THE AMERICAN ECONOMIC REVIEW

VOLUME LIII DECEMBER 1963 NUMBER 5

UNCERTAINTY AND THE WELFARE ECONOMICS OF MEDICAL CARE

By KENNETH J. ARROW*

I. *Introduction: Scope and Method*

This paper is an exploratory and tentative study of the specific differentia of medical care as the object of normative economics. It is contended here, on the basis of comparison of obvious characteristics of the medical-care industry with the norms of welfare economics, that the special economic problems of medical care can be explained as adaptations to the existence of uncertainty in the incidence of disease and in the efficacy of treatment.

It should be noted that the subject is the *medical-care industry,* not *health.* The causal factors in health are many, and the provision of medical care is only one. Particularly at low levels of income, other commodities such as nutrition, shelter, clothing, and sanitation may be much more significant. It is the complex of services that center about the physician, private and group practice, hospitals, and public health, which I propose to discuss.

The focus of discussion will be on the way the operation of the medical-care industry and the efficacy with which it satisfies the needs of society differ from a norm, if at all. The "norm" that the economist usually uses for the purposes of such comparisons is the operation of a competitive model, that is, the flows of services that would be

* The author is professor of economics at Stanford University. He wishes to express his thanks for useful comments to F. Bator, R. Dorfman, V. Fuchs, Dr. S. Gilson, R. Kessel, S. Mushkin, and C. R. Rorem. This paper was prepared under the sponsorship of the Ford Foundation as part of a series of papers on the economics of health, education, and welfare.

offered and purchased and the prices that would be paid for them if each individual in the market offered or purchased services at the going prices as if his decisions had no influence over them, and the going prices were such that the amounts of services which were available equalled the total amounts which other individuals were willing to purchase, with no imposed restrictions on supply or demand.

The interest in the competitive model stems partly from its presumed descriptive power and partly from its implications for economic efficiency. In particular, we can state the following well-known proposition (First Optimality Theorem). If a competitive equilibrium exists at all, and if all commodities relevant to costs or utilities are in fact priced in the market, then the equilibrium is necessarily *optimal* in the following precise sense (due to V. Pareto): There is no other allocation of resources to services which will make all participants in the market better off.

Both the conditions of this optimality theorem and the definition of optimality call for comment. A definition is just a definition, but when the *definiendum* is a word already in common use with highly favorable connotations, it is clear that we are really trying to be persuasive; we are implicitly recommending the achievement of optimal states.[1] It is reasonable enough to assert that a change in allocation which makes all participants better off is one that certainly should be made; this is a value judgment, not a descriptive proposition, but it is a very weak one. From this it follows that it is not desirable to put up with a non-optimal allocation. But it does not follow that if we are at an allocation which is optimal in the Pareto sense, we should not change to any other. We cannot indeed make a change that does not hurt someone; but we can still desire to change to another allocation if the change makes enough participants better off and by so much that we feel that the injury to others is not enough to offset the benefits. Such interpersonal comparisons are, of course, value judgments. The change, however, by the previous argument ought to be an optimal state; of course there are many possible states, each of which is optimal in the sense here used.

However, a value judgment on the desirability of each possible new distribution of benefits and costs corresponding to each possible reallocation of resources is not, in general, necessary. Judgments about the distribution can be made separately, in one sense, from those about allocation if certain conditions are fulfilled. Before stating the relevant proposition, it is necessary to remark that the competitive equilibrium achieved depends in good measure on the initial distribution of purchasing power, which consists of ownership of assets and skills that

[1] This point has been stressed by I. M. D. Little [19, pp. 71-74]. For the concept of a "persuasive definition," see C. L. Stevenson [27, pp. 210-17].

command a price on the market. A transfer of assets among individuals will, in general, change the final supplies of goods and services and the prices paid for them. Thus, a transfer of purchasing power from the well to the ill will increase the demand for medical services. This will manifest itself in the short run in an increase in the price of medical services and in the long run in an increase in the amount supplied.

With this in mind, the following statement can be made (Second Optimality Theorem): If there are no increasing returns in production, and if certain other minor conditions are satisfied, then every optimal state is a competitive equilibrium corresponding to some initial distribution of purchasing power. Operationally, the significance of this proposition is that if the conditions of the two optimality theorems are satisfied, and if the allocation mechanism in the real world satisfies the conditions for a competitive model, then social policy can confine itself to steps taken to alter the distribution of purchasing power. For any given distribution of purchasing power, the market will, under the assumptions made, achieve a competitive equilibrium which is necessarily optimal; and any optimal state is a competitive equilibrium corresponding to some distribution of purchasing power, so that any desired optimal state can be achieved.

The redistribution of purchasing power among individuals most simply takes the form of money: taxes and subsidies. The implications of such a transfer for individual satisfactions are, in general, not known in advance. But we can assume that society can *ex post* judge the distribution of satisfactions and, if deemed unsatisfactory, take steps to correct it by subsequent transfers. Thus, by successive approximations, a most preferred social state can be achieved, with resource allocation being handled by the market and public policy confined to the redistribution of money income.[2]

If, on the contrary, the actual market differs significantly from the competitive model, or if the assumptions of the two optimality theorems are not fulfilled, the separation of allocative and distributional procedures becomes, in most cases, impossible.[3]

The first step then in the analysis of the medical-care market is the

[2] The separation between allocation and distribution even under the above assumptions has glossed over problems in the execution of any desired redistribution policy; in practice, it is virtually impossible to find a set of taxes and subsidies that will not have an adverse effect on the achievement of an optimal state. But this discussion would take us even further afield than we have already gone.

[3] The basic theorems of welfare economics alluded to so briefly above have been the subject of voluminous literature, but no thoroughly satisfactory statement covering both the theorems themselves and the significance of exceptions to them exists. The positive assertions of welfare economics and their relation to the theory of competitive equilibrium are admirably covered in Koopmans [18]. The best summary of the various ways in which the theorems can fail to hold is probably Bator's [6].

comparison between the actual market and the competitive model. The methodology of this comparison has been a recurrent subject of controversy in economics for over a century. Recently, M. Friedman [15] has vigorously argued that the competitive or any other model should be tested solely by its ability to predict. In the context of competition, he comes close to arguing that prices and quantities are the only relevant data. This point of view is valuable in stressing that a certain amount of lack of realism in the assumptions of a model is no argument against its value. But the price-quantity implications of the competitive model for pricing are not easy to derive without major—and, in many cases, impossible—econometric efforts.

In this paper, the institutional organization and the observable mores of the medical profession are included among the data to be used in assessing the competitiveness of the medical-care market. I shall also examine the presence or absence of the preconditions for the equivalence of competitive equilibria and optimal states. The major competitive preconditions, in the sense used here, are three: the *existence* of competitive equilibrium, the *marketability* of all goods and services relevant to costs and utilities, and *nonincreasing returns*. The first two, as we have seen, insure that competitive equilibrium is necessarily optimal; the third insures that every optimal state is the competitive equilibrium corresponding to some distribution of income.[4] The first and third conditions are interrelated; indeed, nonincreasing returns plus some additional conditions not restrictive in a modern economy imply the existence of a competitive equilibrium, i.e., imply that there will be some set of prices which will clear all markets.[5]

The concept of marketability is somewhat broader than the traditional divergence between private and social costs and benefits. The latter concept refers to cases in which the organization of the market does not require an individual to pay for costs that he imposes on others as the result of his actions or does not permit him to receive compensation for benefits he confers. In the medical field, the obvious example is the spread of communicable diseases. An individual who fails to be immunized not only risks his own health, a disutility which presumably he has weighed against the utility of avoiding the procedure, but also that of others. In an ideal price system, there would be a price which he would have to pay to anyone whose health is endangered, a price sufficiently high so that the others would feel compensated; or, alternatively, there would be a price which would be paid to him by others to induce him to undergo the immunization procedure.

[4] There are further minor conditions, for which see Koopmans [18, pp. 50-55].

[5] For a more precise statement of the existence conditions, see Koopmans [18, pp. 56-60] or Debreu [12, Ch. 5].

Either system would lead to an optimal state, though the distributional implications would be different. It is, of course, not hard to see that such price systems could not, in fact, be practical; to approximate an optimal state it would be necessary to have collective intervention in the form of subsidy or tax or compulsion.

By the absence of marketability for an action which is identifiable, technologically possible, and capable of influencing some individual's welfare, for better or for worse, is meant here the failure of the existing market to provide a means whereby the services can be both offered and demanded upon payment of a price. Nonmarketability may be due to intrinsic technological characteristics of the product which prevent a suitable price from being enforced, as in the case of communicable diseases, or it may be due to social or historical controls, such as those prohibiting an individual from selling himself into slavery. This distinction is, in fact, difficult to make precise, though it is obviously of importance for policy; for the present purposes, it will be sufficient to identify nonmarketability with the observed absence of markets.

The instance of nonmarketability with which we shall be most concerned is that of risk-bearing. The relevance of risk-bearing to medical care seems obvious; illness is to a considerable extent an unpredictable phenomenon. The ability to shift the risks of illness to others is worth a price which many are willing to pay. Because of pooling and of superior willingness and ability, others are willing to bear the risks. Nevertheless, as we shall see in greater detail, a great many risks are not covered, and indeed the markets for the services of risk-coverage are poorly developed or nonexistent. Why this should be so is explained in more detail in Section IV.C below; briefly, it is impossible to draw up insurance policies which will sufficiently distinguish among risks, particularly since observation of the results will be incapable of distinguishing between avoidable and unavoidable risks, so that incentives to avoid losses are diluted.

The optimality theorems discussed above are usually presented in the literature as referring only to conditions of certainty, but there is no difficulty in extending them to the case of risks, provided the additional services of risk-bearing are included with other commodities.[6]

However, the variety of possible risks in the world is really staggering. The relevant commodities include, in effect, bets on all possible occurrences in the world which impinge upon utilities. In fact, many of these "commodities," i.e., desired protection against many risks, are

[6] The theory, in variant forms, seems to have been first worked out by Allais [2], Arrow [5], and Baudier [7]. For further generalization, see Debreu [11] and [12, Ch. 7].

simply not available. Thus, a wide class of commodities is nonmarketable, and a basic competitive precondition is not satisfied.[7]

There is a still more subtle consequence of the introduction of risk-bearing considerations. When there is uncertainty, information or knowledge becomes a commodity. Like other commodities, it has a cost of production and a cost of transmission, and so it is naturally not spread out over the entire population but concentrated among those who can profit most from it. (These costs may be measured in time or disutility as well as money.) But the demand for information is difficult to discuss in the rational terms usually employed. The value of information is frequently not known in any meaningful sense to the buyer; if, indeed, he knew enough to measure the value of information, he would know the information itself. But information, in the form of skilled care, is precisely what is being bought from most physicians, and, indeed, from most professionals. The elusive character of information as a commodity suggests that it departs considerably from the usual marketability assumptions about commodities.[8]

That risk and uncertainty are, in fact, significant elements in medical care hardly needs argument. I will hold that virtually all the special features of this industry, in fact, stem from the prevalence of uncertainty.

The nonexistence of markets for the bearing of some risks in the first instance reduces welfare for those who wish to transfer those risks to others for a certain price, as well as for those who would find it profitable to take on the risk at such prices. But it also reduces the desire to render or consume services which have risky consequences; in technical language, these commodities are complementary to risk-bearing. Conversely, the production and consumption of commodities and services with little risk attached act as substitutes for risk-bearing and are encouraged by market failure there with respect to risk-bearing. Thus the observed commodity pattern will be affected by the nonexistence of other markets.

[7] It should also be remarked that in the presence of uncertainty, indivisibilities that are sufficiently small to create little difficulty for the existence and viability of competitive equilibrium may nevertheless give rise to a considerable range of increasing returns because of the operation of the law of large numbers. Since most objects of insurance (lives, fire hazards, etc.) have some element of indivisibility, insurance companies have to be above a certain size. But it is not clear that this effect is sufficiently great to create serious obstacles to the existence and viability of competitive equilibrium in practice.

[8] One form of production of information is research. Not only does the product have unconventional aspects as a commodity, but it is also subject to increasing returns in use, since new ideas, once developed, can be used over and over without being consumed, and to difficulties of market control, since the cost of reproduction is usually much less than that of production. Hence, it is not surprising that a free enterprise economy will tend to underinvest in research; see Nelson [21] and Arrow [4].

The failure of one or more of the competitive preconditions has as its most immediate and obvious consequence a reduction in welfare below that obtainable from existing resources and technology, in the sense of a failure to reach an optimal state in the sense of Pareto. But more can be said. I propose here the view that, when the market fails to achieve an optimal state, society will, to some extent at least, recognize the gap, and nonmarket social institutions will arise attempting to bridge it.[9] Certainly this process is not necessarily conscious; nor is it uniformly successful in approaching more closely to optimality when the entire range of consequences is considered. It has always been a favorite activity of economists to point out that actions which on their face achieve a desirable goal may have less obvious consequences, particularly over time, which more than offset the original gains.

But it is contended here that the special structural characteristics of the medical-care market are largely attempts to overcome the lack of optimality due to the nonmarketability of the bearing of suitable risks and the imperfect marketability of information. These compensatory institutional changes, with some reinforcement from usual profit motives, largely explain the observed noncompetitive behavior of the medical-care market, behavior which, in itself, interferes with optimality. The social adjustment towards optimality thus puts obstacles in its own path.

The doctrine that society will seek to achieve optimality by nonmarket means if it cannot achieve them in the market is not novel. Certainly, the government, at least in its economic activities, is usually implicitly or explicitly held to function as the agency which substitutes for the market's failure.[10] I am arguing here that in some circumstances other social institutions will step into the optimality gap, and that the medical-care industry, with its variety of special institutions, some ancient, some modern, exemplifies this tendency.

It may be useful to remark here that a good part of the preference for redistribution expressed in government taxation and expenditure policies and private charity can be reinterpreted as desire for insurance. It is noteworthy that virtually nowhere is there a system of subsidies that has as its aim simply an equalization of income. The subsidies or other governmental help go to those who are disadvantaged in life by events the incidence of which is popularly regarded as unpre-

[9] An important current situation in which normal market relations have had to be greatly modified in the presence of great risks is the production and procurement of modern weapons; see Peck and Scherer [23, pp. 581-82] (I am indebted for this reference to V. Fuchs) and [1, pp. 71-75].

[10] For an explicit statement of this view, see Baumol [8]. But I believe this position is implicit in most discussions of the functions of government.

dictable: the blind, dependent children, the medically indigent. Thus, optimality, in a context which includes risk-bearing, includes much that appears to be motivated by distributional value judgments when looked at in a narrower context.[11]

This methodological background gives rise to the following plan for this paper. Section II is a catalogue of stylized generalizations about the medical-care market which differentiate it from the usual commodity markets. In Section III the behavior of the market is compared with that of the competitive model which disregards the fact of uncertainty. In Section IV, the medical-care market is compared, both as to behavior and as to preconditions, with the ideal competitive market that takes account of uncertainty; an attempt will be made to demonstrate that the characteristics outlined in Section II can be explained either as the result of deviations from the competitive preconditions or as attempts to compensate by other institutions for these failures. The discussion is not designed to be definitive, but provocative. In particular, I have been chary about drawing policy inferences; to a considerable extent, they depend on further research, for which the present paper is intended to provide a framework.

II. *A Survey of the Special Characteristics of the Medical-Care Market*[12]

This section will list selectively some characteristics of medical care which distinguish it from the usual commodity of economics textbooks. The list is not exhaustive, and it is not claimed that the characteristics listed are individually unique to this market. But, taken together, they do establish a special place for medical care in economic analysis.

A. *The Nature of Demand*

The most obvious distinguishing characteristics of an individual's demand for medical services is that it is not steady in origin as, for example, for food or clothing, but irregular and unpredictable. Medical services, apart from preventive services, afford satisfaction only in the event of illness, a departure from the normal state of affairs. It is hard, indeed, to think of another commodity of significance in the average budget of which this is true. A portion of legal services, devoted to defense in criminal trials or to lawsuits, might fall in this category but the incidence is surely very much lower (and, of course, there

[11] Since writing the above, I find that Buchanan and Tullock [10, Ch. 13] have argued that all redistribution can be interpreted as "income insurance."

[12] For an illuminating survey to which I am much indebted, see S. Mushkin [20].

are, in fact, strong institutional similarities between the legal and medical-care markets.)[13]

In addition, the demand for medical services is associated, with a considerable probability, with an assault on personal integrity. There is some risk of death and a more considerable risk of impairment of full functioning. In particular, there is a major potential for loss or reduction of earning ability. The risks are not by themselves unique; food is also a necessity, but avoidance of deprivation of food can be guaranteed with sufficient income, where the same cannot be said of avoidance of illness. Illness is, thus, not only risky but a costly risk in itself, apart from the cost of medical care.

B. *Expected Behavior of the Physician*

It is clear from everyday observation that the behavior expected of sellers of medical care is different from that of business men in general. These expectations are relevant because medical care belongs to the category of commodities for which the product and the activity of production are identical. In all such cases, the customer cannot test the product before consuming it, and there is an element of trust in the relation.[14] But the ethically understood restrictions on the activities of a physician are much more severe than on those of, say, a barber. His behavior is supposed to be governed by a concern for the customer's welfare which would not be expected of a salesman. In Talcott Parsons's terms, there is a "collectivity-orientation," which distinguishes medicine and other professions from business, where self-interest on the part of participants is the accepted norm.[15]

A few illustrations will indicate the degree of difference between the behavior expected of physicians and that expected of the typical businessman.[16] (1) Advertising and overt price competition are virtually eliminated among physicians. (2) Advice given by physicians as to further treatment by himself or others is supposed to be completely

[13] In governmental demand, military power is an example of a service used only irregularly and unpredictably. Here too, special institutional and professional relations have emerged, though the precise social structure is different for reasons that are not hard to analyze.

[14] Even with material commodities, testing is never so adequate that all elements of implicit trust can be eliminated. Of course, over the long run, experience with the quality of product of a given seller provides a check on the possibility of trust.

[15] See [22, p. 463]. The whole of [22, Ch. 10] is a most illuminating analysis of the social role of medical practice; though Parsons' interest lies in different areas from mine, I must acknowledge here my indebtedness to his work.

[16] I am indebted to Herbert Klarman of Johns Hopkins University for some of the points discussed in this and the following paragraph.

divorced from self-interest. (3) It is at least claimed that treatment is dictated by the objective needs of the case and not limited by financial considerations.[17] While the ethical compulsion is surely not as absolute in fact as it is in theory, we can hardly suppose that it has no influence over resource allocation in this area. Charity treatment in one form or another does exist because of this tradition about human rights to adequate medical care.[18] (4) The physician is relied on as an expert in certifying to the existence of illnesses and injuries for various legal and other purposes. It is socially expected that his concern for the correct conveying of information will, when appropriate, outweigh his desire to please his customers.[19]

Departure from the profit motive is strikingly manifested by the overwhelming predominance of nonprofit over proprietary hospitals.[20] The hospital per se offers services not too different from those of a hotel, and it is certainly not obvious that the profit motive will not lead to a more efficient supply. The explanation may lie either on the supply side or on that of demand. The simplest explanation is that public and private subsidies decrease the cost to the patient in nonprofit hospitals. A second possibility is that the association of profit-making with the supply of medical services arouses suspicion and antagonism on the part of patients and referring physicians, so they do prefer nonprofit institutions. Either explanation implies a preference on the part of some group, whether donors or patients, against the profit motive in the supply of hospital services.[21]

[17] The belief that the ethics of medicine demands treatment independent of the patient's ability to pay is strongly ingrained. Such a perceptive observer as René Dubos has made the remark that the high cost of anticoagulants restricts their use and may contradict classical medical ethics, as though this were an unprecedented phenomenon. See [13, p. 419]. "A time *may come* when medical ethics will have to be considered in the harsh light of economics" (emphasis added). Of course, this expectation amounts to ignoring the scarcity of medical resources; one has only to have been poor to realize the error. We may confidently assume that price and income do have some consequences for medical expenditures.

[18] A needed piece of research is a study of the exact nature of the variations of medical care received and medical care paid for as income rises. (The relevant income concept also needs study.) For this purpose, some disaggregation is needed; differences in hospital care which are essentially matters of comfort should, in the above view, be much more responsive to income than, e.g., drugs.

[19] This role is enhanced in a socialist society, where the state itself is actively concerned with illness in relation to work; see Field [14, Ch. 9].

[20] About 3 per cent of beds were in proprietary hospitals in 1958, against 30 per cent in voluntary nonprofit, and the remainder in federal, state, and local hospitals; see [26, Chart 4-2, p. 60].

[21] C. R. Rorem has pointed out to me some further factors in this analysis. (1) Given the social intention of helping all patients without regard to immediate ability to pay, economies of scale would dictate a predominance of community-sponsored hospitals. (2)

Conformity to collectivity-oriented behavior is especially important since it is a commonplace that the physician-patient relation affects the quality of the medical care product. A pure cash nexus would be inadequate; if nothing else, the patient expects that the same physician will normally treat him on successive occasions. This expectation is strong enough to persist even in the Soviet Union, where medical care is nominally removed from the market place [14, pp. 194-96]. That purely psychic interactions between physician and patient have effects which are objectively indistinguishable in kind from the effects of medication is evidenced by the use of the placebo as a control in medical experimentation; see Shapiro [25].

C. *Product Uncertainty*

Uncertainty as to the quality of the product is perhaps more intense here than in any other important commodity. Recovery from disease is as unpredictable as is its incidence. In most commodities, the possibility of learning from one's own experience or that of others is strong because there is an adequate number of trials. In the case of severe illness, that is, in general, not true; the uncertainty due to inexperience is added to the intrinsic difficulty of prediction. Further, the amount of uncertainty, measured in terms of utility variability, is certainly much greater for medical care in severe cases than for, say, houses or automobiles, even though these are also expenditures sufficiently infrequent so that there may be considerable residual uncertainty.

Further, there is a special quality to the uncertainty; it is very different on the two sides of the transaction. Because medical knowledge is so complicated, the information possessed by the physician as to the consequences and possibilities of treatment is necessarily very much greater than that of the patient, or at least so it is believed by both parties.[22] Further, both parties are aware of this informational inequality, and their relation is colored by this knowledge.

To avoid misunderstanding, observe that the difference in information relevant here is a difference in information as to the consequence of a purchase of medical care. There is always an inequality of information as to production methods between the producer and the purchaser of any commodity, but in most cases the customer may well

Some proprietary hospitals will tend to control total costs to the patient more closely, including the fees of physicians, who will therefore tend to prefer community-sponsored hospitals.

[22] Without trying to assess the present situation, it is clear in retrospect that at some point in the past the actual differential knowledge possessed by physicians may not have been much. But from the economic point of view, it is the subjective belief of both parties, as manifested in their market behavior, that is relevant.

have as good or nearly as good an understanding of the utility of the product as the producer.

D. *Supply Conditions*

In competitive theory, the supply of a commodity is governed by the net return from its production compared with the return derivable from the use of the same resources elsewhere. There are several significant departures from this theory in the case of medical care.

Most obviously, entry to the profession is restricted by licensing. Licensing, of course, restricts supply and therefore increases the cost of medical care. It is defended as guaranteeing a minimum of quality. Restriction of entry by licensing occurs in most professions, including barbering and undertaking.

A second feature is perhaps even more remarkable. The cost of medical education today is high and, according to the usual figures, is borne only to a minor extent by the student. Thus, the private benefits to the entering student considerably exceed the costs. (It is, however, possible that research costs, not properly chargeable to education, swell the apparent difference.) This subsidy should, in principle, cause a fall in the price of medical services, which, however, is offset by rationing through limited entry to schools and through elimination of students during the medical-school career. These restrictions basically render superfluous the licensing, except in regard to graduates of foreign schools.

The special role of educational institutions in simultaneously subsidizing and rationing entry is common to all professions requiring advanced training.[23] It is a striking and insufficiently remarked phenomenon that such an important part of resource allocation should be performed by nonprofit-oriented agencies.

Since this last phenomenon goes well beyond the purely medical aspect, we will not dwell on it longer here except to note that the anomaly is most striking in the medical field. Educational costs tend to be far higher there than in any other branch of professional training. While tuition is the same, or only slightly higher, so that the subsidy is much greater, at the same time the earnings of physicians rank highest among professional groups, so there would not at first blush seem to be any necessity for special inducements to enter the profession. Even if we grant that, for reasons unexamined here, there is a social interest in subsidized professional education, it is not clear why the rate of subsidization should differ among professions. One might ex-

[23] The degree of subsidy in different branches of professional education is worthy of a major research effort.

pect that the tuition of medical students would be higher than that of other students.

The high cost of medical education in the United States is itself a reflection of the quality standards imposed by the American Medical Association since the Flexner Report, and it is, I believe, only since then that the subsidy element in medical education has become significant. Previously, many medical schools paid their way or even yielded a profit.

Another interesting feature of limitation on entry to subsidized education is the extent of individual preferences concerning the social welfare, as manifested by contributions to private universities. But whether support is public or private, the important point is that both the quality and the quantity of the supply of medical care are being strongly influenced by social nonmarket forces.[24,25]

One striking consequence of the control of quality is the restriction on the range offered. If many qualities of a commodity are possible, it would usually happen in a competitive market that many qualities will be offered on the market, at suitably varying prices, to appeal to different tastes and incomes. Both the licensing laws and the standards of medical-school training have limited the possibilities of alternative qualities of medical care. The declining ratio of physicians to total employees in the medical-care industry shows that substitution of less trained personnel, technicians, and the like, is not prevented completely, but the central role of the highly trained physician is not affected at all.[26]

E. *Pricing Practices*

The unusual pricing practices and attitudes of the medical profession are well known: extensive price discrimination by income (with an extreme of zero prices for sufficiently indigent patients) and, formerly, a strong insistence on fee for services as against such alternatives as prepayment.

[24] Strictly speaking, there are four variables in the market for physicians: price, quality of entering students, quality of education, and quantity. The basic market forces, demand for medical services and supply of entering students, determine two relations among the four variables. Hence, if the nonmarket forces determine the last two, market forces will determine price and quality of entrants.

[25] The supply of Ph.D.'s is similarly governed, but there are other conditions in the market which are much different, especially on the demand side.

[26] Today only the Soviet Union offers an alternative lower level of medical personnel, the feldshers, who practice primarily in the rural districts (the institution dates back to the 18th century). According to Field [14, pp. 98-100, 132-33], there is clear evidence of strain in the relations between physicians and feldshers, but it is not certain that the feldshers will gradually disappear as physicians grow in numbers.

The opposition to prepayment is closely related to an even stronger opposition to closed-panel practice (contractual arrangements which bind the patient to a particular group of physicians). Again these attitudes seem to differentiate professions from business. Prepayment and closed-panel plans are virtually nonexistent in the legal profession. In ordinary business, on the other hand, there exists a wide variety of exclusive service contracts involving sharing of risks; it is assumed that competition will select those which satisfy needs best.[27]

The problems of implicit and explicit price-fixing should also be mentioned. Price competition is frowned on. Arrangements of this type are not uncommon in service industries, and they have not been subjected to antitrust action. How important this is is hard to assess. It has been pointed out many times that the apparent rigidity of so-called administered prices considerably understates the actual flexibility. Here, too, if physicians find themselves with unoccupied time, rates are likely to go down, openly or covertly; if there is insufficient time for the demand, rates will surely rise. The "ethics" of price competition may decrease the flexibility of price responses, but probably that is all.

III. *Comparisons with the Competitive Model under Certainty*

A. *Nonmarketable Commodities*

As already noted, the diffusion of communicable diseases provides an obvious example of nonmarket interactions. But from a theoretical viewpoint, the issues are well understood, and there is little point in expanding on this theme. (This should not be interpreted as minimizing the contribution of public health to welfare; there is every reason to suppose that it is considerably more important than all other aspects of medical care.)

Beyond this special area there is a more general interdependence, the concern of individuals for the health of others. The economic manifestations of this taste are to be found in individual donations to hospitals and to medical education, as well as in the widely accepted responsibilities of government in this area. The taste for improving the health of others appears to be stronger than for improving other aspects of their welfare.[28]

In interdependencies generated by concern for the welfare of others there is always a theoretical case for collective action if each participant derives satisfaction from the contributions of all.

[27] The law does impose some limits on risk-shifting in contracts, for example, its general refusal to honor exculpatory clauses.

[28] There may be an identification problem in this observation. If the failure of the market system is, or appears to be, greater in medical care than in, say, food an individual otherwise equally concerned about the two aspects of others' welfare may prefer to help in the first.

B. *Increasing Returns*

Problems associated with increasing returns play some role in allocation of resources in the medical field, particularly in areas of low density or low income. Hospitals show increasing returns up to a point; specialists and some medical equipment constitute significant indivisibilities. In many parts of the world the individual physician may be a large unit relative to demand. In such cases it can be socially desirable to subsidize the appropriate medical-care unit. The appropriate mode of analysis is much the same as for water-resource projects. Increasing returns are hardly apt to be a significant problem in general practice in large cities in the United States, and improved transportation to some extent reduces their importance elsewhere.

C. *Entry*

The most striking departure from competitive behavior is restriction on entry to the field, as discussed in II.D above. Friedman and Kuznets, in a detailed examination of the pre-World War II data, have argued that the higher income of physicians could be attributed to this restriction.[29]

There is some evidence that the demand for admission to medical school has dropped (as indicated by the number of applicants per place and the quality of those admitted), so that the number of medical-school places is not as significant a barrier to entry as in the early 1950's [28, pp. 14-15]. But it certainly has operated over the past and it is still operating to a considerable extent today. It has, of course, constituted a direct and unsubtle restriction on the supply of medical care.

There are several considerations that must be added to help evaluate the importance of entry restrictions: (1) Additional entrants would be, in general, of lower quality; hence, the addition to the supply of medical care, properly adjusted for quality, is less than purely quantitative calculations would show.[30] (2) To achieve genuinely competitive conditions, it would be necessary not only to remove numerical restrictions on entry but also to remove the subsidy in medical education. Like any other producer, the physician should bear all the costs of production,

[29] See [16, pp. 118-37]. The calculations involve many assumptions and must be regarded as tenuous; see the comments by C. Reinold Noyes in [16, pp. 407-10].

[30] It might be argued that the existence of racial discrimination in entrance has meant that some of the rejected applicants are superior to some accepted. However, there is no necessary connection between an increase in the number of entrants and a reduction in racial discrimination; so long as there is excess demand for entry, discrimination can continue unabated and new entrants will be inferior to those previously accepted.

including, in this case, education.[31] It is not so clear that this change would not keep even unrestricted entry down below the present level. (3) To some extent, the effect of making tuition carry the full cost of education will be to create too few entrants, rather than too many. Given the imperfections of the capital market, loans for this purpose to those who do not have the cash are difficult to obtain. The lender really has no security. The obvious answer is some form of insured loans, as has frequently been argued; not too much ingenuity would be needed to create a credit system for medical (and other branches of higher) education. Under these conditions the cost would still constitute a deterrent, but one to be compared with the high future incomes to be obtained.

If entry were governed by ideal competitive conditions, it may be that the quantity on balance would be increased, though this conclusion is not obvious. The average quality would probably fall, even under an ideal credit system, since subsidy plus selected entry draw some highly qualified individuals who would otherwise get into other fields. The decline in quality is not an over-all social loss, since it is accompanied by increase in quality in other fields of endeavor; indeed, if demands accurately reflected utilities, there would be a net social gain through a switch to competitive entry.[32]

There is a second aspect of entry in which the contrast with competitive behavior is, in many respects, even sharper. It is the exclusion of many imperfect substitutes for physicians. The licensing laws, though they do not effectively limit the number of physicians, do exclude all others from engaging in any one of the activities known as medical practice. As a result, costly physician time may be employed at specific tasks for which only a small fraction of their training is needed, and which could be performed by others less well trained and therefore less expensive. One might expect immunization centers, privately operated, but not necessarily requiring the services of doctors.

In the competitive model without uncertainty, consumers are presumed to be able to distinguish qualities of the commodities they buy. Under this hypothesis, licensing would be, at best, superfluous and exclude those from whom consumers would not buy anyway; but it might exclude too many.

D. *Pricing*

The pricing practices of the medical industry (see II.E above) de-

[31] One problem here is that the tax laws do not permit depreciation of professional education, so that there is a discrimination against this form of investment.

[32] To anticipate later discussion, this condition is not necessarily fulfilled. When it comes to quality choices, the market may be inaccurate.

part sharply from the competitive norm. As Kessel [17] has pointed out with great vigor, not only is price discrimination incompatible with the competitive model, but its preservation in the face of the large number of physicians is equivalent to a collective monopoly. In the past, the opposition to prepayment plans has taken distinctly coercive forms, certainly transcending market pressures, to say the least.

Kessel has argued that price discrimination is designed to maximize profits along the classic lines of discriminating monopoly and that organized medical opposition to prepayment was motivated by the desire to protect these profits. In principle, prepayment schemes are compatible with discrimination, but in practice they do not usually discriminate. I do not believe the evidence that the actual scale of discrimination is profit-maximizing is convincing. In particular, note that for any monopoly, discriminating or otherwise, the elasticity of demand in each market at the point of maximum profits is greater than one. But it is almost surely true for medical care that the price elasticity of demand for all income levels is less than one. That price discrimination by income is not completely profit-maximizing is obvious in the extreme case of charity; Kessel argues that this represents an appeasement of public opinion. But this already shows the incompleteness of the model and suggests the relevance and importance of social and ethical factors.

Certainly one important part of the opposition to prepayment was its close relation to closed-panel plans. Prepayment is a form of insurance, and naturally the individual physician did not wish to assume the risks. Pooling was intrinsically involved, and this strongly motivates, as we shall discuss further in Section IV below, control over prices and benefits. The simplest administrative form is the closed panel; physicians involved are, in effect, the insuring agent. From this point of view, Blue Cross solved the prepayment problem by universalizing the closed panel.

The case that price discrimination by income is a form of profit maximization which was zealously defended by opposition to fees for service seems far from proven. But it remains true that this price discrimination, for whatever cause, is a source of nonoptimality. Hypothetically, it means everyone would be better off if prices were made equal for all, and the rich compensated the poor for the changes in the relative positions. The importance of this welfare loss depends on the actual amount of discrimination and on the elasticities of demand for medical services by the different income groups. If the discussion is simplified by considering only two income levels, rich and poor, and if the elasticity of demand by either one is zero, then no reallocation of medical services will take place and the initial situation is optimal. The

only effect of a change in price will be the redistribution of income as between the medical profession and the group with the zero elasticity of demand. With low elasticities of demand, the gain will be small. To illustrate, suppose the price of medical care to the rich is double that to the poor, the medical expenditures by the rich are 20 per cent of those by the poor, and the elasticity of demand for both classes is .5; then the net social gain due to the abolition of discrimination is slightly over 1 per cent of previous medical expenditures.[33]

The issues involved in the opposition to prepayment, the other major anomaly in medical pricing, are not meaningful in the world of certainty and will be discussed below.

IV. *Comparison with the Ideal Competitive Model under Uncertainty*

A. *Introduction*

In this section we will compare the operations of the actual medical-care market with those of an ideal system in which not only the usual commodities and services but also insurance policies against all conceivable risks are available.[34] Departures consist for the most part of

[33] It is assumed that there are two classes, rich and poor; the price of medical services to the rich is twice that to the poor, medical expenditures by the rich are 20 per cent of those by the poor, and the elasticity of demand for medical services is .5 for both classes. Let us choose our quantity and monetary units so that the quantity of medical services consumed by the poor and the price they pay are both 1. Then the rich purchase .1 units of medical services at a price of 2. Given the assumption about the elasticities of demand, the demand function of the rich is $D_R(p) = .14\ p^{-.5}$ and that of the poor is $D_P(p) = p^{-.5}$. The supply of medical services is assumed fixed and therefore must equal 1.1. If price discrimination were abolished, the equilibrium price, $\bar{p}$, must satisfy the relation,

$$D_R(\bar{p}) + D_P(\bar{p}) = 1.1,$$

and therefore $\bar{p} = 1.07$. The quantities of medical care purchased by the rich and poor, respectively, would be $D_R(\bar{p}) = .135$ and $D_P(\bar{p}) = .965$.

The inverse demand functions, the price to be paid corresponding to any given quantity are $d_R(q) = .02/q^2$, and $d_P(q) = 1/q^2$. Therefore, the consumers' surplus to the rich generated by the change is:

$$(1) \qquad \int_{.1}^{.135} (.02/q^2)dq - \bar{p}(.135 - .1),$$

and similarly the loss in consumers' surplus by the poor is:

$$(2) \qquad \int_{.965}^{1} (1/q^2)dq - \bar{p}(1 - .965)$$

If (2) is subtracted from (1), the second terms cancel, and the aggregate increase in consumers' surplus is .0156, or a little over 1 per cent of the initial expenditures.

[34] A striking illustration of the desire for security in medical care is provided by the expressed preferences of *émigrés* from the Soviet Union as between Soviet medical practice and German or American practice; see Field [14, Ch. 12]. Those in Germany preferred the German system to the Soviet, but those in the United States preferred (in a ratio of 3 to 1) the Soviet system. The reasons given boil down to the certainty of medical care, independent of income or health fluctuations.

insurance policies that might conceivably be written, but are in fact not. Whether these potential commodities are nonmarketable, or, merely because of some imperfection in the market, are not actually marketed, is a somewhat fine point.

To recall what has already been said in Section I, there are two kinds of risks involved in medical care: the risk of becoming ill, and the risk of total or incomplete or delayed recovery. The loss due to illness is only partially the cost of medical care. It also consists of discomfort and loss of productive time during illness, and, in more serious cases, death or prolonged deprivation of normal function. From the point of view of the welfare economics of uncertainty, both losses are risks against which individuals would like to insure. The nonexistence of suitable insurance policies for either risk implies a loss of welfare.

B. *The Theory of Ideal Insurance*

In this section, the basic principles of an optimal regime for risk-bearing will be presented. For illustration, reference will usually be made to the case of insurance against cost in medical care. The principles are equally applicable to any of the risks. There is no single source to which the reader can be easily referred, though I think the principles are at least reasonably well understood.

As a basis for the analysis, the assumption is made that each individual acts so as to maximize the expected value of a utility function. If we think of utility as attached to income, then the costs of medical care act as a random deduction from this income, and it is the expected value of the utility of income after medical costs that we are concerned with. (Income after medical costs is the ability to spend money on other objects which give satisfaction. We presuppose that illness is not a source of satisfaction in itself; to the extent that it is a source of dissatisfaction, the illness should enter into the utility function as a separate variable.) The expected-utility hypothesis, due originally to Daniel Bernoulli (1738), is plausible and is the most analytically manageable of all hypotheses that have been proposed to explain behavior under uncertainty. In any case, the results to follow probably would not be significantly affected by moving to another mode of analysis.

It is further assumed that individuals are normally risk-averters. In utility terms, this means that they have a diminishing marginal utility of income. This assumption may reasonably be taken to hold for most of the significant affairs of life for a majority of people, but the presence of gambling provides some difficulty in the full application of this view. It follows from the assumption of risk aversion that if an individual is given a choice between a probability distribution of income, with a given mean m, and the certainty of the income m, he would prefer

the latter. Suppose, therefore, an agency, a large insurance company plan, or the government, stands ready to offer insurance against medical costs on an actuarially fair basis; that is, if the costs of medical care are a random variable with mean m, the company will charge a premium m, and agree to indemnify the individual for all medical costs. Under these circumstances, the individual will certainly prefer to take out a policy and will have a welfare gain thereby.

Will this be a social gain? Obviously yes, if the insurance agent is suffering no social loss. Under the assumption that medical risks on different individuals are basically independent, the pooling of them reduces the risk involved to the insurer to relatively small proportions. In the limit, the welfare loss, even assuming risk aversion on the part of the insurer, would vanish and there is a net social gain which may be of quite substantial magnitude. In fact, of course, the pooling of risks does not go to the limit; there is only a finite number of them and there may be some interdependence among the risks due to epidemics and the like. But then a premium, perhaps slightly above the actuarial level, would be sufficient to offset this welfare loss. From the point of view of the individual, since he has a strict preference for the actuarially fair policy over assuming the risks himself, he will still have a preference for an actuarially unfair policy, provided, of course, that it is not too unfair.

In addition to a residual degree of risk aversion by insurers, there are other reasons for the loading of the premium (i.e., an excess of premium over the actuarial value). Insurance involves administrative costs. Also, because of the irregularity of payments there is likely to be a cost of capital tied up. Suppose, to take a simple case, the insurance company is not willing to sell any insurance policy that a consumer wants but will charge a fixed-percentage loading above the actuarial value for its premium. Then it can be shown that the most preferred policy from the point of view of an individual is a coverage with a deductible amount; that is, the insurance policy provides 100 per cent coverage for all medical costs in excess of some fixed-dollar limit. If, however, the insurance company has some degree of risk aversion, its loading may also depend on the degree of uncertainty of the risk. In that case, the Pareto optimal policy will involve some element of coinsurance, i.e., the coverage for costs over the minimum limit will be some fraction less than 100 per cent (for proofs of these statements, see Appendix).

These results can also be applied to the hypothetical concept of insurance against failure to recover from illness. For simplicity, let us assume that the cost of failure to recover is regarded purely as a money cost, either simply productive opportunities foregone or, more gener-

ally, the money equivalent of all dissatisfactions. Suppose further that, given that a person is ill, the expected value of medical care is greater than its cost; that is, the expected money value attributable to recovery with medical help is greater than resources devoted to medical help. However, the recovery, though on the average beneficial, is uncertain; in the absence of insurance a risk-averter may well prefer not to take a chance on further impoverishment by buying medical care. A suitable insurance policy would, however, mean that he paid nothing if he doesn't benefit; since the expected value is greater than the cost, there would be a net social gain.[35]

C. *Problems of Insurance*

1. *The moral hazard.* The welfare case for insurance policies of all sorts is overwhelming. It follows that the government should undertake insurance in those cases where this market, for whatever reason, has failed to emerge. Nevertheless, there are a number of significant practical limitations on the use of insurance. It is important to understand them, though I do not believe that they alter the case for the creation of a much wider class of insurance policies than now exists.

One of the limits which has been much stressed in insurance literature is the effect of insurance on incentives. What is desired in the case of insurance is that the event against which insurance is taken be out of the control of the individual. Unfortunately, in real life this separation can never be made perfectly. The outbreak of fire in one's house or business may be largely uncontrollable by the individual, but the probability of fire is somewhat influenced by carelessness, and of course arson is a possibility, if an extreme one. Similarly, in medical policies the cost of medical care is not completely determined by the illness suffered by the individual but depends on the choice of a doctor and his willingness to use medical services. It is frequently observed that widespread medical insurance increases the demand for medical care. Coinsurance provisions have been introduced into many major medical policies to meet this contingency as well as the risk aversion of the insurance companies.

To some extent the professional relationship between physician and patient limits the normal hazard in various forms of medical insurance. By certifying to the necessity of given treatment or the lack thereof, the physician acts as a controlling agent on behalf of the insurance companies. Needless to say, it is a far from perfect check; the physicians themselves are not under any control and it may be convenient for them or pleasing to their patients to prescribe more expensive medi-

[35] It is a popular belief that the Chinese, at one time, paid their physicians when well but not when sick.

cation, private nurses, more frequent treatments, and other marginal variations of care. It is probably true that hospitalization and surgery are more under the casual inspection of others than is general practice and therefore less subject to moral hazard; this may be one reason why insurance policies in those fields have been more widespread.

2. *Alternative methods of insurance payment.* It is interesting that no less than three different methods of coverage of the costs of medical care have arisen: prepayment, indemnities according to a fixed schedule, and insurance against costs, whatever they may be. In prepayment plans, insurance in effect is paid in kind—that is, directly in medical services. The other two forms both involve cash payments to the beneficiary, but in the one case the amounts to be paid involving a medical contingency are fixed in advance, while in the other the insurance carrier pays all the costs, whatever they may be, subject, of course, to provisions like deductibles and coinsurance.

In hypothetically perfect markets these three forms of insurance would be equivalent. The indemnities stipulated would, in fact, equal the market price of the services, so that value to the insured would be the same if he were to be paid the fixed sum or the market price or were given the services free. In fact, of course, insurance against full costs and prepayment plans both offer insurance against uncertainty as to the price of medical services, in addition to uncertainty about their needs. Further, by their mode of compensation to the physician, prepayment plans are inevitably bound up with closed panels so that the freedom of choice of the physician by the patient is less than it would be under a scheme more strictly confined to the provision of insurance. These remarks are tentative, and the question of coexistence of the different schemes should be a fruitful subject for investigation.

3. *Third-party control over payments.* The moral hazard in physicians' control noted in paragraph 1 above shows itself in those insurance schemes where the physician has the greatest control, namely, major medical insurance. Here there has been a marked rise in expenditures over time. In prepayment plans, where the insurance and medical service are supplied by the same group, the incentive to keep medical costs to a minimum is strongest. In plans of the Blue Cross group, there has developed a conflict of interest between the insurance carrier and the medical-service supplier, in this case particularly the hospital.

The need for third-party control is reinforced by another aspect of the moral hazard. Insurance removes the incentive on the part of individuals, patients, and physicians to shop around for better prices for hospitalization and surgical care. The market forces, therefore, tend to be replaced by direct institutional control.

4. *Administrative costs.* The pure theory of insurance sketched in Section B above omits one very important consideration: the costs of operating an insurance company. There are several types of operating costs, but one of the most important categories includes commissions and acquisition costs, selling costs in usual economic terminology. Not only does this mean that insurance policies must be sold for considerably more than their actuarial value, but it also means there is a great differential among different types of insurance. It is very striking to observe that among health insurance policies of insurance companies in 1958, expenses of one sort or another constitute 51.6 per cent of total premium income for individual policies, and only 9.5 per cent for group policies [26, Table 14-1, p. 272]. This striking differential would seem to imply enormous economies of scale in the provision of insurance, quite apart from the coverage of the risks themselves. Obviously, this provides a very strong argument for widespread plans, including, in particular, compulsory ones.

5. *Predictability and insurance.* Clearly, from the risk-aversion point of view, insurance is more valuable, the greater the uncertainty in the risk being insured against. This is usually used as an argument for putting greater emphasis on insurance against hospitalization and surgery than other forms of medical care. The empirical assumption has been challenged by O. W. Anderson and others [3, pp. 53-54], who asserted that out-of-hospital expenses were equally as unpredictable as in-hospital costs. What was in fact shown was that the probability of costs exceeding $200 is about the same for the two categories, but this is not, of course, a correct measure of predictability, and a quick glance at the supporting evidence shows that in relation to the average cost the variability is much lower for ordinary medical expenses. Thus, for the city of Birmingham, the mean expenditure on surgery was $7, as opposed to $20 for other medical expenses, but of those who paid something for surgery the average bill was $99, as against $36 for those with some ordinary medical cost. Eighty-two per cent of those interviewed had no surgery, and only 20 per cent had no ordinary medical expenses [3, Tables A-13, A-18, and A-19 on pp. 72, 77, and 79, respectively].

The issue of predictability also has bearing on the merits of insurance against chronic illness or maternity. On a lifetime insurance basis, insurance against chronic illness makes sense, since this is both highly unpredictable and highly significant in costs. Among people who already have chronic illness, or symptoms which reliably indicate it, insurance in the strict sense is probably pointless.

6. *Pooling of unequal risks.* Hypothetically, insurance requires for its full social benefit a maximum possible discrimination of risks. Those

in groups of higher incidences of illness should pay higher premiums. In fact, however, there is a tendency to equalize, rather than to differentiate, premiums, especially in the Blue Cross and similar widespread schemes. This constitutes, in effect, a redistribution of income from those with a low propensity to illness to those with a high propensity. The equalization, of course, could not in fact be carried through if the market were genuinely competitive. Under those circumsances, insurance plans could arise which charged lower premiums to preferred risks and draw them off, leaving the plan which does not discriminate among risks with only an adverse selection of them.

As we have already seen in the case of income redistribution, some of this may be thought of as insurance with a longer time perspective. If a plan guarantees to everybody a premium that corresponds to total experience but not to experience as it might be segregated by smaller subgroups, everybody is, in effect, insured against a change in his basic state of health which would lead to a reclassification. This corresponds precisely to the use of a level premium in life insurance instead of a premium varying by age, as would be the case for term insurance.

7. *Gaps and coverage.* We may briefly note that, at any rate to date, insurances against the cost of medical care are far from universal. Certain groups—the unemployed, the institutionalized, and the aged—are almost completely uncovered. Of total expenditures, between one-fifth and one-fourth are covered by insurance. It should be noted, however, that over half of all hospital expenses and about 35 per cent of the medical payments of those with bills of $1,000 a year and over, are included [26, p. 376]. Thus, the coverage on the more variable parts of medical expenditure is somewhat better than the over-all figures would indicate, but it must be assumed that the insurance mechanism is still very far from achieving the full coverage of which it is capable.

D. *Uncertainty of Effects of Treatment*

1. There are really two major aspects of uncertainty for an individual already suffering from an illness. He is uncertain about the effectiveness of medical treatment, and his uncertainty may be quite different from that of his physician, based on the presumably quite different medical knowledges.

2. *Ideal insurance.* This will necessarily involve insurance against a failure to benefit from medical care, whether through recovery, relief of pain, or arrest of further deterioration. One form would be a system in which the payment to the physician is made in accordance with the degree of benefit. Since this would involve transferring the risks from the patient to the physician, who might certainly have an aversion to bearing them, there is room for insurance carriers to pool the risks,

either by contract with physicians or by contract with the potential patients. Under ideal insurance, medical care will always be undertaken in any case in which the expected utility, taking account of the probabilities, exceeds the expected medical cost. This prescription would lead to an economic optimum. If we think of the failure to recover mainly in terms of lost working time, then this policy would, in fact, maximize economic welfare as ordinarily measured.

3. *The concepts of trust and delegation.* In the absence of ideal insurance, there arise institutions which offer some sort of substitute guarantees. Under ideal insurance the patient would actually have no concern with the informational inequality between himself and the physician, since he would only be paying by results anyway, and his utility position would in fact be thoroughly guaranteed. In its absence he wants to have some guarantee that at least the physician is using his knowledge to the best advantage. This leads to the setting up of a relationship of trust and confidence, one which the physician has a social obligation to live up to. Since the patient does not, at least in his belief, know as much as the physician, he cannot completely enforce standards of care. In part, he replaces direct observation by generalized belief in the ability of the physician.[36] To put it another way, the social obligation for best practice is part of the commodity the physician sells, even though it is a part that is not subject to thorough inspection by the buyer.

One consequence of such trust relations is that the physician cannot act, or at least appear to act, as if he is maximizing his income at every moment of time. As a signal to the buyer of his intentions to act as thoroughly in the buyer's behalf as possible, the physician avoids the obvious stigmata of profit-maximizing. Purely arms-length bargaining behavior would be incompatible, not logically, but surely psychologically, with the trust relations. From these special relations come the various forms of ethical behavior discussed above, and so also, I suggest, the relative unimportance of profit-making in hospitals. The very word, "profit," is a signal that denies the trust relations.

Price discrimination and its extreme, free treatment for the indigent, also follow. If the obligation of the physician is understood to be first of all to the welfare of the patient, then in particular it takes precedence over financial difficulties.

As a second consequence of informational inequality between physician and patient and the lack of insurance of a suitable type, the patient must delegate to the physician much of his freedom of choice.

[36] Francis Bator points out to me that some protection can be achieved, at a price, by securing additional opinions.

He does not have the knowledge to make decisions on treatment, referral, or hospitalization. To justify this delegation, the physician finds himself somewhat limited, just as any agent would in similar circumstances. The safest course to take to avoid not being a true agent is to give the socially prescribed "best" treatment of the day. Compromise in quality, even for the purpose of saving the patient money, is to risk an imputation of failure to live up to the social bond.

The special trust relation of physicians (and allied occuptions, such as priests) extends to third parties so that the certifications of physicians as to illness and injury are accepted as especially reliable (see Section II.B above). The social value to all concerned of such presumptively reliable sources of information is obvious.

Notice the general principle here. Because there are barriers to the information flow and because there is no market in which the risks involved can be insured, coordination of purchase and sales must take place through convergent expectations, but these are greatly assisted by having clear and prominent signals, and these, in turn, force patterns of behavior which are not in themselves logical necessities for optimality.[37]

4. *Licensing and educational standards.* Delegation and trust are the social institutions designed to obviate the problem of informational inequality. The general uncertainty about the prospects of medical treatment is socially handled by rigid entry requirements. These are designed to reduce the uncertainty in the mind of the consumer as to the quality of product insofar as this is possible.[38] I think this explanation, which is perhaps the naive one, is much more tenable than any idea of a monopoly seeking to increase incomes. No doubt restriction on entry is desirable from the point of view of the existing physicians, but the public pressure needed to achieve the restriction must come from deeper causes.

The social demand for guaranteed quality can be met in more than one way, however. At least three attitudes can be taken by the state or other social institutions toward entry into an occupation or toward the production of commodities in general; examples of all three types exist. (1) The occupation can be licensed, nonqualified entrants being simply excluded. The licensing may be more complex than it is in medicine; individuals could be licensed for some, but not all, medical activities, for example. Indeed, the present all-or-none approach could

[37] The situation is very reminiscent of the crucial role of the focal point in Schelling's theory of tacit games, in which two parties have to find a common course of action without being able to communicate; see [24, esp. pp. 225 ff.].

[38] How well they achieve this end is another matter. R. Kessel points out to me that they merely guarantee training, not continued good performance as medical technology changes.

be criticized as being insufficient with regard to complicated specialist treatment, as well as excessive with regard to minor medical skills. Graded licensing may, however, be much harder to enforce. Controls could be exercised analogous to those for foods; they can be excluded as being dangerous, or they can be permitted for animals but not for humans. (2) The state or other agency can certify or label, without compulsory exclusion. The category of Certified Psychologist is now under active discussion; canned goods are graded. Certification can be done by nongovernmental agencies, as in the medical-board examinations for specialists. (3) Nothing at all may be done; consumers make their own choices.

The choice among these alternatives in any given case depends on the degree of difficulty consumers have in making the choice unaided, and on the consequences of errors of judgment. It is the general social consensus, clearly, that the *laissez-faire* solution for medicine is intolerable. The certification proposal never seems to have been discussed seriously. It is beyond the scope of this paper to discuss these proposals in detail. I wish simply to point out that they should be judged in terms of the ability to relieve the uncertainty of the patient in regard to the quality of the commodity he is purchasing, and that entry restrictions are the consequences of an apparent inability to devise a system in which the risks of gaps in medical knowledge and skill are borne primarily by the patient, not the physician.

Postscript

I wish to repeat here what has been suggested above in several places: that the failure of the market to insure against uncertainties has created many social institutions in which the usual assumptions of the market are to some extent contradicted. The medical profession is only one example, though in many respects an extreme one. All professions share some of the same properties. The economic importance of personal and especially family relationships, though declining, is by no means trivial in the most advanced economies; it is based on nonmarket relations that create guarantees of behavior which would otherwise be afflicted with excessive uncertainty. Many other examples can be given. The logic and limitations of ideal competitive behavior under uncertainty force us to recognize the incomplete description of reality supplied by the impersonal price system.

References

1. A. A. Alchian, K. J. Arrow, and W. M. Capron, *An Economic Analysis of the Market for Scientists and Engineers,* RAND RM-2190-RC. Santa Monica 1958.

2. M. ALLAIS, "Géneralisation des théories de l'équilibre économique général et du rendement social au cas du risque," in Centre National de la Recherche Scientifique, *Econometrie,* Paris 1953, pp. 1-20.
3. O. W. ANDERSON AND STAFF OF THE NATIONAL OPINION RESEARCH CENTER, *Voluntary Health Insurance in Two Cities.* Cambridge, Mass. 1957.
4. K. J. ARROW, "Economic Welfare and the Allocation of Resources for Invention," in Nat. Bur. Econ. Research, *The Role and Direction of Inventive Activity: Economic and Social Factors,* Princeton 1962, pp. 609-25.
5. ———, "Les rôle des valeurs boursières pour la répartition la meilleure des risques," in Centre National de la Recherche Scientifique, *Econometrie,* Paris 1953, pp. 41-46.
6. F. M. BATOR, "The Anatomy of Market Failure," *Quart. Jour. Econ.,* Aug. 1958, *72,* 351-79.
7. E. BAUDIER, "L'introduction du temps dans la théorie de l'équilibre général," *Les Cahiers Economiques,* Dec. 1959, 9-16.
8. W. J. BAUMOL, *Welfare Economics and the Theory of the State.* Cambridge, Mass. 1952.
9. K. BORCH, "The Safety Loading of Reinsurance Premiums," *Skandinavisk Aktuariehdskrift,* 1960, pp. 163-84.
10. J. M. BUCHANAN AND G. TULLOCK, *The Calculus of Consent.* Ann Arbor 1962.
11. G. DEBREU, "Une économique de l'incertain," *Economie Appliquée,* 1960, *13,* 111-16.
12. ———, *Theory of Values.* New York 1959.
13. R. DUBOS, "Medical Utopias," *Daedalus,* 1959, *88,* 410-24.
14. M. G. FIELD, *Doctor and Patient in Soviet Russia.* Cambridge, Mass. 1957.
15. MILTON FRIEDMAN, "The Methodology of Positive Economics," in *Essays in Positive Economics,* Chicago 1953, pp. 3-43.
16. ——— AND S. S. KUZNETS, *Income from Independent Professional Practice.* Nat. Bur. Econ. Research, New York 1945.
17. R. A. KESSEL, "Price Discrimination in Medicine," *Jour. Law and Econ.,* 1958, *1,* 20-53.
18. T. C. KOOPMANS, "Allocation of Resources and the Price System," in *Three Essays on the State of Economic Science,* New York 1957, pp. 1-120.
19. I. M. D. LITTLE, *A Critique of Welfare Economics.* Oxford 1950.
20. SELMA MUSHKIN, "Towards a Definition of Health Economics," *Public Health Reports,* 1958, *73,* 785-93.
21. R. R. NELSON, "The Simple Economics of Basic Scientific Research," *Jour. Pol. Econ.,* June 1959, *67,* 297-306.
22. T. PARSONS, *The Social System.* Glencoe 1951.
23. M. J. PECK AND F. M. SCHERER, *The Weapons Acquisition Process: An Economic Analysis.* Div. of Research, Graduate School of Business, Harvard University, Boston 1962.

24. T. C. SCHELLING, *The Strategy of Conflict.* Cambridge, Mass. 1960.
25. A. K. SHAPIRO, "A Contribution to a History of the Placebo Effect," *Behavioral Science,* 1960, *5,* 109-35.
26. H. M. SOMERS AND A. R. SOMERS, *Doctors, Patients, and Health Insurance.* The Brookings Institution, Washington 1961.
27. C. L. STEVENSON, *Ethics and Language.* New Haven 1945.
28. U. S. DEPARTMENT OF HEALTH, EDUCATION AND WELFARE, *Physicians for a Growing America,* Public Health Service Publication No. 709, Oct. 1959.

APPENDIX

On Optimal Insurance Policies

The two propositions about the nature of optimal insurance policies asserted in Section IV.B above will be proved here.

Proposition 1. If an insurance company is willing to offer any insurance policy against loss desired by the buyer at a premium which depends only on the policy's actuarial value, then the policy chosen by a risk-averting buyer will take the form of 100 per cent coverage above a deductible minimum.

Note: The premium will, in general, exceed the actuarial value; it is only required that two policies with the same actuarial value will be offered by the company for the same premium.

Proof: Let W be the initial wealth of the individual, X his loss, a random variable, $I(X)$ the amount of insurance paid if loss X occurs, P the premium, and $Y(X)$ the wealth of the individual after paying the premium, incurring the loss, and receiving the insurance benefit.

$$Y(X) = W - P - X + I(X). \tag{1}$$

The individual values alternative policies by the expected utility of his final wealth position, $Y(X)$. Let $U(y)$ be the utility of final wealth, y; then his aim is to maximize,

$$E\{U[Y(X)]\}, \tag{2}$$

where the symbol, E, denotes mathematical expectation.

An insurance payment is necessarily nonnegative, so the insurance policy must satisfy the condition,

$$I(X) \geq 0 \quad \text{for all} \quad X. \tag{3}$$

If a policy is optimal, it must in particular be better in the sense of the criterion (2), than any other policy with the same actuarial expectation, $E[I(X)]$. Consider a policy that pays some positive amount of insurance at one level of loss, say X_1, but which permits the final wealth at some other loss level, say X_2, to be lower than that corresponding to X_1. Then, it is intuitively obvious that a risk-averter would prefer an alternative policy with the same actuarial value which would offer slightly less protection for losses in the neighborhood of X_1 and slightly higher protection for those in the neighborhood of X_2, since risk aversion implies that the marginal utility

of $Y(X)$ is greater when $Y(X)$ is smaller: hence, the original policy cannot be optimal.

To prove this formally, let $I_1(X)$ be the original policy, with $I_1(X)>0$ and $Y_1(X_1)>Y_2(X_2)$, where $Y_1(X)$ is defined in terms of $I_1(X)$ by (I). Choose δ sufficiently small so that,

(4) $I_1(X) > 0 \quad \text{for} \quad X_1 \leq X \leq X_1 + \delta,$

(5) $Y_1(X') < Y_1(X) \quad \text{for} \quad X_2 \leq X' \leq X_2 + \delta, \quad X_1 \leq X \leq X_1 + \delta.$

(This choice of δ is possible if the functions $I_1(X)$, $Y_1(X)$ are continuous; this can be proved to be true for the optimal policy, and therefore we need only consider this case.)

Let π_1 be the probability that the loss, X, lies in the interval $\langle X_1, X_1+\delta\rangle$, π_2 the probability that X lies in the interval $\langle X_2, X_2+\delta\rangle$. From (4) and (5) we can choose $\epsilon>0$ and sufficiently small so that,

(6) $I_1(X) - \pi_2\epsilon \geq 0 \quad \text{for} \quad X_1 \leq X \leq X_1 + \delta,$

(7) $Y_1(X') + \pi_1\epsilon < Y_1(X) - \pi_2\epsilon$

$$\text{for} \quad X_2 \leq X' \leq X_2 + \delta, \quad X_1 \leq X \leq X_1 + \delta.$$

Now define a new insurance policy, $I_2(X)$, which is the same as $I_1(X)$ except that it is smaller by $\pi_2\epsilon$ in the interval from X_1 to $X_1+\delta$ and larger by $\pi_1\epsilon$ in the interval from X_2 to $X_2+\delta$. From (6), $I_2(X)\geq 0$ everywhere, so that (3) is satisfied. We will show that $E[I_1(X)]=E[I_2(X)]$ and that $I_2(X)$ yields the higher expected utility, so that $I_1(X)$ is not optimal.

Note that $I_2(X)-I_1(X)$ equals $-\pi_2\epsilon$ for $X_1\leq X\leq X_1+\delta$, $\pi_1\epsilon$ for $X_2\leq X \leq X_2+\delta$, and 0 elsewhere. Let $\phi(X)$ be the density of the random variable X. Then,

$$\begin{aligned} E[I_2(X) - I_1(X)] &= \int_{X_1}^{X_1+\delta} [I_2(X) - I_1(X)]\phi(X)dX \\ &\quad + \int_{X_2}^{X_2+\delta} [I_2(X) - I_1(X)]dX \\ &= (-\pi_2\epsilon) \int_{X_1}^{X_1+\delta} \phi(X)dX + (\pi_1\epsilon) \int_{X2}^{X_2+\delta} \phi(X)dX \\ &= -(\pi_2\epsilon)\pi_1 + (\pi_1\epsilon)\pi_2 = 0, \end{aligned}$$

so that the two policies have the same actuarial value and, by assumption, the same premium.

Define $Y_2(X)$ in terms of $I_2(X)$ by (1). Then $Y_2(X)-Y_1(X)=I_2(X)-I_1(X)$. From (7),

(8) $Y_1(X') < Y_2(X') < Y_2(X) < Y_1(X)$

$$\text{for} \quad X_2 \leq X' \leq X_2 + \delta, \quad X_1 \leq X \leq X_1 + \delta.$$

Since $Y_1(X)-Y_2(X)=0$ outside the intervals $\langle X_1, X_1+\delta\rangle$, $\langle X_2, X_2+\delta\rangle$, we

can write,

$$(9)\quad E\{U[Y_2(X)] - U[Y_1(X)]\} = \int_{X_1}^{X_1+\delta} \{U[Y_2(X)] - U[Y_1(X)]\}\phi(X)dX + \int_{X_2}^{X_2+\delta} \{U[Y_2(X)] - U[Y_1(X)]\}\phi(X)dX.$$

By the Mean Value Theorem, for any given value of X,

$$(10)\quad U[Y_2(X)] - U[Y_1(X)] = U'[Y(X)][Y_2(X) - Y_1(X)] = U'[Y(X)][I_2(X) - I_1(X)],$$

where $Y(X)$ lies between $Y_1(X)$ and $Y_2(X)$. From (8),

$$Y(X') < Y(X) \quad \text{for} \quad X_2 \leq X' \leq X_2 + \delta, \quad X_1 \leq X \leq X_1 + \delta,$$

and, since $U'(y)$ is a diminishing function of y for a risk-averter,

$$U'[Y(X')] > U'[Y(X)]$$

or, equivalently, for some number u,

$$(11)\quad \begin{aligned} U'[Y(X')] &> u \quad \text{for} \quad X_2 \leq X' \leq X_2 + \delta, \\ U'[Y(X)] &< u \quad \text{for} \quad X_1 \leq X \leq X_1 + \delta. \end{aligned}$$

Now substitute (10) into (9),

$$E\{U[Y_2(X)] - U[Y_1(X)]\} = -\pi_2\epsilon \int_{X_1}^{X_1+\delta} U'[Y(X)]\phi(X)dX + \pi_1\epsilon \int_{X_2}^{X_2+\delta} U'[Y(X)]\phi(X)dX.$$

From (11), it follows that,

$$E\{U[Y_2(X)] - U[Y_1(X)]\} > -\pi_2\epsilon u\pi_1 + \pi_1\epsilon u\pi_2 = 0,$$

so that the second policy is preferred.

It has thus been shown that a policy cannot be optimal if, for some X_1 and X_2, $I(X_1) > 0$, $Y(X_1) > Y(X_2)$. This may be put in a different form: Let $Y_{\min}$ be the minimum value taken on by $Y(X)$ under the optimal policy; then we must have $I(X) = 0$ if $Y(X) > Y_{\min}$. In other words, a minimum final wealth level is set; if the loss would not bring wealth below this level, no benefit is paid, but if it would, then the benefit is sufficient to bring up the final wealth position to the stipulated minimum. This is, of course, precisely a description of 100 per cent coverage for loss above a deductible.

We turn to the second proposition. It is now supposed that the insurance company, as well as the insured, is a risk-averter; however, there are no administrative or other costs to be covered beyond protection against loss.

Proposition 2. If the insured and the insurer are both risk-averters and there are no costs other than coverage of losses, then any nontrivial Pareto-

optimal policy, $I(X)$, as a function of the loss, X, must have the property, $0<dI/dX<1$.

That is, any increment in loss will be partly but not wholly compensated by the insurance company; this type of provision is known as coinsurance. Proposition 2 is due to Borch [9, Sec. 2]; we give here a somewhat simpler proof.

Proof: Let $U(y)$ be the utility function of the insured, $V(z)$ that of the insurer. Let W_0 and W_1 be the initial wealths of the two, respectively. In this case, we let $I(X)$ be the insurance benefits less the premium; for the present purpose, this is the only significant magnitude (since the premium is independent of X, this definition does not change the value of dI/dX). The final wealth positions of the insured and insurer are:

$$\begin{aligned} Y(X) &= W_0 - X + I(X), \\ Z(X) &= W_1 - I(X), \end{aligned} \tag{12}$$

respectively. Any given insurance policy then defines expected utilities, $u=E\{U[Y(X)]\}$ and $v=E\{V[Z(X)]\}$, for the insured and insurer, respectively. If we plot all points (u, v) obtained by considering all possible insurance policies, the resulting expected-utility-possibility set has a boundary that is convex to the northeast. To see this, let $I_1(X)$ and $I_2(X)$ be any two policies, and let (u_1, v_1) and (u_2, v_2) be the corresponding points in the two-dimensional expected-utility-possibility set. Let a third insurance policy, $I(X)$, be defined as the average of the two given ones,

$$I(X) = (\tfrac{1}{2})I_1(X) + (\tfrac{1}{2})I_2(X),$$

for each X. Then, if $Y(X)$, $Y_1(X)$, and $Y_2(X)$ are the final wealth positions of the insured, and $Z(X)$, $Z_1(X)$, and $Z_2(X)$ those of the insurer for each of the three policies, $I(X)$, $I_1(X)$, and $I_2(X)$, respectively,

$$\begin{aligned} Y(X) &= (\tfrac{1}{2})Y_1(X) + (\tfrac{1}{2})Y_2(X), \\ Z(X) &= (\tfrac{1}{2})Z_1(X) + (\tfrac{1}{2})Z_2(X), \end{aligned}$$

and, because both parties have diminishing marginal utility,

$$\begin{aligned} U[Y(X)] &\geq (\tfrac{1}{2})U[Y_1(X)] + (\tfrac{1}{2})U[Y_2(X)], \\ V[Z(X)] &\geq (\tfrac{1}{2})V[Z_1(X)] + (\tfrac{1}{2})V[Z_2(X)]. \end{aligned}$$

Since these statements hold for all X, they also hold when expectations are taken. Hence, there is a point (u, v) in the expected-utility-possibility set for which $u\geq(\tfrac{1}{2})u_1+(\tfrac{1}{2})u_2$, $v\geq(\tfrac{1}{2})v_1+(\tfrac{1}{2})v_2$. Since this statement holds for every pair of points (u_1, y_1) and (u_2, v_2) in the expected-utility-possibility set, and in particular for pairs of points on the northeast boundary, it follows that the boundary must be convex to the northeast.

From this, in turn, it follows that any given Pareto-optimal point (i.e., any point on the northeast boundary) can be obtained by maximizing a linear function, $\alpha u+\beta v$, with suitably chosen α and β nonnegative and at least one positive, over the expected-utility-possibility set. In other words, a Pareto-optimal insurance policy, $I(X)$, is one which maximizes,

$$\alpha E\{U[Y(X)]\} + \beta E\{V[Z(X)]\} = E\{\alpha U[Y(X)] + \beta V[Z(X)]\},$$

for some $\alpha \geq 0$, $\beta \geq 0$, $\alpha > 0$ or $\beta > 0$. To maximize this expectation, it is obviously sufficient to maximize:

$$\alpha U[Y(X)] + \beta V[Z(X)], \tag{13}$$

with respect to $I(X)$, for each X. Since, for given X, it follows from (12) that,

$$dY(X)/dI(X) = 1, \qquad dZ(X)/dI(X) = -1,$$

it follows by differentiation of (13) that $I(X)$ is the solution of the equation,

$$\alpha U'[Y(X)] - \beta V'[Z(X)] = 0. \tag{14}$$

The cases $\alpha = 0$ or $\beta = 0$ lead to obvious trivialities (one party simply hands over all his wealth to the other), so we assume $\alpha > 0, \beta > 0$. Now differentiate (14) with respect to X and use the relations, derived from (12),

$$dY/dX = (dI/dX) - 1, \qquad dZ/dX = -(dI/dX).$$

$$\alpha U''[Y(X)][(dI/dX) - 1] + \beta V''[Z(X)](dI/dX) = 0,$$

or

$$dI/dX = \alpha U''[Y(X)]/\{\alpha U''[Y(X)] + \beta V''[Z(X)]\}.$$

Since $U''[Y(X)] < 0$, $V''[Z(X)] < 0$ by the hypothesis that both parties are risk-averters, Proposition 2 follows.