Int J Health Care Finance Econ (2006) 6:191–213
DOI 10.1007/s10754-006-9001-8

Price vs. quantity in health insurance reimbursement

Francesca Barigozzi

Received: 19 March 2003 / Accepted: 16 May 2006 /
Published online: 7 October 2006

Abstract While "integrated" systems regulate the *quantity* of health services, "Bismarckian" systems regulate their *price*. This paper compares the consumers' allocations implemented within the two reimbursement systems. In the model, illness has a negative impact on labor productivity while public insurance is financed through income tax. Consumers have private information with respect to a parameter which can be interpreted as heterogeneity either in intensity of their preferences for treatment or in the type of illness. The social planner may be constrained to adopt uniform insurance plans, or may be free to choose self selecting plans. The analysis of uniform plans shows that Bismarckian systems dominate integrated systems from the social welfare point of view; whereas the opposite ranking holds with self-selecting plans.

Keywords Public health insurance · In-kind transfers · Reimbursement insurance · Adverse selection

JEL Classification I11 · I18 · D82 · H42

Introduction

Risk-averse consumers demand health insurance. They insure against the financial risk associated with buying medical care: consumers pay a premium ex-ante and receive reimbursement if illness occurs. This paper studies and compares two alternative health insurance reimbursement structures: *in-kind* reimbursement which is used in integrated systems (such as the NHS in the UK and the HMO in the US) and *reimbursement insurance* which is used in most countries with Bismarckian systems

F. Barigozzi (✉)
University of Bologna, Strada Maggiore 45, 40125 Bologna, Italy
e-mail: barigozz@spbo.unibo.it

or in the case of traditional health insurance.[1] Both methods represent demand-side, cost-containment measures. However, it is obvious that the welfare implications of the two differ significantly. The analysis presented in this paper is of importance because, at a country level, health care expenses are greatly influenced by the way insurance reimbursement affects health care consumption. Moreover, a better understanding of the impact of different reimbursement methods on social welfare can help towards the implementation of appropriate health reforms.

A stylized description of the two reimbursement methods follows. When reimbursement is in-kind, consumers directly receive the medical services they need. Access to care is essentially free, and to prevent excessive demand for care, the health services available to consumers need to be rationed. As a consequence, in-kind reimbursement implies the regulation of the *quantity* of health services provided.[2] Moreover, as for integrated systems, free access to care implies that health services are paid directly by the insurer. Later on integrated systems will be referred to as QR (quantity-regulated) systems.

On the contrary, in the case of reimbursement insurance, the cost of health services is shared by the patient and the insurer. In particular, consumers are free to choose the quantity of health services they desire at the consumption price, which is determined by the coinsurance parameter. Thus, in reimbursement insurance the *price* of health services is regulated. Since cost-sharing acts as a subsidy to health services, consumers do not internalize the entire health care cost, and consequently they demand an excessive quantity of services. This represents the well-known problem of ex-post moral hazard in health insurance. In Bismarckian systems, health services are generally paid for by consumers who, after submitting an insurance claim, receive partial reimbursement from the insurer. Bismarckian systems will hereafter be referred to as PR (price-regulated) systems.

This simple description of quantity- and price-regulated systems obviously does not cover all of their complex features. However, it provides a workable framework within which to compare the different reimbursement methods adopted by the two systems. A further reason for this comparison is the fact that quantity and price regulation are widely employed in diverse European *public* health insurance systems. Before providing some examples of the respective popularity of the two systems in Europe, we should point out that in practice diverse countries adopt different regulation targets for different health services. Thus, at the national level it is difficult to formulate a generalization of the exploited regulation approach. The following countries have free access to care in the form of general practitioners' services:[3] Germany, Greece, Spain, Italy, the Netherlands and the UK. As far as specialists' services are concerned, these are free in Germany, Spain, the Netherlands and the UK. In the

[1] The terminology *in-kind reimbursement* can be found in Besley and Gouveia (1994), while Besley (1988) uses the term *reimbursement insurance* to describe traditional health insurance based on cost-sharing.

[2] Here *health services* indicate all services provided by physicians and the total amount of treatment patients consume for each illness episode.

[3] Note that in the case of free access to care, health services can be deduced as being quantity-regulated. Take, for example, GP services in Italy. Since each consumer subscribes to the patient list of a single GP, he/she only has free access to *his/her* GP's services. If the consumer requires a second opinion, *he/she must pay* another GP (unless he/she decides to leave the former and subscribe to the patient list of the latter). This is not true in the case of price-regulated systems. For each illness, a French consumer can consult every GP in town, and social insurance always reimburses about 70% of the service cost.

case of France, Belgium and Portugal, on the other hand, price regulation is used for both GP's and specialists' services. Hospital care is completely free for all patients in Denmark, Greece, Spain, Italy, Portugal and the UK.[4]

In this paper, health insurance is public and health care expenditure is financed by income tax. This modeling strategy is supported by the fact that direct taxation is the main source of health care financing in many EU Member States. In particular, the UK, Ireland, Portugal, Spain, Denmark, Sweden, Italy and Finland all earmark a part of fiscal revenue for funding health services.[5] Moreover, in order to simplify matters and focus on the problem of cost containment in public health insurance provision, health care is considered the sole form of public expenditure and is financed entirely by income tax. Hence, in the present model, the social planner has two roles: that of public insurer and that of fiscal authority.[6]

The comparison between QR and PR systems is made within a model where consumers have private information with respect to a parameter which can be interpreted as heterogeneity either in intensity of their preferences for treatment or in the type of illness. In the first part of the paper, the insurance plan is constrained to be uniform in the sense that consumers' heterogeneity is not taken into account. In the case of uniform plans we show that PR systems dominate QR ones from a social welfare point of view. This result depends on the fact that, while uniform QR systems constrain both ill consumer types to the same quantity of health services, in PR ones, consumers choose the preferred quantity of health services given the coinsurance parameter. Moreover, in PR systems, health services are subsidized, and this makes the uniform allocation closer to the utilitarian optimum. The second part of the paper looks at self-selecting plans, i.e. allocations where consumers can choose insurance plans which take into account their health services preference or need. In this case, we show that QR systems dominate PR ones. In fact, by rationing health services within QR systems the social planner is able to partially prevent patients from mimicking. In particular, a QR system corresponds to the *direct mechanism*, which implements the optimal incentive compatible allocation.

The well-known debate on price versus quantity regulation begun by Weitzman (1974) is clearly of importance here. Moreover, the present paper also borrows ideas from economic studies of moral-hazard in health insurance, one of the seminal contributions being that of Zeckhauser (1970). With respect to this, PR systems represent a particular case of the more general reimbursement schedule analyzed in his paper. As for QR systems, reference is made to the literature on in-kind transfers and optimal taxation (including, among others, Cremer & Gahvari, 1997), which analyzes the self-selecting property of in-kind transfers in second-best economies. Finally, the paper deals with income taxation under uncertainty and it is then related to the vast literature in which taxation is used to insure consumers against various types of wage and health risks (see, for example, Cremer & Gahvari, 1995; Varian, 1980). However, our analysis differs from all previous studies inasmuch as it formulates an institutional comparison between two alternative methods of reimbursing health expenses,

[4] See Le Grand and Mossialos (1999) pages 75–83. Their survey of cost sharing for GP and specialist visits and for in-patient care refers to 1996.

[5] See again Le Grand and Mossialos (1999) for a general discussion of sources of finance for health care expenditure in Europe.

[6] Public health insurance was analyzed for the first time by Blomquist and Horn (1984), who showed that, when individuals differ in their earning ability and in the probability of falling ill, a public health insurance funded with linear income tax proves an efficient method of redistributing welfare.

based on quantity and price regulation respectively. Such an analysis seems to cover unexplored terrain, to the best of our knowledge.

The plan of the paper is as follows. The following section describes the model and its assumptions. Section "The structure of alternative insurance plans and the utilitarian optimum" analyzes the utilitarian optimum. In Section "Uniform plans", *uniform* insurance plans are characterized and then compared. In Section "Self-selecting plans", *self-selecting* plans are analyzed and compared. Finally, Section "Conclusion" offers our conclusion.

The model

Consumers' earning ability is normalized to equal the wage rate, and captures their health status. Illness occurs with probability p. With probability $1 - p$ consumers are healthy and their (marginal) labor productivity is w. When ill, consumers lose their earning ability and productivity falls to zero.

Consumers' preferences are state-dependent and separable. Utility is determined by aggregate consumption, labor supply and the consumption of health services as follows:[7]

$$U_1 (C, L) \qquad = u (C_1) - v (L) \tag{1}$$

$$U_2^j (C, X) = u (C_2) - H + \theta^j \phi (X) \tag{2}$$

where the subscript $i = 1, 2$ indicates health status: 1 represents good health, while 2 corresponds to illness. C is an aggregated consumption good taken as numeraire, X is health care consumption and L is labor supply.[8] H is a fixed utility loss which occurs in the case of illness and can be partially recovered through health care consumption. The term $\theta^j \phi (X)$ indicates the benefit from health care consumption, where $\phi (X)$ is health improvement from treatment. As for the parameter $\theta^j, j = l, h$, it can represent either intensity of preferences for health services or the type of illness; in both cases $0 < \theta^l < \theta^h$. In the former interpretation, the parameter θ allows the consideration of different attitudes towards health services. In particular, for the same illness episode, it describes the heterogeneity in the propensity for health service consumption which normally characterizes patients.[9] Whereas, with the latter interpretation, for a given level of health services X, marginal productivity of treatment depends on θ. In particular, the higher is θ, the higher illness is responsive to treatment. Later on in the paper, θ will be interpreted as intensity of preferences for health services. With probability μ_l consumers have low preference for health care consumption (they are low-types), while, with probability $\mu_h = 1 - \mu_l$ they have high preference for health care consumption (they are high-types).

[7] Preferences are similar to that considered in Blackorby and Donaldson (1988) where ill people consume a composite commodity together with health care. The difference with respect to Blackorby and Donaldson is that in the current paper healthy peoples' utility is also affected by labor supply.

[8] With state-dependent utilities, the ill consumers' consumption bundle can be separated into aggregated consumption C_2 and treatment X.

[9] For example, Chernew, Encinosa, Hirth (2000) emphasize the heterogeneity of preferences for health treatment as the cause of variations in expenditure.

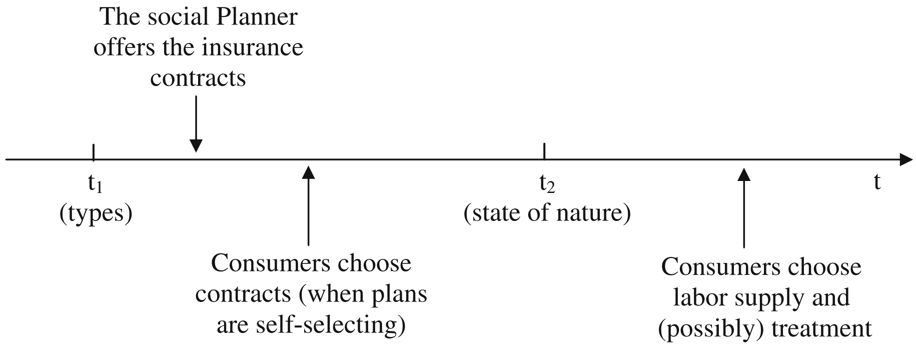
The social Planner offers the insurance contracts
t1 (types)
t2 (state of nature)
t
Consumers choose contracts (when plans are self-selecting)
Consumers choose labor supply and (possibly) treatment

Fig. 1 The timing of the model

Standard hypotheses regarding utility functions hold: $u'(C) > 0, u''(C) < 0; v'(L) > 0, v''(L) > 0, \phi'(X) > 0$ and $\phi''(X) < 0$. Moreover, $H \geq \theta^j \phi(X)$, $\forall j = l, h$ and $\forall X$, so that a consumer's utility is always higher when he/she is in good health than when ill.

Aggregated consumption C_1 always corresponds to labor income wL minus the insurance premium. On the contrary, the structure of aggregated consumption C_2 depends on the type of reimbursement considered. In particular, it depends on whether the social planner (as in the case of quantity regulation), or both the social planner and consumers (as in the case of price regulation), pay for health services.[10] Such a point will be clarified in the next section, where the structure of QR and PR systems will be presented in greater detail.

The social planner will be concerned with making comparisons of utility levels across consumer types. Thus, full comparability of consumer utility is assumed.

The timing of the model is as follows: at t_1 (interim) consumers learn of their type and at t_2 (ex-post) the health-risk is realized and consumers learn of their state of health too. As Fig. 1 shows, the social planner decides the insurance plan at the interim stage, while consumers choose ex-post either labor supply or health services (the latter only in the case of price regulation). In the case of self-selecting plans, at the interim stage the social planner offers consumers the menu of contracts for the two types, and consumers choose the contract they prefer.

The described model can be used to analyze both the case of a public provider vertically integrated with the public insurer (as in those integrated systems based on quantity regulation) and the case of a private provider in a competitive market (as in the case of Bismarckian systems based on price regulation). Assuming a linear technology, the unitary cost of health care is constant. Thus we can say that consumers and the public insurer are faced with the same health service price (q). Moreover, when consumers are reimbursed through cost-sharing (PR system), they choose the quantity of health services they require. This implies that the provider behaves as a perfect agent for its patients.[11]

Consumers may privately know both their marginal labor productivity w (capturing the health status) and their type θ^j (high/low propensity for health service consumption). The paper investigates the welfare implications of integrated and Bismarckian

[10] Note that, since consumers have no exogenous revenue, health insurance also plays the role of a *disability insurance*. In fact, in the case of illness, consumers have no resources to devote either to treatment or to aggregate consumption.

[11] A certificate of illness from the physician is necessary in order to obtain insurance reimbursement.

systems by analyzing and comparing ill consumers' allocations. In order to obtain clear results, we assume that health status is observable. Hence, in the present model, θ^j is consumers' private information, while marginal labor productivity is common knowledge. This implies that a healthy consumer cannot mimic illness by choosing $L = 0$: the provider always behaves as a perfect agent for the insurer too. This assumption is made in order to sharpen the analysis, although it is wholly plausible in integrated health systems.[12] On the contrary, when the provider is a private agent in a competitive market, imposing the observability of the health status is equivalent to assuming that collusion between patient and physician is impossible.[13,14] A justification for this assumption relies on the legislation which in several countries allows an employer to check that the absent employee is not mimicking illness.[15] On the contrary, note that, when θ represents the type of illness, the social planner is not fully informed as for the health status; in fact he is able to detect whether one consumer is ill, but he cannot discern whether the illness is treatment-intensive or not.

Unlike marginal labor productivity, preferences for health services are not observable by the insurer (or by the physician), and the public insurer may be constrained to adopt uniform plans or may be free to choose self-selecting ones.

The structure of alternative insurance plans and the utilitarian optimum

The social planner maximizes the utilitarian social welfare function $SW = \mu_l EU(\theta^l) + \mu_h EU(\theta^h)$, where $EU(\theta^l)$ is low-type and $EU(\theta^h)$ is expected utility of high-type consumers. The two expected utilities are, respectively, multiplied by the proportion of low-type and high-type consumers in the population.[16]

In this section the social planner observes consumer preferences for health services. Thus, the utilitarian optimum can be implemented. It can be decentralized by a contract contingent upon both the health status and the preference for health services; that is, by a plan characterized by four non-uniform monetary transfers $(P^j, R^j), j = l, h$.[17]

[12] In a different context, Chernew et al. (2000) assume that individuals have *observable* severe diseases and unobservable preferences for alternative treatments. They justify this assumption (on page 589) on the grounds of vertical integration.

[13] This simplification allows to make the analysis treatable. See Alger and Ma (2003) for a model which considers collusion between patient and physician.

[14] This model does not consider the much-debated problem of physician's "induced demand" which can be of importance in Bismarckian systems with price regulation. Anyway, induced demand turns out to increase consumption of health services in PR systems. In other words, it induces overconsumption of care exactly as ex-post moral hazard does. Thus, such a phenomenon could be implicitly taken into account by increasing the level of ex-post moral hazard in the model, i.e. the sensibility of demand for health services to cost-sharing. The results provided by the model are robust to this extension.

[15] It is a well known fact that some physicians play the role of public inspectors for the social insurance institution. Social insurance assures a partial wage provision in the case of absence from work due to illness. This monitoring activity consists of a medical check on employees on sick leave.

[16] Considering a large number of consumers, μ_j is equivalent, ex-post, to the proportion of the j-type.

[17] The utilitarian optimum specifies all the terms of healthy and ill consumers' utility $(C_i^j, L^j, X^j), j = l, h$ and $i = 1, 2$. However the choice of X^j and L^j can be decentralized because consumers face efficient prices w and q. Thus, the utilitarian optimum is obtained by offering the contract $(P^j, R^j), j = l, h$ and letting consumers choose (ex-post) either labor supply or treatment quantity.

Consumption in the case of the two states of health is:

$$C_1^j = wL - P^j, \quad j = l, h,$$
$$C_2^j = R^j - qX^j, \quad j = l, h,$$

where P^j is the premium paid by healthy consumers, and R^j is the reimbursement in the case of illness for the two consumer types.

The social planner solves:

$$\left\{ \begin{array}{ll} \underset{P^j, R^j}{\text{Max}} & \Sigma_{j=l,h}\, \mu_j \left\{(1-p)\left[u(wL^j - P^j) - v(L^j)\right]\right. \\ & \qquad \left. +p\left[u\left(R^j - qX^j\right) - H + \theta^j \phi\left(X^j\right)\right]\right\} \\ & \\ \text{s.t.:} & (1-p)\sum_{j=l,h} \mu_j P^j = p \sum_{j=l,h} \mu_j R^j. \end{array} \right. \tag{P1}$$

Note that the expected contribution of healthy consumers is equivalent to the expected cash transfers for ill consumers.[18]

The full-insurance result follows from FOCs:[19]

$$C_1^l = C_1^h = C_2^l = C_2^h = C. \tag{3}$$

Moreover:

$$L = L(w, P): \quad wu'(C) = v'(L) \tag{4}$$
$$X^j = X(\theta^j, R^j, q): \quad \theta^j \phi'\left(X_2^j\right) = qu'\left(C\right), \quad j = l, h. \tag{5}$$

For both labor supply and health care quantity marginal benefit equals marginal cost when aggregated consumption is optimal. Since healthy consumers are characterized by the same wage rate w and full-insurance for aggregate consumption is optimal (see note 19), both types provide the same labor supply L. This implies that lump-sum taxes imposed on healthy consumers are equal: $P^l = P^h = P$. As a consequence, the utilitarian optimum can be achieved by setting only three non-uniform monetary transfers: (P, R^l, R^h).

In state of health 2, as expected, $X^h > X^l$ and $R^h > R^l$: high type consumers receive a higher monetary transfer and buy a greater quantity of health services.

In Fig. 2, ill consumers' allocation is shown when preference for health services is observable. The slope of the low-type utility function is higher than the high-type one: in fact, $\frac{\mathrm{d}C}{\mathrm{d}X} = -\theta^j \frac{\phi'(X^j)}{u'(C^j)}$.

The rest of this section details the structure of reimbursement in QR and PR systems given full information. The two systems will be specifically treated in the case of asymmetric information in sections "Quantity-regulated systems" and "Price-regulated systems".

[18] Because of the way the parameter θ^j enters the utility functions, social welfare is increasing with treatment propensity. In other words, high-type consumers have the highest weight in this economy since they benefit more from health care consumption.

[19] Separability between aggregate consumption and treatment consumption is assumed. Thus illness does not alter the marginal utility of income even if utilities are state-dependent. As a consequence, full insurance is still optimal and it obviously concerns only aggregate consumption.

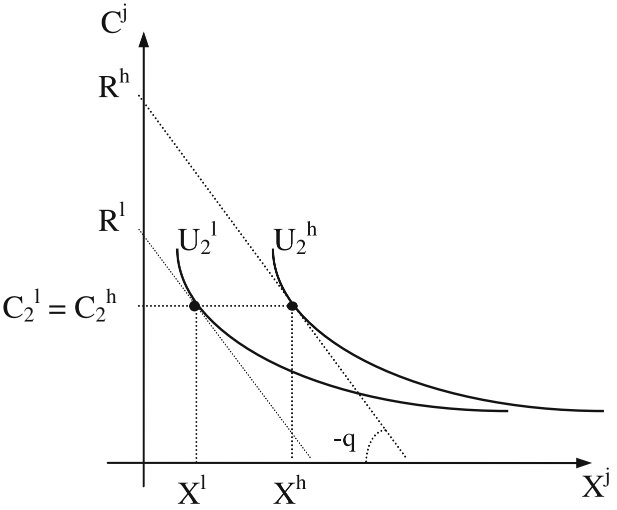

Fig. 2 Ill consumers' allocation in the utilitarian optimum

Quantity-regulated insurance systems. When reimbursement is quantity regulated, the social planner uses rationing and provides a ceiling to the available amount of services $\bar{X}$ as a measure against overconsumption induced by free access to health care. Since the wage rate is observable, the social planner will always use a monetary transfer contingent upon ill-health. Moreover, with full information on preferences, such a monetary transfer and the ceiling to health services will be contingent upon the ill consumer type. As a consequence, QR plans are characterized by $(P^{\mathrm{QR}}, R^{\mathrm{QR}j}, \bar{X}^j), j = l, h$: that is, by three monetary transfers and two care-packages. Health care is free and utility is increasing in health care, then patients entirely consume amount $\bar{X}^j$. The previous interpretation of QR systems is rather stylized; nevertheless, it renders the model solvable and represents a good approximation of reality in many situations.

Consumption in the case of illness is:

$$C_2^{\mathrm{QR}j} = R^{\mathrm{QR}j}, \quad \text{with } X^j = \bar{X}^j,$$

where $j = l, h$. Note that ill consumers' aggregate consumption is exactly equivalent to the transfer $R^{\mathrm{QR}j}$, while health care consumption corresponds to the package of care $\bar{X}^j$. Obviously, with full information this does not represent a constraint for consumers because $R^{\mathrm{QR}j}$ and $\bar{X}^j$ is the amount consumers would have chosen given efficient prices 1 and q.

Price-regulated insurance systems. In PR systems, consumers pay a part of health care costs. The linear[20] cost-sharing parameter is denoted by α^j $(j = l, h)$, and the contract is characterized by $(P^{\mathrm{PR}}, R^{\mathrm{PR}j}, \alpha^j), j = l, h$: that is, by three monetary transfers and two coinsurance parameters. Consumers choose the preferred health care quantity at price $\alpha^j q$.

[20] Bismarckian insurance systems are generally characterized by linear cost-sharing parameters. However the insurer *could ex-post verify* health care consumption since reimbursement is based on the provider's bill. Thus more complex, non-linear mechanisms could be implemented (although these mechanisms are usually not employed), and the resulting analysis would be considerably more complicated. (See Blomquist, 1997 for a model with non-linear health insurance)

Consumption in the case of illness is:

$$C_2^{\mathrm{PR}j} = R^{\mathrm{PR}j} - \alpha^j q X^j$$

Remark 1 Given full information, both quantity- and price-regulated systems permit the implementation of the utilitarian optimum.[21]

In fact, in both QR and PR systems, the social planner can use two additional "instruments" with regard to program P1, (respectively, $\bar{X}^j$ in QR and α^j in PR, $j = l, h$) so that it may perform at least as well as in the utilitarian optimum. Clearly, when monetary transfers contingent upon health services preference are available, such additional instruments are of no use. In the case of PR systems, with full information the social planner clearly sets $\alpha^j = 1$ so that prices are not distorted.

It is obvious from Fig. 2 that, when θ^j is not observable, the utilitarian optimum cannot be implemented because low-type consumers would imitate high-type ones. The optimal allocation of resources is not envy-free.

Dealing with low-type incentive constraints, the public insurer has a choice of two kinds of insurance plan: uniform or self-selecting policies.

Which kind of allocation does the social planner implement in the real world: the uniform or the self-selecting one? Does the social planner discriminate reimbursement according to patients' tastes? Whenever consumers have access to different qualities of health service or to different treatment options, the answer is "yes". One simply has to consider the availability of single room hospitalization in National-Health-Service type organizations; or those cases in which patients' preferences are a determinant of treatment choice, and plans allow enrollees options between diverse health service paths.[22] On the other hand, there are many situations in which the insurance plan is uniform; that is, where the same reimbursement is provided to all consumers with the same illness without taking account of different preferences for health services. In the case of uniform reimbursement, the possibility of satisfying different propensities for health services is generally left to private insurance: individuals with a high preference for health services can buy *supplementary* insurance or *opt-out* of the public sector (such behavior is not explicitly analyzed here).

Since uniform and self-selecting plans are equally plausible, the paper investigates them both: uniform allocations are treated in section "Uniform plans", whereas self-selecting ones are analyzed in section "Self-selecting plans".

Uniform plans

With uniform policies, the same contract is offered to low- and high-type consumers. This implies that both types of consumers are subject to the same budget constraint, both when healthy and when ill. Given that labor productivity observability implies that reimbursement can be contingent upon health status, public insurance will always reimburse ill consumers in the form of a monetary transfer (but, it will clearly not discriminate reimbursement according to propensity for health service consumption).

[21] This remark confirms the views of Arrow (1963) when he says that, in a hypothetically perfect market, the diverse methods of covering health care expenses should be equivalent. (Arrow, 1963, page 962)

[22] In the US, managed care organizations broadly rely on shared decision-making (SDM) and disease carve-out programs to facilitate the inclusion of patients' preferences in the decision-making process (see Chernew et al., 2000).

Quantity-regulated systems

With the uniformity constraint, the plan is characterized by two monetary transfers and by a package of care: $(P^{\mathrm{QR}}, R^{\mathrm{QR}}, \bar{X})$.

Individuals' consumption in the case of the two states of health is:

$$\begin{aligned} C_1^{\mathrm{QR}} &= wL - P^{\mathrm{QR}} \\ C_2^{\mathrm{QR}j} &= R^{\mathrm{QR}}, \quad \text{with } X^{\mathrm{QR}j} = \bar{X},\ j = l, h. \end{aligned}$$

With respect to the utilitarian optimum, two constraints have been added: $R^h = R^l = R^{\mathrm{QR}}$ and $\bar{X}^l = \bar{X}^h = \bar{X}$, so that the allocation of ill consumers is completely determined.

The healthy consumers' program is:

$$\max_L \ u(wL - P^{\mathrm{QR}}) - v(L).$$

Then labor supply is defined according to the following equation:

$$L^* = L(w, P^{\mathrm{QR}}): \quad wu'(C_1) = v'(L). \tag{6}$$

The public insurance program is:

$$\left\{ \begin{aligned} &\underset{P^{\mathrm{QR}}, R^{\mathrm{QR}}, \bar{X}}{\text{Max}} \quad (1-p)\left[u(wL^* - P^{\mathrm{QR}}) - v(L^*)\right] \\ &\qquad\qquad +p\left[u\left(R^{\mathrm{QR}}\right) - H + \theta_M \phi\left(\bar{X}\right)\right] \\ &\text{s.t.:} \qquad (1-p)\,P^{\mathrm{QR}} = p\left(R^{\mathrm{QR}} + q\bar{X}\right), \end{aligned} \right. \tag{P2}$$

where $\theta_M = \sum_{j=l,h} \mu_j \theta^j$. In fact the social planner maximizes the utility of the θ_M-type consumer. Not surprisingly, from FOCs with respect to P^{QR} and R^{QR} the full-insurance condition is verified:

$$C_1^{\mathrm{QR}} = C_2^{\mathrm{QR}} = \bar{C}. \tag{7}$$

Moreover, the package of health services is determined according to the following equation:

$$\bar{X} = X\left(\theta_M, q, \bar{C}\right): \qquad \theta_M \phi'\left(\bar{X}\right) = qu'(\bar{C}). \tag{8}$$

Clearly neither type of ill consumer receives the optimal quantity of health services (determined by Eq. 5). Such a uniform contract imposes the same allocation $\left(\bar{C}, \bar{X}\right)$ to both types of ill consumer and their utility is: $U_2^j = u(\bar{C}) - H + \theta^j \phi(\bar{X}),\ j = l, h$. So that: $U_2^h - U_2^l = \phi\left(\bar{X}\right)\left(\theta^h - \theta^l\right) > 0$. This inequality shows that high-type utility is still greater than low-type utility: as in the utilitarian optimum, low-type consumers are characterized by a lower utility level.

It is clear that, if there is no heterogeneity ($\theta^l = \theta^h$), the utilitarian optimum can be obtained.

Price-regulated systems

The uniform plan is characterized by two monetary transfers and by a coinsurance parameter: $\left(P^{\mathrm{PR}}, R^{\mathrm{PR}}, \alpha\right)$. Individuals' consumption, given the two states of health, is:

$$\begin{aligned} C_1^{\mathrm{PR}} &= wL - P^{\mathrm{PR}} \\ C_2^{\mathrm{PR}j} &= R^{\mathrm{PR}} - \alpha q X^j \qquad j = l, h. \end{aligned}$$

With regard to the utilitarian optimum, two uniformity constraints have been added: $R^h = R^l = R^{\mathrm{PR}}$ and $\alpha^h = \alpha^l = \alpha$.

Healthy consumers' decisions are the same as before, and Eq. 6 still holds with P^{PR} instead of P^{QR}. In the case of illness, on the other hand, consumers solve:

$$\max_{X^j} \; u\left(R^{\mathrm{PR}} - \alpha q X^j\right) - H + \theta^j \phi\left(X^j\right).$$

As a consequence, the chosen quantity of health services is such that:

$$X^{*j} = X(\theta^j, \alpha, q, R^{\mathrm{PR}}): \quad \theta^j \phi'\left(X^j\right) = \alpha q u'\left(C_2^j\right), \quad j = l, h. \tag{9}$$

The public insurance program is:

$$\left\{ \begin{aligned} &\underset{P^{\mathrm{PR}}, R^{\mathrm{PR}}, \alpha}{\mathrm{Max}} \quad (1-p)\left[u(wL^* - P^{\mathrm{PR}}) - v(L^*)\right] \\ &\qquad\qquad + p_{\sum_{j=l,h}} \mu_j \left[u\left(R^{\mathrm{PR}} - \alpha q X^{*j}\right) - H + \theta^j \phi\left(X^{*j}\right)\right] \\ &\text{s.t.:} \qquad (1-p)\, P^{\mathrm{PR}} = p\left[(1-\alpha)\, q \sum_{j=l,h} \mu_j X^{*j} + R^{\mathrm{PR}}\right]. \end{aligned} \right. \tag{P3}$$

From FOCs with respect to P^{PR} and R^{PR} we get the following equation:

$$E\left[u'(C_2^{\mathrm{PR}})\right] = u'(C_1^{\mathrm{PR}})\left[1 + (1-\alpha) q E\left[\frac{\partial X}{\partial R^{\mathrm{PR}}}\right]\right], \tag{10}$$

where $E\left[\dfrac{\partial X}{\partial R^{\mathrm{PR}}}\right] = \mu_l \dfrac{\partial X^l}{\partial R^{\mathrm{PR}}} + \mu_h \dfrac{\partial X^h}{\partial R^{\mathrm{PR}}}$. By totally differentiating Eq. 9, we get $\dfrac{\partial X^j}{\partial R^{\mathrm{PR}}} > 0$, so that $E\left[\dfrac{\partial X}{\partial R^{\mathrm{PR}}}\right] > 0$.

From FOC with respect to the coinsurance parameter α it follows:

$$\left[-(1-\alpha)\, E\left[\frac{\partial X}{\partial \alpha}\right] + E(X)\right] u'\left(C_1^{\mathrm{PR}}\right) = E\left[X u'(C_2)\right], \tag{11}$$

where $E\left[\dfrac{\partial X}{\partial \alpha}\right] = \mu_l \dfrac{\partial X^l}{\partial \alpha} + \mu_h \dfrac{\partial X^h}{\partial \alpha}$, $E(X) = \mu_l X^l + \mu_h X^h$ and $E\left[X u'(C_2)\right] = \mu_l X^l u'\left(C_2^l\right) + \mu_h X^h u'\left(C_2^h\right)$. By totally differentiating Eq. 9, we get $\dfrac{\partial X^j}{\partial \alpha} < 0$, so that $E\left[\dfrac{\partial X}{\partial \alpha}\right] < 0$. Moreover, $E\left[X u'(C_2)\right] = \mathrm{cov}\left[X, u'(C_2)\right] + E(X)\, E\left[u'(C_2)\right]$.

One initial observation is that α is always different from 1 with a positive level of heterogeneity.[23] In particular, we shall show later, α is always lower than 1. On the contrary, when there is no heterogeneity it is optimal to set $\alpha = 1$, and the utilitarian optimum is obtained.

The interpretation of Eq. 11 is as follows: the left-hand side represents the consumers' marginal cost, while the right-hand side represents the consumers' marginal benefit from a negative variation in α (a fall in health service price). When α decreases, consumers' out-of-pocket expenses decrease as well, while insurance reimbursement expenses increase. As a consequence, insurance premiums must increase. Marginal cost is measured by the marginal variation in the insurance premium (in brackets) multiplied by the marginal utility of consumption given the healthy state. In fact the premium is paid by healthy consumers. On the right-hand side, the positive income effect from a negative variation in α is measured by the product of health service quantity and the marginal utility of consumption in the case of illness. Mean values appear because a uniform plan is considered.

From Eqs. 10 and 11, the optimal coinsurance parameter can be written as:

$$\alpha = 1 + \frac{\operatorname{cov}\left[X, u'(C_2)\right]}{E\left[u'(C_2)\right] E\left[\dfrac{\partial X}{\partial \alpha}\right] + qE\left[Xu'(C_2)\right] E\left[\dfrac{\partial X}{\partial R^{\mathrm{PR}}}\right]}. \tag{12}$$

Remark 2 In uniform, price-regulated systems, health services are always subsidized.

Proof Let us consider Eq. 12. $\operatorname{cov}\left[X, u'(C_2)\right]$ is positive. The denominator on the r.h.s. of 12 can be rewritten as: $\mu_h^2 u'(C_2^h)\left(\dfrac{\partial X^h}{\partial \alpha} + qX^h \dfrac{\partial X^h}{\partial R^{\mathrm{PR}}}\right) + \mu_l^2 u'(C_2^l)$ $\left(\dfrac{\partial X^l}{\partial \alpha} + qX^l \dfrac{\partial X^l}{\partial R^{\mathrm{PR}}}\right) + \mu_l\mu_h u'(C_2^h)\left(\dfrac{\partial X^l}{\partial \alpha} + qX^h \dfrac{\partial X^l}{\partial R^{\mathrm{PR}}}\right) + \mu_l\mu_h u'(C_2^l)\left(\dfrac{\partial X^h}{\partial \alpha} + qX^l \dfrac{\partial X^h}{\partial R^{\mathrm{PR}}}\right)$, where $\dfrac{\partial X^j}{\partial \alpha} + qX^j \dfrac{\partial X^j}{\partial R^{\mathrm{PR}}}$, $j = l, h$, corresponds to compensated demand for health services with respect to health service price, which is negative. It follows that the entire previous expression is negative too. Thus the denominator on the r.h.s. of 12 is negative. This implies that $\alpha < 1$. □

The covariance with respect to X and $u'(C_2)$ reflects the objective of risk sharing: the more consumers are risk averse, the larger is covariance and the more health services are subsidized. At the same time, the coinsurance parameter α negatively affects the mean derivative with respect to α of health service demand (the term $E\left[\dfrac{\partial X}{\partial \alpha}\right]$). This term can be seen as a measure of moral hazard. In other words, $\frac{\partial X}{\partial \alpha}$ is related to the price elasticity of demand for health services, so that Eq. 12 recalls the inverse elasticity rule in Ramsey taxation: the more health service demand is inelastic, the more health services are subsidized.[24] Ill consumers' allocations in uniform PR systems are shown in Fig. 3.

[23] In fact, from Eq. 11, $\alpha = 1$ implies $u'(C_1) = \dfrac{E\left[Xu'(C_2)\right]}{E(X)}$; and from Eq. 10, $\alpha = 1$ implies $u'(C_1) = E\left[u'(C_2)\right]$. This means that $E\left[u'(C_2)\right] E(X) = E\left[Xu'(C_2)\right]$, which is impossible because C_2 also depends on X.

[24] Results from the RAND Health Insurance Experiment show that health care price elasticities fall within the range [–0.1, –0.2] (See Manning et al. 1987).

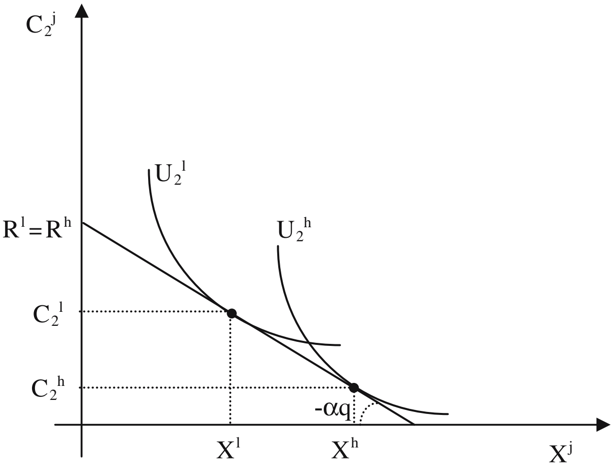

Fig. 3 Ill consumers' allocation with uniform plans in price regulated systems

Comparing the alternative uniform plans

The comparison between QR and PR uniform plans is not a trivial one. In fact, on the one hand, in QR systems ill consumers with heterogeneous preferences are constrained to the uniform rationing of health care consumption. On the other hand, in PR systems health services are subsidized and the moral hazard problem arises. The following result holds:

Proposition 1 *With heterogenous consumers (i.e. $\theta^l \neq \theta^h$) and uniform plans, price-regulated systems dominate quantity-regulated ones.*

Proof This proof is based on the comparison between the allocation of ill consumers implemented in uniform QR and PR systems and the same allocation implemented in the uniform "purely cash" insurance plan (P, R). (i) First of all, let us consider the comparison between the QR plan, characterized by the instruments $(P^{\mathrm{QR}}, R^{\mathrm{QR}}, \bar{X})$, and the "purely cash" one. In the "purely cash" plan, given the monetary transfer R, ill consumers choose the preferred quantity of treatment under undistorted prices 1 and q. Thus, high-type consumers buy more treatment than low-type ones. In QR systems, on the contrary, the package of health services $\bar{X}$ constrains both ill consumer types to consume an amount of health services which does not correspond to the preferred one (see Eq. 8). Thus, from a social welfare point of view, the uniform, "purely cash" insurance plan (P, R) strictly dominates the QR system $(P^{\mathrm{QR}}, R^{\mathrm{QR}}, \bar{X})$. (ii) Let us now compare the PR plan $(P^{\mathrm{PR}}, R^{\mathrm{PR}}, \alpha)$, and the "purely cash" one. In PR systems, given health services' price αq, each ill consumer type obtains the preferred quantity of health care. Moreover, Remark 2 proves that the *optimal* coinsurance parameter is different from one, which in turn implies that distorting health service price is welfare improving. Thus, from a social welfare point of view, the PR system $(P^{\mathrm{PR}}, R^{\mathrm{PR}}, \alpha)$ strictly dominates the uniform "purely cash" insurance plan (P, R).

The previous discussion shows that insurance plans can be ranked as follows: uniform PR systems dominate uniform "purely cash" plans which, in turn, dominate uniform QR systems. □

To get an intuition of this result, note that parameter α modifies the cost of health services so that a positive effect on social welfare is obtained. In particular, the subsidy imposed on health care partially mitigates the distortion induced by the uniformity constraint. The slope of the ill consumers' budget constraint increases: health care becomes relatively cheaper. Given that high-type consumers are characterized by a higher propensity towards health service consumption, such a change in relative prices implies that high-type consumers are relatively better off, while ill consumers' allocation moves closer to the utilitarian optimum.

As a final remark on the comparison of the two systems, let us consider the effect of consumers' heterogeneity on their relative performance. As it has been previously discussed, the two systems are equivalent with no heterogeneity (i.e. $\theta^l = \theta^h$), this result together with Proposition 1 allow to state that

Corollary 1 *With uniform plans, consumers' heterogeneity impairs quantity-regulated systems comparatively more than price-regulated ones.*

The intuition for this result is a direct consequence of imposing less flexibility and choice within quantity-regulated systems as compared to price-regulated ones where consumers have the possibility to partially adapt their decisions to their preferences.

Interestingly enough, this effect of heterogeneity will be dramatically affected when considering self-selecting plans, as it is shown in the next section.

Self-selecting plans

When dealing with self-selecting allocations, the previous ranking of QR and PR systems may be substantially affected. One important point is that the rationing of the quantity of health services in QR systems now becomes a useful instrument. By directly providing free health services, the social planner can observe the quantity of treatment consumed by ill individuals. In PR systems, on the contrary, given the linear coinsurance parameter α^j, consumers choose the preferred quantity of health services. As stated in note 20, in the real world, the opportunity to ex-post verify the quantity of health services consumed by patients is not "exploited" by the public insurer. Thus, in practice, in PR systems, the quantity of health services is not observable and cheating on health care consumption arises.

The social planner's programs addressed in this section are standard cases of mechanism design under adverse selection. In order to define the optimal mechanism for both insurance schemes, we are going to use the well known Revelation Principle.[25] Hence, the direct mechanisms by which consumers (truthfully) announce their type θ and the insurer offers an allocation which specifies all the relevant variables in the contractual relationship with consumers, will be considered.

The social planner's optimal allocation attainable *within each* reimbursement scheme will be analyzed. In particular the instruments available to the social planner will be $(P^{\mathrm{QR}j}, R^{\mathrm{QR}j}, \bar{X}^j)$, $j = l, h$, in the case of QR systems, and $(P^{\mathrm{PR}j}, R^{\mathrm{PR}j}, \alpha^j)$, $j = l, h$, in the case of PR systems.

[25] Myerson (1979), among others.

Quantity-regulated systems

As explained in Section "The structure of alternative insurance plans and the utilitarian optimum", when access to care is free, $\bar{X}^j$ is always entirely consumed. This implies that the social planner can observe health service consumption, as well as ill consumers' aggregated consumption. Moreover, marginal labor productivity w is observable too, and the premiums $P^{\mathrm{QR}j}$ can be chosen by the social planner to induce the desired labor supply and aggregated consumption in the healthy state. As a consequence, the contracts proposed by the social planner to low- and high-type consumers are, respectively, $(C_1^{\mathrm{QR}l}, L^l, C_2^{\mathrm{QR}l}, \bar{X}^l)$ and $(C_1^{\mathrm{QR}h}, L^h, C_2^{\mathrm{QR}h}, \bar{X}^h)$. It is interesting to note that QR systems represent the unconstrained direct mechanism in that, given the agent's type announcement, all the relevant variables are chosen by the social planner. As a consequence, we may deduce that the QR optimal allocation corresponds to the allocation which weakly dominates the others.

In order to have consumers truthfully report their type, the social planner has to maximize his objective function (also) under the incentive compatibility constraints. The social planner's program is, thus:

$$\left\{\begin{array}{ll} \underset{C_i^{\mathrm{QR}j}, L^{\mathrm{QR}j}, \bar{X}^j}{\mathrm{Max}} & \sum_{j=l,h_j} \mu \left\{(1-p)\left[u(C_1^{\mathrm{QR}j}) - v(L^{\mathrm{QR}j})\right]\right. \\ & \left. \quad + p\left[u\left(C_2^{\mathrm{QR}j}\right) - H + \theta^j \phi\left(\bar{X}^j\right)\right]\right\} \\ \text{s.t.:} & (1-p)\sum_{j=l,h} \mu_j \left(wL^{\mathrm{QR}j} - C_1^{\mathrm{QR}j}\right) = p \sum_{j=l,h} \mu_j \left(C_2^{\mathrm{QR}j} + q\bar{X}^j\right) \qquad (\gamma) \\ & (1-p)\left[u(C_1^{\mathrm{QR}l}) - v(L^{\mathrm{QR}l})\right] + p\left[u\left(C_2^{\mathrm{QR}l}\right) - H + \theta^l \phi\left(\bar{X}^l\right)\right] \\ & \geq (1-p)\left[u(C_1^{\mathrm{QR}h}) - v(L^{\mathrm{QR}h})\right] + p\left[u\left(C_2^{\mathrm{QR}h}\right) - H + \theta^l \phi\left(\bar{X}^h\right)\right] \qquad (\lambda), \end{array}\right. \tag{P4}$$

where $\gamma \neq 0$ and $\lambda \geq 0$ represent, respectively, the budget constraint Lagrange multiplier and the incentive constraint Khun Tucker multiplier.[26]

Lemma 1 and Fig. 4 describe the structure of the self-selecting allocation in QR systems.

Lemma 1 *The optimal self-selecting allocation in quantity-regulated systems is such that: (i) in good health, low-type consumers are better off, in fact $C_1^{QRl} \geq C_1^{QRh}$ and $L^{QRl} \leq L^{QRh}$ hold. Labor supplies are not distorted. (ii) When consumers are ill, high-types are better off with $C_2^{QRl} \geq C_2^{QRh}$ and $X^{QRl} \leq X^{QRh}$. There is no distortion for low-type consumers, while high-type consumers are forced to consume too many health services and too little aggregate consumption.*

Proof See appendix □

The incentive constraint is binding in expected terms, and the optimal allocation is such that utility is higher for low-type consumers in good health and for high-type

[26] As stated at the end of section "The structure of alternative insurance plans and the utilitarian optimum", low-type consumers are the mimickers. Standard mechanism design techniques with discrete types (see Fundenberg & Tirole (1991), pages 246–250) prove that it is optimal to make the mimickers' incentive compatibility constraints binding. As shown in the proof of Lemma 1, the incentive compatibility constraint of the high-type is always satisfied by the solution to program P4. It is then irrelevant and thus omitted.

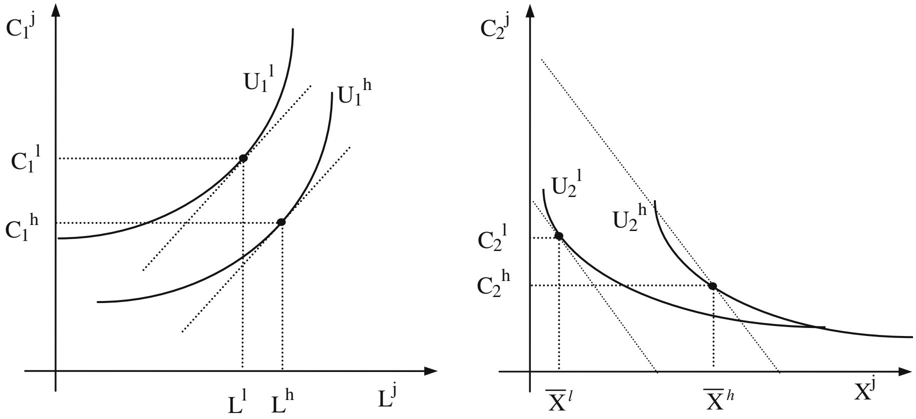

Fig. 4 On the left the allocation of healthy consumers and on the right the allocation of ill consumers with self-selecting plans in price-regulated systems

ones in illness. Since low-type consumers are better off when healthy, distortions in the illness state between treatment and aggregate consumption of high-type consumers can be reduced, while a lower degree of inefficiency is required to prevent mimicking. Note that when ill, high-type consumers are better off as in the case of the utilitarian optimum (see Fig. 2). The difference with respect to the utilitarian optimum is in the allocation for consumers in good health: here, high-type individuals consume less and supply more labor than low-types ones.

As stated before, QR systems correspond exactly to the direct mechanism in this adverse selection setting. Consumers announce their type and receive the second-best allocation. All the other relevant decisions are taken by the social planner.

Price-regulated systems

Within price regulated systems, the contracts proposed by the social planner to low- and high-type consumers are respectively $(P^{\mathrm{PR}l}, L^l, R^{\mathrm{PR}l}, \alpha^l)$ and $(P^{\mathrm{PR}h}, L^h, R^{\mathrm{PR}h}, \alpha^h)$. In other words, for each ill consumer type, the contract specifies a monetary transfer and a subsidy for health services, that is, *a budget constraint*. In this adverse selection problem, the PR insurance plan constitutes an indirect mechanism. In fact, consumers make their decisions after the social planner has chosen the terms of the insurance contract (the lump-sum premium for the healthy and the budget constraint for the ill), and no additional communication with the social planner occurs.

Since it can be assumed that the quantity of health services cannot be observed by the social planner (see the discussion at the beginning of this section), ill consumers will choose this quantity according to Eq. 13:

$$X^{*j} = X\left(\theta^j, R^{\mathrm{PR}j}, \alpha^j, q\right): \quad \theta^j \phi'(X) - \alpha^j q u'\left(R^{\mathrm{PR}j} - \alpha^j q X\right) = 0, \tag{13}$$

while mimickers will choose the quantity of preferred health services according to Eq. 14:

$$X^{*jk} = X\left(\theta^j, R^{\mathrm{PR}k}, \alpha^k, q\right): \quad \theta^j \phi'(X) - \alpha^k q u'\left(R^{\mathrm{PR}k} - \alpha^k q X\right) = 0, \tag{14}$$

where *j* is the true type and *k* is the feigned type.

The social planner's program is thus:

$$\begin{cases} \underset{P^{\mathrm{PR}j},L^{\mathrm{PR}j},R^{\mathrm{PR}j},\alpha^j}{\mathrm{Max}} \quad \sum_{j=l,h} \mu_j \left\{(1-p)\left[u(wL^{\mathrm{PR}j} - P^{\mathrm{PR}j}) - v(L^{\mathrm{PR}j})\right] \right. \\ \qquad\qquad\qquad \left. +p\left[u\left(R^{\mathrm{PR}j} - \alpha^j qX^{*j}\right) - H + \theta^j \phi\left(X^{*j}\right)\right]\right\} \\ \text{s.t.:} \quad (1-p)\sum_{j=l,h}\mu_j P^{\mathrm{PR}j} = p\sum_{j=l,h}\mu_j\left(R^{\mathrm{PR}j} + \left(1-\alpha^j\right)qX^{*j}\right) \qquad (\gamma) \\ (1-p)\left[u(wL^{\mathrm{PR}j} - P^{\mathrm{PR}j}) - v(L^{\mathrm{PR}j})\right] + p\left[u\left(R^{\mathrm{PR}j} - \alpha^j qX^{*j}\right) + \theta^j\phi\left(X^{*j}\right)\right] \geq \\ (1-p)\left[u(wL^{\mathrm{PR}k} - P^{\mathrm{PR}k}) - v(L^{\mathrm{PR}k})\right] + p\left[u\left(R^{\mathrm{PR}k} - \alpha^k qX^{*jk}\right) + \theta^j\phi\left(X^{*jk}\right)\right] \quad (\lambda_j), \end{cases} \tag{P5}$$

Where $\gamma \neq 0$ and $\lambda_j \geq 0, j = l,h$, represent respectively the budget constraint Lagrange multiplier and the incentive constraint Khun Tucker multipliers.

Since the main focus here is on ranking quantity- and price-regulated systems, a few properties of self-selecting allocations in price-regulated systems are briefly discussed when, as in program P4, low-types mimic high-types. Let us consider healthy consumers. One can show that low-type ones are better off: $u\left(C_1^{\mathrm{PR}l}\right) - v(L^{\mathrm{PR}l}) \geq u\left(C_1^{\mathrm{PR}h}\right) - v(L^{\mathrm{PR}h})$. In particular $C_1^{\mathrm{PR}l} \geq C_1^{\mathrm{PR}h}$ and $L^{\mathrm{PR}l} \leq L^{\mathrm{PR}h}$ hold, labor supplies are not distorted and $P^{\mathrm{PR}l} \leq P^{\mathrm{PR}h}$. The previous allocation for healthy consumers can also be obtained in QR systems, the difference here being that the optimal labor supply is "decentralized" through the lump-sum premium $P^{\mathrm{PR}j}$. On the contrary, when consumers are ill, it is easy to show that the inequality $C_2^{\mathrm{PR}l} \geq C_2^{\mathrm{PR}h}$ holds, and as far as cost-sharing parameters are concerned, that the price of health services for low-type consumers is not distorted, i.e. $\alpha^l = 1$, whereas health services consumed by high-types are subsidized, i.e. $\alpha^h < 1$.[27] Finally, when ill, low-type consumers would prefer the budget constraint for high-type consumers $\left(R^{\mathrm{PR}h}, \alpha^h\right)$ to their own: $u\left(C_2^{\mathrm{PR}l}\right) + \theta^l\phi\left(X^{*l}\right) < u\left(C_2^{\mathrm{PR}lh}\right) + \theta^l\phi\left(X^{*lh}\right)$.

What is more important here, the next Lemma provides information about the effect of delegating the choice of allocations to consumers, which is a characteristic of price-regulated systems.

Lemma 2 *Within price-regulated systems, the optimal self-selecting allocation obtained in quantity-regulated systems cannot be implemented because it is not incentive compatible.*

Proof As discussed before, by optimally choosing the premium $P^{\mathrm{PR}j}$, the self-selecting allocation for healthy consumers obtained with program P4 can also be implemented in PR systems. Thus, only the self-selecting allocation for ill consumers is considered here. Let us suppose the social planner offers ill consumers the contracts $\left(R_0^{\mathrm{PR}l}, \alpha_0^l\right)$ and $\left(R_0^{\mathrm{PR}h}, \alpha_0^h\right)$, where $R_0^{\mathrm{PR}l} \equiv R^{\mathrm{QR}l} + \alpha_0^l q\bar{X}^l$, $R_0^{\mathrm{PR}h} \equiv R^{\mathrm{QR}h} + \alpha_0^h q\bar{X}^h$, α_0^l such

[27] In fact the FOC with respect to α^l is $X^{*l}\left[\mu_l u'\left(C_2^{\mathrm{PR}l}\right) + \lambda u'\left(C_2^{\mathrm{PR}l}\right) - \gamma\mu_l\right] + \gamma\mu_l\frac{\partial X^{*l}}{\partial\alpha^l}\left(1-\alpha^l\right) = 0$ and substituting 23 with $C_2^{\mathrm{PR}l}$ istead of $C_2^{\mathrm{QR}l}$, $\alpha^l = 1$ holds. Similarly, rearranging the FOCs for α^h and $R^{\mathrm{PR}h}$ it follows $\alpha^h = 1 + \lambda u'\left(C_2^{\mathrm{PR}lh}\right)\left(X^{*h} - X^{*lh}\right) / \left(\gamma\mu_h\frac{\partial X^{*h}}{\partial\alpha^h}\right)$ where $\frac{\partial X^{*h}}{\partial\alpha^h} < 0, \gamma > 0$ and, from 13 and 14, $X^{*h} - X^{*lh} > 0$, so that finally $\alpha^h < 1$.

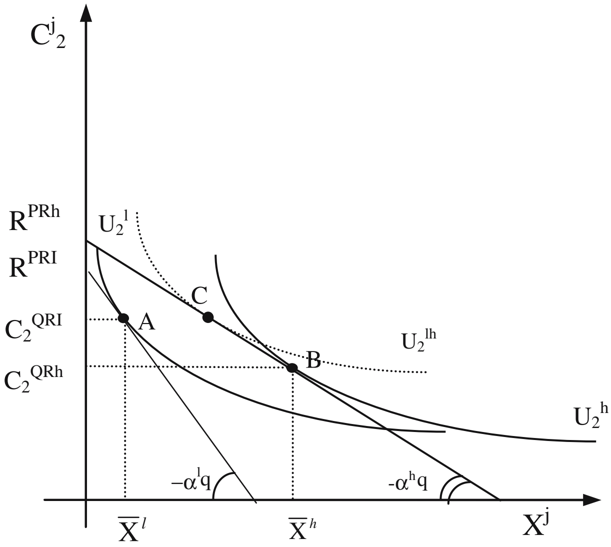

Fig. 5 Illustration of lemma 3

that $\theta^l\phi'\left(\bar{X}^l\right) = \alpha_0^l qu'\left(R^{\mathrm{QR}l}\right)$, α_0^h such that $\theta^h\phi'\left(\bar{X}^h\right) = \alpha_0^h qu'\left(R^{\mathrm{QR}h}\right)$, and where the bundles $(C_2^{\mathrm{QR}l} = R^{\mathrm{QR}l}, \bar{X}^l)$ and $(C_2^{\mathrm{QR}h} = R^{\mathrm{QR}h}, \bar{X}^h)$ correspond to the optimal quantity-regulated allocation as defined by Lemma 1. $(C_2^{\mathrm{QR}l}, \bar{X}^l)$ and $(C_2^{\mathrm{QR}h}, \bar{X}^h)$ are represented respectively by points A and B in Fig. 5. According to Lemma 1, aggregate and health service consumption by low-type consumers are not distorted, thus $\alpha_0^l = 1$. Moreover, $\theta^h\phi'\left(\bar{X}^h\right) < qu'\left(R^{\mathrm{QR}h}\right)$ holds; that is, high-type consumers are forced to overconsume health services. This implies that $\alpha_0^h < 1$. We show below that low-type consumers would not choose the QR allocation $(C_2^{\mathrm{QR}l}, \bar{X}^l)$ but would prefer a bundle on high-type consumers' budget constraint $\left(R_0^{\mathrm{PR}h}, \alpha_0^h\right)$.

From Lemma 1, low-type consumers prefer bundle B to bundle A. In B, by definition, $\theta^h\phi'\left(\bar{X}^h\right) = \alpha_0^h qu'\left(C_2^{\mathrm{QR}h}\right)$; as a consequence the following must be true:

$$\theta^l\phi'\left(\bar{X}^h\right) < \alpha_0^h qu'\left(C_2^{\mathrm{QR}h}\right) \tag{15}$$

that is, in B, given the price of health services $\alpha_0^h q$, the quantity of health services is too high and aggregated consumption too low, for low-type consumers. Let us now consider bundle C, where the low-type consumers' indifference curve is tangent to the budget constraint $\left(R_0^{\mathrm{PR}h}, \alpha_0^h\right)$. In C, the quantity of health services X^{*lh} and aggregated consumption C_2^{*lh} are such that:

$$\theta^l\phi'\left(X^{*lh}\right) = \alpha_0^h qu'\left(C_2^{*lh}\right), \tag{16}$$

where $C_2^{*lh} = R_0^{\mathrm{PR}h} - \alpha_0^h qX^{*lh} = R^{\mathrm{QR}h} + \alpha_0^h q\bar{X}^h - \alpha_0^h qX^{*lh} = R^{\mathrm{QR}h} + \alpha_0^h q(\bar{X}^h - X^{*lh}) > R^{\mathrm{QR}h} = C_2^{\mathrm{QR}h}$. By comparing Eqs. 15 and 16 it is clear that low-type consumers prefer bundle C to bundle B. Since $B \succ A$ and $C \succ B$, as a result of transitivity, low-type consumers will choose bundle C instead of bundle A: the quantity-regulated, self-selecting allocation is not incentive compatible. □

The result in the previous lemma naturally leads to the final comparison between quantity- and price-regulated systems discussed in the next section.

Comparing the alternative separating plans

As previously discussed, of all admissible price-regulated systems, the "real world" only offers those characterized by *linear* cost sharing. Lemma 2 shows that the second-best optimal allocation obtained by the direct QR mechanism cannot be implemented using such linear-PR systems. The achievement of this result is in fact, more general and pertains to the impossibility of reproducing the second-best welfare of the direct mechanism using linear-pricing systems, as the following proposition states.

Proposition 2 *With heterogenous consumers (i.e. $\theta^l \neq \theta^h$) and self-selecting plans, if the social planner is constrained to use linear cost-sharing, quantity-regulated systems dominate price-regulated ones.*

This result is an immediate consequence of the constraints imposed by the linear cost-sharing rule. As a matter of fact, the second-best allocation could be obtained in PR systems too, provided that the social planner uses *non-linear* cost-sharing parameters. In fact, if the social planner offers one single non-linear schedule $C_2^{\mathrm{PR}}(X)$ (i.e. a non-linear tariff),[28] instead of two linear schedules $C_2^{\mathrm{PR}l}(X) = R^{\mathrm{PR}l} - \alpha^l qX$ and $C_2^{\mathrm{PR}h}(X) = R^{\mathrm{PR}h} - \alpha^h qX$ (i.e. the two budget constraints in the previous section), then one can apply the Taxation Principle of mechanism design literature (Rochet, 1985). This Principle guarantees that the maximum welfare obtainable using direct mechanisms (in this case the QR system) can also be achieved by means of optimal non-linear tariffs (in this case the PR system). However, it is also generally acknowledged that complex non-linear tariffs are difficult to employ, and so linear pricing mechanisms are used instead. Interestingly, this reasoning and the result in Proposition 2 may help to explain why systems based on rationing of health services are so commonly encountered.

Lastly, it is worthwhile to conclude this section discussing the relative impact of heterogeneity on the performance of the two systems. Clearly, the two are equivalent with no heterogeneity (i.e. $\theta^l = \theta^h$) so that, with Proposition 2 we can state the following.

Corollary 2 *With self-selecting plans, consumers' heterogeneity impairs price-regulated systems comparatively more than quantity-regulated ones.*

Here, the intuition for this result is a consequence of the relative better performance of quantity-regulated systems as opposed to price-regulated ones in inducing consumers selection. Finally, note that this result reverses what has been obtained in uniform plans analysis (see Corollary 1). This means that with heterogeneity the welfare improvement that can be obtained with self-selecting plans as opposed to uniform ones has a stronger impact on quantity rather than price-regulated systems.

[28] Note that this schedule is the non-linear equivalent of the pooling schedule $C_2^{\mathrm{PR}}(X) = R^{\mathrm{PR}} - \alpha X$ analyzed in section "prices regulated systems", when uniform PR plans were considered.

Conclusion

This work presents an institutional comparison of alternative health insurance schemes: that is, of quantity- and price-regulated ones.

In this model, health insurance is public, illness has a negative impact on labor productivity, and consumers are heterogeneous with regard to the intensity of their preferences for health services (or with regard to the type of illness). The public insurer is fully informed of consumers' labor marginal productivity, but cannot observe preferences for health services (or the type of illness). As a consequence, low-type consumers imitate high-type ones in order to receive higher reimbursement, while public insurance may be constrained to adopt uniform insurance plans or may be free to choose self-selecting plans.

The first part of the work analyzes uniform plans: the same reimbursement is paid to both types of ill consumers. The main result is that price-regulated systems dominate quantity-regulated ones. In the second part of the paper, self-selecting plans are analyzed. Intuitively speaking, the rationale for quantity-regulated schemes should be stronger in such a case: by rationing health services, the public insurer should be able to partially prevent mimicking. Results seem to confirm this intuition: the quantity-regulated scheme corresponds to the direct mechanism and is not dominated by any other insurance scheme. If cost-sharing parameters have to be linear, the quantity-regulated scheme is always going to dominate the price-regulated scheme. Thus this model shows that, when the public insurer uses self-selecting plans, directly controlling quantity is better than regulating treatment (linear) price.[29] With this respect it has also been shown that heterogeneity has a different impact on the ranking between the two systems depending on the possibility to discriminate among consumers' type. In fact, with uniform plans, heterogeneity negatively affects quantity-regulated systems more than price-regulated ones, whilst the opposite holds in the case of self-selecting plans.

Acknowledgements I thank Alberto Bennardo, Giacomo Calzolari, Umberto Galmarini, Alessandro Lizzeri, Eric Malin, Alessandro Pavan, François Salanié, and in particular Helmuth Cremer, for their helpful comments. An anonymous referee made vital suggestions as to how to improve the paper. I gratefully acknowledge the TMR ERBFMBICT950278 grant I received during my stay at the University of Toulouse.

Appendix

Proof of Lemma 1

Part (i). Let us consider program P4: FOC with respect to $C_1^{\mathrm{QR}l}$ gives:

$$\frac{\mu_l + \lambda}{\mu_l} u'\left(C_1^{\mathrm{QR}l}\right) - \gamma = 0, \tag{17}$$

where $\frac{\mu_l+\lambda}{\mu_l} \geq 1$ because $\lambda \geq 0$. Eq. 17 implies that $\gamma > 0$ and that:

$$u'\left(C_1^{\mathrm{QR}l}\right) \leq \gamma. \tag{18}$$

[29] As an anonymous referee remarked, this could partially explain the increase in managed care in the USA.

FOC with respect to $L^{\mathrm{QR}l}$ gives:

$$\frac{\mu_l+\lambda}{\mu_l} v'\left(L^{\mathrm{QR}l}\right) - w\gamma = 0. \tag{19}$$

such that, not surprisingly, from 17 and 19: $v'(L^{\mathrm{QR}l}) = wu'\left(C_1^{\mathrm{QR}l}\right)$: low-types' labor supply is not distorted.

FOC with respect to $C_1^{\mathrm{QR}h}$ gives:

$$\frac{\mu_h-\lambda}{\mu_h} u'\left(C_1^{\mathrm{QR}h}\right) - \gamma = 0, \tag{20}$$

where $\frac{\mu_h-\lambda}{\mu_h} \leq 1$ because $\lambda \geq 0$. Eq. 20 implies that:

$$u'\left(C_1^{\mathrm{QR}h}\right) \geq \gamma. \tag{21}$$

From 18 and 21 $u'\left(C_1^{\mathrm{QR}h}\right) \geq u'\left(C_1^{\mathrm{QR}l}\right)$ holds, such that $C_1^{\mathrm{QR}l} \geq C_1^{\mathrm{QR}h}$: low-types' aggregate consumption when in good health is higher than that of high-types.

FOC with respect to $L^{\mathrm{QR}h}$ gives:

$$\frac{\mu_h-\lambda}{\mu_h} v'\left(L^{\mathrm{QR}h}\right) - w\gamma = 0 \tag{22}$$

such that, not surprisingly, from 20 and 22 $v'(L^{\mathrm{QR}h}) = wu'\left(C_1^{\mathrm{QR}h}\right)$: high-types' labor supply is not distorted. By totally differentiating equation $v'(L^{\mathrm{QR}j}) = wu'\left(C_1^{\mathrm{QR}j}\right)$ with respect to $L^{\mathrm{QR}j}$ and $C_1^{\mathrm{QR}j}$, it follows that $\frac{\mathrm{d}L}{\mathrm{d}C_1} < 0$. Since $C_1^{\mathrm{QR}l} \geq C_1^{\mathrm{QR}h}$, the latter inequality implies that $L^{\mathrm{QR}h} \geq L^{\mathrm{QR}l}$. Clearly the following inequality is verified: $u\left(C_1^{\mathrm{QR}l}\right) - v(L^{\mathrm{QR}l}) \geq u\left(C_1^{\mathrm{QR}h}\right) - v\left(L^{\mathrm{QR}h}\right)$.

Part (ii). With regard to ill consumers, FOC with respect to $C_2^{\mathrm{QR}\,l}$ of program P4 gives:

$$\frac{\mu_l+\lambda}{\mu_l} u'\left(C_2^{\mathrm{QR}l}\right) - \gamma = 0. \tag{23}$$

Again Eq. 23 implies that: $u'\left(C_2^{\mathrm{QR}l}\right) \leq \gamma$.

From FOC with respect to $\bar{X}^l$:

$$\frac{\mu_l+\lambda}{\mu_l} \theta^l \phi'\left(\bar{X}^l\right) - q\gamma = 0. \tag{24}$$

From 23 and 24 it follows that $\theta^l \phi'\left(\bar{X}^l\right) - qu'\left(C_2^{\mathrm{QR}l}\right) = 0$. As a consequence, aggregate and treatment consumption of low-type ill consumers are not distorted.

From FOC with respect to $C_2^{\mathrm{QR}h}$:

$$\frac{\mu_h-\lambda}{\mu_h} u'\left(C_2^{\mathrm{QR}h}\right) - \gamma = 0. \tag{25}$$

Again, Eq. 25 implies that: $u'\left(C_2^{\mathrm{QR}h}\right) \geq \gamma$. Thus, comparing aggregate consumption of low- and high-type consumers: $C_2^{\mathrm{QR}h} \leq C_2^{\mathrm{QR}l}$.

Given that, when healthy, low-type consumers are better off than high-type consumers and that the incentive constraint is binding, the following holds:

$$u\left(C_2^{\mathrm{QR}h}\right) - H + \theta^h\phi\left(\bar{X}^h\right) \geq u\left(C_2^{\mathrm{QR}h}\right) - H + \theta^l\phi\left(\bar{X}^h\right) \geq u\left(C_2^{\mathrm{QR}l}\right) - H + \theta^l\phi\left(\bar{X}^l\right).$$

That is, high-type consumers are better off. Moreover rearranging the binding incentive constraint and given that $C_2^{\mathrm{QR}h} \leq C_2^{\mathrm{QR}l}$, it follows that $\bar{X}^h > \bar{X}^l$. This monotonicity condition implies that the incentive compatibility constraint for high-type consumers is always verified by the solution, and omitting it in program P4 is without loss of generality.

FOC with respect to $\bar{X}^h$ yields:

$$\mu_h\theta^h\phi'\left(\bar{X}^h\right) - \lambda\theta^l\phi'\left(\bar{X}^h\right) - \mu_h q\gamma = 0. \tag{26}$$

Using Eq. 26 and $\mu_h qu'\left(C_2^{\mathrm{QR}h}\right) - \lambda qu'\left(C_2^{\mathrm{QR}h}\right) - \mu_h q\gamma = 0$ (which is FOC with respect to $C_2^{\mathrm{QR}h}$ with each term multiplied by q) and solving as in Stiglitz (1987), page 1005, we find that: $\theta^h\phi'\left(\bar{X}^h\right) < qu'\left(C_2^{\mathrm{QR}h}\right)$. This proves that high-type ill consumers are forced to consume too many health services and too little aggregate consumption.

References

Alger, I., & Ma, C. A. (2003). Moral hazard, insurance, and some collusion. *Journal of Economic Behavior and Organization, 50*(3), 225–247.

Arrow, K. J. (1963). Uncertainty and the welfare economics of medical care. *The American Economic Review, 53*, 941–938.

Besley, T. J. (1991). The demand for health care and health insurance. In A. McGuire, Fenn P., & K. Mayhew (Eds.), *Providing health care: the economics of alternative systems of finance and delivery*, Oxford University Press.

Besley, T. J. (1988). Optimal reimbursement health insurance and the theory of Ramsey Taxation. *Journal of Health Economics, 7*, 321–336.

Besley, T., & Gouveia, M. (1994). Alternative systems of health care provision. *Economic Policy*, October, 203–258.

Blackorby, C., & Donaldson, D. (1988). Cash versus kind, self-selection, and efficient transfers. *The American Economic Review, 78*(4), 691–700.

Blomquist, A. (1997). Optimal non-linear health insurance. *Journal of Health Economics, 16*, 303–321.

Blomquist, A., & Horn, H. (1984). Public health insurance and optimal income taxation. *Journal of Public Economics, 24*, 353–371.

Chernew, M. E., Encinosa, W. E., & Hirth, R. A. (2000). Optimal health insurance: The case of observable, severe ilness. *Journal of Health Economics, 19*, 585–609.

Cremer, H., & Gahvari, F. (1995). Uncertainty, optimal taxation and the direct versus indirect tax controversy. *The Economic Journal, 105*, 1165–1179.

Cremer, H., & Gahvari, F. (1997). In-kind transfers, self-selection and optimal tax policy. *European Economic Review, 41*, 97–114.

Fundember, D., & Tirole, J. (1991). *Game theory*, The MIT Press.

Le Grand, J., & Mossialos E. (Eds.), (1999). *Health care and cost containment in the European Union*, Ashgate.

Manning, W. G., Newhouse, J. P., Duan, N., Keeler, E. B., Liebowitz, A., & Marquis, M. S. (1987). Health insurance and the demand for health care; evidence from a randomized experiment. *American Economic Review, 77*, 251–77.

Myerson, R. (1979). Incentive compatibility and the bargaining problem. *Econometrica, 47*, 61–73.

Rochet, J. C. (1985). The taxation principle and multitime Hamilton-Jacobi equations. *Journal of Mathematical Economics, 14*, 113–128.

Stiglitz, J. E. (1987). Pareto efficient and optimal taxation and the new new welfare economics. In A. J. Auerbach & Feldstain (Eds.), *Handbook of public economics* (Vol. 2). Amsterdam: North-Olland.

Varian, H. R. (1980). Redistributive taxation as social insurance. *Journal of Public Economics, 14*, 49–68.

Weitzman, M. L. (1974) Prices vs. Quantities. *Review of Economic Studies, 41*(4), 477–491.

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