H_\theta(p)-H_\theta(0)]q,
$$

$$
Y(\theta,q,p) = S'(p) + h_\theta(p)(1-q).
$$

By our technical assumptions, T and Y are continuous functions. Moreover, since both functions are defined on the compact domain $\Theta\times[0,1]\times\left[p^N,p^F\right]$, they are uniformly continuous. Hence, for any $\varepsilon>0$, there exists $\delta>0$ such that:

$$
\int_{\underline{\theta}}^{\bar{\theta}}\int_0^1 |T(\theta,q,p)-T(\theta,q,p')|\,d\pi(\theta,q) \le \sup_{(\theta,q)\in\Theta\times[0,1]} |T(\theta,q,p)-T(\theta,q,p')| < \varepsilon,
$$

$$
|\lambda|\int_{\underline{\theta}}^{\bar{\theta}}\int_0^1 |Y(\theta,q,p)-Y(\theta,q,p')|\,d\pi(\theta,q) \le \sup_{(\theta,q)\in\Theta\times[0,1]} M|Y(\theta,q,p)-Y(\theta,q,p')|) < \varepsilon,
$$

for all $|p-p'|<\delta$. Therefore, $\mathcal{L}_p(\pi,(v,\lambda),\cdot)$ is uniformly continuous and the family $\{\mathcal{L}_p(\pi,(v,\lambda),\cdot)\}_{(\pi,(v,\lambda))}$ is uniformly equicontinuous. Hence, $\{\mathcal{L}(\pi,(v,\lambda),\cdot)\}_{(\pi,(v,\lambda))}$ is equidifferentiable.

Since $\Delta(\Theta\times[0,1])$ and $L^1(f)\times[\underline{\lambda},\overline{\lambda}]$ are second-countable topological spaces, all assumptions of Theorem 4 of Milgrom and Segal (2002) are satisfied. Thus:

$$
\frac{\partial \Pi(p)}{\partial p} = \Pi'(p) = \mathcal{L}_p(\pi^*,(v^*,\lambda^*),p).
$$

To conclude the proof, we simplify $\mathcal{L}_p$ and express it in terms of x^*. Since π^* satisfies BP, $\pi^*(\cdot\mid\theta)\equiv\frac{\pi^*(\theta,\cdot)}{f(\theta)}:[0,1]\to[0,1]$ is a well-defined cumulative distribution function. Moreover, by

Corollary 1, any $q \in (0,1)$ in the support of $\pi^*(\cdot \mid \theta)$ must satisfy the first-order condition:

$$J(\theta, b_\theta(q)) + \lambda^* = 0.$$

Then:

$$\mathcal{L}_p(\pi^*, (v^*, \lambda^*), p) = \int_{\underline{\theta}}^{\bar{\theta}} \left[(1 - \pi^*(0 \mid \theta)) J(\theta, p) h_\theta(p) - [H_\theta(p) - H_\theta(0)] \int_0^1 (1 - \pi^*(q \mid \theta))\, dq \right] f(\theta)\, d\theta$$
$$- \lambda^* \Big(S'(p) + \int_{\underline{\theta}}^{\bar{\theta}} h_\theta(p) \pi^*(0 \mid \theta) f(\theta)\, d\theta \Big).$$

Recall:

$$x^*(\theta, b_\theta(q)) = 1 - \pi^*(q \mid \theta).$$

Then, changing variables $b = b_\theta(q)$, we get:

$$\mathcal{L}_p(\pi^*, (v^*, \lambda^*), p) = \int_{\underline{\theta}}^{\bar{\theta}} x^*(\theta, p) J(\theta, p) h_\theta(p) f(\theta)\, d\theta - \int_{\underline{\theta}}^{\bar{\theta}} \int_0^p x^*(\theta, b) h_\theta(b)\, db\, f(\theta)\, d\theta$$
$$- \lambda^* \Big(S'(p) + \int_{\underline{\theta}}^{\bar{\theta}} h_\theta(p) [1 - x^*(\theta, p)] f(\theta)\, d\theta \Big).$$

■

Proof of Corollary 5. Let $\beta > \beta'$ and p_β be a solution to the second-stage problem $P2$ under β, and let $p_{\beta'}$ denote the highest price that solves $P2$ under β'. By Corollary 4, p_β exists. Moreover, since the set of maximizers in the second-stage problem $P2$ is nonempty and compact (also by Corollary 4), $p_{\beta'}$ is well defined.

Suppose, for the sake of contradiction, that $p_\beta > p_{\beta'}$. Let Π_β and $\Pi_{\beta'}$ denote the insurer's profit functions under β and β', respectively. Then,

$$\Pi_\beta(p_\beta) - \Pi_\beta(p_{\beta'}) = \int_{p_{\beta'}}^{p_\beta} \Pi'_\beta(p)\, dp \geq 0,$$

since p_β is a maximizer of Π_β.

However, by Theorem 3, we know that $\Pi'_\beta(p) \leq \Pi'_{\beta'}(p)$ for all p. This inequality holds because the only difference between the two derivatives is:

$$-g(H_{\underline{\theta}}(p + \beta) \leq -g(H_{\underline{\theta}}(p + \beta').$$

Therefore, we obtain:

$$\Pi_{\beta'}(p_\beta) - \Pi_{\beta'}(p_{\beta'}) = \int_{p_{\beta'}}^{p_\beta} \Pi'_{\beta'}(p)\, dp \geq \int_{p_{\beta'}}^{p_\beta} \Pi'_{\beta}(p)\, dp \geq 0,$$

which contradicts the fact that $p_{\beta'}$ is the highest maximizer of $\Pi_{\beta'}$. ■

Proof of Proposition 5. We begin by recalling key properties that any optimal multiplier $\lambda^*(p)$ and optimal deductible $D(\theta; p)$ must satisfy. Define:

$$\hat{J}(\theta, b) = J(\theta, b) + p,$$

which is independent of p; since $J(\theta, \cdot)$ is strictly increasing, so is $\hat{J}(\theta, \cdot)$.

Let:

$$q(\theta; p) = 1 - H_\theta(D(\theta; p) \mid p).$$

By Complementary Slackness (Corollary 1), every optimal $q(\theta; p)$ lies in the set *Supp* and must satisfy:

- **Interior case** ($q(\theta; p) \in (0, 1)$, i.e., $D(\theta; p) \in (0, p)$):

$$\phi(\theta, q(\theta; p)) = J(\theta, b_\theta(q(\theta; p))) = \hat{J}(\theta, D(\theta; p)) - p = -\lambda^*(p).$$

- **Upper boundary** ($q(\theta; p) = 0$, i.e., $D(\theta; p) = p$):

$$\hat{J}(\theta, p) \leq -\lambda^*(p) + p.$$

- **Lower boundary** ($q(\theta; p) = 1$, i.e., $D(\theta; p) = 0$):

$$\hat{J}(\theta, 0) \geq -\lambda^*(p) + p.$$

Since $\hat{J}(\theta, \cdot)$ is strictly increasing and continuous, it is invertible. Therefore, we can define the function:

$$D\left(\theta; \left(p, \hat{\lambda}\right)\right) = \min\left\{p, \max\left\{0, \hat{J}^{-1}\left(\theta, -\hat{\lambda}\right)\right\}\right\}.$$

Then, from the first-order conditions stated above, it is easy to see that the optimal deductible for risk type θ is given by:

$$D(\theta;p) = D\left(\theta;\left(p,\hat{\lambda}^*(p)\right)\right),$$

where $\hat{\lambda}^*(p) = \lambda^*(p) - p$.

Step 1: Monotonicity of the adjusted multiplier. We will prove that $\hat{\lambda}^*(\cdot)$ is strictly increasing. Assume, for contradiction, that $\hat{\lambda}^*(p') \leq \hat{\lambda}^*(p)$ for some $p' > p$.

1. **Deductibles must be weakly larger at the higher price.** This immediately follows from the fact that $D(\theta;(p,\cdot))$ is non-increasing, since the inverse function $\hat{J}^{-1}(\theta,\cdot)$ is strictly increasing, and from the fact that $D\left(\theta;\left(\cdot,\hat{\lambda}\right)\right)$ is non-decreasing. Thus:

$$D(\theta;p) = D\left(\theta;\left(p,\hat{\lambda}^*(p)\right)\right) \leq D\left(\theta;\left(p',\hat{\lambda}^*(p')\right)\right) = D(\theta;p').$$

2. **Demand–supply contradiction.**

   Since $D(\theta;p') \geq D(\theta;p)$, we have:

$$\mathcal{D}(p') = \int_{\underline{\theta}}^{\bar{\theta}} \left[1 - H_\theta(D(\theta;p'))\right] f(\theta)\,d\theta \leq \int_{\underline{\theta}}^{\bar{\theta}} \left[1 - H_\theta(D(\theta;p))\right] f(\theta)\,d\theta = \mathcal{D}(p).$$

   But supply is strictly increasing, so $S(p') > S(p)$. Since the market clears at p (i.e., $\mathcal{D}(p) = S(p)$), we get:

$$S(p') > \mathcal{D}(p'),$$

   which contradicts market clearing.

Thus, $\hat{\lambda}^*(p') > \hat{\lambda}^*(p)$. Since $\lambda^*(p') = \hat{\lambda}^*(p') + p'$, we conclude:

$$\lambda^*(p') - \lambda^*(p) = \hat{\lambda}^*(p') - \hat{\lambda}^*(p) + (p' - p) > p' - p > 0,$$

so $\lambda^*(\cdot)$ is also strictly increasing.

Step 2: Monotonicity of the deductible.

- **Interior case** ($0 < D(\theta;p) < p$): In this case:

$$D(\theta;p) = D\left(\theta;\left(p,\hat{\lambda}^*(p)\right)\right) = \hat{J}^{-1}\left(\theta, -\hat{\lambda}^*(p)\right).$$

Since $\hat{J}^{-1}(\theta, \cdot)$ is strictly increasing, it follows that:

$$\hat{J}^{-1}\left(\theta, -\hat{\lambda}^*(p')\right) < \hat{J}^{-1}\left(\theta, -\hat{\lambda}^*(p)\right),$$

which leads to:

$$D(\theta; p') = D\left(\theta; \left(p, \hat{\lambda}^*(p')\right)\right) = \hat{J}^{-1}\left(\theta, -\hat{\lambda}^*(p')\right) < D(\theta; p).$$

- **Boundary case** ($D(\theta; p) = 0$): In this case:

$$\hat{J}^{-1}\left(\theta, -\hat{\lambda}^*(p')\right) < \hat{J}^{-1}\left(\theta, -\hat{\lambda}^*(p)\right) \leq 0,$$

which allows us to conclude that $D(\theta; p') = 0$.

Hence, $D(\theta; \cdot)$ is strictly decreasing whenever $D(\theta; p) \in (0, p)$, and weakly decreasing when $D(\theta; p) = 0$.

Step 3: Uniqueness of the multiplier. Define the following demand function, which depends not only on p but also on $\hat{\lambda}$:

$$\mathcal{D}\left(p, \hat{\lambda}\right) = \int_{\underline{\theta}}^{\bar{\theta}} \left[1 - H_\theta\left(D\left(\theta; (p, \hat{\lambda})\right)\right)\right] f(\theta)\, d\theta.$$

Since $H_\theta(\cdot)$ is strictly increasing, it follows that $\mathcal{D}(p, \cdot)$ is non-decreasing. Moreover, whenever $\hat{\lambda} > \hat{\lambda}'$ and there exists a set of types with positive measure such that $D\left(\theta; (p, \hat{\lambda})\right) < D\left(\theta; (p, \hat{\lambda}')\right)$, we have $\mathcal{D}\left(p, \hat{\lambda}\right) > \mathcal{D}\left(p, \hat{\lambda}'\right)$.

Now fix $p \in (p^N, p^F)$ and suppose that $\hat{\lambda} > \hat{\lambda}'$. Since $\hat{J}^{-1}(\cdot, -\hat{\lambda})$ is continuous, two cases arise: Either, for all θ, $D\left(\theta; (p, \hat{\lambda})\right) = D\left(\theta; (p, \hat{\lambda}')\right) = 0$ (or equal to p), in which case $\mathcal{D}(p, \hat{\lambda}) \neq S(p)$, so the market does not clear. Or there exists a positive mass of types for which $D\left(\theta; (p, \hat{\lambda})\right) < D\left(\theta; (p, \hat{\lambda}')\right)$, in which case $\mathcal{D}(p, \hat{\lambda}) > \mathcal{D}(p, \hat{\lambda}')$.

Together with the continuity of $\mathcal{D}(p, \cdot)$, this implies that if there exists a multiplier $\hat{\lambda}$ such that the market clears at price p, then this multiplier must be unique. Since we know that at $\hat{\lambda}^*(p)$, the market clears, i.e., $\mathcal{D}\left(p, \hat{\lambda}^*(p)\right) = S(p)$, we conclude that $\hat{\lambda}^*(p)$ is unique.

Step 4: Continuity of the multiplier. Let $p_0 \in \left(p^N, p^F\right)$, and suppose, towards a contradiction, that $\hat{\lambda}^*$ is discontinuous from the left at p_0:

$$\hat{\lambda}^-(p_0) := \lim_{p \to p_0^-} \hat{\lambda}^*(p) < \hat{\lambda}^*(p_0).$$

Since the function $\mathcal{D}$ (as defined in Step 3) is continuous in all its arguments, and S is also continuous, it follows that:

$$\lim_{p \to p_0^-} \left[\mathcal{D}(p, \hat{\lambda}^*(p)) - S(p)\right] = \mathcal{D}\left(p_0, \hat{\lambda}^-(p_0)\right) - S(p_0).$$

But since $\mathcal{D}(p, \hat{\lambda}^*(p)) = S(p)$ for all p, the left-hand side is zero. Therefore,

$$\mathcal{D}(p_0, \hat{\lambda}^-(p_0)) = S(p_0),$$

which contradicts the uniqueness of $\hat{\lambda}^*(p_0)$ established in Step 3.

A similar argument applies to rule out discontinuity from the right.

Finally, since $\hat{\lambda}(\cdot)$ is continuous, it follows that $\lambda(\cdot)$ is also continuous.

Step 5: Continuity of the deductibles. The continuity of $D(\theta; \cdot)$ follows immediately from the fact that $D(\theta; (p, \cdot))$ is continuous and that $\hat{\lambda}(\cdot)$ is continuous. ■

Proof of Lemma 9. We have three possibilities:

1. $\lambda^*(p) \geq 0$ for all p. In this case, since $\lambda^*(\cdot)$ is strictly increasing (see Proposition 5), it follows that $\lambda^*(p) > 0$ for all $p > p^N$, and the unique price-taking benchmark is p^N. Indeed, at p^N, the optimal deductible for all types is $D(\theta; p^N) = p^N$. Then, by the first-order conditions (see the proof of Proposition 5), we have:

$$\hat{J}(\theta, p^N) \leq -\lambda^*(p^N) + p^N \leq 0 + p^N,$$

   so $\lambda^*(p^N) = 0$ is also an optimal multiplier for price p^N.

2. $\lambda^*(p) \leq 0$ for all p. In this case, $\lambda^*(p) < 0$ for all $p < p^F$, and the unique price-taking benchmark is p^F. Indeed, at p^F, the optimal deductible for all types is $D(\theta; p^F) = 0$. Then,

by the first-order conditions, we have:

$$\hat{J}(\theta, 0) \geq -\lambda^*(p^F) + p^F \geq 0 + p^F,$$

it follows that $\lambda^*(p^F) = 0$ is also an optimal multiplier for price p^F.

3. $\lambda^*(p^N) < 0$ and $\lambda^*(p^F) > 0$. In this case, since $\lambda^*(\cdot)$ is strictly increasing and continuous (see Proposition 5), the Intermediate Value Theorem implies that there exists a unique $p^U \in (p^N, p^F)$ such that $\lambda^*(p^U) = 0$.

■

Proof of Proposition 6. Suppose that $p \leq p^U$, that SE_θ satisfies condition H_θ-regularity, and that $D(\theta; p) \in [0, p)$.

Step 1. Monotonicity of the multiplier. By Proposition 5, $\lambda^*(\cdot)$ is strictly increasing. Hence, for all $p < p^U$ we have

$$-\lambda^*(p) > -\lambda^*(p^U) = 0.$$

Step 2. Characterization of the cost effect. By complementary slackness (Corollary 1), $D(\theta; p)$ satisfies

$$J(\theta, D(\theta; p)) \;\geq\; -\lambda^*(p),$$

with equality if $D(\theta; p) \in (0, p)$ and inequality if $D(\theta; p) = 0$. Substituting this expression into the definition of J yields the following representation for the cost effect:

$$1 - H_\theta(D(\theta; p)) \;\leq\; SE_\theta(D(\theta; p)) + [D(\theta; p) - p + \lambda^*(p)]\, h_\theta(D(\theta; p)).$$

Step 3. Comparison of forces. The difference between the surplus extraction force and the cost effect is therefore

$$\begin{aligned} SE_\theta(p) - \big(1 - H_\theta(D(\theta; p))\big) \;&\geq\; \underbrace{SE_\theta(p) - SE_\theta(D(\theta; p)) + [p - D(\theta; p)]\, h_\theta(D(\theta; p))}_{\geq 0 \text{ since } SE_\theta \text{ satisfies } H_\theta\text{-regularity}} - \underbrace{\lambda^*(p)\, h_\theta(D(\theta; p))}_{\geq 0 \text{ since } -\lambda^*(p) \geq 0 \text{ for } p \leq p^U} \\ &\geq 0, \end{aligned}$$

with strict inequality whenever $p < p^U$.

Step 4. Positivity of the profit derivative. Substituting this inequality into the expression for $\Pi'(p)$ gives

$$\Pi'(p) \geq \int_{\underline{\theta}}^{\bar{\theta}} \big[\underbrace{SE_\theta(p) - SE_\theta(D(\theta;p)) + [p - D(\theta;p)]\, h_\theta(D(\theta;p))}_{\geq 0 \text{ by } H_\theta\text{-regularity}}\big] f(\theta)\, d\theta$$
$$\underbrace{-\, \lambda^*(p) \left[S'(p) + \int_{\underline{\theta}}^{\bar{\theta}} h_\theta(D(\theta;p))\, f(\theta)\, d\theta\right]}_{\geq 0 \text{ since } -\lambda^*(p) \geq 0 \text{ for } p \leq p^U} \geq 0,$$

with strict inequality for $p < p^U$.

Step 5. Existence of a maximizer greater or equal than the benchmark. Since $\Pi'(p) \geq 0$ for all $p \leq p^U$, there exists an optimal $p^* \geq p^U$ that solves the second-stage problem $P2$. ■

E Technical Proofs in Section 3.6

The proof of Theorem 2 is a consequence of the following three lemmas:

Lemma 10 *The value of the primal problem OT equals the value of the dual problem D.*

Proof. Let V be given by:

$$V = \sup_{\pi \in \mathcal{M}^+(\Theta \times [0,1])} \int_{\underline{\theta}}^{\bar{\theta}} \int_0^1 \Phi(\theta, q)\, d\pi(\theta, q) + A(\pi),$$

where $\mathcal{M}^+(\Theta \times [0,1])$ is the set of positive measures over $\Theta \times [0,1]$, and A is given by:

$$A(\pi) = \inf_{(v,\lambda) \in L^1(f) \times \mathbb{R}} \int_{\underline{\theta}}^{\bar{\theta}} v(\theta) \left[f(\theta)\, d\theta - \int_0^1 d\pi(\theta, q)\right] + \int_{\underline{\theta}}^{\bar{\theta}} \int_0^1 \lambda\left([H_\theta(p) - H_\theta(0)]q - RS(p)\right) d\pi(\theta, q),$$

where $L^1(f)$ denotes the set of all measurable functions $v : \Theta \to \mathbb{R}$ that satisfy:

$$\int_{\underline{\theta}}^{\bar{\theta}} |v(\theta)| f(\theta)\, d\theta < \infty.$$

Note that if π satisfies BP and EQ, $A(\pi) = 0$. However, if π does not satisfy BP or EQ, we can find a function v or multiplier λ such that $A(\pi) = -\infty$. This implies that the value of the primal problem OT is equal to V. We can also express V as:

$$V = \sup_{\pi \in \mathcal{M}^+(\Theta \times [0,1])} \inf_{(v,\lambda) \in L^1(f) \times \mathbb{R}} \int_{\underline{\theta}}^{\bar{\theta}} v(\theta) f(\theta)\, d\theta + \int_{\underline{\theta}}^{\bar{\theta}} \int_0^1 \left[\Phi(\theta, q) + \lambda\left([H_\theta(p) - H_\theta(0)]q - RS(p)\right) - v(\theta)\right] d\pi(\theta, q).$$

Here, it holds that $\sup \inf = \inf \sup$ (for a rigorous argument see the proof of Theorem 5.10 in Villani et al. (2009)), which yields:

$$V = \inf_{(v,\lambda) \in L^1(f) \times \mathbb{R}} \int_{\underline{\theta}}^{\bar{\theta}} v(\theta) f(\theta)\, d\theta + B(v, \lambda),$$

where:

$$B(v, \lambda) = \sup_{\pi \in \mathcal{M}^+(\Theta \times [0,1])} \int_{\underline{\theta}}^{\bar{\theta}} \int_0^1 \left[\Phi(\theta, q) + \lambda\left([H_\theta(p) - H_\theta(0)]q - RS(p)\right) - v(\theta)\right] d\pi(\theta, q).$$

Observe that if $v(x) < \Phi(\theta, q) + \lambda\left([H_\theta(p) - H_\theta(0)]q - RS(p)\right)$ for some (θ, q), we can find a positive measure π such that $B(v, \lambda) = +\infty$. On the other hand, if $v(x) \geq \Phi(\theta, q) + \lambda\left([H_\theta(p) - H_\theta(0)]q - RS(p)\right)$ for all (θ, q), then $B(v, \lambda) = 0$. This implies that the value of V is equal to the value of the dual problem D, which shows that there is zero duality gap. ∎

Lemma 11 *A solution to the primal problem OT exists.*

Proof. Let Π denote the set of probability measures on $\Theta \times [0, 1]$ that satisfy conditions BP and EQ. Formally, Π is given by:

$$\Pi = \left\{ \pi \in \Delta\left(\Theta \times [0, 1]\right) \,\middle|\, \pi \text{ satisfies } BP \text{ and } EQ \right\}.$$

We first show that Π is compact under the weak topology on $\Delta\left(\Theta \times [0, 1]\right)$. For this, we invoke **Prokhorov's theorem**, which states:

Let Y be a Polish space, and let $\Delta(Y)$ denote the space of probability measures on Y. A set $P \subset \Delta(Y)$ is precompact under the weak topology if and only if it is tight; that is, for any $\varepsilon > 0$, there exists a compact set $K_\varepsilon \subset Y$ such that $\mu(K_\varepsilon) > 1 - \varepsilon$ for all $\mu \in P$.

Since $\Theta \times [0, 1]$, equipped with the Euclidean metric, is a compact metric space, any collection of probability measures in $\Delta\left(\Theta \times [0, 1]\right)$ is tight. Indeed, for any $\varepsilon > 0$, we can set $K_\varepsilon = \Theta \times [0, 1]$. By

Prokhorov's theorem, it follows that Π is precompact. To conclude that Π is compact, it remains to show that Π is closed.

Recall that a sequence of probability measures $\{\pi_n\}_{n=1}^{\infty} \subset \Pi$ converges to π^* in the weak topology if, for every continuous function $g : \Theta \times [0,1] \to \mathbb{R}$:

$$\lim_{n\to\infty} \int_{\underline{\theta}}^{\bar{\theta}} \int_0^1 g(\theta, q)\, d\pi_n(\theta, q) = \int_{\underline{\theta}}^{\bar{\theta}} \int_0^1 g(\theta, q)\, d\pi^*(\theta, q).$$

In particular, since H_θ is continuous in θ, we must have that if $\{\pi_n\}_{n=1}^{\infty} \to \pi^*$, then:

$$\lim_{n\to\infty} \int_{\underline{\theta}}^{\bar{\theta}} \int_0^1 [H_\theta(p) - H_\theta(0)]q\, d\pi_n(\theta, q) = \int_{\underline{\theta}}^{\bar{\theta}} \int_0^1 [H_\theta(p) - H_\theta(0)]q\, d\pi^*(\theta, q).$$

Using the fact that the following condition holds for all n:

$$\int_{\underline{\theta}}^{\bar{\theta}} \int_0^1 [H_\theta(p) - H_\theta(0)]q\, d\pi_n(\theta, q) = RS(p),$$

it follows that π^* must satisfy EQ. A similar argument shows that if $\{\pi_n\}_{n=1}^{\infty} \to \pi^*$, then π^* must also satisfy BP. Thus, $\pi^* \in \Pi$, proving that Π is closed.

Finally, compactness of Π implies the existence of a solution to OT. Indeed, let $\{\pi_n\}_{n=1}^{\infty}$ be any maximizing sequence, which admits a cluster point $\pi^* \in \Pi$. By our technical assumptions, Φ is a bounded and continuous function, then it can be written as the limit of a non-increasing sequence $\{\Phi_k\}_{k=1}^{\infty}$ of bounded continuous functions. By invoking successively the monotone convergence theorem, the fact that π^* is a cluster point, the inequality $\Phi_k(\theta, q) \geq \Phi(\theta, q)$ and the maximizing property of $\{\pi_n\}_{n=1}^{\infty}$, we obtain:

$$\begin{aligned}
\int_{\underline{\theta}}^{\bar{\theta}} \int_0^1 \Phi(\theta, q)\, d\pi^*(\theta, q) &= \lim_{k\to\infty} \int_{\underline{\theta}}^{\bar{\theta}} \int_0^1 \Phi_k(\theta, q)\, d\pi^*(\theta, q) \\
&\geq \lim_{k\to\infty} \liminf_{n\to\infty} \int_{\underline{\theta}}^{\bar{\theta}} \int_0^1 \Phi_k(\theta, q)\, d\pi_n(\theta, q) \\
&\geq \liminf_{n\to\infty} \int_{\underline{\theta}}^{\bar{\theta}} \int_0^1 \Phi(\theta, q)\, d\pi_n(\theta, q) \\
&= \sup_{\pi \in \Pi(f,p)} \int_{\underline{\theta}}^{\bar{\theta}} \int_0^1 \Phi(\theta, q) d\pi(\theta, q).
\end{aligned}$$

Therefore, π^* is a solution to OT. ■

Lemma 12 *A solution to the dual problem D exists.*

Proof. Let π^* be any solution to the primal problem OT, whose existence is guaranteed by Lemma 11. We can express the value of OT as:

$$V = \int_{\underline{\theta}}^{\bar{\theta}} \int_0^1 \Phi(\theta, q)\, d\pi^*(\theta, q) = \int_{\underline{\theta}}^{\bar{\theta}} \int_0^1 \Phi(\theta, q)\, d\pi^*(q \mid \theta) f(\theta)\, d\theta,$$

where $\pi^*(q \mid \theta) = \frac{\pi^*(\theta,q)}{f(\theta)}$ is the conditional distribution of q given θ. Indeed, since π^* satisfies BP, we know that $\pi^*(\cdot \mid \theta) \in \Delta([0,1])$. For the rest of the proof, with a slight abuse of notation, we denote by π the collection of conditional distributions, $\{\pi(\cdot \mid \theta)\}_{\theta\in\Theta}$, where each $\pi(\cdot \mid \theta) \in \Delta([0,1])$.

We now define two key functions:

$$y(\pi) = \int_{\underline{\theta}}^{\bar{\theta}} \int_0^1 \Phi(\theta, q)\, d\pi(q \mid \theta) f(\theta)\, d\theta, \quad g(\pi) = \int_{\underline{\theta}}^{\bar{\theta}} \int_0^1 \left([H_\theta(p) - H_\theta(0)]q - RS(p)\right)\, d\pi(q \mid \theta) f(\theta)\, d\theta.$$

From the definitions of y and g, we immediately have the following reformulation of the original problem:

$$V = \max_{\pi} \quad y(\pi) \quad \text{s.t} \quad g(\pi) = 0,$$

and π^* solves this optimization problem:

$$\pi^* \in \arg\max_{\pi} \quad y(\pi) \quad \text{s.t} \quad g(\pi) = 0.$$

By standard Lagrangian duality, we can argue that there exists a multiplier $\lambda^* \in \mathbb{R}_+$ such that the constrained optimization problem can be reformulated as:

$$V = \max_{\pi} \quad y(\pi) + \lambda^* g(\pi),$$

$$\pi^* \in \arg\max_{\pi} \quad y(\pi) + \lambda^* g(\pi).$$

To rigorously justify the existence of λ^*, we use the separating hyperplane theorem. Define the following sets in $\mathbb{R}^2$:

$$\begin{aligned} A &= \left\{(r,z) \;\middle|\; r \leq y(\pi) \text{ and } z = g(\pi) \text{ for some } \pi \in \Delta([0,1])^{\Theta}\right\}, \\ B &= \{(r,0) \mid r \geq V\}. \end{aligned}$$

Since both y and g are linear in π, the sets A and B are convex. Furthermore, A contains no interior points of B. Therefore, by the separating hyperplane theorem, there exists a non-zero vector $(r_0, \lambda^*) \in \mathbb{R}^2$ such that:

$$r_0 r_1 + \lambda^* z_1 \leq r_0 r_2,$$

for all $(r_1, z_1) \in A$ and $(r_2, 0) \in B$. We now show that $r_0 > 0$.

Case 1 $r_0 < 0$: In this case, we can find a point $(r_2, 0) \in B$ with sufficiently large r_2 such that the inequality $r_0 r_1 + \lambda^* z_1 > r_0 r_2$ holds, leading to a contradiction.

Case 2 $r_0 = 0$: Here, $\lambda^* \neq 0$. Since, for all p, g can be positive for some π and negative for other, i.e. it is always possible to have excess demand or excess supply. Then, A contains two points, (r, z) and (r', z'), such that $z > 0$ and $z' < 0$. However, if $\lambda^* > 0$ ($\lambda^* < 0$), then $\lambda^* z \leq 0$ ($\lambda^* z' \leq 0$), leading to a contradiction.

Thus, we must have $r_0 > 0$, and without loss of generality, we can normalize $r_0 = 1$. Therefore, we have:

$$V = \max_{(r,z)\in A} \quad r + \lambda^* z = \max_{\pi} \quad y(\pi) + \lambda^* g(\pi),$$

which proves the existence of λ^*.

We now exploit some properties of π^*. Since π^* maximizes $y(\pi) + \lambda^* g(\pi)$, for f-almost every θ we must have that:

$$\int_0^1 \left[\Phi(\theta, q) + \lambda^* \left([H_\theta(p) - H_\theta(0)]q - RS(p)\right)\right] \, d\pi^*(q \mid \theta),$$

is equal to:

$$\max_{q\in[0,1]} \quad \Phi(\theta, q) + \lambda^* \left([H_\theta(p) - H_\theta(0)]q - RS(p)\right).$$

The maximization problem is well-defined because $\Phi(\theta, \cdot)$ is a continuous function and we are maximizing over the compact set $[0,1]$. Then, define the function $v^* : \Theta \to \mathbb{R}$ by:

$$v^*(\theta) = \int_0^1 \left[\Phi(\theta, q) + \lambda^* \left([H_\theta(p) - H_\theta(0)]q - RS(p)\right)\right] \, d\pi^*(q \mid \theta).$$

Then, the pair (v^*, λ^*) satisfies SP. Finally, observe that:

$$V = \int_{\underline{\theta}}^{\bar{\theta}} v^*(\theta) f(\theta) \, d\theta.$$

This shows that (v^*, λ^*) achieves the same value as the primal problem OT, and by Lemma 10, we conclude that (v^*, λ^*) is a solution to the dual problem D. ■

Proof of the Necessity Part of Corollary 1. First, since (v^*, λ^*) is a solution of the dual problem D it immediately follows that for $f-$almost every θ we must have:

$$v^*(\theta) = \max_{q \in [0,1]} \quad \Phi(\theta, q) + \lambda^* \left([H_\theta(p) - H_\theta(0)]q - RS(p)\right),$$

where the maximum is well-defined because $\Phi(\theta, \cdot)$ is a continuous function and we are maximizing over the compact set $[0,1]$.

The above implies that if any joint distribution π satisfies conditions BP and EQ, but there is a positive measure of risk types θ for which $\pi(\cdot \mid \theta)$ is not supported on $Supp$, then the following inequality holds:

$$\begin{aligned}
\int_{\underline{\theta}}^{\bar{\theta}} v^*(\theta) f(\theta) \, d\theta &> \int_{\underline{\theta}}^{\bar{\theta}} \int_0^1 \left[\Phi(\theta, q) + \lambda^* \left([H_\theta(p) - H_\theta(0)]q - RS(p)\right)\right] \, d\pi(q \mid \theta) f(\theta) \, d\theta \\
&= \int_{\underline{\theta}}^{\bar{\theta}} \int_0^1 \Phi(\theta, q) \, d\pi(q \mid \theta) f(\theta) \, d\theta \\
&\quad + \lambda^* \int_{\underline{\theta}}^{\bar{\theta}} \int_0^1 \left([H_\theta(p) - H_\theta(0)]q - RS(p)\right) \, d\pi(q \mid \theta) f(\theta) \, d(\theta) \\
&= \int_{\underline{\theta}} \int_0^1 \Phi(\theta, q) \, d\pi(q, \theta).
\end{aligned}$$

By Theorem 2, this implies that π is not a solution of the primal problem OT. ■

Proof of the Sufficiency Part of Corollary 1. Suppose that π satisfies BP, EQ and for $f-$almost every θ, the conditional distribution $\pi(\cdot \mid \theta)$ is supported on $Supp$. Then, the following equality holds:

$$\begin{aligned}
\int_{\underline{\theta}}^{\bar{\theta}} v^*(\theta) f(\theta)\, d\theta &= \int_{\underline{\theta}}^{\bar{\theta}} \int_0^1 \left[\Phi(\theta, q) + \lambda^* \left([H_\theta(p) - H_\theta(0)]q - RS(p)\right)\right]\, d\pi(q \mid \theta) f(\theta)\, d\theta \\
&= \int_{\underline{\theta}}^{\bar{\theta}} \int_0^1 \Phi(\theta, q)\, d\pi(q \mid \theta) f(\theta)\, d\theta \\
&\quad + \lambda^* \int_{\underline{\theta}}^{\bar{\theta}} \int_0^1 \left([H_\theta(p) - H_\theta(0)]q - RS(p)\right)\, d\pi(q \mid \theta) f(\theta)\, d(\theta) \\
&= \int_{\underline{\theta}} \int_0^1 \Phi(\theta, q)\, d\pi(q, \theta).
\end{aligned}$$

By Theorem 2, it follows that π is a solution to the primal problem OT. ■