Optimal Rationing within a Heterogeneous Population

Philippe Choné, Stéphane Gauthier

Université Paris 1 – Panthéon - Sorbonne –

Documents de Travail du
Centre d'Economie de la Sorbonne

CES Working Papers

Optimal Rationing within a Heterogeneous Population

Philippe CHONÉ, Stéphane GAUTHIER

2014.32

Maison des Sciences Économiques, 106-112 boulevard de L'Hôpital, 75647 Paris Cedex 13
http://centredeconomiesorbonne.univ-paris1.fr/
ISSN : 1955-611X

Optimal Rationing within a Heterogeneous Population

Philippe Choné*
CREST-ENSAE

Stéphane Gauthier†
PSE and University of Paris 1

January 1, 2014

Abstract

A government agency delegates to a provider (hospital, medical gatekeeper, school, social worker) the decision to supply a service or treatment to individual recipients. The agency does not perfectly know the distribution of individual treatment costs in the population. The single-crossing property is not satisfied when the uncertainty pertains to the dispersion of the distribution. We find that the provision of service should then be distorted upwards relative to efficiency when the (first-best) efficient number of recipients is sufficiently high.

JEL classification numbers: I18, D82, D61

Keywords: Rationing, screening, universal coverage, upward distortion, Spence-Mirrlees condition

*ENSAE, 3 avenue Pierre Larousse, 92245 Malakoff; Phone: +33(0)141175390; E-mail address: philippe.chone@ensae.fr.

†MSE, 106-112 bd de l'Hôpital, 75013 Paris, France; Phone: +33(0)144078289; E-mail address: stephane.gauthier@univ-paris1.fr.

1 Introduction

We consider a government agency in charge of supplying a service or treatment to a population of potential recipients. We assume that the cost and the benefit of the treatment vary across individuals. Efficiency requires that the treatment is provided to an individual if and only if social benefit exceeds social cost for that individual.

The individual heterogeneity in cost and benefit may in part be explained by recipients' characteristics that are observed by the agency. To some extent, the agency can thus use "rationing by denial", whereby the treatment is denied to individuals with unfavorable cost-benefit ratio. In many instances, however, there remains substantial unobserved heterogeneity conditional on observable variables. The agency has then to leave the selection of recipients to the discretion of the service provider, i.e. rely on "rationing by selection", see [3]. In practice, the supply decision is indeed delegated to the provider for various services, e.g. medical or social care, after-school or training programs.

Under rationing by selection, the agency first determines the total number of recipients. This step requires knowing the distribution of unobserved heterogeneity among potential recipients. Perfect knowledge of the distribution is commonly assumed in the literature. For instance, the assumption is made in [5] and [4], with the former article considering many treatment varieties and the latter investigating provider altruism –two issues not addressed here.

The present note places the emphasis on the uncertainty about the distribution of unobserved heterogeneity. The composition of the population of recipients addressed by a given provider depends on local conditions, creating significant variation in the underlying distributions across providers. The effect of differences in composition on the cost and benefit distributions is difficult to estimate empirically (see [6] for a recent survey).

Accordingly, we relax the assumption that the agency perfectly knows the cost and benefit distribution in the population of potential recipients. We find that the uncertainty about the distribution of heterogeneity causes the number of recipients to be distorted relative to efficiency. The direction of the distortion depends on whether the uncertainty pertains to the mean or to the dispersion of individual treatment costs. In the former case, the usual Spence-Mirrlees condition is satisfied and the distortion is necessarily downwards: The number of treated recipients is lower than recommended by first-best efficiency. In the latter case, the Spence-Mirrlees condition no longer holds. We find that the first-best optimum governs the the sign of distortion in the second-best program: The distorsion is upwards when the first-best number of treated recipients is sufficiently high. Uncertainty about the cost dispersion then pushes towards universal coverage policies.

2 Model

A population of recipients, whose size is normalized to one, is eligible for a treatment supplied by a single provider. Recipients are indexed by two nonnegative real numbers b and c, with distribution function $\Phi(b, c)$. The treatment of a type (b, c) recipient yields benefit b to the recipient and costs c to the provider. The corresponding net social benefit is $b - (1 + \lambda)c$, where λ is the (exogenously given) marginal cost of public funds.

Assumption 1. *The (expected) net social benefit of treatment for a given cost level c, $\mathbb{E}(b \mid c) - (1 + \lambda)c$, is a non-increasing function of cost c with a unique zero, denoted by c^{**}.*

Assumption 1 holds true in particular when the expected benefit decreases with cost, a case often considered in the literature, e.g., in [2] and [4]. [5] provides health related examples where this assumption is relevant. Under this assumption, the first-best requires to treat recipients with cost $c \leq c^{**}$. Denoting by F the marginal distribution of individual treatment costs in the population of recipients, the first-best optimal number of treated recipients is $n^{**} = F(c^{**})$.

In this paper, we assume that the agency relies on rationing by selection: the agency observes the number n of recipients but not their individual characteristics. The treatment decision is delegated to a single provider who observes the individual characteristics of recipients. The agency offers a take or leave contract specifying the number of recipients that must be treated by the provider and a compensating transfer T. The utility of the provider when she accepts the contract is $U(n, T) = T - C(n)$, where $C(n)$ represents the aggregate cost of treating n recipients.

Given (n, T), utility maximization requires that the least costly recipients be treated in priority. The provider's cost of treating n recipients, $0 \leq n \leq 1$, is therefore given by

$$C(n) = \int_0^{F^{-1}(n)} c \,\mathrm{d}F(c).$$

The marginal cost is $C'(n) = F^{-1}(n)$, i.e., the cost of the marginal treated recipient is $F^{-1}(n)$, the nth-percentile of the distribution F. It follows that the cost function $C(n)$ is convex in n.

The net social benefit of treating n recipients is given by $S(n) = B(n) - (1+\lambda)C(n)$, where

$$B(n) = \int_0^{F^{-1}(n)} \mathbb{E}(b \mid c) \,\mathrm{d}F(c).$$

represents the (expected) the gross social benefit. Under Assumption 1, the net social benefit function $S(n)$ is concave in n, reaching its maximum at $n^{**} = F(c^{**})$. When the agency knows the distribution of individual recipients characteristics Φ, it can choose the number of recipients n and the transfer T that maximize

$$B(n) + U(n,T) - (1+\lambda)T = S(n) - \lambda U(n,T)$$

subject to the provider's participation constraint $U(n,T) \geq 0$. The solution to this maximization problem is the first-best optimum, $n = n^{**}$ and $U = U^{**} = 0$.

Under Assumption 1, the provider treats in priority the recipients with highest expected social values: The social net benefit and the provider's private objective are 'aligned.' Hence, the agency can achieve the first-best optimum if it knows the distribution of individual recipients characteristics: From the distribution function Φ, the agency infers the optimal number n^{**} and the corresponding cost, $C(n^{**})$. It is then sufficient to ask the provider to treat n^{**} recipients and reimburse $C(n^{**})$.

3 Unknown cost distribution

We now relax the assumption that the agency knows the distribution of individual characteristics of recipients. We suppose that the agency only knows that the distribution function is Φ_i $(i = H, L)$, associated with marginal cost distribution F_i, with probability π_i $(\pi_L + \pi_H = 1)$. The true distribution function is private information to the provider. We refer to i as the provider type. Hence, the agency is faced with two types of provider, themselves confronted with different populations of recipients and marginal distributions of individual costs. Provider i has cost function $C_i(n)$ with $C_i'(n) = F_i^{-1}(n)$, and the net social benefit of having n recipients treated by provider i is

$$S_i(n) = \int_0^{F_i^{-1}(n)} [\,\mathbb{E}_i(b \mid c) - (1+\lambda)c\,] \,\mathrm{d}F_i(c),$$

where the conditional expectation $\mathbb{E}_i$ is taken under provider i's distribution Φ_i. Assumption 1 is supposed to hold for the two provider types: The first-best cost threshold for provider i is c_i^{**} such that $\mathbb{E}_i(b \mid c_i^{**}) = (1+\lambda)c_i^{**}$, where $\mathbb{E}_i(b \mid c_i^{**})$ represents the mean treatment benefit for provider i in the population of recipients whose treatment cost is c_i^{**}. The corresponding first-best number of recipients is $n_i^{**} = F_i(c_i^{**})$.

Appealing to the revelation principle, the agency offers a menu (n_i, T_i), $i = H, L$, maximizing

$$\sum_i \pi_i \left[S_i(n_i) - \lambda U_i(n_i, T_i)\right]$$

subject to the provider's participation constraints $U_i(n_i, T_i) = T_i - C_i(n_i) \geq 0$ and the incentive constraints $U_i(n_i, T_i) \geq U_i(n_j, T_j)$ for all $i, j = H, L$.

We examine how the solution to this problem departs from the first-best optimum when the marginal distribution F_L stochastically dominates the distribution F_H.

3.1 Local distortions

Suppose first that F_L first-order stochastically dominates F_H: $F_L(c) \leq F_H(c)$ for all c, with strict inequality on a non-degenerated interval. In this case, for all $n > 0$, provider H's marginal cost stands below provider L's:

$$C_H'(n) = F_H^{-1}(n) < C_L'(n) = F_L^{-1}(n),$$

and hence $C_H(n) < C_L(n)$, since $C_H(0) = C_L(0)$. The first-best menu $(n_i^{**}, C_i(n_i^{**}))$ is not incentive compatible: Provider H chooses $(n_L^{**}, C_L(n_L^{**}))$ and earns $C_L(n_L^{**}) - C_H(n_L^{**}) > 0$. In order to reduce the informational rent $C_L(n_L) - C_H(n_L)$ of the type H provider, the agency distorts the number of recipients treated by type L with respect to the first-best n_L^{**}. The usual single-crossing condition is satisfied:

$$\left.\frac{\partial T}{\partial n}\right|_{U_H} = C_H'(n) < C_L'(n) = \left.\frac{\partial T}{\partial n}\right|_{U_L} \qquad \text{for all } n.$$

Therefore a lower rent requires a lower number of recipients treated by provider L. This is the standard pattern in adverse selection problems.

Proposition 1. *Suppose that F_L first-order stochastically dominates F_H. At the second-best optimum, provider H (provider L) treats the optimal number of (too few) recipients and earns a positive (zero) profit.*

Proof. Since $C_L(n) > C_H(n)$, the type H incentive constraint is binding, $U_H^* = C_L(n_L^*) - C_H(n_L^*)$, and $U_L^* = 0$. The optimal number n_L^* satisfies the first-order condition

$$\pi_L S_L'(n_L^*) - \lambda\pi_H[C_L'(n_L^*) - C_H'(n_L^*)] = 0. \tag{1}$$

Thus $S_L'(n_L^*) > 0 = S'(n_L^{**})$, and $S(n)$ being single-peaked, $n_L^* < n_L^{**}$. Type H treats $n_H^* = n_H^{**}$ recipients and earns $U_H^* = C_L(n_L^*) - C_H(n_L^*) > 0$. $\square$

The main result of the paper concerns the case where F_H is a mean-preserving spread of F_L: Both distributions F_H and F_L have the same mean,

$$C_L(1) = \int_0^\infty c\,\mathrm{d}F_H(c) = \int_0^\infty c\,\mathrm{d}F_L(c) = C_H(1),$$

and F_L second-order stochastically dominates F_H,

$$\int_0^c F_L(c)\,\mathrm{d}c \le \int_0^c F_H(c)\,\mathrm{d}c \qquad \text{for all } c.$$

Assumption 2. *The two distribution functions cross only once, at some individual treatment cost denoted $\hat{c}$.*

The single-crossing property does not hold, but Assumption 2 restricts the pattern of violation of this assumption, delimitating two different regions:

$$\left.\frac{\partial T}{\partial n}\right|_{U_H} = C_H'(n) < C_L'(n) = \left.\frac{\partial T}{\partial n}\right|_{U_L} \iff n < \hat{n}, \tag{2}$$

with $\hat{n} = F_L(\hat{c}) = F_H(\hat{c})$. In the spirit of [1], we use this partition to handle the adverse selection problem.[1]

Proposition 2. *Suppose that F_H is a mean-preserving spread of F_L and that Assumption 2 holds. Then, provider's H incentive constraint is the only one to be binding at the optimum. This provider treats n_H^{**} recipients.*

The agency has a local incentive to distort the number of recipients treated by provider L upwards from n_L^{**} if and only if $n_L^{**} > \hat{n}$.

Proof. By Assumption 2, $F_H(c) > F_L(c)$ for $c < \hat{c}$ and $F_H(c) < F_L(c)$ for $c > \hat{c}$. Using the expression of the marginal costs, we find that $C'_H(n) < C'_L(n)$ for $n < \hat{n}$ and $C'_H(n) > C'_L(n)$ for $n > \hat{n}$. Since $C_H(0) = C_L(0)$ and $C_H(1) = C_L(1)$, we have

$$C_H(n) \leq C_L(n)$$

for all n, with equality for $n = 0$ and $n = 1$. The cost difference $C_L(n) - C_H(n)$ first increases, then decreases as n rises, achieving its maximum at $\hat{n} = F_H(\hat{c}) = F_L(\hat{c})$. The efficient provider, provider H, is the one faced with more dispersed individual treatment costs.

It follows that at the second-best optimum, H's incentive constraint must be binding, $U_H^* = C_L(n_L^*) - C_H(n_L^*)$, and $U_L^* = 0$. The number of recipients treated by provider H is undistorted. The optimal number n_L^* of treated by the L type maximizes

$$K(n) \equiv \pi_L S_L(n) + \lambda \pi_H [C_H(n) - C_L(n)]. \tag{3}$$

The first derivative of this function at n_L^{**} is $\lambda \pi_H [C'_H(n_L^{**}) - C'_L(n_L^{**})]$, since $S_L(n_L^{**}) = 0$. By (2), it is positive if and only if $n_L^{**} > \hat{n}$. $\square$

Figures 1 and 2 describe how to derive the aggregate cost function from the marginal distribution of individual cost. By Assumption 2 the marginal distributions intersect at $\hat{c}$. From the second order stochastic dominance property we also know that $F_H(c)$ is above (below) $F_L(c)$ for c less (greater) than $\hat{c}$. Therefore, the slope of the cost function for provider H is less (greater) than the slope of the cost function for provider L when n is less (greater) than $\hat{n}$. This change in the way the slopes are ordered implies a failure of the usual single-crossing condition. However, given that both providers

[1] In our setup, however, $C'_H(n) - C'_L(n)$ is non-monotonic in n, hence [1]'s Assumption A2 does not hold.

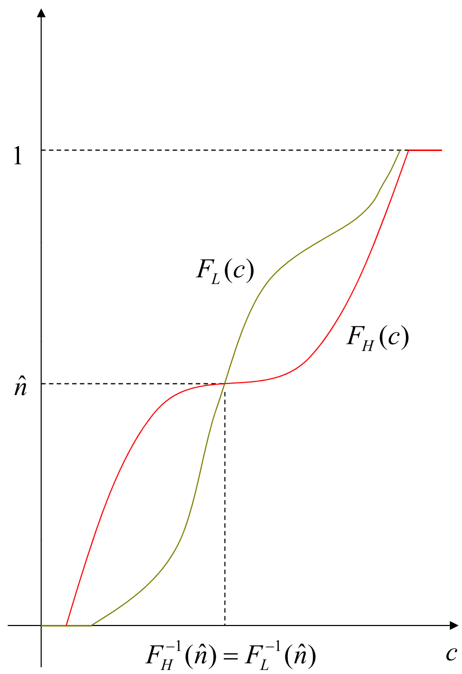

Figure 1: Individual cost cdf

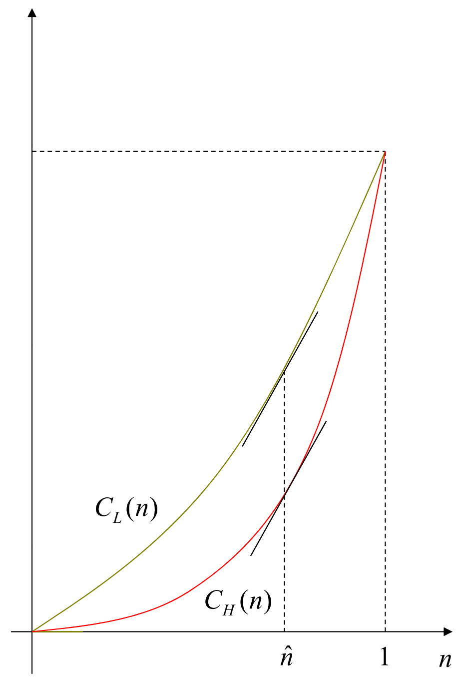

Figure 2: Aggregate cost function

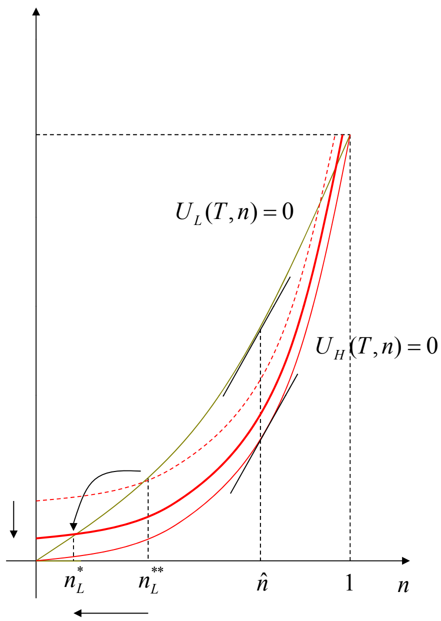

Figure 3: Downward distortion

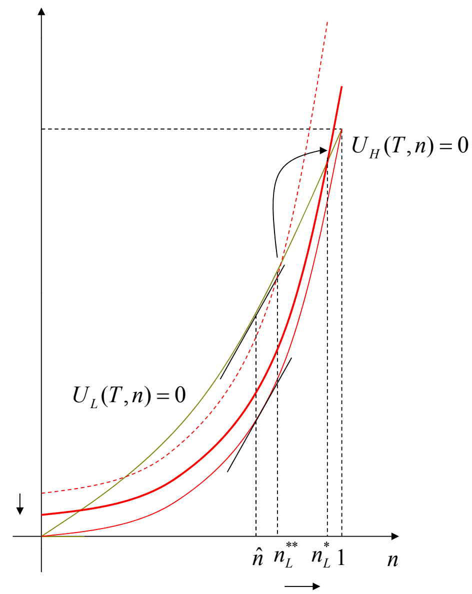

Figure 4: Upward distortion

have the same aggregate costs at $n \in \{0,1\}$, provider's H aggregate cost must be lower for all $n \in (0,1)$.

The intuition is simple: When n recipients are treated, the individual treatment costs are concentrated around the mean $C_L(1)$ for provider L and are more dispersed for provider H. The latter provider treats in priority the least costly recipients, which gives him a cost advantage relative to provider L, that advantage being maximal for $n = \hat{n}$. Though the single-crossing condition is not met, type L (H) is unambiguously the 'inefficient' ('efficient') type, i.e., $C_L(n) \geq C_H(n)$ for all n, thus allowing us to handle with incentive constraints.

Figures 3 and 4 illustrate how the failure of the Spence-Mirrlees condition can be exploited by the agency to reduce the type H informational rent $C_L(n_L) - C_H(n_L)$. The iso-utility curves are increasing and convex in n. In the first-best, both types would earn zero utility: The corresponding iso-utility curves pass through the origin $n = T = 0$. If the first-best contracts were offered, type H would imitate type L by choosing to treat n_L^{**} recipients. Provider H would then earn a positive profit $T_L^{**} - C_H(n_L^{**}) = C_L(n_L^{**}) - C_H(n_L^{**}) > 0$ (dotted iso-utility curve). In order to lower this rent, the agency has a local incentive to reduce the number of treated by the L type when $n_L^{**} < \hat{n}$, as is the case in Figure 3. On the contrary, $n_L^{**} > \hat{n}$ in Figure 4, and the agency then has a local incentive to increase the number of treated above n_L^{**}.

Thus, unlike the main strand of the literature, the sign of the local distorsion from the first-best optimum in the second-best setup are driven by the number of treated in the first-best problem: A high number of treated in the first-best yields an upward distorsion.

3.2 A global result

Previous results describe the pattern of local incentives to distort the number of treated recipients from the first-best. However it could be that the global optimum involves to treat $n_L^* < n_L^{**}$ recipients in the case where n_L^{**} is above $\hat{n}$. To go beyond the local result of Proposition 2, we introduce

Assumption 3. *The distributions F_L and F_H are symmetric around $\hat{c}$, i.e. $F_i(c) + F_i(2\hat{c} - c) = 1$ for all c and $i = H, L$.*

Under Assumptions 2 and 3, the distributions are equal to one half when they cross: $\hat{n} = F_H(\hat{c}) = F_L(\hat{c}) = 1/2$.

Proposition 3. *Suppose that Assumptions 2 and 3 hold, F_H is a mean-preserving spread of F_L, and n_L^{**} is larger than 1/2. Then the number of recipients treated by provider L is distorted upwards if $S_L(1) > S_L(1 - n_L^{**})$.*

Proof. By Assumption 3, the rent $U_H(n_L) = C_L(n_L) - C_H(n_L)$ is symmetric around its global maximum $\hat{n} = 1/2$, i.e. $U_H(n) = U_H(1 - n)$ for all n. Recall that the second-best number n_L^* of recipients treated by provider L maximizes the function $K(n)$ given by (3). We want to show that $n_L^* > n_L^{**}$ when $n_L^{**} > 1/2$. The proof proceeds in three steps:

1. Let $\tilde{n}_L = 1 - n_L^{**} < 1/2$. Since $U_H(n) = U_H(1 - n)$ and S_L increases below n_L^{**}, the maximum of K on $[\tilde{n}_L, n_L^{**}]$ is achieved above 1/2.

2. Since S_L is increasing and U_H is decreasing on the interval $[1/2, n_L^{**}]$, K is increasing on this interval, implying that the maximum of K on $[\tilde{n}_L, n_L^{**}]$ is achieved at n_L^{**}.

3. By symmetry of U_H and monotonicity of S_L on the intervals $[0, n_L^{**}]$ and $[n_L^{**}, 1]$, we have

$$K(n) - K(1-n) = S_L(n) - S_L(1-n) \geq S_L(1) - S_L(1-n_L^{**}) > 0$$

for all $n \geq n_L^{**}$. It follows that the maximum of $K(n)$ on $[0,1]$ is achieved at the right of n_L^{**}.

By Proposition 2, the maximum is achieved strictly above n_L^{**}. □

This global result is closely reminiscent of Proposition 2: The number of recipients is distorted upwards from the first-best optimum if the agency prefers the inefficient provider (the one faced with less dispersed individual treatment costs) to treat a sufficiently high number of recipients.

The new global condition in Proposition 3 depends on whether treating all the population is socially preferred to treating only a fraction $1 - n_L^{**}$. It should be satisfied in practice: if the first-best optimum recommends to treat $n_L^{**} = 80\%$ of the population, the agency should prefer providing universal coverage to treating only $1 - n_L^{**} = 20\%$ of the population.[2]

References

[1] Araujo, A., and H. Moreira (2010): "Adverse selection problems without the Spence-Mirrlees condition," *Journal of Economic Theory*, 145(3).

[2] De Fraja, G. (2000): "Contracts for health care and asymmetric information," *Journal of Health Economics*, 19, 663–677.

[3] Klein, R., and J. Mayblin (2012): "Thinking about rationing," Discussion paper, The King's Fund.

[4] Makris, M., and L. Siciliani (2013): "Optimal incentive schemes for altruistic providers," *Journal of Public Economic Theory*, 15(5), 675–699.

[2] This condition follows from Assumption 3. Global results involving an upwards distortion obtain for asymmetric distributions when, for high values of n ($n > \hat{n}$), the first-best surplus for the inefficient provider is sufficiently high, and the difference between the aggregate cost functions of both providers decreases sufficiently with n.

[5] Malcomson, J. (2005): "Supplier Discretion over Provision:Theory and an Application to Medical Care," *RAND Journal of Economics*, 36(2).

[6] Street, A., D. Scheller-Kreinsen, A. Geissler, and R. Busse (2010): "Determinants of hospital costs and performance variation: Methods, models and variables for the EuroDRG project," Working Papers in Health Policy and Management 3.