1997

Monopolistic Competition and Supply-Side Cost Sharing in the Physician Services Market

Åke Blomqvist

ISSN:0318-725X
ISBN:0-7714-1975-9

RESEARCH REPORT 9705

Monopolistic Competition and Supply-Side Cost Sharing in the Physician Services Market

by

Ake Blomqvist

April 1997

Department of Economics
Social Science Centre
University of Western Ontario
London, Ontario, Canada
N6A 5C2
econref@sscl.uwo.ca

Monopolistic competition and supply-side cost sharing in the physician services market

Åke Blomqvist*

February 15 1997

Abstract

The interaction of insurance and the market for physician services is considered in a model where imperfectly informed consumers rely on doctors for advice on the utilization of services and there is monopolistic competition among physicians on the basis of price and the quality of their advice. Equilibria under fee for service and capitation are compared, and I analyze the use of a system of capitation and partial supply-side cost sharing to attain an outcome superior to either pure fee for service or pure capitation.

1 Introduction

The design of efficient incentive structures in health care revolves around several kinds of information problems. One, the information asymmetry between health insurers, on the one hand, and patients and those who treat them on the other, is the reason why conventional health insurance takes the form of a subsidy for health services utilization rather than state-contingent

*Department of Economics, University of Western Ontario, London, Ontario, Canada N6A 5C2. E-mail: blomqvist@uwo.ca. I wish to thank Ig Horstmann and Al Slivinski for helpful discussions, and the Social Science and Humanities Research Council of Canada for financial assistance. Remaining errors and shortcomings are my own.

lump-sum contracts, and thus gives rise moral hazard problem and the question of the optimum degree of patient cost-sharing.[1] Another is the information asymmetry between patients and the doctors regarding what constitutes appropriate treatment in different situations. This problem has been extensively discussed in the literature on supplier-induced demand (SID).[2]

In recent years, the search for efficient contract structures has focussed on the idea that the problems caused by these information problems can be reduced through insurance that involve some form of "supply-side cost sharing", in the terminology of Ellis and McGuire (1993). As Newhouse (1996, p. 1237) has noted, the simplest form of supply-side cost sharing is a linear combination of capitation in which the incentive to economize on the use of resources is shifted to the provider (physician), and fee for service under which the patient's incentive to economize depends on the degree of coinsurance.[3]

Existing attempts at formal analysis of supply-side cost sharing have relied on models with somewhat special features, such as incorporating ethical or altruistic elements in the providers' objective functions, abstracting from the incentive effects of different insurance contracts on patients, or assuming away the information asymmetry between patient and provider.[4] In this

[1]Two classic analytical papers on this are Pauly (1968) and Zeckhauser (1970). For empirical evidence and quantitative estimates of the associated efficiency loss, see Feldstein (1973), Manning *et al* (1987), Newhouse *et al.* (1995), Feldman and Dowd (1991), and Blomqvist (forthcoming).

[2]A good discussion is in Pauly (1980); see also the survey in Pauly (1986). For formal models of SID, see e.g., Satterthwaite (1979), or Dranove (1988). Although the present paper differs from Dranove's model of a single physician, it draws on his analysis in other respects.

[3]Under capitation, patients choose a single physician or provider organization such as a Health Maintenance Organization or group practice as their sole provider during the contract period; the providers are then remunerated on the basis of the number of patients on their list, regardless of the volume of service they provide to each patient. The term "coinsurance" refers to the patient's share of the fee charged, under the insurance contract.

[4]In the influential model of Ellis and McGuire (1986), patients play a completely passive role, but doctors care about patients' welfare; the same is true in the paper by Wedig (1993). In Ellis and McGuire (1990), patients are assumed to be fully informed, and the conflict between the incentives on doctors and patients under different payement systems is modelled as a bargaining process. In Ma and McGuire (1995), the treatment decision is formally modelled as a game between a single doctor and a single patient (who is fully informed) when certain aspects of the transaction are "noncontractible".

paper, I consider supply-side cost sharing in a model of monopolistic competition where all agents respond to incentives in a purely self-interested manner, and both kinds of information asymmetry are present. Using this model, I demonstrate the potential efficiency problems that arise under conventional insurance and remuneration of doctors through fee for service, on the one hand, and under a system of capitation, on the other. I then show that there exists a form of supply-side cost sharing through which it is possible to attain an efficient equilibrium where patients are fully insured and the utilization of health services is at its Pareto-efficient level.[5]

2 Insurance and information asymmetry under fee for service

I first consider the case of fee for service. To make the model as simple as possible, I postulate a world in which the only health services are physician services. Every consumer faces the same probability density function $f(\theta)$ for the illness severity parameter θ if he falls ill.[6] Given that he goes to the i'th doctor, the consumer's expected utility is:

$$E = \int [U_C(y - m - \sigma p_i \cdot z_i(\theta)) + U_H(\, z_i(\theta) - \theta)]\, f(\theta) d\theta \qquad (1)$$

At the policy level, the idea of supply-side cost sharing has been at the center of the debate concerning how to pay physicians under the U.S. Medicare plan, in the context of the proposals for a "resource-based relative value scale" (RBRVS). For a discussion, see, e.g., Pauly (1991).

[5]Selden (1990) reaches a similar conclusion in a model that constitutes an elegant generalization of Ellis and McGuire (1986). Although this paper is very close in spirit to Selden's, it differs from it by modelling the service utilization decision explicitly from maximizing behaviour on the part of both doctors and patients, and by the assumption of monopolistic competition among doctors. For an approach involving a different type of incentive scheme see Blomqvist (1991).

By using a model with identical consumers, I abstract from a third form of information problem which is important in practice, that arising from imperfectly observable differences in patients' risk of illness. For a discussion of the relevance of supply-side cost sharing in addressing this type of problem, see Newhouse (1996).

[6]I assume that the period under consideration is long enough so that everyone falls ill and goes to see a doctor.

/here y is income (exogenously given), m is the premium on his health ısurance, σ is the share of the cost of health care payable by the consumer i.e., the insurer's share is $(1-\sigma)$), p_i is the price per unit of health services harged by the i'th doctor, and $z_i(\theta)$ is the quantity of services she provides ı state of the world θ. For simplicity, the patient's utility function when ill .as been made separable in consumption and health services, respectively.)octors are supposed to differ in "practice style", and patients differ in terms f which style they prefer. In writing down (1), I implicitly assume that he patient has already chosen the doctor whose practice style most closely orresponds to his preferences.

For a given value of σ, and if θ were observable, the consumer would hoose z_i so as to maximize (1), given θ, yielding the first-order condition

$$\sigma p_i \cdot U_C'(\theta) = U_H'(\theta) \tag{2}$$

vhere U_i', $i = C, H$, denote derivatives; I denote the solution to (2) as $z^*(\sigma p, \theta)$. The patient's utility function is assumed known to the doctor. However, I ıssume that θ, and hence $z^*(\sigma p, \theta)$, are only imperfectly observable, by both)atient and doctor. In particular, I assume that each patient's estimate of :he best level of treatment in state of the world θ is $z_p = z^*(\sigma p, \theta) + \epsilon_p$ where $\epsilon_p \sim N(0, v_p)$. The treatment recommended by the doctor, z_d, is given by $z_d = z^*(\sigma p, \theta) + \epsilon_d + \delta$, where every doctor's $\epsilon_d \sim N(0, v_d)$ and δ is a parame:er that measures the extent to which the doctor systematically recommends)vertreatment (or undertreatment) of her patients, relative to her estimate)f what would be in their best interest given the signal that she has received :egarding illness severity.[7]

In this environment, patients and doctors play the following game. First, ɜach doctor announces the price she charges per unit of service. Given the price and their preferences regarding each doctor's practice style (assumed to be observable *ex ante*), patients decide which doctor they will go to for a diagnosis. Following this, each patient and his doctor receive the noisy signal on which they base their estimate of z^*. The patient then compares the treatment z_d recommended by the doctor with his own estimate z_p of z^*. If the absolute value of the random variable $z_d - z_p \equiv x$ is less than or equal to some critical value ψ, he accepts the doctor's recommendation

[7]For simplicity, I assume that the value of δ is independent of the signal the doctor receives.

and receives z_d. On the other hand, if $\mid z_d - z_p \mid > \psi$, the patient concludes that either the doctor he has chosen has inadequate diagnostic skills, or else systematically overtreats (undertreats) her patients. Given this, he does not accept the doctor's recommendation, but instead goes to another one. For simplicity, I assume that when doing so, he chooses which new doctor to go to, at random, and that he accepts whatever treatment the second doctor recommends.

Instead of modelling explicitly the matching between individual doctors and patients, I will simply assume that the distributions of patient preferences and physician characteristics are such that the patients' choice of doctors can be represented as the solution to the following maximization problem for a hypothetical representative patient:[8]

$$\underset{s_i}{Max} \int [U_C(y - m - \sigma \cdot T(\theta)) + U_H(h(\theta) - \theta)] f(\theta) d\theta \tag{3}$$

where

$$T(\theta) = \sum_{i=1}^{N} p_i s_i \hat{z}(\theta), \; h(\theta) = \left[\sum_{i=1}^{N} (s_i \hat{z}(\theta))^{\beta} \right]^{\frac{1}{\beta}} \tag{4}$$

with s_i denotes the the i'th doctor's share of the total patient population, and $\hat{z}(\theta)$ is the common treatment recommendation the patient expects from each doctor in state of the world θ; it is supposed to be the same for each doctor because the patient has no information on which to base a prediction of an individual doctor's recommendation strategy. With $\hat{z}(\theta)$ being the same for each i, (4) is equivalent to $T(\theta) = \hat{z}(\theta) \sum p_i s_i$, $h(\theta) = \hat{z}(\theta) \left[\sum s_i^{\beta} \right]^{\frac{1}{\beta}}$.

If all doctors charge the same price $p_i = p$ per unit of z, the solution to this problem is $s_i = s$, $\forall i$, that is, each doctor ends up being chosen by the same number of patients. To derive the elasticity of demand for a single s_i with respect to p_i, I assume that each individual s_i is small enough so that a change in a single s_i alone has a negligible impact on the aggregate amount of health services $h(\theta)$. The demand function for each s_i can then be approximated by finding the solution to the problem of minimizing $T(\theta)$ holding $h(\theta)$ constant. Solving this problem yields $p_i/p_j = (s_i/s_j)^{\beta - 1}$, $\forall j \neq i$, so that the elasticity of s_i with respect to its own price p_i is $\varepsilon \equiv (\beta - 1)^{-1}$.

[8] My approach is patterned on the model used by Dixit and Stiglitz (1977) in their paper on monopolistic competition and product diversity. For notational simplicity, the patient population is normalized to one.

Each doctor has the same utility function which depends positively on her income, and negatively on her workload. It is given by

$$W = W(C, L) \tag{5}$$

where C is the doctor's income, L is her workload, $W_C > 0$, $W_L < 0$, $W_{CC} < 0$, $W_{LL} < 0$, $W_{CL} < 0$. The doctors' objective is to maximize expected utility, but for simplicity I assume that each has a large enough number of patients so that expected utility is just the value of her utility function at the expected value of its arguments. Specifically, let $z^* = E(z^*(\theta))$, and let the expected number of units of services provided to each patient she treats be denoted by z_i, given by $z_i = E(z^* + \epsilon_d + \delta_i) = z^* + \delta_i$. Further, let $\bar{s}_i$ denote the average number of patients that she will treat. It will be given by

$$\bar{s}_i = s_i \cdot (1 - P_l) + P_g \tag{6}$$

where P_g is the number of patients she gains at the diagnosis stage as those who leave other doctors choose her as their doctor, while P_l is the proportion of her initial patient stock that she loses as $\mid x \mid$ exceeds its critical value when she makes her recommendation.

Given these definitions, the doctor's problem can be written as

$$\underset{\delta_i, p_i}{Max}\ W(p_i z_i \cdot \pi \bar{s}_i,\ z_i \cdot \pi \bar{s}_i) \tag{7}$$

where it is understood that each one takes the price and recommendation strategies of other doctors as given.

Note that since z^* is the same constant for each doctor, one can treat either δ_i or z_i as the choice variable. Since the problem is symmetric, in equilibrium each doctor will follow the same strategy, so that from now on I will generally suppress the subscript i.

As shown in the Appendix, the first-order conditions corresponding to (7) can be written

$$pW_C \cdot (1 + \frac{1}{\varepsilon}) + W_L = 0 \tag{8}$$

$$1 + \frac{z}{\bar{s}} \cdot \frac{d\bar{s}}{d\delta} = 0 \tag{9}$$

where the partial derivatives W_C and W_L are the doctor's marginal utility of income and disutility of work, and where I have used $d\bar{s}/d\delta = d\bar{s}/dz$. In this

paper, I will not concern myself with what restrictions on functional forms and parameter values are necessary to guarantee that equilibria exist, but simply assume that they do.

To interpret the expression $(z/\bar{s})(d\bar{s}/d\delta)$ in (8), recall that a patient will leave his doctor and go to another randomly selected one if the absolute value of $x \equiv z_d - z_p$ is greater than a critical value ψ. Since both z_d and z_p are normally distributed with means δ and 0 respectively, x is normally distributed with mean δ and, assuming the random terms ϵ_d and ϵ_p are independent, variance $v_x = v_d + v_p$. For a given recommendation strategy δ, the proportion P_l of the initial patient stock s_i that a doctor will lose at the diagnosis stage will then be given by

$$P_l = \int_{\frac{\psi-\delta}{\sqrt{v_x}}}^{+\infty} h(x)dx + \int_{-\infty}^{-\frac{\psi+\delta}{\sqrt{v_x}}} h(x)dx \tag{10}$$

where $h(x)$ is the density of a standard normal variable with zero mean and unit variance, and the first term on the right-hand side corresponds to those who leave because they think the doctor has overestimated the illness severity parameter, while the second term represents those who leave because they think she has underestimated it. The relation $P_l(\delta)$ will have a shape such as shown in Figure 1; it is identical to the power function for the standard test of the hypothesis $E(x) = 0$ for a normal variable with variance v_x and critical values ψ, $-\psi$. Its derivative is

$$\frac{dP_l}{d\delta} = \frac{1}{\sqrt{v_x}}\left[h(\frac{\psi-\delta}{\sqrt{v_x}}) - h(\frac{\psi+\delta}{\sqrt{v_x}})\right] \equiv g(\delta;\, \psi, v_x) \tag{11}$$

where I have used the symmetry of the normal distribution. Note that $g(0;\, \psi, v_x) = 0$ and $sgn(g(.) = sgn(\delta)$, and that for any ψ, there is a $\delta(\psi) > \psi > 0$ such that $g(.)$ reaches a minimum at $\delta^l = -\delta(\psi)$ and a maximum at $\delta^u = \delta(\psi)$. (This is evident from Figure 1.) For δ such that $\delta^l \leq \delta \leq \delta^u$, $\partial g(.)/\partial\delta > 0$.

Figure 1 about here.

Because the model is symmetric, each doctor will use the same strategy in equilibrium. With P_l being the same for each doctor, and since those that leave their first doctor choose their second one at random, it follows that $P_g = s \cdot P_l$, so that in equilibrium, $\bar{s} = s$. Since $d\bar{s}/d\delta = -s \cdot dP_l/d\delta$, (9) can be rewritten

$$1 - z \cdot g(z - z^*;\, \psi, v_x) = 0 \tag{12}$$

Because g is not monotonic in δ, it is possible that 12 has more than one solution. In the following, I will always focus on equilibria for which ($\delta^l < \delta < \delta^u$).

The properties of g now imply the following:

Proposition 1. If there is monopolistic competition among doctors and diagnostic information is imperfect, payment of doctors through fee for service implies $\delta = z - z^* > 0$. That is, there is demand creation in the sense that doctors on average recommend a treatment level that exceeds what would be in the patients' best interest, given the insurance coverage they have.

Figure 2 about here.

The equilibrium configuration of the model can be illustratred in a diagram in $\{p, z\}$-space such as Figure 2, where F^1 and F^2 are the loci of points along which (8) and (9) , respectively, are satisfied. As for the case of a competitive labour supply curve, the slope of F^1 depends on the balance of a substitution and an income effect, and it may have a backward-bending portion such as in the figure. In the following, I will assume that equilibrium always occurs on the upward-sloping portion where the substitution effect dominates the income effect. As shown in the Appendix, the locus F^2 is downward-sloping. The following comparative-statics results are now easily derived (see the Appendix):

Proposition 2. An increase in $\mid \varepsilon \mid$ shifts F^1 to the right and therefore reduces the unit price p and increases average treatment intensity z.

Proposition 3. An increase in the patient's share σ of the cost of treatment shifts F^2 to the left and therefore reduces both p and z.

Proposition 4. Provided $(\psi - \delta)/\sqrt{v_x} > \sqrt{.5}$ in equilibrium, an increase in ψ shifts F^2 to the right and therefore raises both p and z.

Proposition 5. If ψ is proportional to $\sqrt{v_x}$, the extent of demand creation $z - z^*$ will go to zero as v_x goes to zero.

In the foregoing discussion, the coinsurance parameter σ and the health insurance premium m were both taken as exogenous. Since σ in part determines the amount of services the patients will utilize, a more complete

specification should include a requirement that the premium has to be large enough to cover the insurer's expected benefit payments and costs. If one assumes that adminstrative costs are zero and the insurance market is competitive, this constraint would take the form

$$m = (1 - \sigma)pz \tag{13}$$

which becomes another condition, in addition to (8) and (9), for equilbrium.[9]

A competitive insurance market would also offer patients a choice of policies with different coinsurance parameters. For the case of perfectly observable illness severity parameters and competitively supplied medical services, the choice of insurance policy can be modelled by assuming that the consumer chooses the policy which maximizes his expected utility. The solution to this problem is that policy which best balances the consumer's gain from more complete insurance against the efficiency loss associated with the tendency for insured patients to overuse services (the moral hazard effect).[10]

In the present model, the situation is more complicated because of the unobservability of the illness severity parameter, and also because medical services are supplied in a monopolistically competitive market. For these reasons, little can be said in general about the efficiency properties of a competitive insurance market. However, two observations may be made. First, in comparison with the case of perfectly observable θ, the presence of supplier-induced demand in this model would tend to magnify the overuse of medical services in comparison with the efficient level. Second, because there is monopolistic competition among doctors, the social opportunity cost of medical services (that is, the opportunity cost of the physicians' time) is less than the price p of z. By itself, this would cause a tendency for *too few* health services to be used.

[9]Strictly speaking, imposing (13) implies that the variable z^* should be written $z^*(\sigma p, \theta, m) = z^*(\sigma p, \theta, pz)$. Although this would not affect the first-order conditions (since both doctors and patients take m as given), it would affect the model's equilibrium solution and hence the comparative statics. However, it is easy to show that as long as it remains true that $dz^*/dp = \partial z^*/\partial p + (\partial z^*/\partial m)(dm/dp)$ is negative, and $dz^*/dz = (\partial z^*/\partial m)(dm/dz)$ is negative, the comparative statics results will not be qualitatively affected. Both conditions are reasonable.

[10]As noted above, the classic reference on this is Zeckhauser (1970). For an extension to non-linear insurance, see Blomqvist (forthcoming).

3 Information asymmetry under capitation and mixed payment systems

Even a second-best optimal conventional insurance contract represents an uneasy compromise, in the sense that it offers consumers less than full insurance on the one hand, but still implies a tendency toward inefficiently high utilization of health services, on the other. The large-scale shift toward different forms of insurance involving "managed care" that has occurred in the United States in recent years, and the use of alternative institutional mechanisms to reimburse providers in other countries, can both be seen as attempts to improve on this compromise. What all these arrangements have in common is closer contractual links between insurers and providers than in conventional insurance, and more emphasis on incentives on physicians (providers) than on users in trying to limit the cost of health services; that is, on "supply side", rather than "demand side", cost sharing.

In the two types of arrangements that I consider in this section, I assume that there is no demand-side cost sharing, that is, that there are no out-of-pocket charges (user fees) to patients for the services provided. Both are variants of the system of capitation, in which providers receive all or part of their income in the form of a fixed amount per unit of time for each patient that is formally signed up with their practice. In return, physicians are responsible for providing services to the patients on their list[11]

Since patients are free to choose which doctor to sign up with, physicians in a capitation system compete for patients just as under conventional insurance. In real-world capitation systems, however, the capitation payments are not made directly by the patients to doctors, but instead through a third-party insurer. That is, the patient pays an insurance premium (or taxes, in a public system), and the doctors receive their capitation payments from the insurer. In these circumstances, any price competition among doctors is indirect, in the sense that doctors negotiate about capitation rates with insurers, and patients can choose only among those doctors with whom their insurer has a contract. Under competitive conditions, however, insurance premia will indirectly reflect the capitation charges the insurers have negotiated with the

[11] Capitation is also sometimes referred to as "prepayment"; that is, when the patient receives services, there is no out-of-pocket charge since the services have "already been paid for" through the capitation charge.

doctors, so that doctors who are willing to accept low capitation rates will indirectly attract patients who are looking for low-cost insurance. In the formal analysis, I will disregard the distinction between indirect and direct price competition and model the market as if each physician were to collect her capitation charges directly from her patients. Thus, I model capitation as an insurance contract with a coinsurance parameter σ of zero, and an insurance premium m_i equal to the fixed amount each patient pays the doctor per unit of time.

I first consider the case of pure capitation, where the doctor is paid entirely through capitation. The game played between patients and doctors is similar to that in the fee for service case, in that patients initially choose which doctor to sign up with on the basis of the capitation charge she announces, and her practice style. In the second stage, the patient and doctor receive noisy signals regarding θ, and the patients compare the doctors' treatment recommendation z_d with their own estimate z_p of the treatment level $z^*(\sigma p, \theta) = z^*(0, \theta)$ which would be optimal from their point of view.[12] If $\mid z_d - z_p \mid \leq \psi$, they stay with the doctor they originally chose, while if $\mid z_d - z_p \mid > \psi$, they switch to another. As in the fee for service case, I assume that those who leave one doctor choose which other one to go to, at random. In the final stage, the doctor receives the capitation charge from her remaining patients (including those who have come to her after having left other doctors), and provides the treatment that she recommended at the diagnostic stage.

By analogy with the fee for service case, I again assume that the matching between patients and doctors can be represented by maximizing a representative patient's expected utility function, with patients expecting each doctor to provide the same level of treatment $z(\theta)$ in state of the world θ. Under capitation, this expected utility can be written as

$$E = U_C(y - \sum m_i s_i) + \int U_H(h(\theta) - \theta) f(\theta) d\theta \qquad (14)$$

where

$$h(\theta) = \hat{z}(\theta) \left[\sum (s_i)^\beta\right]^{\frac{1}{\beta}}$$

[12]Here and throughout the paper, I assume that even at zero out-of-pocket cost per unit ($\sigma = 0$), there are other non-pecuniary costs of utilizing health services, such that $z^*(0, \theta)$ is less than some upper bound z^u, for any θ.

as in the fee for service case. As before, if each s_i is small enough to have no more than a negligible impact on the aggregate quantity of health services $h(\theta)$, the demand function for s_i can be approximated by minimizing $\sum m_i s_i$ subject to $h(\theta) = const.$, which again yields an own-price elasticity given by $\varepsilon = (\partial s_i/\partial m_i)(s_i/m_i) = (\beta - 1)^{-1}$.

The doctor's problem would now take the form (again suppressing subscripts):

$$\underset{m\,,\,z}{Maximize}\; W(m \cdot \bar{s},\ \bar{s}z), \tag{15}$$

subject to (6).

The first-order conditions (8) and (9) in the previous section will now be replaced by (see the Appendix):

$$m \cdot W_C \cdot (1 + \frac{1}{\varepsilon}) + z \cdot W_L = 0 \tag{16}$$

$$(1 + \varepsilon) - z \cdot g(z - z^*, \psi, v_x) = 0 \tag{17}$$

Since $\varepsilon < -1$, (17) implies that $g(z - z^*, \psi, v_x) < 0$. Given the properties of g established above, we have the following:

Proposition 6. Under capitation, we will have $z - z^* < 0$; that is, there is "negative supply inducement".

The following comparative statics propositions are established in the Appendix.

Proposition 7. An increase in $\mid \varepsilon \mid$ reduces both m and z.

Proposition 8. Provided $(\psi + \delta)\sqrt{v_x} > \sqrt{.5}$ in equilibrium, an increase in ψ reduces both m and z.

Proposition 9. If ψ is proportional to $\sqrt{v_x}$, the extent of (negative) demand creation $z - z^*$ will go to zero as v_x goes to zero.

From a patient's point of view, an advantage with a system of capitation is that it represents "complete" insurance: Under capitation, illness does not entail any incremental out-of-pocket cost and therefore does not force a reduction in consumption. At the same time, the absence of cost-sharing obviously has a tendency to cause an overutilization of health services, since

$z^*(0,\theta) > z^*(\sigma p,\theta)$ for any $\sigma p > 0$ and any θ.[13] Whether the net equilibrium value of z under capitation will be larger or smaller than in a system with a given degree of patient cost-sharing under fee for service, thus depends on the balance of two effects: The average difference between $z^*(0,\theta)$ and $z^*(\sigma p,\theta)$, on the one hand, and the difference between the amounts of supply inducement $z - z$, on the other.

The phenomenon of negative supply inducement under capitation arises because, when the doctor's income consists solely of capitation payments paid by the insurer, her short-term marginal benefit from providing each patient with more services is negative, once the patient has signed up: An extra unit of service generates no extra income, but has an opportunity cost in terms of the doctor's time. This raises the question whether it might be possible to counteract this incentive by a system of supply-side cost sharing under which each doctor's income depends *both* on the number of patients signed up on her list, *and* on the number of units of services $z(\theta)$ rendered to each one.

I considering such a system, I continue to assume that the patient pays no user charge, only a premium m. However, only a portion of m now goes directly to the doctor; an amount r is kept by the insurer, to finance the

[13] Although patient charges in capitation systems usually are zero, there is no inherent reason why this need be the case. It is possible to imagine a capitation plan with some degree of demand-side cost-sharing. In such a plan, the patients pay both a premium and, in addition, a user fee c for each unit of z. If the capitation amount per patient paid to the doctor is m, the actuarially fair premium paid by the patient to the insurer would then be $m - c \cdot z$, and the amount z^*demanded by a perfectly informed patient would be $z^*(c,s) < z^*(0,s)$.

From the viewpoint of economic efficiency, this type of demand-side cost sharing entails some degree of loss since it implies a transfer from the "well" state to the "ill" state for the representative consumer; that is, it implies less complete insurance. On the other hand, since it will lead to a reduction in z, it will reduce the loss from overutilization if z was above its efficient level to begin with.

If demand-side cost sharing were efficient, however, one would expect that capitation plans with zero user charges would find it difficult to compete. The fact that actual capitation plans like HMOs in fact typically impose no user charges can, therefore, be taken as tentative support for the idea that the SID-effect is significant in reality, so that the incentive on patients to overutilize services is less important than the tendency toward negative SID under capitation.

For a different way of modelling HMO contracts than that used here, see Baumgardner (1991).

payment to doctors of p per unit of service z rendered to each patient. The doctor's income would be $\bar{s}\cdot(m-r+pz)$, where p now is a parameter specified in the doctor's contract with the insurer. Note that if insurance is actuarially fair, one will have $r = p \cdot z$ in equilibrium.

Given this arrangement,[14] the doctor's maximization problem may be written:

$$\underset{m,z}{Maximize}\ W(\bar{s}\cdot(m-r+pz,\ \bar{s}z) \tag{18}$$

After rearrangement, the first-order conditions for a maximum can be written:

$$mW_C \cdot \left(1+\frac{1}{\varepsilon}\right) + zW_L = 0 \tag{19}$$

$$(-\varepsilon)\cdot\frac{pz}{m} + (1+\varepsilon) - z\cdot g(z-z^*;\psi,v_x) = 0 \tag{20}$$

Note that with $p = 0$, (19) and (20) are identical to the first-order conditions (16) and (17) in the previous section. Moreover, one can verify that, as p (and thus r) increases and the amount $m - r$ retained by the doctor as a pure capitation element approaches zero, the solution to (19) and (20) will approach the solution to (8) and (9) for the fee for service case, if the coinsurance parameter σ is set equal to zero.

Consider finally Figure 3, where the equilibria for the cases $p = 0$ and $m = r = pz$ are shown. The preceding arguments imply that at the former, there is negative supplier-induced demand, whereas at the latter it is positive. By varying the parameter p, any solution intermediate between xx and xx can be attained.

From an economic point of view, a solution of particular interest is that value of z which, loosely speaking, represents the socially efficient treatment

[14] An arrangement of this kind could be seen as falling somewhere between two types of insurance forms that currently exist in the United States: On the one hand, a Health Maintenance Organization organized as an "independent practice association" in which participating physicians continue to practice independently, but are paid by the HMO on the basis of capitation, and, on the other hand, a Preferred Provider Organization in which patients are restricted to doctors on the insurer's list and participating physicians are paid on the basis of fee for service, at rates agreed upon between the doctors and the insurer. A similar system, albeit on a small scale, has been used in the U.K. where the general practitioners with whom patients have to be signed up to get access to the National Health Service get most of their income through capitation, but are paid on the basis of fee for service for certain specific types of care (for example, some forms of prevention) that the NHS wants to encourage (see, e.g., OECD (1992)).

intensity at which the representative doctor's marginal rate of substitution of income for leisure $-W_L/W_C$ equals the representative patient's marginal rate of substutition of income for physician services U_H'/U_C'. If one supposes that this treatment intensity (call it z^{opt}) is larger than z^{CAP}, the relatively low equilibrium value under pure capitation, and smaller than z^{FFS}, the equilibrium value under fee for service, we have

Proposition 10. When $z^{CAP} < z^{opt} < z^{FFS}$, there exists a mixed payment system with $p > 0$ which results in the economically efficient treatment intensity z^{opt}.

The ratio pz/m associated with this solution can be interpreted as the efficient degree of supply side cost sharing. Note also that since patients are fully insured (that is, there is no demand side cost sharing) at this equilibrium, it yields a higher level of welfare for the representative patient than a fee for service equilibrium with the second-best optimal value of the coinsurance parameter σ.

Figure 3 about here

4 Conclusion.

In this paper, I have analyzed a model which combines three features that most analysts agree are important in health services markets: first, heterogeneity and monopolistic competition among providers (physicians); second, information asymmetry between providers on the one hand and users on the other; and third, the presence of insurance and a potential moral hazard effect (which also arises because of information asymmetry, in this case between insurers and providers). While each of these elements have been discussed by others, the implications of their presence together have not, to my knowledge, been previously considered in the context of a single formal model. The main conclusion is that in the environment described by the model, the most efficient payment system is one that involves zero demand side cost sharing, and a degree of supply side cost sharing that depends on the intensity of price competition (the elasticity of demand for the individual provider's services), and on the parameters representing the degree of information asymmetry.

From an empirical point of view, the result that a mixed payment system yields higher welfare than either pure fee for service or pure capitation is

consistent with both the declining role of conventional insurance and fee for service payment, on the one hand, and the failure of HMOs (which in some respects come closest to the pure capitation model) to capture a growing share of the market, in the United States' system of private insurance, on the other. As described in, e.g., Phelps (1992, 320-3), the fastest-growing forms of insurance are prepayment plans such as Independent Practice Associations (IPAs), Preferred Provider Organizations (PPOs), second-opinion plans, and other forms of "managed-care" insurance which Phelps describes as "sort of a halfway house between the pure HMO and a fee-for-service system" (1992, p. 320). For purposes of empirical implementation, however, one would probably want to extend the model to take account of the fact that a substantial capitation element in paying providers may create problems of risk discrimination and adverse selection (as discussed in Newhouse (1996)), and also that real-world insurance plans cover the costs of other inputs in the production of health (such as hospital services), as well as those provided by physicians.

5 Appendix 1: Derivations

The first-order conditions for the doctor's problem in the fee for service case can be written:

$$\frac{\partial W}{\partial p} = W_C \cdot \bar{s}z + \frac{\partial \bar{s}}{\partial p} \cdot (W_C \cdot pz + W_L \cdot z) \equiv F_1 = 0 \tag{21}$$

$$\frac{\partial W}{\partial z} = \bar{s} \cdot (W_C \cdot p + W_L) \cdot (1 + \frac{z}{\bar{s}} \frac{\partial \bar{s}}{\partial z}) \equiv F_2 = 0 \tag{22}$$

Totally differentiating 8 and 9 with respect to p, z, ε, σ and ψ yields:

$$\begin{bmatrix} F_p^1 & F_z^1 \\ F_p^2 & F_z^2 \end{bmatrix} \begin{bmatrix} dp \\ dz \end{bmatrix} = \begin{bmatrix} -F_{|\varepsilon|}^1 d \mid \varepsilon \mid \\ -F_\sigma^2 d\sigma - F_\psi^2 d\psi \end{bmatrix} \tag{23}$$

where

$$F_p^1 = R \cdot (W_C + W_{CC} p\bar{s}z) + W_{LC}\bar{s}z$$

$$F_z^1 = \bar{s}p^2 R W_{CC} + p\bar{s}(R+1)W_{LC} + \bar{s}W_{LL} < 0,$$

and $R \equiv (1 + \frac{1}{\varepsilon})$. The slope of F^1 in Figure 2 in the text is $-F_z^1/F_p^1$; thus the assumption that equilibria occur on the upward-sloping portion of F^1 implies $F_p^1 > 0$ (since $F_z^1 < 0$).

Differentiation of (9) in the text produces $F_p^2 = \sigma z(\partial g(.)/\partial\delta)dz^*/dp < 0$ and $F_z^2 = -g(.) - z(\partial g(.)/\partial\delta) < 0$. (This implies that the slope of F^2 in Figure 2 is negative.[15]) From (8), $F_{|\varepsilon|}^1 = W_C \cdot \mid \varepsilon \mid^{(-2)} > 0$. Proposition 2 then follows through straightforward application of Cramer's rule. Proposition 3 is implied by $F_\sigma^2 < 0$. (Note that $F_\sigma^2 = (p/\sigma)F_p^2$).

To establish Proposition 4, note that the sign of $F_\psi^2 = -z(\partial g(.)/\partial\psi)$ is the opposite of the sign of $h'(u_1) - h'(u_0)$ where h is the standard normal distribution and $u_1 = (\psi - \delta)/\sqrt{v_x}$, $u_0 = (\psi + \delta)/\sqrt{v_x}$. Since $h'(u)$ reaches a minimum at $u = \sqrt{.5}$, the assumption $u_1 = (\psi - \delta)/\sqrt{v_x} > \sqrt{.5}$ guarantees that $h'(u_0) < h'(u_1) < 0$, so $\partial g/\partial\psi < 0$, which in turn implies $F_\psi^2 > 0$. Proposition 4 then follows by Cramer's rule. To establish Proposition 5, finally, note that for any $\sqrt{\varepsilon}$, no matter how small, (11) equals zero at $\delta = 0$, but $\partial(zg(\delta;\psi,v_x))/\partial\delta$ goes to infinity as v_x goes to zero.

I consider the pure capitation and the mixed models together since capitation is just the special case of the mixed model when $p = 0$. The first-order conditions (19) and (20) for the mixed case can be written:

$$\frac{\partial W}{\partial m} = \bar{s}W_C + \frac{\partial s}{\partial m}(W_C \cdot (m - r + pz) + zW_L) \equiv G^1 = 0 \qquad (24)$$

$$\frac{\partial W}{\partial z} = p\bar{s}W_C + \bar{s}W_L + \frac{\partial \bar{s}}{\partial z}(mW_C + zW_L) \equiv G^2 = 0 \qquad (25)$$

To derive (19), one uses the actuarial fairness condition $r = pz$ and $\partial\bar{s}/\partial m = \partial s/\partial m = s\varepsilon/m$ in (24) and simplifies; as before, $R = (1 + \varepsilon^{-1})$. To obtain (20), one substitutes $W_L = -mW_CR/z$ from (19) into (25) and simplifies (noting that $(1 - R) = -1/\varepsilon$), using (11). Totally differentiating (19) and (20) produces

$$\begin{bmatrix} G_m^1 & G_z^1 \\ G_m^2 & G_z^2 \end{bmatrix} \begin{bmatrix} dm \\ dz \end{bmatrix} = \begin{bmatrix} -G_{|\varepsilon|}^1 d \mid \varepsilon \mid \\ (1 - \frac{pz}{m})d \mid \varepsilon \mid -G_\psi^2 d\psi + G_p^2 dp \end{bmatrix}$$

The term G_m^1 is identical to F_p^1 in the fee for service case except that m replaces pz; it is positive for the same reason that F_p^1 is. $G_z^1 = \bar{s}m \cdot R \cdot (pW_{CC} +$

[15]This will still be true if the actuarial-fairness condition $m = (1 - \sigma)pz$ is explicitly imposed on the model, under the reasonable conditions noted in footnote 9 in the text. Note that with this condition, one would have $\partial g/\partial z = (\partial g/\partial z) \cdot (1 - (\partial z^*/\partial m)(dm/dz))$.

$W_{LC}) + \bar{s}z \cdot (W_{LC}p + W_{LL}) + W_L < 0$. The slope of G^2 is $-G_z^2/G_m^2$, where $G_m^2 = - \mid \varepsilon \mid pz/m^2 \leq 0$, with equality in the pure capitation case where $p = 0$ The sign of G_z^2 is ambiguous in general. However, it must be negative in the interval $\{z_0, z_1\}$ defined by the solutions to $1 = z_1 g(z_1 - z^*; \psi, v_x)$ and $1 + \varepsilon = z_0 g(z_0 - z^*; \psi, v_x)$ corresponding to the pure fee for service and capitation cases, respectively. It follows that G^2 in Figure XXX is downward-sloping; it is vertical at the left end of the interval where $p = 0$.

Proposition 7 in the text follows from Cramer's rule, with $G_{|\varepsilon|}^1$ positive and G_m^2 equal to zero because p equals zero in the pure capitation case. Proposition 8 is the converse of Proposition 4: As in the latter, an increase in ψ raises the required absolute value of δ in equilibrium, except that with $\delta = z - z^*$ being negative in the capitation case, this requires a smaller z, not a larger one as in the capitation case. Also note that with $\delta < 0$, it is now the term $(\psi + \delta)/\sqrt{v_x}$ that is closer to the critical value for the derivative of the standard normal distribution. Proposition 9 is proved the same way as Proposition 5. Finally, although Proposition 10 follows directly from continuity, one can also use Cramer's rule with $G_p^2 = \varepsilon z/m$ to establish $dz/dp > 0$.

6 References

Baumgardner, James R. (1991), "The Interaction between forms of insurance contracts and types of technological change in health care", Rand Journal, 22, 36-53

Blomqvist, Åke (1991),"The doctor as double agent: information asymmetry, health insurance, and medical care", Journal of Health Economics, 10, 411-32

Blomqvist, Åke (1997), "Optimal non-linear health insurance", Journal of Health Economics, forthcoming

Dixit, Avinash K., and Joseph E. Stiglitz (1977), "Monopolistic competition and optimum product diversity", American Economic Review, 67, 297-308

Dranove, David (1988), "Demand inducement and the physician-patient relationship", Economic Inquiry, 26, 281-98

Ellis, Randall P., and Thomas McGuire (1986), "Provider response to prospective payment: Cost sharing and supply", Journal of Health Eco-

nomics, 5, 129-51

Ellis, Randall P., and Thomas McGuire (1993), "Supply-side and demand-side cost sharing in health care", Journal of Economic Perspectives, 7, 135-51

Feldman, Roger, and Bryan Dowd (1991), "A new estimate of the welfare loss of excess health insurance", American Economic Review, 81, 297-301

Feldstein, Martin (1973), "The welfare loss of excess health insurance", Journal of Political Economy, 81, 251-80

Ma, Ching-to Albert, and Thomas McGuire (1995), "Optimal health insurance and provider payment", Boston University, Industry Studies WP #59

Manning, Willard G., *et. al* (1987), "Health insurance and the demand for medical care: Evidence from a randomized experiment", American Economic Review, 77, 251-77

Newhouse, Joseph P., *et al.* (1995), Free for all? Lessons from the Rand health insurance experiment. Cambridge, Mass.: Harvard University Press.

Newhouse, Joseph P. (1996), "Reimbursing health plans and health providers: Selection vs. efficiency in production", Journal of Economic Literature, 34, 1236-63.

Organization for Economic Cooperation and Development (1992), The reform of health care: A comparative analysis of seven OECD countries. Health Policy Studies No. 2. Paris: OECD.

Pauly, Mark V. (1968), "The economics of moral hazard", American Economic Review, 58, 231-7.

Pauly, Mark V. (1980), Doctors and their workshops. Chicago: University of Chicago Press.

Pauly, Mark V. (1986), "Taxation, health insurance, and market failure in the medical economy", Journal of Economic Literature, 87-99

Pauly, Mark V. (1991), "Fee schedules and utilization", in H. E. Frech III, ed., Regulating doctors' fees. Washington, D. C.: American Entreprise Institute Press, pp. 288-305

Phelps, Charles E. (1992), Health economics. New York: Harper Collins.

Satterthwaite, Mark A. (1979), "Consumer information, equilibrium, industry price, and the number of sellers", Bell Journal of Economics, 10, 583-502

Selden, Thomas M. (1990), "A model of capitation", Journal of Health Economics, 9, 397-409

Zeckhauser, Richard (1970), "Medical insurance: A case study of the tradeoff between risk spreading and appropriate incentives", Journal of Economic Theory, 2, 10-26

Wedig, Gerard J. (1993), "Ramsey pricing and supply-side incentives in physician markets", Journal of Health Economics, 12, 365-84

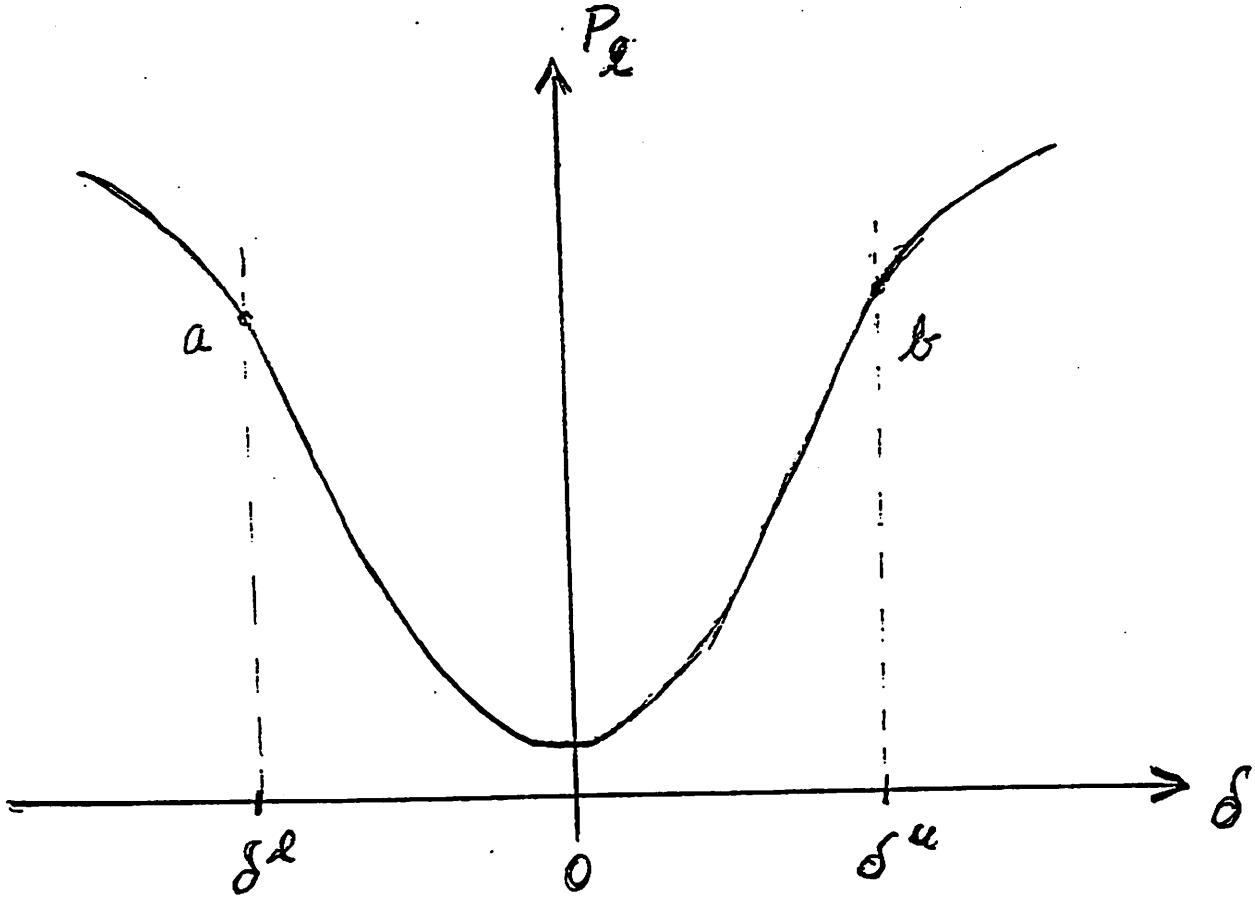

Figure 1

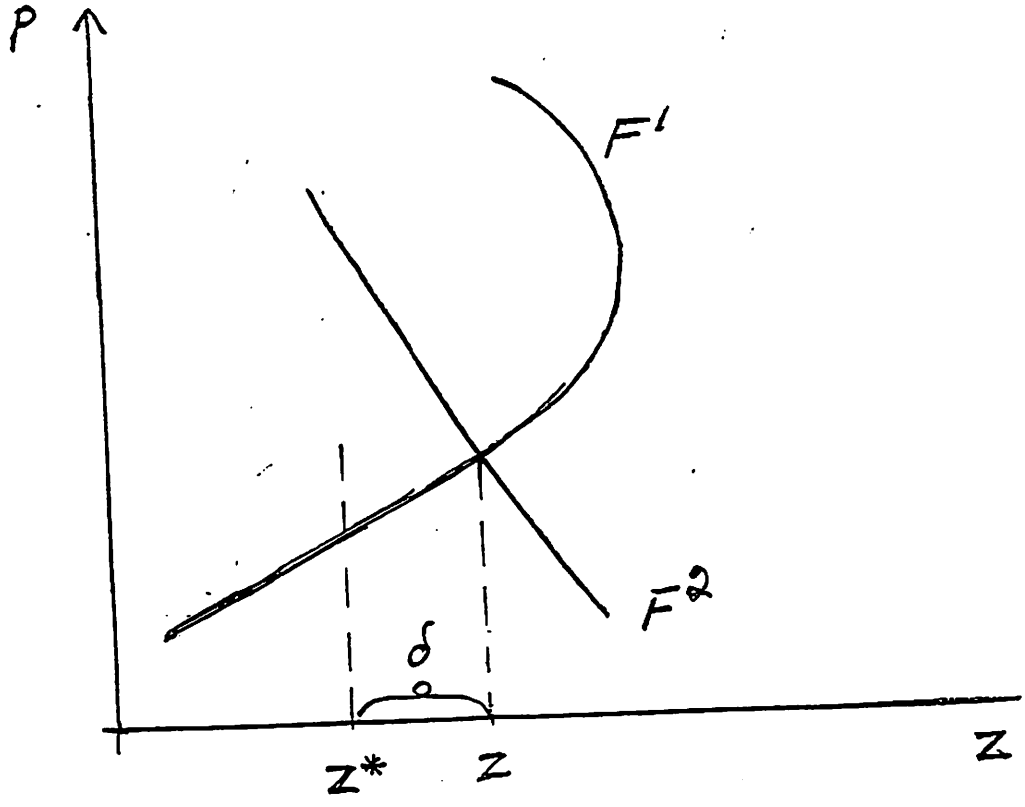

Figure 2.

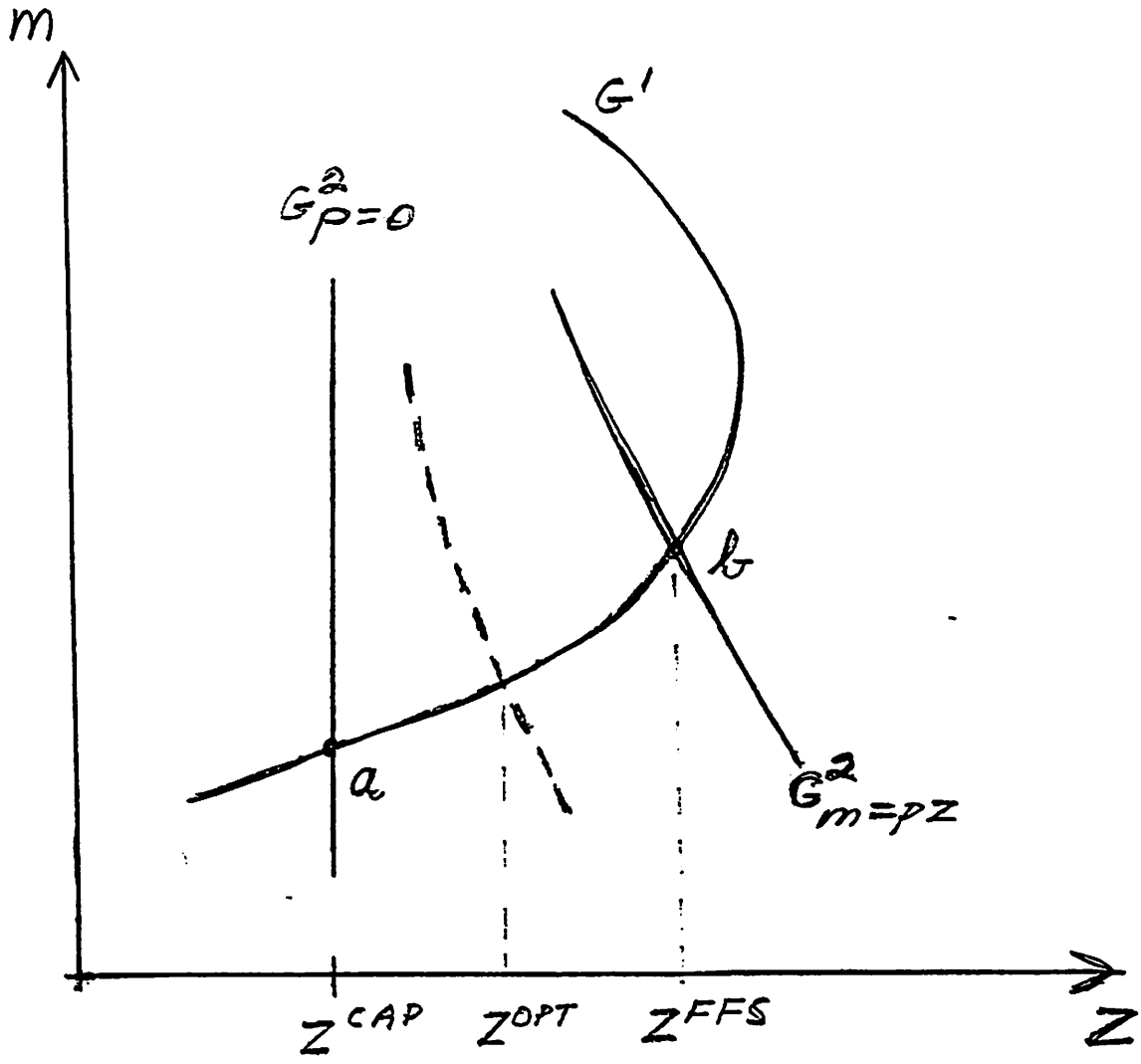

Figure 3.