Volume Title: Doctors and Their Workshops: Economic Models of Physician Behavior

Volume Author/Editor: Mark Pauly

Volume Publisher: University of Chicago Press

Volume ISBN: 0-226-65044-8

Volume URL: http://www.nber.org/books/paul80-1

Publication Date: 1980

Chapter Title: Physicians and Hospitals

Chapter Author: Mark Pauly

Chapter URL: http://www.nber.org/chapters/c11522

Chapter pages in book: (p. 17 - 24)

2 Physicians and Hospitals

Many physicians do not produce the bulk of their output in their own offices, where they pay the outlay costs of inputs used. Instead, much of their output is produced in the hospital, where they neither have ownership rights nor are directly responsible for paying the cost of hospital inputs.

The preceding chapter suggested that, if physicians were income maximizers, we should expect them to use an efficient combination of inputs, whether or not they made explicit outlays for those inputs. If they were utility maximizers, no alternative institutional arrangement could improve efficiency. Yet it is widely suggested that hospitals are inefficient, even though the empirical (as opposed to anecdotal) evidence for this contention is weak, and there is some evidence that very little of the interhospital variation in hospital costs can be attributed to inefficiency.[1] It is also not generally recognized that efficiency cannot be judged solely by the costs the hospital pays, but must also include consideration of the cost of inputs supplied by physicians. Consequently, it remains to investigate whether there are any reasons to suppose that physicians will not, individually and collectively, use hospital inputs efficiently.

There are two broad classes of reasons we will consider. Both reasons are based on distortions in the prices physicians-as-agents face for hospital inputs. Those distortions may arise either from (1) customary forms of hospitalization insurance or (2) from imperfect cooperation of physicians within the hospital, aggravated by imperfect pricing of hospital services. Each of these reasons will be considered in turn.

Hospitalization Insurance

Typical hospitalization insurance in the United States makes payments to hospitals which depend on the hospital costs or bills incurred by

insureds. It is easy to see intuitively that this arrangement will induce physicians to choose higher levels of costly hospital inputs than would occur without insurance, as long as that cost is associated with improvement in patient well-being. Because the user cost of an increment in quality will have been reduced by insurance, patients, or their physicians acting as agents, will tend to choose higher quality. Because there may be a time cost associated with increased quantity (in the sense of patient-days), but no time cost for additional quality or services during a day the patient is already in the hospital, one would expect the response of quality to insurance to be greater than the response of quantity. Of course, increased quality will show up as higher costs or charges per unit output. In demanding higher quality for his patients, each physician is acting as their true agent, even though the result for all patients of all physicians is likely to be inefficient. This type of "moral hazard" has generally been recognized as an important determinant of hospital cost inflation, and does not require further discussion here.[2] Virtually complete insurance coverage also reduces the incentive for patient or physician to search for lower cost or more efficient hospitals, and can thereby contribute to inflation.[3]

There is another less obvious but possibly important effect of typical hospitalization insurance. Such insurance will tend to encourage excessive substitution of hospital inputs for physician inputs. This oversubstitution not only causes the hospital unit cost to rise, but total costs per unit (over hospitals and physicians) rise as well: there will be overuse of hospital inputs relative to physician inputs, as compared to the cost-minimizing level.

To see this, suppose that, in the absence of insurance, the equilibrium gross price of hospital output at a given hospital is $\bar{P}_T$, and the quality $\bar{Q}$. $\bar{P}_T$ is the sum of the hospital price $\bar{P}_H$ and the physician price $\bar{P}_M$. Assume for simplicity that the price $\bar{P}_T$ is competitively determined; permitting the extent of competition to vary with the level of insurance coverage would unnecessarily complicate the problem. Suppose there is an opportunity cost C of physician time spent in the hospital; this could be either the physician's income from providing ambulatory care or a money measure of his value of leisure. Suppose the only hospital input is represented by H, which is available at a constant unit cost W. Finally, suppose that the hospital price P_H is set equal to average cost: $P_H = WH/Q$. Holding output constant at, say, $\bar{Q}$, it is possible to substitute H for M within the limits given by

$$\frac{\partial Q}{\partial H}\, dH - \frac{\partial Q}{\partial M}\, dM = 0$$

Cost minimization for a given output implies that

$$\frac{W}{\partial Q/\partial H} = \frac{C}{\partial Q/\partial M}$$

Let H^* and M^* be the levels of M and H which satisfy this equality. It follows that if H is increased by a small amount ΔH, and M is reduced to $M^* - \Delta M$, Q constant, then

$$W\Delta H \simeq C\Delta M$$

The reduction in physician opportunity cost would approximately equal the increase in hospital cost.

Now suppose that there exists an insurance which covers $\lambda(0 > \lambda > 1)$ of P_H. The insured pays $(1-\lambda)(P_H) + P_M$ for a unit of hospital care instead of $\bar{P}_H + \bar{P}_M$. If, after the provision of insurance, output is to be held constant at $\bar{Q}$, the physician price must rise by λP_H to keep price constant at $\bar{P}_T$. Physician income will then increase by $\bar{Q}$ (λP_H).

Suppose H is increased by ΔH and M decreased by ΔM as before. The rise in the use of hospital inputs raises P_H by $W\Delta H/Q$. In order to keep the user price constant at $\bar{P}_T$, the physician price will have to fall by $(1-\lambda)W\Delta H/Q$, and physician gross revenues by $(1-\lambda)W\Delta H$. The reduction in physician opportunity cost is $C\Delta M$. Since initially $C\Delta M \simeq W\Delta H$, it follows that $C\Delta M > (1-\lambda)W\Delta H$. That is, the decline in physician costs $(-C\Delta M)$ exceeds the decline in physician gross revenues $\{(1-\lambda)W\Delta H\}$. So physician net income will increase if the level of physician input is reduced while the level of hospital inputs is increased. Of course, equilibrium output will also change, and may be accompanied by changes in input ratios, but it will still be true that, whatever level of output is produced, it will be produced with relatively more than the cost-minimizing amount of H. With hospitalization insurance, substitution of hospital for physician inputs increases total cost, but it also increases total revenue by several times the increase in total cost, enough to offset the increase in total cost. Put still another way, hospital insurance reduces the user price of hospital inputs below their "true" market price, and so leads to the use of relatively more of them.

Physician Fee Insurance

Where hospitalization insurance is present, insurance to cover physician charges for in-hospital physician services is also typically found. This result should not be surprising: if the loss from consuming one more unit of hospital output is the total price $P_M + P_H$, then one would expect to find both parts of the hospital bill covered by insurance. Would insurance coverage of the physician's fee offset the incentive to

overuse of hospital inputs? To answer this question, we need to consider two kinds of physician fee insurance.

1. "Indemnity" insurance. Many physician insurance policies pay the entire physician's bill or a maximum dollar amount, whichever is less. Sometimes the maximum is set by an explicit schedule of maximum fees, although more recently it has been set at some percentile of a screen of reasonable and customary fees. We shall consider first the situations in which the full maximum amount is paid.

Under such an arrangement, the insurance payment is independent of both the level of the physician's fee (once it is at or above the maximum) and the way the physician uses his time. It is obvious that such insurance coverage will not affect the relative use of inputs, since the insurance payment does not depend on the amount of inputs used.

2. Proportional coinsurance. When the fee is below the fee schedule maximum, then insurance coverage may have some effect if the insurance pays some fraction $\gamma \leq 1$ of the fee. If γ is less than one, holding P_T constant implies that P_M is increased to $\frac{1}{(1-\gamma)} P_M$ when insurance is obtained. Ideally, P_T could still be held constant by raising P_M by this amount, reducing P_H, and continuing to use the cost-minimizing combination of inputs. However, the hospital would then sustain a deficit, and there might be practical problems getting physicians to underwrite this deficit.

If the hospital is constrained to charge a breakeven price, physician insurance may provide an incentive to reduce hospital inputs. Reducing hospital inputs and raising the physician's fee (and his time inputs) may lead to higher physician net incomes as long as the rise in the physician fee exceeds the opportunity cost of the extra physician time.

The conclusion is that *some* physician insurance schemes may produce an offsetting effect. But since in any market area many persons with hospital insurance will not have physician insurance with the proportional coinsurance type of coverage, the effect of hospital insurance on input combinations will not be fully offset.[4]

Imperfect Pricing, Imperfect Cooperation, and the Size Principle

If all hospital services were sold at the marginal cost of the inputs used to produce those services, and if insurance were not present, physicians would have an incentive to minimize total costs. The physician would know that any extra hospital inputs he ordered would show up on a hospital bill to his patient, which would mean less that he could collect. The cost-minimizing solution would be chosen by the physician because prices would play the role of coordinator. In such a situation, the physician would not treat the hospital as a rent-free workshop, despite the nominal separation of physician and hospital billing.

But there are reasons to expect that hospitals cannot or will not price every dimension of their service at its marginal cost, and so will offer another incentive to the physician to depart from the cost-minimizing input combination. First, pricing at marginal cost may involve enormous administrative problems in monitoring every extra nursing or housekeeping minute devoted to each patient. And second, even if pricing at marginal cost were feasible, it might cause the hospital to violate the zero profit constraint when marginal and average costs diverge. In particular, if there are services which involve high initial fixed costs, pricing at marginal cost may not permit the hospital to cover costs and still provide all services which generate consumers' surplus.

Even with average cost pricing, there will still be some incentive for the physician to keep costs down, since the individual physician will bear some fraction of any addition to total costs he might cause. But the larger the number of physicians over whose patients these costs are spread, the smaller will be the share and the smaller the incentive for each individual physician. This "size principle" has been extensively discussed in the literature.[5]

One would expect physicians collectively to try to institute some means of enforcing cooperation. This might take the form of rules, committee structures, moral suasion, and so on. They may delegate some of this task to the hospital's lay administration or to chiefs of the medical service. One would also expect physicians to sort themselves according to their responsiveness to incentives, or their degree of cooperativeness. But eventually there will be some departure from the perfectly cooperative, cost-minimizing solution, if only because coordination is itself costly.

This departure from cost-minimization will take two forms. Both of them will involve reduction in the amount of physician inputs, but they differ with regard to the type of input whose amount is altered. The physician provides two inputs to the medical care process: his own time and what one might call "effort" or care in directing the production process. Imperfect cooperation can affect the amounts of both of these inputs.

When the physician or his patient bears only a partial share of the benefit from his being careful and being concerned about the costliness of treatment procedures, and when such effort or care involves disutility, one would expect to observe less effort and higher hospital costs than when the physician can receive the full reward from "effort." The physician's physical time input will also be lower, and hospital costs higher, when the physician cannot capture all of the benefit from devoting his own time to hospital care. If he values his time in other activities, one would expect him to order hospital substitutes for it to a greater extent as the fraction he receives of the benefit from the increased productivity of that time input becomes smaller. The result would be an overuse of

hospital input relative to physician input. Because hospitals price many services on an average cost basis, as noted above, one would therefore expect some overuse of hospital inputs (relative to physician inputs) to occur.

Because effort or care cannot be measured directly, the first effect is much more difficult to determine.[6] However, the second type of overuse is one that can be detected directly by measuring physician time input. It is this second type that will be investigated in the empirical work in the following chapter.

This discussion is based on a model of the hospital in which physicians control the hospital, and induce it to operate so as to maximize their individual utilities, even though they may be constrained by problems of size and coordination. The alternative "hospital administration utility-maximization" models developed by Feldstein and Newhouse make no direct prediction about physician-hospital input ratios, since they do not treat nonsalaried physician time as a productive input.[7] The result of overuse of hospital relative to physician input would, however, be consistent with their kind of theory if higher "quality" were equated with more hospital and less physician input. Whether hospital administrators, who run the hospital according to the theory, actually judge quality in this way or not is unknown.

Measuring Input Substitution in Practice

The way in which inputs are to be measured when estimating a production function depends upon the use to which the results are to be put. In engineering, where it is the purely technical relationships that are of interest, the most appropriate measure would be some index of homogeneous productive effort. If, on the other hand, one is interested in the behavioral response of the system, the appropriate measure is the level of input that can be manipulated by the decision-maker. In more concrete terms, whether one wishes to measure labor input by minutes actually worked at various tasks or by hours of work for which full wages are received depends upon whether or not a feasible control mechanism exists for monitoring, controlling, and paying for only minutes of actual work. Even this states the matter too simply, since what is feasible may often be too costly, and the actual methods of control and reimbursement (and hence the actual allocation of effort) may vary widely among occupations, firms, or skill levels.

All this discussion is by way of elaborate rationalization for the use, in the production function estimates that follow, of the number of physicians available to provide care, rather than actual hours worked, as a measure of physician input. The concrete reason for this procedure is the unavailability of hours-worked data, but the reason for continuing

with the analysis is that, at the present time, the most that might be manipulable from a public policy viewpoint is the number of physicians in an area or on a hospital staff, *not* the number of hours the physician spends at the hospital. In general, the kind of question to be posed is: If one pours additional physicians into a hospital's catchment area, or places additional physicians on its staff, what effect will this have, *ceteris paribus*, on the hospital's output? Put another way, the question is that of how physicians affect hospital productivity.

While allegations of physician overuse of hospital inputs are common, concrete descriptions of the form this overuse might take are less common. The possibility of overuse is most transparent for hospital-employed physicians: they can be substituted for the time of private practice physicians, and one suspects that there is not an off-setting diminution in fees charged by the private practice physician. A similar argument might be made with respect to nurses; they can perform actions which can save the attending physician the time and trouble of making a visit to the patient. The argument that when nurses are scarce, physicians will end up making more visits is a little weaker, but perhaps plausible, especially if one adds the notion of "highly skilled" nurses.

Another way in which physicians substitute for hospital inputs has been suggested by Martin Feldstein: "By increasing the number of doctors, for instance, a hospital may be able to shorten the length of patient stay and thus decrease the input of beds for a given output."[8] While Feldstein's study referred to hospitals in the United Kingdom, the same sort of reasoning might be applied to hospitals in the United States, and to other hospital inputs (nurses, for instance) as well. The explanation here would seem to be that either rate of recovery or delay in performing procedures can be affected by the number of physicians. Where the number of medical staff members are few, rounds may be less frequent, and patients may have to wait in bed for the physician to come by and order procedures, perform operations, or sign discharge forms. I shall later try to determine whether any effect of physicians on output does come through on effect on length of stay.

It should be emphasized that, inasmuch as the estimates to be presented are production function estimates, they do not bear directly on the question of whether physicians can create demand for hospital services. As in other production function studies, we do not ask whether the output should have been produced, or why it was produced; we only ask about the relationship between outputs and inputs. These results do, however, bear indirectly on the question of demand creation, in that they indicate the maximum extent of demand creation, at least as far as physicians are directly concerned. In other words, they indicate the maximum amount of increase in output that could be attributed to demand creation by additional physicians when all hospital inputs are held

constant. However, since inputs and outputs are not measured in per capita terms, the results are not directly comparable to those from demand studies.

One difficulty with the empirical analysis is that scale or hospital size is likely to be positively correlated with the number of physicians and with the level of physician input. Hospitals with larger medical staffs will have more physician time input available, but more difficulty in coordinating it. In such a case it is not possible to get separate estimates of "true" economies of scale *and* the effects of the size principle. If it were possible to measure physician time directly, and observe situations in which different numbers of physicians provide the same amount of time, then one could get a separate estimate of the effect of the size principle.

Conclusion

The following chapter presents production function estimates intended to measure the existence and extent of underuse of physician input. Unfortunately, available data do not permit an explanation of the cause of departures from optimal use; we do not have measures of the extent of insurance coverage or of departures from marginal cost pricing.