

Insurance and Incentives for Medical Innovation

Alan M. Garber*

Department of Veterans Affairs, Stanford University, and NBER

Charles I. Jones

U.C. Berkeley and NBER

and

Paul M. Romer

Stanford University and NBER

February 20, 2006 — Version 1.0

This paper studies the interactions between health insurance and the incentives for innovation. Although we focus on pharmaceutical innovation, our discussion applies to other industries producing novel technologies for sale in markets with subsidized demand. Standard results in the growth and productivity literatures suggest that firms in many industries may possess inadequate incentives to innovate. Standard results in the health literature suggest that health insurance leads to the overutilization of health care. Our study of innovation in the pharmaceutical industry emphasizes the interaction of these incentives. Because of the large subsidies to demand from health insurance, limits on the lifetime of patents and possibly limits on monopoly pricing may be necessary to ensure that pharmaceutical companies do not possess excess incentives for innovation.

*Prepared for RAND/NIH Workshop on the Economic Consequences of Medical Research, June 1, 2005, RAND Corporation, Arlington, VA. We are grateful to Dean Scrimgeour for excellent research assistance. Garber's research is supported in part by the Department of Veterans Affairs, the National Institute on Aging, and an Investigators Award in Health Policy Research from the Robert Wood Johnson Foundation. Jones also thanks the Toulouse Network on Information Technology for research support.

1. INTRODUCTION

When a market suffers from a distortion that cannot be removed, second-best policy prescriptions may run counter to standard first-best results. This kind of distortion is inevitably present in any market where new goods are continually being introduced. Innovation is associated with a fundamental nonconvexity that renders infeasible the standard first-best market equilibrium based on price-taking competition. In the market that we consider here, the market for pharmaceuticals, this nonconvexity is resolved by granting monopoly power to the suppliers of new compounds.

It should come as no surprise that intuitions honed on first-best analysis are a poor guide to the welfare analysis of this market. When consumers purchase pharmaceuticals, moral hazard — the insurance subsidy that enables patients to pay only a fraction of the cost of pharmaceutical products they consume (Pauly 1968) — creates a distortion that can improve welfare. Intuitively, monopoly markups lead to suboptimally low consumption of pharmaceuticals, while a coinsurance rate of less than 100% leads to excessive consumption. Under some circumstances, these effects may offset one another. This point is not new. A similar observation was made by Crew (1969) more than 35 years ago. Nevertheless, the difference between the welfare effect that insurance subsidies have on spending for pharmaceuticals as opposed to spending on other types of health services seems not to have received the attention it deserves. Nor has the potential it creates for excessive innovation in the pharmaceutical industry.

In this paper, we study the optimal provision of medical goods such as pharmaceuticals. We break our analysis into two parts. We use the phrase “static efficiency” to characterize departures from the optimal utilization of a drug that has already been developed, and “dynamic efficiency” to characterize the degree to which innovators have the correct incentives to incur the fixed cost necessary to introduce a new compound.

We start with the ex post analysis of static efficiency. After a pharmaceutical firm has introduced a new compound, we look for a coinsurance rate that achieves efficient utilization of the drug. Next we observe that if the development cost is known only to the firm, the firm will have the right ex ante incentives to develop the drug if its anticipated profits equal the consumer surplus the drug will generate.

In the full equilibrium that we describe, all consumers pay for a pharmaceutical insurance policy, and a sick consumer pays a low out-of-pocket cost for the pharmaceutical. To facilitate the analysis and link it to data on the incidence of a medical condition in the population, we specify a distribution that characterizes the various benefits different people receive from the pharmaceutical. For example, people with other risk factors for heart disease and higher initial levels of blood cholesterol may receive greater benefits from a cholesterol-reducing drug than consumers at lower risk. It is easiest to interpret the demand function implied by this distribution function if we assume that consumers make a zero-one decision about whether to take one standardized course of treatment with the pharmaceutical, and we rely on this simplifying assumption to describe our results. The results can, however, readily be extended to the case where consumers also decide what quantity of the drug to consume. Faced with this simple zero-one decision, consumers purchase the drug if their out-of-pocket cost is less than or equal to the consumer surplus they receive.

Because we specify the demand for the good in terms of a distribution function for its benefits, we can highlight a useful connection between the hazard rate for consumer benefits and the elasticity of demand for the good. This allows us to link the hazard rate to the simple monopoly price that a patent-owning firm would charge, and to characterize the profits the firm receives relative to the consumer surplus it generates in terms of the behavior of the hazard rate. In particular, consider the case where the coinsurance rate

is set to achieve static efficiency. In this case, as long as the hazard rate does not fall too quickly (and certainly for a constant or increasing hazard), profits will exceed consumer surplus. Dynamic efficiency then requires a restriction on monopoly power, such as finite patent life.

These results are not mere curiosities. In many years, pharmaceuticals have been the fastest-growing portion of health care expenditures, and accounted for about 11.5% of U.S. health expenditures in 2005 according to Heffler, Smith, Keehan, Borger, Clemens and Truffer (2005). In the rich countries of the world, they are arguably the most important conduit for the practical application of fundamental biomedical advances. The pharmaceutical industry may also be a prototype industry for an economy that relies increasingly on private firms to develop and put into use fundamental scientific and technological knowledge.

These results are also relevant to current policy debates. The movement away from low coinsurance rates and toward systems in which consumers pay the full cost of a medical treatment (as embodied in the recently introduced Health Savings Accounts) might be welfare-improving for traditional services that are priced near marginal cost. It may, however, harm social welfare if applied to purchases of new drugs.

Gaynor, Haas-Wilson and Vogt (2000), in an extension of the Crew (1969) analysis, have shown that health insurance provided by a competitive market can achieve static efficiency when the copayment for each good or service is set optimally. Taking the price of medical care as exogenous, they show that price reductions, at least when price is above marginal cost, increase welfare. Their analysis does not consider whether the resulting profits offer appropriate rewards for innovation. Our analysis considers an alternative benchmark where monopoly firms set medical prices given an exogenous coinsurance rate. We study the optimal coinsurance rate and incentives for innovation in this framework.

In a recent paper Lakdawalla and Sood (2005) consider dynamic efficiency. They view the health insurance contract as a two-part tariff and show that efficient innovation may result. In particular, the health insurance premium may extract expected consumer surplus from patients while the coinsurance rate ensures optimal utilization. The innovator captures consumer surplus through a competitive insurance industry that pays a two-part tariff to access the innovation. This elegant result illustrates that efficient innovation may result from nonlinear pricing arrangements. Our analysis complements this approach by examining what happens to utilization and incentives to innovate when monopoly innovators can only charge linear prices. We show that when coinsurance rates ensure optimal utilization, incentives for innovation are often excessive.

2. THE BASIC MODEL

The economy contains a collection of people indexed by i on the interval $[0, 1]$. A fraction of the population, s , becomes sick with an illness that can be treated by the single pharmaceutical product that is the focus of the analysis here. Everyone has the same ex ante risk of becoming ill. Those who become sick suffer a loss of utility, denoted by δ_i . For these individuals, utility takes the form

$$U = u(c_i) - \delta_i. \quad (1)$$

The first term in this expression is the utility from consuming standard goods. The second term, δ_i , represents the disutility the person suffers from being sick.

The drug that can treat this condition does not produce the same expected relief in every individual. We assume that consumers know this conditional expected benefit, which we denote by v_i . If a sick consumer uses the phar-

maceutical and consumes other goods at the level c_i , her utility will be

$$U = u(c_i) - \delta_i + v_i.$$

For a curative treatment, $\delta_i = v_i$, an exceptional circumstance that is not required for the model. The fraction s also could equal 1, and v_i could exceed δ_i (which can be zero). Thus the formulation is consistent with a pharmaceutical that positively affects well-being in the absence of disease. The random variable v_i is distributed among the sick according to a distribution function $F(v)$.¹ The utility loss term δ_i is given exogenously for consumer i . Because this disutility term enters utility in an additively separable fashion, we can neglect it in analyzing consumer i 's decisions and in calculating aggregate welfare. This means, in particular, that the distribution of δ_i does not affect anything that follows.

As noted above, we assume for the sake of simplicity that consumers can choose whether or not to receive a common, single dose of the drug in question. To focus only on the risk associated with being sick (or more precisely, on the risk that someone will have an opportunity to give up some income to benefit from a treatment) and to suppress the risk that individuals will receive different levels of income, we assume that all individuals have the same initial income \bar{y} . (A more realistic but more complicated approach would be to assume that income risk is not insurable.)

Our description of the supply side is as simple as possible. The pharmaceutical firm must incur a fixed cost W to develop the drug. This cost will matter only when we consider the ex ante analysis of the decision to introduce the drug in Section 4 below. After developing the drug, the firm can produce each additional dose at a constant cost of w per consumer.

2.1. The First Best

¹Chang and Kim (2003) use a related approach to generate an aggregate labor supply curve based on heterogeneous individual labor productivities.

The first-best optimum is the solution to the social planner's maximization problem:

$$\Omega^* = \max_{v^*} u(\tilde{c}) + s \int_{v^*}^{\infty} vF'(v)dv \quad (2)$$

subject to

$$\tilde{c} = \bar{y} - s(1 - F(v^*))w. \quad (3)$$

Note that we have dropped the δ_i terms because they only add a constant term

$$\int_0^1 \delta_i di$$

to the objective function. By integrating utility from standard goods over all consumers we obtain the single term $u(\tilde{c})$. Because preferences are additively separable, equating the marginal utility of consumption across consumers requires that all people have the same consumption in the first-best equilibrium. With these simplifications, social welfare equals the sum of utility from other goods, plus the benefits from treating people who have a benefit level greater than v^* , minus the neglected δ_i terms.

The constraint in the optimization problem shows that \tilde{c} equals per capita income minus the cost of treating people who are sicker than a cutoff value v^* . A fraction s of all individuals become sick. A fraction $1 - F(v^*)$ of these sick individuals will receive a benefit from treatment that exceeds v^* . The per-person cost of treatment is w .

The first-order condition for this social optimization problem implies that the optimal choice of the cutoff value v^* satisfies

$$v^* = u'(\tilde{c})w. \quad (4)$$

This has a natural interpretation in terms of benefit and cost. The variable w represents the cost in goods of an additional treatment. The expression on the

right-hand side, $u'(\tilde{c})w$, represents the cost in foregone utility of an additional treatment. The first-order condition says that the social planner should first allocate treatment to the patients who receive the greatest benefit and continue only until the marginal benefit v^* equals the marginal cost.

3. EQUILIBRIUM WITH COINSURANCE PAYMENTS

If each consumer's benefit v_i were observable, there would be no problem implementing this first best optimum. If the patient (and the patient's doctor) can observe v_i but the insurer cannot, it is impossible to support this equilibrium. Any patient with $v_i > 0$ will report a value greater than v^* .

Suppose therefore that the insurance contract pays a fraction $1 - \lambda$ of the cost of the drug when someone buys, leaving a required coinsurance payment of λp for the individual. For the moment, we assume the coinsurance rate λ is exogenously given. Later on, we will study the optimal setting of the coinsurance rate (for example, by the government for publicly-provided health plans or through regulation). If λp is small relative to total income and consumption, the loss in utility associated with the payment λp will be small, and the equilibrium with coinsurance payments will provide almost full insurance.

We assume that a single pharmaceutical firm has the exclusive right to produce the pharmaceutical that treats the condition in question. The monopolist observes the total demand curve for the drug and selects the monopoly price at which the drug is to be sold to consumers.

3.1. The Consumer's Problem

Each person who contracts the disease makes the following calculation. A sick person has income $y \equiv \bar{y} - (1 - \lambda)px(\lambda p) + x(\lambda p)(p - w)$. The first term of this equation is the endowment, \bar{y} . The second term is the actuarially fair insurance premium that covers the per person cost of insurance, where p is the price of the drug and x is the total quantity demanded at that price.

Finally, we assume the drug companies are owned by the agents. The last term is equal to each person's share of the ex post profits earned by the firm.

A patient will purchase the drug if

$$u(y - \lambda p) + v_i \geq u(y).$$

Rewriting, person i purchases the drug whenever

$$v_i \geq u(y) - u(y - \lambda p) \approx u'(c)\lambda p \quad (5)$$

where $c = y - \lambda p$ is the consumption of someone who purchases the drug. This approximation is valid as long as λp is small relative to y , which is something that we will need to check in equilibrium. Note that λp need not be small when λ is.

Let q denote the out-of-pocket payment that the consumer makes, $q = \lambda p$. Then we can write the demand curve as

$$x(q) = s(1 - F(u'(c)q)). \quad (6)$$

Faced with a direct cost q , all sick people with a benefit greater than the cutoff value $\bar{v} = u'(c)q$ purchase the drug.

3.2. The Monopoly Problem

We assume that the producer cannot engage in price discrimination. The monopolist faces a standard profit maximization problem with constant marginal cost except that in this case, the monopolist receives payments from both the consumer and the insurance company. When the consumer pays q , the monopolist receives $p = \frac{q}{\lambda}$:

$$\max_q \pi \equiv \left(\frac{q}{\lambda} - w\right)x(q). \quad (7)$$

The first order condition for this problem yields the equation

$$\left(\frac{q}{\lambda} - w\right)x'(q) + \frac{1}{\lambda}x(q) = 0.$$

Multiplying through by $\frac{\lambda}{x(q)}$ yields

$$\frac{q}{x(q)}x'(q) - \frac{\lambda w}{q} \frac{q}{x(q)}x'(q) + 1 = 0.$$

If we let ϵ stand for the elasticity of the consumer demand curve,

$$\epsilon \equiv -x'(q) \frac{q}{x(q)}.$$

the first order condition for the optimal payment by the consumer q^* reduces to

$$\frac{q^*}{\lambda} = \frac{\epsilon}{\epsilon - 1} \cdot w.$$

If we write this in terms of the payment p received by the monopolist, this takes on the more familiar form,

$$p^* = \frac{\epsilon}{\epsilon - 1} \cdot w, \tag{8}$$

where ϵ is evaluated at the price $q^* = \lambda p^*$ paid by the consumer.

For the demand curve given above in equation (6), the derivative $x'(q)$ can be written as

$$x'(q) = -sF'(u'(c)q)u'(c).$$

The expression for ϵ can then be written as

$$\epsilon = h(\bar{v})u'(c)\lambda p, \tag{9}$$

where $h(v) \equiv F'(v)/(1 - F(v))$ is the hazard rate corresponding to the distribution F and $\bar{v} \equiv u'(c)\lambda p$ is the cutoff level for purchasing the drug. (Henceforth, it will be easiest to work with the price p the monopolist receives rather than the price q the consumer pays.)

In our application, the hazard rate $h(v)$ has the following interpretation. Suppose we take the set of people whose health benefit from treatment is at least v and consider a small interval $(v, v + \mu)$. The proportion of this group

with a benefit inside this small interval is equal to $h(v)\mu$. Roughly speaking, as v increases, $h(v)$ tells us the rate at which people drop out of the group of people who have a value greater than or equal to v . (Using traditional language, this group would be called “survivors,” but in our pharmaceutical application, this term instead means the people gain a benefit greater than the level in question.)

The key result here is that the elasticity of demand for the drug is this hazard rate, multiplied by the marginal utility of consumption and by the out-of-pocket cost to the consumer.

Combining this expression for the demand elasticity with the monopoly markup rule in equation (8) yields a more revealing expression for the monopoly price received by the pharmaceutical manufacturer:

$$p^* = w + \frac{1}{h(\bar{v})u'(c)\lambda}. \quad (10)$$

This is still an implicit equation for the solution since both \bar{v} and c depend on the monopoly price. As we will see, it nevertheless offers insight into the pricing problem of the firm. Other things equal, the markup increases with a decrease in the hazard rate, a decrease in λ , and an increase in income that lowers the marginal utility of consumption.

The profits of the monopolist are given by

$$\pi(\lambda) = \frac{x(\lambda p)}{h(\bar{v})u'(c)\lambda}. \quad (11)$$

In this expression, x is the number of treatments sold, and the inverse of $h(\bar{v})u'(c)\lambda$ is the profit margin earned on each treatment.

3.3. Social Welfare and Consumer Surplus

As shown in Appendix A, social welfare in this equilibrium is approximately

$$\Omega(\lambda p) \approx u(\bar{y}) + CS(\lambda p), \quad (12)$$

where

$$CS(\lambda p) \equiv s \int_{\bar{v}}^{\infty} v F'(v) dv - u'(\bar{y}) w x(\lambda p). \quad (13)$$

The first equation says that social welfare is the sum of two terms. The first $u(\bar{y})$ is the utility consumers would receive from consuming all of their income. The second term is a consumer surplus-like measure, CS . CS , in turn, depends on two terms. The first is the total utility gain generated by drug treatment, the sum of the v_i 's across all people. Subtracted from this is the total cost of producing these treatments, $w x(\lambda p)$, which is converted into its utility equivalent by multiplying by the marginal utility of consumption.

In this expression, we have neglected the loss in utility associated with the incomplete insurance caused by the coinsurance payment λp . This loss is second order in λp and for small value of λp it can be neglected. We will focus instead on two other losses. The first is the static efficiency loss that arises when the price of the drug to the consumer is too high. The second is the dynamic efficiency loss that can arise when a drug with a value CS that is greater than its development cost W is not introduced, and the mirror image case, when a drug with a value CS that is less than W is introduced.

4. EFFICIENT UTILIZATION AND THE INCENTIVE TO INNOVATE

In equation (4), we showed that the efficient utilization of the drug requires all people with health benefits $v_i \geq v^* \equiv u'(\tilde{c})w$ to receive treatment where \tilde{c} is the common consumption level. In our equilibrium with coinsurance, the treatment cutoff value is $\bar{v} \equiv u'(c)\lambda p$ where c is the consumption for someone who is sick and makes the coinsurance payment.

First notice that \tilde{c} , c , and y are related by

$$\tilde{c} = \bar{y} - w x(w).$$

$$y = \bar{y} - x(\lambda p)(\lambda p - w).$$

$$c = y - \lambda p.$$

In particular, as long as \bar{y} is sufficiently large compared to λp , we have

$$u'(\bar{c}) \approx u'(y) \approx u'(c) \approx u'(\bar{y}). \quad (14)$$

(These approximations would also hold exactly if the utility function were linear). From now on, we make the assumption that this approximation holds, so that the marginal utility of consumption is roughly the same whether one buys the drug or not.

Under this assumption, efficient utilization of the drug requires

$$\bar{v} = v^* \iff \lambda p = w. \quad (15)$$

That is, all people with health benefits at least as large as v^* need to get the drug. This requires the intuitive condition that the price faced by the consumer, λp , be equal to the marginal cost of producing the drug, w . As long as this is true, the economy can achieve full static efficiency.

What about the incentives for the pharmaceutical company to introduce the drug in the first place? The pharmaceutical company will consider the expected present discounted value of profits that can be earned when deciding whether to develop and introduce a new drug. From the social welfare measure given above in equation (12), though, we saw that social welfare depends on the consumer surplus-like measure, CS .

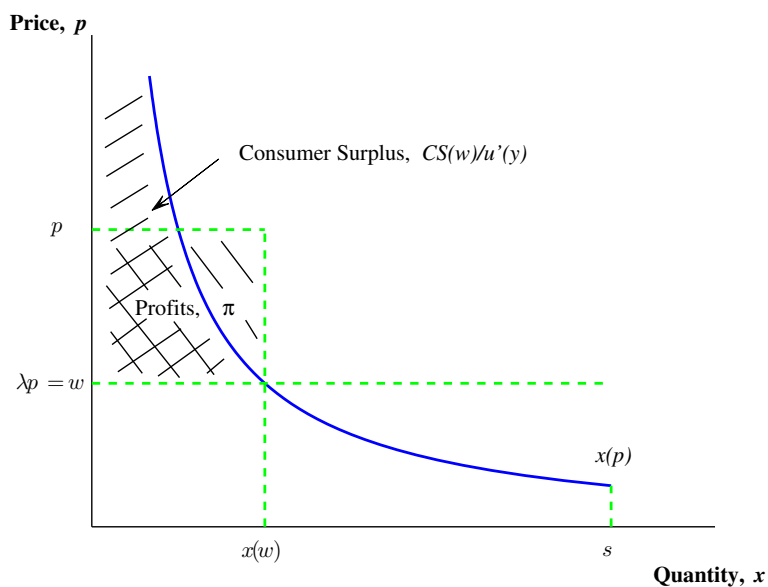
A useful measure to keep in mind when discussing the incentives to innovate, then, is the ratio of profits to consumer surplus, which we will call ρ :

$$\rho \equiv \frac{\pi(\lambda)}{CS(\lambda p)/u'(\bar{y})}. \quad (16)$$

In this expression, we've divided CS by the marginal utility of consumption to convert it from units of utility into units of consumption.

Graphically, Figure 1 shows the profits and consumer surplus for the special case in which the utilization of the drug is efficient, i.e. when $\lambda p = w$. Notice

FIGURE 1. Profits and Consumer Surplus with Efficient Utilization



that, in contrast to the standard monopoly case with no subsidy to demand, the profit rectangle here includes a portion outside the consumer surplus. This raises the possibility — which is confirmed in several examples below — that the profits earned by the pharmaceutical company can exceed the consumer surplus associated with the introduction of the drug treatment.

5. EXAMPLES

5.1. A First Example: the Pareto Distribution

As a simple illustrative example, suppose that treatment benefits are distributed according to the Pareto distribution:

$$F(v) = 1 - \left(\frac{v}{v_0}\right)^{-\alpha}, \quad \alpha > 1, \quad (17)$$

on the interval $[v_0, \infty)$. The Pareto distribution often appears as a good description of the upper tail of the income distribution, which has a “thick” tail.

To see what this means, suppose we consider that part of the population with incomes greater than some cutoff y . Now ask what fraction of this rich population has incomes that exceed y by more than 25%. For the Pareto distribution, this fraction is invariant to the level of the cutoff y , whereas for distributions with thinner tails, the fraction declines as the income cutoff rises. Saez (2001) shows that this invariance accurately characterizes U.S. incomes in the early 1990s between \$100,000 and \$30 million. Assuming a Pareto distribution for treatment benefits, then, assumes that there is a long, thick tail of people with very high benefit levels.

When treatment benefits obey a Pareto distribution, the results from the model are very simple. First, the demand curve for the drug exhibits a constant price elasticity:

$$x(\lambda p) = sv_0^\alpha (u'(\bar{y})\lambda p)^{-\alpha}. \quad (18)$$

Second, the monopoly price is given by a constant markup over marginal cost, where the markup is determined by the constant elasticity of demand:

$$p^* = \frac{\alpha}{\alpha - 1} \cdot w. \quad (19)$$

The monopoly price increases with marginal cost and decreases in the elasticity of demand.

Finally, the ratio of the monopoly profit to consumer surplus is given by²

$$\rho(\lambda) = \frac{1}{1 + \alpha \left(\frac{\alpha}{\alpha - 1} \lambda - 1 \right)}. \quad (20)$$

This formula shows that the ratio of profits to consumer surplus depends on the product of the monopoly markup $\mu \equiv \alpha/(\alpha - 1)$ and the coinsurance rate λ . In particular, a lower coinsurance rate tends to raise the ratio of profits

²Notice that this expression is only defined for $\lambda > \left(\frac{\alpha - 1}{\alpha} \right)^2$. If $\lambda p < w$ by too much, which occurs at this point, then the consumer surplus shrinks to zero: the cost of producing the drug exactly offsets the benefit received.

TABLE 1.
The Ratio of Profits to Consumer Surplus, Pareto Example

Markup μ	Implied α	— The Coinsurance Rate λ —			
		0.20	0.33	0.50	1.00
2	2.00	...	3.00	1.00	0.33
3	1.50	2.50	1.00	0.57	0.25
5	1.25	1.00	0.55	0.35	0.17
10	1.11	0.47	0.28	0.18	0.09

Note: The table reports values of $\rho(\lambda)$ computed according to equation (20). The markup μ is $\alpha/(\alpha - 1)$.

to consumer surplus, as the lower coinsurance rate increases the quantity of drugs sold over which the monopolist earns a markup.

Table 1 shows some values of $\rho(\lambda)$ for some parameter values. There are two things to note about the table. First, the ratio of profits to consumer surplus falls as consumers are forced to pay the entire price of the drug. It also falls as the demand curve becomes less elastic (leading to a higher markup). Second, in extreme cases in which the demand elasticity is relatively high and the coinsurance rate is low, profits can exceed consumer surplus. In this case, the increase in demand associated with the low coinsurance rate leads to “excess” profits. The incentives for a pharmaceutical company to introduce a new drug then exceed what is optimal. This case turns out to be more worthy of consideration than one might have thought, as we will see again in the examples that follow.

Finally, consider the case in which the coinsurance rate λ is chosen to deliver the efficient utilization of the treatment. Recall that the condition for efficient utilization is $\lambda p = w$, so that the first-best level of utilization is achieved by setting $\lambda^* = \frac{\alpha-1}{\alpha}$. In this case, $\rho(\lambda^*) = 1$. That is, the profits earned by the monopolist are exactly equal to the consumer surplus.

So for a Pareto distribution of treatment benefits, we have these striking results. The government or a private insurer can achieve the first best level

of utilization of a drug by setting $\lambda = \frac{\alpha-1}{\alpha}$ so that $\lambda p = w$. Consumers will compare the marginal cost of the drug with its benefits to them, leading to efficient utilization. Moreover, if marginal cost is small compared to average income, the losses from incomplete insurance will be relatively small. Finally, in this equilibrium, a drug manufacturer with an infinitely lived patent will have the correct incentives to introduce a new drug. Profits earned equal consumer surplus, both at a point in time and therefore in present discounted value. If this present value exceeds the cost of introducing the drug, the firm will introduce it. Otherwise, it will not.

5.2. A Second Example: the Exponential Distribution

Suppose now that the values for the treatment benefit v have an exponential distribution,

$$F(v) = 1 - e^{-\alpha v} \quad (21)$$

on the interval $[0, \infty)$. The demand function in this case is also an exponential function

$$x(\lambda p) = e^{-\alpha u'(\bar{y})\lambda p}.$$

This distribution has a constant hazard rate

$$h(v) = \alpha,$$

so the monopoly price is an additive markup over marginal cost:

$$p = w + \frac{1}{\alpha u'(\bar{y})\lambda}. \quad (22)$$

Notice that this expression implies that the treatment cutoff is bounded from below:

$$\bar{v} = u'(\bar{y})\lambda p = u'(\bar{y})\lambda w + \frac{1}{\alpha}.$$

This suggests an important difference between the exponential case and the Pareto case. For the exponential distribution, the mean value of v is $\frac{1}{\alpha}$. The

payment (in utility units) that the consumer makes will be larger than the mean value of the benefit from treatment, no matter how small λ is. Reductions in λ induce offsetting increases in p . In the Pareto case, efficiency required that λp be equal to marginal cost. If marginal cost is low compared to income, then consumer payments will be also. Here, λp is always greater than the mean benefit from treatment. This may be large compared to income, so the assumption that λp is small may not apply. The coinsurance payment may be too large to achieve an adequate level of risk sharing.

For the exponential distribution, the ratio of profits to consumer surplus is given by

$$\rho(\lambda) = \frac{1}{\lambda} \cdot \frac{1}{2 - \alpha(1 - \lambda)u'(\bar{y})w}. \quad (23)$$

At its upper bound $\lambda = 1$, the incentive for innovation is too small by a factor of $1/2$. As λ decreases, though, the ratio of profits to consumer surplus increases. As λ approaches a lower level $1 - \frac{2}{\alpha u'(\bar{y})w}$, the ratio goes to infinity as net benefits go to zero.

To ensure that the treatment will be utilized efficiently, λ needs to be chosen so that $\lambda p = w$. Using the monopoly price from equation (22), this implies a copayment of

$$\lambda^* = 1 - \frac{1}{\alpha u'(\bar{y})w}. \quad (24)$$

At this first-best level of treatment, the ratio of profits to consumer surplus is

$$\rho(\lambda^*) = \frac{1}{\lambda^*} > 1. \quad (25)$$

That is, at the efficient level of treatment, the profits to the pharmaceutical company exceed the consumer surplus, suggesting the possibility of an excessive incentive to innovate.

How is this possible? Recall that the monopolist is earning profits on every treatment sold. Because of the monopoly markup, the efficient utilization of

the drug requires a copayment rate that is less than one, and this subsidy leads to a high level of demand and large profits, as shown in Figure 1 earlier.

What remedies are available in this case? Suppose the government sets an upper bound on the price of the drug that is equal to m times w . Then a copay $\lambda = \frac{1}{m}$ will achieve efficient utilization of the drug, so \bar{v} will be equal to $u'(\bar{y})w$. If m is large, the government will be able to achieve a high degree of risk sharing through a low copay.

With a choice of $\lambda = \frac{1}{m}$, ρ will be equal to m , which may still induce excessive efforts by firms to introduce new drugs. In this case, there may be a case for a finite patent length and competitive production thereafter, with no insurance coverage for off-patent drugs, which sell at marginal cost.

Society will receive the full present discounted value of the net benefits from the drug, but the manufacturer will receive only a fraction of the present discounted value of the profits. If the patent life is chosen so that this fraction is equal to $\lambda = \frac{1}{m}$, then firms will have the correct incentives to introduce new drugs. So the combination of a price ceiling that is large compared to marginal cost and a finite patent life can achieve a result that is very close to the first-best social optimum. Because the firm captures the entire present discounted value of consumer surplus in this case, it will clearly prefer this equilibrium with a price ceiling to an unregulated equilibrium with an infinitely lived patent and no insurance. Of course, from a private point of view, the best outcome for the firm would be to have insurance with a low coinsurance rate and no price ceiling.

5.3. Linear Demand

Finally, suppose that the treatment benefits are uniformly distributed on the interval $[0, \mu]$. Then the distribution function takes the form $F(v) = \frac{v}{\mu}$ and

the demand function is (approximately) given by

$$x(\lambda p) = 1 - \frac{u'(\bar{y})\lambda p}{\mu}.$$

That is, the uniform distribution corresponds to the case of linear demand.

The hazard rate $h(v) = \frac{1}{\mu-v}$ is increasing in v , and the monopoly price can be written as

$$p = \frac{1}{2} \left(w + \frac{\mu}{\lambda u'(\bar{y})} \right).$$

Once again, the payment in utility units $u'(\bar{y})\lambda p$ that consumers make for the drug is bounded from below by the mean benefit in the population from the drug, $\frac{\mu}{2}$.

For this distribution, the ratio of profits to our consumer surplus-like measure is given by

$$\rho(\lambda) = \frac{1}{\lambda} \cdot \frac{2(\mu - w\lambda u'(\bar{y}))}{3\mu - wu'(\bar{y})(4 - \lambda)}. \quad (26)$$

In the case of no coinsurance, where $\lambda = 1$, this ratio is $\rho(1) = 2/3$, so profit falls short of consumer surplus, as one would expect.

What if λ is set to deliver efficient utilization of the drug? Solving the condition $\lambda p = w$ for λ yields

$$\lambda^* = 2 - \frac{\mu}{u'(\bar{y})w}.$$

With this value for λ , the expression for ρ reduces to

$$\rho(\lambda^*) = \frac{2u'(\bar{y})w}{2u'(\bar{y})w - \mu} > 1.$$

In the range of values for which marginal cost is less than the maximum possible benefit and in which it is possible to achieve optimal utilization of the drug,

$$u'(\bar{y})w \in (\mu/2, \mu).$$

Therefore the ratio of profits to consumer surplus varies between 2 and ∞ when the treatment is efficiently utilized.

As in the case of the exponential demand, the combination of an upper bound on the price charged by the monopolist, together with a finite lived patent, can nearly achieve the first-best social optimum.

6. ROBUSTNESS OF THE RESULTS

The examples we provide in this paper illustrate that efficient utilization — allocating the drug to those people who derive a benefit at least as great as marginal cost — does not inevitably lead to appropriate incentives to develop the drug. In particular, the profits earned by the pharmaceutical company setting a monopoly price can exceed the consumer surplus created by the introduction of the drug, leading to a failure of dynamic efficiency. The subsidy to demand inherent in the low copayment leads to excess profits in two of our three examples, and to profits that exactly equal consumer surplus in the third.

Are these results peculiar to the distributions that we used in these examples? Setting the coinsurance rate to achieve static efficiency does not lead to profits in excess of the consumer surplus for *every* distribution of treatment benefits $F(v)$. To see this point, it is helpful to refer back to Figure 1. The monopoly price p is determined by the local elasticity of demand at the point $x(w)$. In contrast, the consumer surplus depends on the shape of the demand function at all prices above w . For the Pareto, exponential, and uniform distributions, the local shape and the global shape of the demand function are connected in such a way that $\rho(\lambda^*) \geq 1$.

For the Pareto distribution, the implied demand function exhibits constant elasticity. The exponential and uniform distributions exhibit large demand elasticities at high prices and are inelastic at low prices. Intuitively, this is why profits can exceed consumer surplus: the monopoly price is based on

the portion of the demand curve with a low price elasticity, generating high profits.

This suggests that a demand curve that had the opposite shape — inelastic at high prices and very elastic at low prices — would lead to profits less than consumer surplus when the treatments are efficiently utilized. Interestingly, this is *not* what conventional distributions like those in our examples yield. In those cases, the percentage change in demand associated with a one percent decline in price is large when demand is low (i.e. at high prices) and small when demand is high (at low prices).³

Appendix B shows how these results can be extended to more general demand curves and more general underlying distributions. As long as the elasticity of demand is a decreasing function of the price, static efficiency is associated with profits that are too large relative to consumer surplus. In terms of the underlying distribution of benefits from the innovation, this decreasing elasticity corresponds to a hazard rate of the distribution $F(v)$ that falls no faster than $1/v$.

Our work shows that static and dynamic efficiency are unlikely to be achieved simultaneously by any simple or generic rule. Optimal strategies depend upon the shape of the demand curve, which in turn reflects the distribution of benefit, and which changes as substitutes and complements become available. Static efficiency may be achievable if the insurance market is competitive, as Gaynor and colleagues have argued (Gaynor et al. 2000). They show that marginal cost pricing leads to a welfare gain over monopoly pricing, as long as copayments for each product or service are set optimally. Their purely static analysis does not consider incentives to innovate and other aspects of dynamic

³Still one can concoct a mechanical counterexample: consider the case where treatment benefits have a Pareto distribution for most people, but for some small fraction of people there is a positive mass at a high benefit level. This will not change the monopoly price or profits, but it will increase consumer surplus. Since the straight Pareto case involved $\rho(\lambda^*) = 1$, this richer example will deliver $\rho(\lambda^*) < 1$.

efficiency. Static efficiency in our model also assumes that copayments are set optimally for a price-taking insurer; the results would differ if, for example, the insurer as well as the pharmaceutical company had substantial market power. In that case, the copayment would be set at a level that takes into account the price response of the monopolist pharmaceutical company.

It is unlikely that the insurance arrangements common today result in optimal copayments; there is typically a single copayment level for drugs in government programs, and usually three tiers of copayments in commercial health insurance plans. An insurer's response to the imposition of a high price for a drug is usually to move the drug to a higher tier. Benefit based copayments are a proposed solution to better align cost sharing to benefits, and if successfully implemented might lead to utilization patterns closer to the optimum (Fendrick, Smith, Chernew and Shah 2001, Goldman, Joyce and Karaca-Madic 2006, Rosen, Hamel, Weinstein, Cutler, Fendrick and Vijan 2005).

We have also assumed that the pharmaceutical producer cannot price discriminate. Perfect price discrimination leads to an excess quantity of consumption, rather than the competitive equilibrium, due to the insurance subsidy. Furthermore, the monopolist gaining all of the surplus would receive much larger profits than those considered in our examples. Thus, with price discrimination, profits are likely to exceed consumer surplus for a broad range of distributions of benefit, leading to a failure of dynamic efficiency. Shorter patent life could offset the excessive rewards for innovation, though such a change would not correct overutilization occurring during the period of monopoly.

7. CONCLUSIONS

This paper illustrates that efficient utilization — allocating the drug to those people who derive a benefit at least as great as marginal cost — does not

inevitably lead to appropriate incentives to develop the drug. Static efficiency, in this formulation, is obtained by the interaction of a monopoly with a market that has subsidized demand. A fractional coinsurance rate can cause excessive use, but it also helps avoid underuse.

However, with efficient utilization — static efficiency — the profits earned by the pharmaceutical company setting a monopoly price can exceed the consumer surplus created by the introduction of the drug, leading to a failure of dynamic efficiency. The subsidy to demand inherent in the low copayment leads to excess profits in many cases. The resulting dynamic inefficiency raises the possibility that finite patent lives could be welfare improving by reducing excessive innovation. The examples also suggest that an upper bound on the price received by the manufacturer may in some cases be required to ensure that revenues are not too large in relation to the benefits consumers receive.

In practical terms, these results imply that if price increases too rapidly as the coinsurance rate declines, insurers may not be able to lower the out-of-pocket costs sufficiently to get high benefit patients to consume the drug. For some distributions of benefit, it may also be very difficult to keep the high-benefit patients adequately insured without paying for the treatment of many low-benefit patients. Moreover, with low coinsurance rates, the incentives to innovate can easily be too great, at least in the case of infinite patent length. Thus we expect that in the typical case, achieving both efficient utilization and dynamic optimality requires a combination of a low coinsurance rate and variation in patent duration.

In practice, however, simple use of either of these instruments may not lead to optimality. For example, pricing should change as demand changes, due perhaps to the entry of new competitors, and an administered price that does not change correctly can lead to the wrong incentives. Similarly, optimal patent duration would vary with the drug, yet rules that permit product-specific patent life, and that preserve optimality properties in the presence of uncer-

tainty, would require fundamental changes in the law. In some extreme cases, a very high upper bound on a drug's price — one that could be much higher than the price a monopolist would charge in a market with a 100% coinsurance rate — may help keep the price from growing without bound as the coinsurance rate gets smaller.

The growing, though still limited, use of consumer-directed health plans and health savings accounts is a sign of generally greater cost sharing. In such plans, the coinsurance rate is 100% below a (high) deductible. Such plans are likely to exacerbate the tendency to consume less than the optimal quantity of monopoly-produced drugs, since price then equals marginal benefit to the patient, but exceeds marginal cost.

Unlike traditional health insurance and high cost-sharing plans like consumer-directed plans, managed care seeks to limit quantities consumed more directly. To the extent that such plans are successful, they steer allocation of drugs toward patients who derive the greatest benefit. Yet such plans are likely to exacerbate the underconsumption of monopoly products, leading to both static and dynamic inefficiency.

An important avenue for future research is to examine pricing strategies and the incentives for innovation empirically. For example, Philipson and Jena (2005) study the ratio of consumer surplus to profits in treating HIV/AIDS. They argue that pharmaceutical innovators have captured only about 5% of the total surplus, suggesting that the incentives for innovation in this particular market may be far too small. As our results (and those of Lakdawalla and Sood (2005)) suggest, this result would not be expected if pharmaceutical firms were free to behave as monopolists who can choose a profit maximizing price to charge consumers covered by health insurance. Prices may be constrained by monopolistic competition between alternative treatments that are close substitutes, an effect that is not captured in our simple one-good model. In such a situation, each product might receive profits equal to the incremental

surplus generated by the new product, even if the profits for all drugs in the class are less than the consumer surplus that results from the availability of the entire class of drugs. In this case, a small ratio of profits to surplus for the class does not mean that the ratio will be small at the level of any single drug, and it is not clear that the incentives to develop new drugs would be inadequate. Alternatively, profits may be low relative to surplus in markets characterized by significant monopsony power, yielding insufficient incentives to innovate.

Empirical work should guide the development of more complete models of the policies that some governments use to limit the profits of pharmaceutical companies. For example, there may be merit to industry claims that aggressive government purchasing schemes pursued in the European Union may go too far in the direction of limiting the profits earned on important new drugs. Such claims would need to be tested empirically. Optimal policy requires balancing the possibility of prices that are too low to encourage innovation and prices that are too high. We find that in the absence of monopsony, price controls, and other tools to limit demand, a system of medical insurance that relies on low consumer coinsurance payments creates incentives for a monopoly provider of a pharmaceutical to charge far more for its product than it otherwise would, and possibly to receive excessive rewards for innovation.

APPENDIX A

Deriving Social Welfare in Equilibrium

To compute the value of social welfare, recall that $x \equiv x(\lambda p)$ denotes the measure of people who purchase the treatment in equilibrium. Then, ignoring the exogenous δ_i terms, social welfare (treating each person equally) is

$$\Omega(\lambda p) = (1 - x)u(y) + xu(y - \lambda p) + s \int_{\bar{v}}^{\infty} vF'(v)dv.$$

Social welfare is the sum of three terms: utility from those who do not purchase, consumption utility from those who do purchase the treatment, and the treatment benefit for these people, i.e. those with benefits larger than \bar{v} .

Assuming λp is small relative to y , we can use the substitution $u(y - \lambda p) \approx u(y) - u'(y)\lambda p$ to get

$$\Omega(\lambda p) \approx u(y) - xu'(y)\lambda p + s \int_{\bar{v}}^{\infty} vF'(v)dv.$$

Next, recall that the relationship between equilibrium income y and the endowment \bar{y} is

$$y = \bar{y} + x(\lambda p - w).$$

That is, income adjusts for the copayment and the profits of the drug company. Substituting this expression in for y we get

$$\Omega(\lambda p) \approx u(\bar{y}) - u'(\bar{y})xw + s \int_{\bar{v}}^{\infty} vF'(v)dv.$$

This is the expression given in the text in equations (12) and (13).

APPENDIX B

Additional Results

The following discussion shows how our results related to the ratio of profits to consumer surplus generalizes to other distributions. Details of the proofs are available from the authors.

PROPOSITION B.1. *Consider our model with demand $x(q)$. Suppose this demand function exhibits a price elasticity $\epsilon(q) \equiv -\frac{dx}{dq} \frac{q}{x}$ that is weakly decreasing in the consumer price q . Then $\rho(\lambda^*) \geq 1$, with strict inequality if $\epsilon(q)$ decreases strictly over some interval.*

This proposition expresses a general result for the ratio of profits to consumer surplus in terms of the elasticity of the demand function. In the text,

we showed that the constant elasticity demand function leads profits to equal consumer surplus when the treatment is utilized efficiently, and gave two examples (exponential and linear demand) when profits exceeded consumer surplus. This proposition generalizes these examples by showing that if the elasticity is decreasing, we get the result that profits will exceed consumer surplus.¹ In fact, this proposition generalizes to any demand function that lies below an artificial constant elasticity demand function, where the artificial constant elasticity is equal to the actual elasticity at the equilibrium price.

We can then extend this result for demand elasticities to the underlying distribution that generates the demand function. The key is to note that, as shown in equation (9), the elasticity of demand is

$$\epsilon = h(v)v.$$

A decreasing elasticity (as a function of price) ends up corresponding to a hazard rate that falls no faster than $1/v$. So as a direct corollary of our proposition, our model delivers $\rho(\lambda^*) \geq 1$ as long as the hazard rate $h(v)$ falls no faster than $1/v$. Any underlying distribution with a constant or increasing hazard, then, will generate the result that profits exceed consumer surplus when utilization is efficient.

REFERENCES

- Chang, Yongsung and Sun-Bin Kim, "From Individual to Aggregate Labor Supply: A Quantitative Analysis based on a Heterogeneous Agent Macroeconomy," July 2003. Federal Reserve Bank of Richmond Working Paper 03-05.
- Crew, Michael, "Coinsurance and the welfare economics of medical care," *American Economic Review*, December 1969, 59 (5), 906–908.
- Fendrick, A.M., D.G. Smith, M.E. Chernew, and S.N. Shah, "A benefit-based copay for prescription drugs: patient contribution based on total benefits, not drug acquisition cost," *American Journal of Managed Care*, 2001, 7, 861–867.

¹This result can be proved graphically by drawing the standard consumer surplus and profit picture on a log-log scale, i.e. with log price on the vertical axis and log quantity on the horizontal axis.

- Gaynor, Martin, Deborah Haas-Wilson, and William B. Vogt, "Are Invisible Hands Good Hands? Moral Hazard, Competition, and the Second-Best in Health Care Markets," *Journal of Political Economy*, 2000, 108, 992–1005.
- Goldman, D.P., G.F. Joyce, and P. Karaca-Madic, "Varying pharmacy benefits with clinical status: the case of cholesterol-lowering therapy," *American Journal of Managed Care*, 2006, 12, 17–28.
- Heffler, S., S. Smith, S. Keehan, C. Borger, M.K. Clemens, and C. Truffer, "Trends: U.S. Health Spending Projections For 2004-2014," *Health Affairs*, February 23 2005, pp. W5 74– W5 85. Accessed on web at <http://content.healthaffairs.org/cgi/reprint/hlthaff.w5.74v1, 6/18/2005>.
- Lakdawalla, Darius and Neeraj Sood, "Insurance and Innovation in Health Care Markets," September 2005. NBER Working Paper 11602.
- Pauly, Mark V., "The Economics of Moral Hazard: Comment," *American Economic Review*, June 1968, 58, 531–537.
- Philipson, Tomas J. and Anupam B. Jena, "Who Benefits from New Medical Technologies? Estimates of Consumer and Producer Surpluses for HIV/AIDS Drugs," December 2005. NBER Working Paper 11810.
- Rosen, A.B., M.B. Hamel, M.C. Weinstein, D.M. Cutler, A.M. Fendrick, and S. Vijan, "Cost effectiveness of full Medicare coverage of angiotensin-converting enzyme inhibitors for beneficiaries with diabetes," *Annals of Internal Medicine*, 2005, 143, 89–99.
- Saez, Emmanuel, "Using Elasticities to Derive Optimal Tax Rates," *Review of Economic Studies*, 2001, 68, 205–229.