

THEORETICAL ISSUES IN MEDICAL MALPRACTICE

Steven Shavell

from The Economics of Medical Malpractice

edited by Simon Rottenberg, published by

American Enterprise Institute, Washington,

D. C., 1978

Introduction

This paper is concerned with medical malpractice as a problem in normative economics. Accordingly, it focuses on the connection between malpractice and two principal determinants of individual welfare, these being compensation for medical accidents and the quality and cost of medical services. It is assumed that individual utility depends on the probabilities of various alternative states of health and on the level of income which would be enjoyed in each possible state of health. These variables are known once the two determinants are specified. The accidents of special interest are iatrogenic, those which may be imputed to medical treatment itself rather than to a prior medical condition. By medical accident is meant any adverse medical event, whether or not it is in some sense "accidental" in nature. So far as the second determinant is concerned, emphasis is placed on the role of nonmarket institutions in assuring the quality of services. Assuring quality, in turn, is viewed as a dual problem: the typical physician¹ must be given the correct incentives to take care in the provision of medical services; and relatively unskilled physicians must be compelled or induced to get further training or to limit the scope of their activities. While there are other factors which clearly affect the quality and cost of medical services—consumer demand, the supply of new physicians, and the level of medical knowledge and technology—these are not emphasized.

The first two sections of the paper review theoretical issues pertinent to compensation and to assuring the quality of medical services. Then alternative approaches to liability for medical accidents are briefly

I wish to thank Kenneth Arrow, Ralph Berry, Guido Calabresi, Richard Danzig, Peter Diamond, Ted Frech, Victor Fuchs, Jerry Green, Robert Keeton, A. Mitchell Polinsky, Henry Steiner, and Richard Zeckhauser for comments and the National Science Foundation and the American Enterprise Institute for research support.

¹ The word *physician* is used for convenience; it is recognized that other providers (nurses, hospitals) of medical services are involved in the problems of malpractice.

compared; these are patient liability, strict liability, no-fault plans, and the negligence system (including proposals which would modify but not radically reform the system). The final section states the major conclusion of the paper—that special reliance on the courts for compensation of medical accidents is inappropriate—and also comments on the problem of assuring the quality of medical services.

The performance of the price system under perfect conditions is taken as a point of departure and standard for comparison. This is because under perfect conditions the price system results in an optimal outcome in Pareto's sense—it is impossible to simultaneously improve everyone's well-being. Moreover, given a suitable reallocation of ownership of resources, any Pareto optimal outcome can be achieved with the price system; complaints about economic outcomes are therefore reduced to complaints about the distribution of initial resources, notably income.² The conditions for an optimal outcome that are, perhaps, most important for the case of medical malpractice pertain to information about risks and the quality of medical services.³ If information about both were perfect, there would be little difficulty with either compensation or assuring quality. Individuals would be aware of the true risks of medical accidents and would provide themselves with adequate insurance protection (given their income).⁴ Individuals would recognize the true characteristics of medical care and would pay for services only in accordance with quality. Of course, much of the discussion here has to do with the consequences of individuals' lack of information about risks and the quality of medical services.

Compensation for Medical Accidents

In order to isolate issues related to compensation,⁵ the probability of medical accidents and thus physician behavior is taken as fixed in this part of the paper.

² See for example Kenneth J. Arrow and Frank Hahn, *General Competitive Analysis* (San Francisco: Holden Day, 1971).

³ See Kenneth J. Arrow, "Uncertainty and the Welfare Economics of Medical Care," *American Economic Review*, vol. 53 (December 1963), pp. 941-73. Throughout the paper the words *optimal*, *ideal*, and *appropriate* mean Pareto optimal.

⁴ There are other obstacles to optimal insurance coverage which would also be eliminated in a hypothetical world of perfect information. These are discussed below.

⁵ R. Keston, "Compensation for Medical Accidents," *University of Pennsylvania Law Review*, vol. 121 (1973), pp. 590-617, contains an interesting discussion of the general subject of this section.

A Guide to Optimal Compensation. Whether publicly or privately provided, compensation for a medical accident ought to reflect the insurance coverage against the accident that a rational, well-informed individual with a socially acceptable level of income would have bought.⁶ That is, the determination of optimal compensation ought to reflect a hypothetical decision about the purchase of insurance.⁷ This point of view certainly does not imply that actual compensation should necessarily be left up to individuals and their insurance companies. As suggested in the introduction and stressed below, individuals may not make well-informed decisions about insurance (or may not have socially appropriate incomes).

In the next several subsections, aspects of the insurance purchase decision when made under perfect market conditions are considered. Then characteristics of the decision when made under more realistic circumstances are discussed.

Optimal Insurance Coverage versus Compensation for Economic Losses versus Compensation to Make a Person Whole. When buying insurance, an individual considers the cost of coverage, the likelihood of an accident, and—what is of particular importance here—the utility he would derive from income if he did not have an accident as compared to the utility he would derive from income if he did have an accident.⁸

Two examples make it clear that the level of coverage which an individual would select does not necessarily correspond to either of two perhaps appealing notions of optimal compensation: (1) the amount of purely economic damages, that is, forgone earnings plus costs of remedial treatment, and (2) the amount required to make a person whole, to restore him to his initial level of well-being, supposing that this were possible. Consider first an individual's decision to buy protection against

⁶ Compensation is taken to include the amount received from *all* sources, that is, from private insurance, social insurance, out-of-court settlements, and court awards. By rights, the definition of ideal compensation for a medical accident involving a particular individual should reflect the arrangements for insurance that all interested parties (for example, employer, dependents) would wish to make; but this consideration is ignored for simplicity's sake.

⁷ The justification for this is that (as stated above) under ideal conditions reliance on consumer preferences and on the price system results in a Pareto optimal outcome.

⁸ The discussion in this subsection is most closely related to K. Arrow, "Optimal Insurance and Generalized Deductibles," *Scandinavian Actuarial Journal* (1974), pp. 1-42 and especially, Richard Zeckhauser, "Coverage for Catastrophic Illness," *Public Policy*, vol. 21 (Spring 1973), pp. 149-72; see also P. Cook and D. Graham, "The Demand for Insurance and Protection: The Case of Irreplaceable Commodities," in *Quarterly Journal of Economics*, vol. 61, no. 1 (February 1977), pp. 143-56. Appendix 1 contains a precise version and a proof of the statements of the subsection.

a permanently disabling medical accident. Assume that the (discounted) costs of treatment, chronic care, and therapy would be \$200,000 and that of forgone earnings, another \$200,000; assume further that an additional \$100,000 would fully compensate pain and suffering. Finally, suppose that the individual has no dependents. In fact, this individual might buy, say, only \$275,000 coverage, for money might have very little use to him when disabled yet be very valuable in his current state. In this case, therefore, optimal insurance coverage would not equal the economic damages, much less make the individual whole.

Now consider an individual who is about to buy coverage against the event that his appendix will have to be surgically removed. Assume that the sum of lost wages and medical expenses would be \$2,000 and that \$1,000 more for pain and suffering would be adequate to make him as well-off as before the contemplated event. If insurance is sold on an actuarially fair basis, this individual might buy only \$2,000 coverage because, while he knows that surgery would be unpleasant and could be fully redressed by receipt of an additional \$1,000, it may be that the utility he would derive from having the extra \$1,000 after surgery is not sufficient to make his paying a higher premium worthwhile. (Equivalently, his need for money after surgery and after the \$2,000 compensation for economic losses might be no different from what it had been beforehand.) In this instance, optimal insurance coverage would equal economic damages but fall short of making the individual whole.

In general, optimal insurance coverage would fail to make an individual whole when the marginal utility of income (net of payment of the insurance premium) if he does not have an accident exceeds the marginal utility of income (at the level which would make him whole) if he does have an accident. One guesses investigation would reveal that for most individuals and most accidents, optimal insurance coverage would not be sufficient to make a person whole and would in fact provide little or no benefits beyond the purely economic; furthermore, for accidents which do not involve permanent disability, optimal insurance coverage would probably equal economic losses (see Appendix 1). Of course, in theory almost any amount of insurance coverage could turn out to be optimal.

Cause as a Determinant of Optimal Compensation. Under perfect market conditions optimal insurance coverage would not ordinarily depend on the cause of a medical accident. Consider first the case of accidents due to factors (iatrogenicity, physician negligence) beyond the control of the patient. In this case, if the cost of each dollar of

coverage per unit of probability of an accident does not depend on the cause of the accident, a rational individual would purchase the same amount of coverage for each cause. For example, suppose that a particular accident can occur in one of two ways, A and B, the latter being twice as likely as the former. Then if the premium for coverage applying in the event of B is twice that for coverage in the event of A, the amount of protection purchased for each would be identical.⁹

The example of an injured patient who wishes greater compensation because he wants to punish a negligent physician may not contradict this assertion. The patient's desire for *retribution* may not reflect an intrinsic need for additional *compensation*. His need for retribution might be satisfied by the physician's paying a fine to the state. It is only because the current negligence system is used that the compensation awarded to a patient and the monetary penalty paid by a physician (or his insurer) are the same.

However, the assumption that costs per unit of probability are independent of the cause is not always true. Suppose that a particular medical accident can be caused either by a factor which is random over the population or by an equally probable factor (an epidemic) which affects simultaneously a large group of individuals. Then the risk bearer, whether an insurance company or government, might have to charge a higher premium for insurance against the second factor. Nevertheless, in most cases considered below the premise about costs may be taken for granted.

Now consider the case of accidents in which the patient might have played a contributory role. In this case too it can be shown that optimal insurance coverage would not depend on cause, so long as the insurer could costlessly determine the actions of the patient (as would be true under perfect market conditions). This is explained in part below.

Obstacles to Compensation. Administrative costs are an obstacle to compensation which would be faced even under perfect market conditions.¹⁰ Two types of administrative costs are noted here (see Appen-

⁹ This is easily verified. Let the concave von Neumann-Morgenstern utility functions U and V give, respectively, the utility of wealth if there was no accident and if there was an accident. Let l be the monetary loss suffered in an accident, p_A and p_B be the probabilities of the accident's being caused by the (exclusive) events A and B , π the cost per dollar of coverage per unit of probability of the accident, and c_A and c_B the levels of coverage in the event of A and B . Then if y is initial wealth, c_A and c_B would be chosen to maximize expected utility $(1-p_A-p_B)U(y-\pi(p_Ac_A+p_Bc_B)) + p_A V(y-\pi(p_Ac_A+p_Bc_B)) + p_B V(y-\pi(p_Ac_A+p_Bc_B)) - l + c_A$ independent of π , p_A , and p_B .

¹⁰ On administrative costs, see D. S. Lees and R. G. Rice, "Comment," *American Economic Review*, vol. 55 (1965), pp. 140-54; V. F. Bowland, "Comment,"

dix 2). The first are fixed costs: the costs of adding an individual's name to the list of insureds and the costs of collecting premiums. It is easy to show that given fixed costs, if it is optimal to purchase any coverage at all, it will be optimal to purchase full coverage (for once one has decided to purchase positive coverage and the fixed costs therefore have to be paid, additional coverage is purchased on actuarially fair terms—in which case full coverage is appropriate). But it will be optimal to purchase coverage only if potential losses are sufficiently high or if they are sufficiently uncertain. The second type of administrative cost is that of processing claims and of making payments. Given this type of cost, the optimal insurance policy is either full coverage against all losses or a kind of deductible (the use of deductibles reduces the number of claims), according to which approximately full coverage is paid if any coverage at all is paid.

There does not seem to be any general reason to believe that the public or the private sector has a special cost advantage. Cases need to be evaluated on an individual basis, taking into account at least three factors: (1) Economies of scale obviously exist in the provision of insurance. One suspects that these may often be fully enjoyed by numerous competing firms. Otherwise, efficiency would require that only a few firms service the entire market, and regulation might be needed to force them to pass on cost reductions.¹¹ (2) "Acquisition costs," the costs of selling insurance, face both public and private insurers, though the proportion of these which may be attributed to informative selling effort may well be higher for public insurers. (3) On the other hand, there is the often-heard argument that the private sector has a greater desire and a better opportunity (especially with respect to its personnel policy) to achieve efficiency.

Several well-known obstacles to compensation which arise under less than perfect market conditions need to be mentioned. It will become obvious that these obstacles are due in one way or another to lack of information. The first obstacle, called the *moral hazard*, is the tendency of insurance to reduce the incentive of the insured to avoid losses.¹²

American Economic Review, vol. 55 (1965), pp. 1172-73; Peter Diamond, "Insurance: Theoretic Aspects of Workers' Compensation," mimeo. (Cambridge, Mass.: MIT, 1975); Arrow, "Uncertainty and Welfare"; and Appendix 2 of this paper.

¹¹ Depending on one's view, of the effectiveness of the regulators and of their ability to ascertain true costs, the regulated insurer may or may not sell coverage at rates achievable by a public insurer.

¹² See Arrow, "Uncertainty and Welfare"; Mark Pauly, "The Economics of Moral Hazard: Comment," *American Economic Review*, vol. 58 (1968), pp. 531-37; Michael Spence and Richard Zeckhauser, "Insurance, Information, and Individual Action," *American Economic Review*, vol. 61 (1971), pp. 380-87; and Steven Shavell, "On Moral Hazard and Insurance," mimeo. (Cambridge, Mass.: Harvard University, 1977).

Two types of moral hazard are of potential interest. One is exemplified by the case of a person with insurance who takes less care to avoid a loss than an uninsured counterpart. This happens because the insurer is unable to observe costlessly how careful the insured person is; otherwise the insurer would certainly make the premium or the insurance payment depend on the level of care, thereby giving the insured an incentive to be careful. Insurers have partially offset this kind of moral hazard by spending to acquire information about the level of care or by using deductible and coinsurance schemes. Both practices give the insured an incentive to be careful.

In theory, this type of moral hazard alters the previous conclusion that cause is irrelevant to compensation for a medical accident. Suppose that an accident (paralysis) may occur in either of two ways (disease, surgical error). Suppose further that if it occurs in the first way, it is possible that the injured individual may have played a contributory role (failed to seek medical attention when symptoms first appeared). Then, because of the moral hazard, deductibles or coinsurance may be desirable. (Alternatively, the insurer may choose to offset the moral hazard by acquiring information about the level of care and then conditioning the premium or insurance payment on the observation of care; in this case if the information was imperfect, partial coverage would still be desirable.) Optimal compensation, therefore, may be less if the accident occurs in the first way than if it occurs in the second way. However, in this paper we will assume that individuals have little influence over the probability of occurrence of a medical accident. This reflects the admittedly casual judgment that in the cases of greatest empirical interest, the question of a contributory role of the individual is not important.

The other type of moral hazard is exemplified by the case of a person with health insurance who (together with his physician) spends more on treatment of a given medical condition than an uninsured individual with a similar condition. This happens because the insurer is unable to observe perfectly the true medical condition; otherwise the insurer would link the premium or the payment to the actual condition. Insurers have attempted to offset this kind of moral hazard in ways similar to those mentioned above.

A second obstacle to compensation is that it may be difficult or impossible for the insurer to distinguish between individuals as to risk. In this circumstance both relatively low- and relatively high-risk individuals have to pay equally for equal amounts of coverage. If risks are pooled, then, the usual consequence is that low-risk individuals pay too much and buy too little insurance (as compared to the optimum under

perfect market conditions), while high-risk individuals pay too little and buy too much.¹³

A third obstacle to compensation is that an individual's perception of the risk he faces may be inaccurate.¹⁴ If he overestimates the risk, his purchase of coverage will be too great, while if he underestimates the risk, his coverage will be too small. (A closely related difficulty is that, because an individual may find it distressing to contemplate the possibility of a medical accident, he may buy on an "irrational" basis.)

This problem may be remedied in several ways by the public sector; the motive and scope for action by private insurers seems limited. On the one hand, individuals can be educated to the true nature of the risk, but this may be very expensive and only partially attainable. Whatever the value of such education, the private sector is unlikely to invest in it sufficiently. An individual insurance firm would find it difficult to appropriate the benefits of its investment, for a better informed customer could just as well buy from the firm's competitors. Although a collective effort on the part of private insurers to inform the public of risks (at least of higher risks) might be worthwhile, such an effort could be hampered by the problem of "free riders."

Rather than attempting to educate individuals to the truth, the public sector can provide compensation on a compulsory basis or subsidize the purchase of insurance. (The case considered here is that individuals purchase too little insurance. If they bought too much, the consequences would be less serious; in any event, in this case a tax would be appropriate.) Compulsory protection alone may fail to individualize adequately compensation; and, ideally, compensation ought to be individualized, especially with respect to differences in forgone earnings associated with differences in income earning capacity. However, there are two reasons why compulsory protection may fail to or, as a practical matter, ought not to individualize compensation. First, it may be difficult to determine a good approximation to optimal coverage on an individual basis. Second, it may be difficult to arrange to individualize payments (implicit premiums) to finance the compulsory compensation system; and if payments are not individualized, benefits should not be individualized. Court awards in malpractice cases are financed by physicians and, therefore, ultimately by patients. But the extent to which physicians usually charge patients higher fees

when the latter would have wished for greater insurance coverage is unclear.

Subsidization alone would probably help to individualize compensation but still fail to induce individuals—particularly those who were overly optimistic about risk or financially constrained—to purchase correct coverage. Therefore, a compromise between compulsory and subsidized insurance seems attractive.

Court Awards and Insurance Purchases as Indicators of Optimal Compensation. To determine optimal compensation, it is unlikely that one should look to actual court awards. These awards, made after the fact, are apparently often based on the doctrine of making a person whole, a doctrine which, as we have seen, is not justifiable by economic theory. The awards may occasionally be based on payment of economic losses alone, another faulty criterion. Moreover, particularly in cases of malpractice, there is the suspicion that awards may sometimes be inflated to punish the physician or to ensure that the plaintiff is left with his rightful share after deduction of lawyers' fees. Also, juries are sometimes swayed by emotions in deciding upon an award.

Neither is it clear that the record of actual insurance purchases would furnish the right information. As mentioned above, if individuals misperceive the risks of adverse medical events or if they find the act of insuring against these risks psychologically burdensome, they will not purchase the correct amount of coverage. In addition, if premiums exceed by a large amount actuarial costs (through lack of competition or inordinately high acquisition costs), individuals will not purchase the correct amount of coverage.

Summary. (1) Compensation (the sum of accident benefits from all sources) for a medical accident ought to reflect the amount of insurance coverage an individual (with a socially appropriate income) would have bought against the accident if he were acting rationally and were cognizant of the relevant risks.

(2) Optimal insurance coverage may exclude attributions to pain and suffering and may fail to restore economic losses (costs of treatment plus forgone earnings).

(3) Furthermore, optimal insurance coverage ordinarily depends only on the circumstances of an injured individual. Causal factors such as iatrogenicity or physician negligence are irrelevant to optimal coverage.

(4) There are four major obstacles to compensation: administrative costs, moral hazard (reduced incentive for insureds to avoid loss), pooling of unequal risks, and misperception of risks.

¹³ See, for example, Arrow, "Uncertainty and Welfare," and Michael Rothschild and Joseph Stiglitz, "Equilibrium in Competitive Insurance Markets," *Quarterly Journal of Economics*, vol. 90 (November 1976), pp. 629-50.

¹⁴ See Michael Spence, "Consumer Misperceptions, Product Failure, and Producer Liability," mimeo. (Stanford, Calif.: Stanford University, 1974).

(5) Misperception of risks can be remedied in several ways (education, compulsory or subsidized insurance) by the public sector. The motive and scope for action by private insurers in this regard seems limited.

(6) Neither court awards nor the actual record of individual insurance purchases may be good indicators of optimal compensation.

Assuring the Quality of Medical Services

Failure of Market Forces to Assure Quality. Demand and supply factors, technology, and certain nonmarket institutions of the medical sector jointly determine the quality and price of medical services. It has been argued that these nonmarket institutions are required because the operation of the price system alone would not be expected to result in a completely satisfactory outcome.¹⁵ It is true that under ideal conditions, market forces would lead to provision of medical services of the correct quality and cost (as well as to an appropriate matching of patients to physicians), for patients would demand those services which best met their needs, taking quality and cost into account. However, in fact patients are presumed to be unable to judge adequately the quality of medical services. Society has therefore resorted to nonmarket institutions to guarantee quality, namely, to educational standards, licensure requirements, and a variety of professional and legal incentives. (Such nonmarket institutions also restrict supply of physicians and therefore tend to raise wages.) Furthermore, the medical profession has developed a set of ethical norms which help to maintain public trust in the quality of medical services.

Nonmarket Influences on Competence and Care. Emphasis in this subsection is on the effects of nonmarket institutions on the competence of and the care taken by physicians; there is no discussion of the effect of nonmarket institutions on other relevant variables (such as medical technology or the number of new physicians). *Competence* refers here to the skill and knowledge of a physician, reflecting both his training and his native ability, while *care* refers to the effort and time taken in the provision of medical services by a physician of given competence.

Consider first the effects of educational standards and tests for licensure. These do not usually raise levels of care, as the term has been defined, but they do influence competence. They encourage

physicians to acquire new skills and they eliminate those who fail to meet prescribed standards. However, use of standards does not affect the competence of different physicians in the same way, for only those who consider failure a real prospect are (other things equal) most highly motivated to acquire new skills, and only those who are least able actually fail. This is not to deny the existence of other reasons for performing well nor is it to rule out the potential for graded licensing.

Nonmarket incentives—threats of financial or professional sanctions based on the performance or outcome of medical activity, along with positive incentives—influence both competence and care. Incentives clearly increase care; this is the subject of the next subsection. They also tend to improve competence in several ways: (1) physicians may acquire new skills, (2) they may decide to avoid certain medical activities altogether, and (3) they may be prevented from engaging in certain activities as a result of sanctions. Items (1) and (2) depend on the extent to which physicians actually assess their competence and act on that evaluation. Item (3) does not require such cooperation from the physician but has the drawback that it often applies only after harm has been done. Also, inasmuch as evidence of incompetence may appear rarely, sanctions may have to be imposed on the basis of incomplete information. It may be difficult to distinguish the unlucky from the incompetent. Presumably, evidence may appear rarely either because situations which truly test skills are infrequent or because it is difficult to obtain credible information.

The difference between the effects on competence and care of educational standards and tests for licensure on the one hand and of incentives on the other has relevance for the design of policy. For example, if physicians are on the whole sufficiently competent and most cases of malpractice are due to lack of care, stricter incentives might be necessary. But if under existing incentives adequate care is usually taken and most cases of malpractice are due to the incompetence of a small subgroup of physicians, then stricter licensure rather than stricter incentives might be advisable. Determination of whatever is the best policy is likely to be sensitive to the relative importance of the problems of adequacy of care and incompetence.

Incentives and Care. The discussion of nonmarket incentives and care logically belongs in the previous subsection, but because of its length, it is presented separately. It is assumed here that physicians are of equal competence.

Penalties (or rewards) under an incentive system may be made conditional on the *outcome* of a medical activity or on the degree of

¹⁵ The point of view summarized in this paragraph is elaborated in Arrow's important paper, "Uncertainty, and Welfare."

care taken during the activity, that is, on *process*. For example, a system under which surgeons had to pay damages whenever a patient died on the operating table (regardless of care taken), would be a purely outcome-oriented incentive system. A system under which surgeons' activities were under constant observation and fines imposed whenever care was insufficient (regardless of outcome) would be a purely process-oriented incentive system. A negligence system, under which inquiries were made into a surgeon's activity only after a patient had died on the operating table, and fines were imposed when care was found to have been insufficient, would be an incentive system that was both outcome and process oriented.

*Outcome versus process-oriented incentive systems.*¹⁶ One justification for outcome-oriented incentive systems (when there is no question of contributory negligence) runs as follows. Suppose that whenever there is an accident, a penalty equal to the social loss incurred is imposed on the physician. Since the physician will then weigh the costs of taking care against the social loss which he would pay in the event of an accident, he will make a socially optimal decision. (If costs of taking care are relatively high compared to the social loss, the physician will take relatively little care, and so forth.) Furthermore, all that the designer of the incentive system needs to know is the magnitude of the social loss. There is no necessity for him to know the costs the physician bears in taking care, the technical relationship between care and the probability of an accident, or the actual level of care taken in a particular incident.

This justification for an outcome-oriented incentive system overlooks the effect of risk aversion on the part of the physician. Assume for now that the physician cannot buy insurance against a penalty he may have to pay. If he is averse to risk, he will either take more care than is desirable or he will insist on extra compensation for bearing risk, charging more for his services. By contrast, an ideal incentive system would induce the physician to take the optimal level of care without imposing risk.¹⁷

A perfect process-oriented incentive system can accomplish this. Suppose that the penalty is imposed only if the actual level of care is

below the optimal level. Then physicians would take the optimal level of care, would not fear the imposition of a penalty, and therefore would bear no risk. In other words, a perfectly accurate and costlessly operated process-oriented incentive system is better than an outcome-oriented system; information about the level of care is valuable.¹⁸

Now consider the possibility that physicians can purchase insurance against penalties imposed under an outcome-oriented incentive system. If the insurer observes the level of care, the insurance policy would depend on care, in effect making the incentive system process oriented. If the insurer does not observe the level of care, then the insurance coverage purchased will be only partial. This is because the insurer will face a problem of moral hazard of the first type outlined above.¹⁹ With incomplete coverage, then, physicians would still be subject to some risk and a process-oriented incentive system would still be desirable.

Under a perfect process-oriented system, it is clear that there would be no motive to buy insurance against liability. Of course, in the usual case, when the process-oriented system makes use of inaccurate information, this statement no longer holds.

There are, however, important qualifications to be made about the case for a process-oriented incentive system. First, information may be costly to gather and evaluate, as is particularly true of information about the care exercised by a physician. Consequently, it may be advantageous to use a process-oriented system only part of the time—for instance, when there is, before acquisition of information about a case, some basis for thinking that the care taken was insufficient. Presumably, this is what is accomplished under a negligence system, since investigation of care is made only if there is a suit.

Second, information may be inaccurate. If information about care is not precise, a physician who in fact takes the optimal level of care is subjected to the risk of penalty since his observed level may appear too low.²⁰

Third, the standard of care (below which a penalty is imposed) may not be correctly chosen or consistently applied under a process-oriented incentive system. Choice of the correct standard requires knowledge not only of the loss imposed by physicians, but also of the costs physicians face in taking care and of the technical relationship between care taken and the quality of medical services. Therefore, the

¹⁸ This fact is equivalent to the statement that the (first) problem of moral hazard is eliminated when the insurance payment can be made conditional on the level of care.

¹⁹ For simplicity's sake, it is assumed that the insurer does not invest in information about care. If the insurer does, the argument is more complicated.

²⁰ See Shavell, "Moral Hazard."

¹⁶ The following is based in part on Steven Shavell, "Accidents, Liability, and Insurance," mimeo. (Cambridge, Mass.: Harvard University, 1977).

¹⁷ This would clearly be a superior situation, for either the potential injurer would be made better off by no longer having to bear risk, or else he would no longer need to be compensated for bearing the risk. Note that under the conditions most congenial to the market (in this case, perfect information about the quality of medical services), the physician is induced (by the prospect of a lower price) to take the right amount of care without having to bear risk.

opportunities for error and for inconsistent application of standards by the agents of the incentive system are probably substantial.

These last two qualifications provide different reasons why risks would be imposed on physicians, diluting the advantage of the process-oriented system and making it resemble an outcome-oriented system. If risks are imposed on physicians under a process-oriented system, the previous observation that they would not purchase insurance obviously fails to apply. The insurance coverage purchased, however, would be incomplete for the same reason that it would be under an outcome-oriented system. This does not necessarily imply that insurance against court awards would be incomplete, for there are costs other than the award itself which may not be insured.

Bias in patient selection. A problem which would arise, especially under an outcome-oriented incentive system, is a tendency toward bias in patient selection. Suppose that a physician has more detailed knowledge of the medical risks facing a patient than does the agency administering the incentives. Then the penalties imposed by the agency for adverse medical outcomes cannot be made to depend on all the circumstances which the physician recognizes as influencing risk. Consequently, a physician would (other things equal) choose to treat low-risk patients rather than high-risk ones. For example, surgeons might select patients at least partly on the basis of the risk of adverse outcomes. This criterion for patient selection is not generally desirable; for instance, high-risk patients might on average have a greater need for treatment than low-risk patients. In addition, physicians might needlessly expend medical resources in determining risks even though the level of risk would not always change the choice of treatment.

Patient-selection bias appears most likely under outcome-oriented incentive systems. Under process-oriented incentive systems, information concerning the true condition (risk) of the patient might be adduced by the agency administering the incentives; penalties could be made to depend on the condition of the patient. Knowing this, a physician would have less reason to select on the basis of risk. But the information acquired might be incomplete or inaccurate; patient-selection bias is therefore likely to remain under a process-oriented incentive system.

It must be admitted, however, that physicians could charge more to high-risk than to low-risk patients. This would not only lessen the selection bias but also could do positive good by increasing (from what may be a suboptimal level) the general recognition of patient risk.²¹ Yet it would often be difficult to arrange differential charges for medical

²¹ I owe this point to Peter F. Aronson.

treatment. First, the fact that patients usually prepay medical expenses through insurance plans would hinder the individualization of charges. And, second, the fact that patients would frequently be unable to verify easily their supposed risk category would make it hard for a physician to charge high-risk patients more.

Defensive medicine. Defensive medicine is the use of medical resources beyond the point that is justified by an evaluation of true social costs and benefits. As mentioned above, three factors can lead under a process-oriented incentive system to the imposition of risk on providers—and thus to unduly conservative action, defensive medicine. The three factors are the use of inaccurate information, the inconsistent application of standards, and difficulties in determining correct standards. The current negligence system is subject to these factors.

It also appears that the risks of penalty to physicians under the current negligence system go beyond the settlements which may be made with patients. The risks include the opportunity cost of time lost in handling claims and damage from loss of reputation. Physicians typically bear these risks. Because of the problems of moral hazard, the potential insurability of the financial risks is limited. The non-financial risks cannot be insured.

Two additional factors deserve mention. First, certain types of medical care (such as diagnostic tests or the administration of drugs) are readily observable. Realizing this, physicians may overemploy such types of care. Second, many individuals' medical expenses are covered by health insurance plans, suggesting an overuse of medical resources to begin with—the second moral hazard discussed above. In other words, the current negligence system may exacerbate an existing pattern of overuse of certain medical resources.²²

Summary. (1) Nonmarket institutions (educational standards, licensure requirements, incentive systems) are needed to assure the quality of medical services because patients are presumably ill equipped to judge quality on their own.

(2) The quality of medical services depends on the competence of physicians and on the effort or care they exercise. The choice of policy depends on the relative importance of the problems of incompetence and adequacy of care.

²² It is unclear whether one such resource is physician time per patient. For example, if under a health insurance plan a physician earns more over a given period the more patients he sees (the quality of services being constant), he might spend less time per patient than would be optimal.

(3) Educational standards and tests for licensure affect and distinguish competence but have little influence on the care taken by a physician of given competence.

(4) Incentive systems (threats of financial or professional sanctions) affect both competence and care. Competence is affected ex ante and ex post; care is affected primarily ex ante. Ex ante effects depend on physicians' recognition of their own characteristics and of the incentives. Ex post effects do not depend on this, but there are other problems with such effects.

(5) Penalties under an incentive system may be dependent on the outcome of medical activity or on the procedure or process itself. The advantage of a process-oriented incentive system over an outcome-oriented system is that the former reduces the risk borne by physicians since penalties are imposed only when there is negligence. (But the possibilities for physicians to insure themselves against the penalties of an incentive system must be taken into account.) However, the cost of employing a process-oriented incentive system suggests that it might be advantageous to use it only part of the time. In addition, inaccurate information, inconsistent application of standards, and difficulties in determining appropriate standards seriously limit the attractiveness of a process-oriented system by imposing risk on physicians.

(6) There are two additional potential problems with incentive systems. The first, patient-selection bias, arises because physicians may often be able to differentiate patients on the basis of risk of an adverse medical event and may shun high-risk patients. Patient-selection bias would be more significant under an outcome-oriented system. The second problem, defensive medicine, may be caused by the general factors which promote risk for physicians under incentive systems, but there are other special predisposing factors including (1) the physician's inability to insure against certain losses (such as damage to reputation), (2) the tendency for courts to focus attention on the use of certain medical resources (such as diagnostic tests) which may be easily observed, thus leading physicians to overemploy such resources, and (3) a pattern of overuse of certain types of medical resources initially (because of patient health insurance).

Approaches to Liability for Medical Accidents

Several approaches to liability for medical accidents are briefly considered here both as a means of compensation and as a means for assuring the quality of medical services. Of course, only if individuals

would otherwise be underinsured does it make sense to judge a system of liability on how well it serves as a means of compensation.

Patient Liability, Strict Liability, and No-Fault Plans. By definition, patient liability has no role as a means of compensation. However, in regard to assuring the quality of medical services, it should be recognized that patient liability does not preclude the use of incentives, including financial penalties, to affect physician behavior.

As a means of compensation, strict liability has a serious drawback. The drawback is that compensation is given only if the accident in question is determined to be iatrogenic, that is, caused by medical treatment but not necessarily by negligent medical treatment.²³ This determination, which would often be costly and time consuming, has little value from the point of view of achieving optimal compensation, for it has been argued that the latter does not depend on the cause of an accident. As a means of assuring the quality of medical services, strict liability has the general disadvantages of any outcome-oriented incentive system, namely patient-selection bias, the imposition of risk on physicians, and defensive medicine.

A no-fault plan is probably best viewed as a system of strict liability. Under such a plan, physicians would contribute to a fund from which individuals would be compensated for a predetermined class of iatrogenic injuries. Thus, as with strict liability, compensation is not given for all accidents and there might be substantial operating costs owing to the possibility of having to decide a large number of cases in which it would be difficult to determine iatrogenicity. Proposed no-fault plans differ somewhat from strict liability with regard to assuring the quality of medical services. Under one no-fault plan, that of Senators Inouye and Kennedy, a mechanism would be established for collecting information about medical accidents. This information would supposedly be used to influence physician behavior, but few details are given.²⁴ Under another no-fault plan called Medical Adversity Insurance, the use of coinsurance and especially experience-rating would be emphasized.²⁵

²³ Should the death of the patient during surgery be attributed to medical treatment itself or to a prior medical condition? See, more generally, Robert Keeton, "Compensation for Medical Accidents," *University of Pennsylvania Law Review*, vol. 121 (1973), pp. 590-617 and Richard Epstein, "Medical Malpractice: The Case for Contract," *American Bar Foundation Research Journal* (1976), pp. 87-149.

²⁴ S. 215, 94th Congress (1975) and H.R. 4881, 94th Congress.

²⁵ See Clark Havighurst, "Medical Adversity Insurance—Has Its Time Come?" *Duke Law Journal* (1975), pp. 1233-80. The usefulness of experience-rating appears to be limited. Experience-rating has two effects. (1) It supplies an incentive to take care. In this respect the advantage it has over coinsurance is not

The Negligence System. As a means of compensation, the negligence system suffers from much the same drawback as strict liability. Under the negligence system compensation is not given to all injured individuals, but only to those who establish (or make a credible threat to establish) physician negligence; and the substantial costs of attempting to establish negligence are of dubious benefit in achieving optimal compensation.

As a means of assuring the quality of medical services, the current negligence system has two major disadvantages. First, it appears to be subject to the general factors mentioned above associated with process-oriented incentive systems which impose risks on physicians and which therefore contribute to the problems (including patient-selection bias and defensive medicine) associated with outcome-oriented incentive systems. Popular wisdom regards the risks faced by physicians as important, and the strength of demand for malpractice insurance confirms the belief. Second, medically irrelevant characteristics of the physician and the socioeconomic status and personality type of the patient are apparently important indicators of the likelihood of a malpractice claim.

Modifications of the negligence system. Many suggestions for mitigating the malpractice problem would modify the current negligence system. Most of these suggestions aim to reduce the risks and insurance costs borne by physicians.²⁶ Before we discuss several of the suggestions it may be appropriate to make a few remarks about rates, claims, and settlements. Rates are based on both experienced costs and projected costs. Projected costs are of particular importance since claims on present occurrences may easily be made as many as five or more years into the future. Experienced costs are determined by the volume of claims and by the expense per claim, comprising legal costs and settlements. Legal costs and the size distribution of settlements influence the motive to make claims and therefore the volume of claims.

(1) **Ceiling on malpractice awards:** This would reduce malpractice insurance rates in three ways: by lowering the size of settlements, by reducing the motive to make claims and therefore the number of claims, and by making prediction of costs easier and consequently by permitting insurers to reduce allowances for risk which they would otherwise build

into rates. A ceiling on malpractice awards is, however, undesirable from the point of view of providing compensation since it would deny coverage to those in greatest need. A better approach would be to place a floor on awards and to base the size of awards in a predetermined way on the circumstances of injured individuals.

(2) **Prohibition of contingent fees for lawyers:**²⁷ The effect of prohibiting the use of contingent fees on the number of claims is unclear. Such a prohibition would shift the direct financial risk of losing a case to individuals. If lawyers are better able than individual patients to bear this risk, the typical patient would be less inclined to make a claim if his legal fees were not contingent, and hence the number of claims would fall. Present use of contingent fees suggests that lawyers are in fact the better risk bearers, but there are other (not necessarily competing) explanations. The prohibition would also diminish the incentive of lawyers to expend effort to win cases, further reducing the inclination of individuals to make claims. Because lawyers would not need to win their cases to collect fees, however, they would be more willing to accept tenuous claims, implying an increase in the number of claims.²⁸ The net effect on the number of claims is therefore ambiguous. Other things equal, the average size of claims would decline, since lawyers would have less reason to press large claims over small ones. However, as the total value of claims is determined by the total number and average size of claims, it does not seem possible to conclude that the total value of claims would fall.

The desirability of prohibiting contingent fees is likewise unclear. Consider first the issue, compensation. After prohibition, compensation net of lawyers' fees would be lower for those getting small amounts and higher for those getting large amounts than with contingent fees. While on average this change might be considered beneficial (those bearing greater losses being helped at the expense of those with lesser), the indeterminacy of the change in the total value of claims makes it impossible to conclude whether prohibition would make better or worse a problem of underinsurance against medical accidents.

Consider now the issue of assuring the quality of medical services. In this regard, the question is whether prohibiting contingent fees would lead to more rational use of the negligence system.²⁹ Ideally, the negli-

²⁷ Under a contingent fee plan, a lawyer is paid only if his client actually collects damages. See Rader's paper in this volume.

²⁸ It is fair to ask why tenuous claims might not be accepted today on a non-contingent fee basis. The answer is probably that the refusal of a lawyer to accept a case today on a contingent fee basis is a strong signal to the potential client that his claim is not a good one.

²⁹ I plan to investigate the question in subsequent work.

clear. (2) Experience-rating also discriminates against physicians with bad records. Because these physicians may be unlucky rather than incompetent and there is no reason to penalize the unlucky, the optimal structure of experience-rating would strongly depend on the distribution of competence among physicians. The dangers of miscalculation would therefore be significant.

²⁶ Suggestions concerning experience-rating, arbitration, and narrowing of defendant liability are not discussed here; see for example Epstein, "Medical Malpractice."

gence system ought to be employed only if the social costs of making an additional claim are less than or equal to the social benefits. While the social costs include the costs of running the courts and the legal expenses of both plaintiff and defendant, in the United States the private cost to the plaintiff usually includes only that of his own counsel. The social benefits of making another claim reside in the claim's effect in assuring quality. The relationship between the social benefits and the private benefits (the expected award or out-of-court settlement) is not clear. If it turns out, as I would suppose, that there are in fact too many claims, and if prohibition would reduce the number of claims, prohibition would have to be considered as beneficial in assuring the quality of services.

(3) Subsidization of small claims: Subsidization of small claims would no doubt increase the total number of claims. The objective of subsidization appears to be the provision of compensation through the medium of courts to a class of individuals who might otherwise be underinsured. This is an expensive way for society to buy insurance. As to assuring the quality of medical services, subsidization appears to be undesirable, in view of the considerations discussed in the previous paragraph.

(4) Use of "claims made" versus "occurrence" insurance.³⁰ The St. Paul Fire and Marine Insurance Company has introduced a "claims made" medical malpractice insurance policy. Suppose that in year one physician treats a patient who decides in year two to sue for malpractice because he discovers evidence of malpractice only in year two. Under an "occurrence" policy, the usual kind of medical malpractice insurance policy, the physician would have purchased at the beginning of year one coverage against all claims—whenever they might be made—on accidents occurring during year one. Thus, his insurance would apply to the suit in question even though it was brought in year two. Under claims-made policy, however, the physician's coverage purchased at the beginning of year one would have protected him only against claims made that year. But he would also have had the option of purchasing the beginning of year two coverage protecting him against claims made during that year for occurrences in year one. The rate he would pay year two would depend on the insurer's projection of costs at the beginning of that year. Thus, with claims made, the insurer would

³⁰ Although this is a modification of the terms of insurance rather than a modification of the negligence system, it seems appropriate to mention it here. See St. Paul Fire and Marine Insurance Company, "Preserving a Medical Malpractice Insurance Marketplace: Problems and Remedies," January 1975.

not have to bear the risk of unexpected changes in average claim costs for periods greater than one year; this risk would be borne by physicians. On the other hand, physicians would still be insured against the risk they fear most, that of having to pay a large award. If insurers are particularly averse to fluctuations in malpractice underwriting profits, then claims-made policies appear to offer a better means of risk sharing than occurrence policies.³¹

Concluding Remarks

The discussion of this paper has suggested that it would be undesirable to place special reliance on the courts for compensation of victims of medical accidents. This conclusion is based on the argument that compensation should ideally depend only on the circumstances of victims and the nature of their injuries, not on physician negligence or iatrogenicity. But whether compensation is given to victims by the courts does depend on the latter factors. Moreover, it was claimed that when compensation is given by the courts, the principles used in determining the size of awards are not generally appropriate. Additionally, inquiries into medical accidents by the courts are typically much more costly than would be the mere provision of insurance benefits to victims.

With regard to the goal of assuring the quality of medical services, this paper distinguished problems related to competence of physicians from problems related to adequacy of care taken by physicians. It was pointed out that educational standards and tests for licensure affect primarily the competence of physicians, while the risk of legal and professional sanctions following medical accidents provides incentives which affect not only competence but also adequacy of care. These incentives were discussed from a general theoretical perspective, and alternative approaches to liability for medical accidents were compared. However, the relative lack of information about the problem of assuring the quality of medical services made it difficult to conclude that one approach was better than another.

³¹ The St. Paul company points out that an advantage of such policies would relieve young physicians of having to pay large premiums when they were starting their careers. Epstein in "Medical Malpractice" suggests that retiring physicians would be hurt by such policies since they would have to pay premiums after the end of their careers. This overlooks the fact that such premiums could be financed out of savings from premiums under occurrence policies which would have been paid before retirement. On the other hand, the consequences of unexpectedly early retirement would be more serious, imposing extra risk on physicians. It should be mentioned that there are apparent accounting advantages to claims-made policies; insurers' required reserves are less under claims-made than under occurrence policies.

Appendix A: Optimal Insurance Coverage, Economic Losses, and Making a Person Whole

The discussion of the relationship among optimal insurance coverage, economic losses, and making a person whole (fully compensating pain and suffering as well as economic losses) makes reference to a simple example of an insurance decision; the example is similar to the one in an article by P. Cook and D. Graham.³² Let p be the (fixed) probability of an accident, l the associated economic (monetary) loss, $U(y)$ the von Neumann-Morgenstern utility of wealth y if there is no accident, and $V(y)$ utility of wealth if there is an accident. Assume that in either state the individual is risk averse (U, V are concave), that the accident itself is not desirable (so $U(y) > V(y)$), and that insurance coverage is sold on an actuarially fair basis. Then the individual would select a level of coverage c to maximize his expected utility

$$(1-p)U(y-pc) + pV(y-pc-l+c). \quad (1)$$

(The premium for coverage c is pc .) The first order condition (which in this case determines the optimal coverage c^*) of course equates marginal utilities of wealth in the two states,

$$U'(y-pc^*) = V'(y-pc^*-l+c^*). \quad (2)$$

Let $\tilde{c}$ be the amount of coverage such that the person would be made whole (assuming $\tilde{c}$ exists), that is

$$U(y-p\tilde{c}) = V(y-p\tilde{c}-l+\tilde{c}). \quad (3)$$

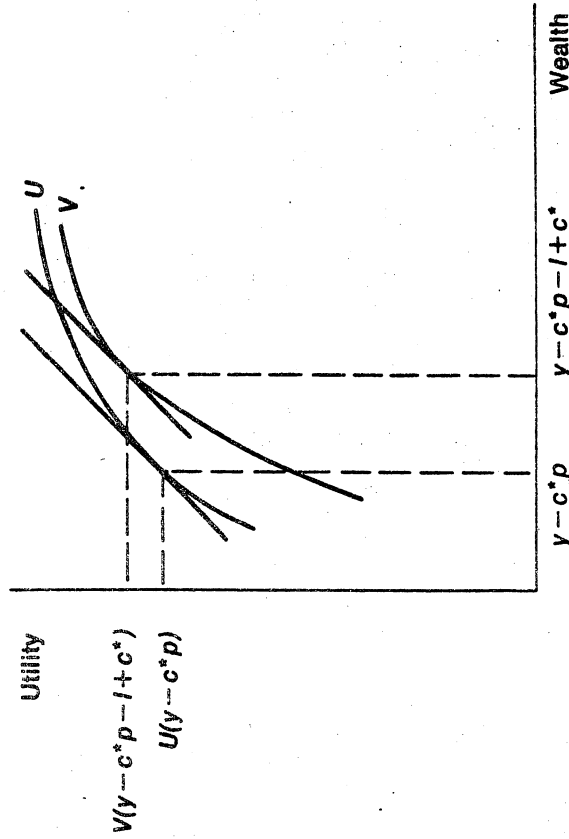
Note that $\tilde{c} > l$ since $U(y) > V(y)$ for any y . (The amount $\tilde{c} - l$ compensates pain and suffering.) Then it follows (from (2) and the concavity of U and V) that

$$c^* < \tilde{c} \text{ if and only if } U'(y-p\tilde{c}) > V'(y-p\tilde{c}-l+\tilde{c}). \quad (4)$$

Moreover, if $U'(y) > V'(y)$ for all y (marginal utility of wealth is lower after an accident at identical levels of wealth), (2) shows that $c^* < l$, optimal coverage is below economic losses and certainly the individual is not made whole. If $U'(y) = V'(y)$ for all y , $c^* = l$, coverage equals economic losses and again the individual is not made whole. If $U'(y) < V'(y)$ for all y , $c^* > l$, coverage exceeds economic loss—there is compensation for pain and suffering—and the person may or may not be made whole (it is possible that he would actually be better off after the accident and compensation). This is illustrated in Figure 1 and Figure 2. Obviously, cases in which U' and V' are not so simply related are also possible.

³² P. Cook and D. Graham, "The Demand for Insurance and Protection: The Case of Irreplaceable Commodities," *Quarterly Journal of Economics*, vol. 61, no. 1 (February 1977) p. 143-56.

Figure 1
INDIVIDUAL (MORE THAN) MADE WHOLE

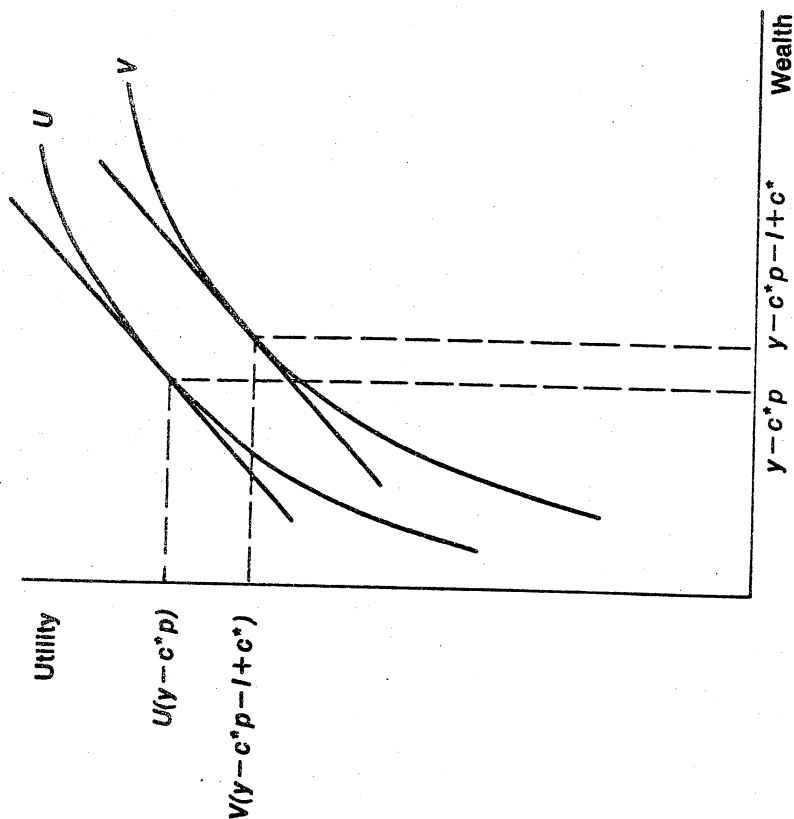


If an accident does not result in a permanent disability—would not result in any alteration of life style (given the same level of income)—the natural assumption would seem to be $U'(y) = V'(y)$; coverage should therefore equal economic losses and nothing should be given for pain and suffering. Only if the accident does result in lasting changes would one expect $U'(y)$ and $V'(y)$ to differ. I suspect that if the disability is not severe, so that income can still be enjoyed, it *might* be that $V'(y) > U'(y)$, coverage would exceed economic damages but would not make the individual whole in the cases which seem plausible to me. If the disability is severe (for example, the well-known case of Karen Quinlan, who is in an essentially vegetative state), $V'(y) < U'(y)$ would seem typical and coverage would fall short of even economic losses.

Appendix B: Administrative Costs of Insurance and Optimal Insurance Policies

Two types of administrative costs are considered: (1) fixed costs of opening a policy and of collecting premiums and (2) claims-processing costs. The problem is to describe optimal insurance policies given these

Figure 2
INDIVIDUAL NOT MADE WHOLE



administrative costs. The relationship between this problem and a similar one discussed by K. J. Arrow³³ is discussed below. Denote by

- | | | |
|--------------|-----------------|-------------------------------------|
| $x_i > 0$ | $i=1, \dots, n$ | possible losses |
| $p_i > 0$ | $i=1, \dots, n$ | probability of loss of x_i |
| $p_0 > 0$ | | probability of no loss |
| $q_i \geq 0$ | $i=1, \dots, n$ | payment by insurer if loss is x_i |
| π | | insurance premium |
| $a \geq 0$ | | fixed costs of opening a policy |
| $b \geq 0$ | | claims-processing costs |
| y | | initial wealth |
| $U(\cdot)$ | | utility of wealth |

³³ Arrow, "Uncertainty and Welfare."

It is assumed that individuals act so as to maximize the expected utility of final wealth, that they always value increases in wealth ($U' > 0$), and that they are risk averse ($U'' < 0$). It is also assumed that $U(\cdot)$ is defined only over nonnegative levels of wealth (and therefore that $x_i \leq y, i=1, \dots, n$).

An insurance policy is then specified by (π, q) , with $q = (q_1, \dots, q_n)$. It is assumed that the premium equals the insurer's expected costs, comprising both expected direct payments to the insured and expected administrative costs. Therefore, if $I = \{i | q_i > 0\}$, the index set of losses for which positive insurance has been arranged, then

$$\begin{aligned} \pi &= a + \sum_{i \in I} p_i (q_i + b) & \text{if } I \neq \phi \\ &= 0 & \text{if } I = \phi \end{aligned} \quad (1)$$

$I = \phi$ means no losses are positively insured—no policy is purchased. Note that the claims-processing cost b is assumed to be independent of the size of a claim.³⁴ The expected utility associated with a policy (π, q) is

³⁴ The importance of this assumption is that claims-processing costs do not go to zero with the size of the claim.

$$p_0 U(y - \pi) + \sum_{i=1}^n p_i U(y - \pi - x_i + q_i). \quad (2)$$

The policy which will be purchased—the *optimal insurance policy*—maximizes (2) over (π, q) subject to (1) and $q_i \geq 0, i=1, \dots, n$.

PROPOSITION 1: If the only administrative costs are the fixed costs of opening a policy ($a > 0, b=0$) and it is optimal to purchase positive coverage against some loss, the optimal insurance policy will in fact give full coverage against all losses.

The reason for this is clear. Once the decision has been made to buy a policy, fixed costs have to be paid. Then coverage can be bought on the margin on actuarially fair terms, in which case full coverage is optimal.

Proof: Since $I \neq \phi$ and $b=0$, (1) is just

$$\pi = a + \sum_{i=1}^n p_i q_i \quad (3)$$

and the appropriate Lagrangean to consider is³⁵

$$L(\pi, q) = p_0 U(y - \pi) + \sum_{i=1}^n p_i U(y - \pi - x_i + q_i) - \lambda (a + \sum_{i=1}^n p_i q_i - \pi). \quad (4)$$

³⁵ In this case and those considered below, the constraint qualification (guaranteeing that the solution must be a critical point of the Lagrangean) obviously holds. In this case we have not explicitly incorporated the constraints that the q_i are nonnegative since it turns out that without doing so the q_i obey the constraints at the optimum.

At the optimum

$$0 = \frac{\partial L}{\partial q_j} = p_j U'(y - \pi - x_j + q_j) - \lambda p_j \quad (5)$$

or

$$U'(y - \pi - x_j + q_j) = \lambda. \quad (6)$$

Also

$$0 = \frac{\partial L}{\partial \pi} = -p_0 U'(y - \pi) - \sum_{i=1}^n p_i U'(y - \pi - x_i + q_i) + \lambda \quad (7)$$

so (using (6) and $p_0 = 1 - \sum_{i=1}^n p_i$),

$$U'(y - \pi) = \lambda. \quad (8)$$

Since $U'' < 0$, $y - \pi = y - \pi - x_j + q_j$ therefore $q_j = x_j$, which completes the proof.

The advantage in terms of expected utility of insuring is therefore

$$U(y - a - \sum_{i=1}^n p_i x_i) - [p_0 U(y) + \sum_{i=1}^n p_i U(y - x_i)]. \quad (9)$$

If a declines, (9) increases and is positive for all a sufficiently small. If x_i increases (to a maximum of y , by assumption), (9) increases but need not become positive. If p_0 is sufficiently close to 1, (9) is negative. However, as p_0 approaches zero, (9) is positive unless a is sufficiently large or $n=1$. Summing up, it will be optimal to purchase (full) coverage if administrative costs are sufficiently low. High losses make insuring more advantageous. If the probability of loss is sufficiently close to zero, no coverage will be bought, but if the probability of loss is close to (or equals) one, it may still be optimal to purchase (full) coverage.

PROPOSITION 2. Suppose that there are claims-processing costs and, possibly, fixed costs of opening a policy ($a \geq 0$, $b > 0$). Then if it is optimal to purchase positive coverage against some loss, the optimal insurance policy will either provide (i) full coverage against all losses or (ii) no coverage against some losses, in which case the policy will be a kind of deductible.

Note: Both (i) and (ii) are possible whether or not $a=0$. In (ii) the deductible policy referred to is such that if $q_i > 0$, then $q_i = x_i - d$ where $d > 0$. However, it is suggested that in an important case d is "close" to zero, that is, the optimal policy is approximated by a policy which gives full coverage against losses whenever the optimal policy gave positive coverage.

The explanation for this proposition is similar to that for the first. If it is optimal to positively insure against all of the x_i , then expected

claims-processing costs (and any fixed costs) may be regarded as fixed costs. In this case the explanation for Proposition 1 applies and the optimal policy fully insures all losses. On the other hand, suppose that it is optimal to positively insure only some losses. Then optimality requires that for these losses, post-insurance-payment marginal utilities of wealth must be constant, that is $x_i - q_i = d$ for $i \in I$. Why the deductible d is in fact positive is explained below; it is not because, on the margin, coverage is purchased at actuarially unfair rates for, on the margin, coverage is purchased at fair rates.

Two factors are shown below to militate against positively insuring a loss: low magnitude of loss and high probability of loss.³⁶ It is not worth insuring small losses, for as loss size goes to zero, the implicit "loading" (due to claims-processing costs) grows arbitrarily large. It is not worth insuring extremely likely losses, for if one is relatively confident of suffering a loss, he may as well shoulder the risk himself and avoid the claims-processing costs. However, as a practical matter, the first factor is probably the more important.

Proof: Suppose first that it is optimal to insure positively against each loss, so $I = \{1, \dots, n\}$. Then (1) becomes

$$\pi = a + b \sum_{i=1}^n p_i + \sum_{i=1}^n p_i q_i \quad (10)$$

which is of the same form as (3). Hence, by the proof to Proposition 1, $q_i = x_i$, $i=1, \dots, n$.

On the other hand, suppose that it is optimal to insure positively only some losses, so $\phi \neq I \neq \{1, \dots, n\}$. Then the appropriate Lagrangean may be written as³⁷

$$\begin{aligned} L(\pi, q) = & p_0 U(y - \pi) + \sum_{i \notin I} p_i U(y - \pi - x_i) + \sum_{i \in I} p_i U(y - \pi - x_i + q_i) \\ & - \lambda(a + \sum_{i \in I} p_i(q_i + b) - \pi). \end{aligned} \quad (11)$$

At the optimum, for $j \in I$

$$0 = \frac{\partial L}{\partial q_j} = p_j U'(y - \pi - x_j + q_j) - \lambda p_j \quad (12)$$

or

$$U'(y - \pi - x_j + q_j) = \lambda. \quad (13)$$

Hence (as $U'' < 0$), $x_j - q_j$ is equal to a constant, say d , for $j \in I$. We wish to show $d > 0$. At the optimum

³⁶ The second factor is noted in Diamond, "Insurance Theoretic Aspects."

³⁷ The constraints q_i nonnegative, $i \in I$, have not been introduced for the reason cited in footnote 39.

$$0 = \frac{\partial L}{\partial \pi} = -p_0 U'(y-\pi) - \sum_{i \neq l} p_i U'(y-\pi-x_i) \quad (14)$$

$$- \sum_{i \neq l} p_i U'(y-\pi-x_i+q_i) + \lambda.$$

Substituting (13) and solving for λ ,

$$\lambda = (1 - \sum_{i \neq l} p_i)^{-1} [p_0 U'(y-\pi) + \sum_{i \neq l} p_i U'(y-\pi-x_i)] \quad (15)$$

$$> (1 - \sum_{i \neq l} p_i)^{-1} [p_0 U'(y-\pi) + \sum_{i \neq l} p_i U'(y-\pi)] = U'(y-\pi).$$

Thus $d > 0$, which completes the proof.

The first line of (15) says that at the optimum the marginal utility of wealth after insured losses ($i \neq l$) is a (conditional probability) weighted average of the marginal utilities of wealth after uninsured losses or no loss. This indicates why the deductible d is positive if it is optimal not to insure at all against certain x_i (a zero deductible would make the marginal utility of wealth after insured losses equal to that if there were no loss, and thus less than the average of marginal utilities after uninsured losses or no loss).

However, in one important case the optimal policy might be closely approximated by fully insuring any loss which is positively insured, that is, by setting $q_i = x_i$ for $i \neq l$ (even through $l \neq \{1, \dots, n\}$). This case is that in which the probability of no loss is large compared to the probability of uninsured losses, for then, by (15), λ is approximately $U'(y-\pi)$. In any event, by (15) λ is always less than $\max_{i \neq l} U'(y-\pi-x_i)$, implying that the deductible d is less than the largest uninsured loss. (This would unfortunately increase the incentive to overstate losses.)

Now briefly consider the two factors mentioned above which may induce an individual to leave a loss x_i uninsured. First, if an x_i is sufficiently low, the optimal coverage $q_i = 0$: Suppose that $q_i > 0$ is optimal. Then the premium may be written as $\pi = z + p_j(b+q_i)$. Then expected utility may be written

$$p_0 U(y-\pi) + p_l U(y-\pi-x_l+q_l) + \sum_{i \neq j} U(y-\pi-x_i+q_i) \quad (16)$$

$$= p_0 [U(y-z) - p_j(b+q_i) U'(y-z) + e(y-z, -p_j(b+q_i))] \\ + p_l [U(y-z) - (p_j(b+q_i) - x_l + q_l) U'(y-z) \\ + e(y-z, -(p_j(b+q_i) - x_l + q_l))] + \sum_{i \neq j} U(y-\pi-x_i+q_i),$$

where $e(s, t) = U(s+t) - (U(s) + tU'(s))$. Now write expected utility for the policy modified only in that $q_i = 0$ (and the premium is accordingly reduced to z):

$$p_0 U(y-z) + p_l U(y-z-x_l) + \sum_{i \neq j} p_i U(y-z-x_i+q_i) \quad (17)$$

$$> p_0 U(y-z) + p_l U(y-z-x_l) + \sum_{i \neq j} p_i U(y-\pi-x_i+q_i) \\ = p_0 U(y-z) + p_l [U(y-z) - x_l U'(y-z) + e(y-z, -x_l)] \\ + \sum_{i \neq j} p_i U(y-\pi-x_i+q_i).$$

To establish the claim, it is shown that if x_i is sufficiently low, the expression written after the equality sign in (17) exceeds the expression written after the equality sign in (16). Subtracting the latter from the former,

$$p_l [e(y-z, -x_l) + (p_j(b+q_i) + q_l) U'(y-z)] \quad (18)$$

$$- e(y-z, -p_j(b+q_i) - x_l + q_l)] \\ + p_0 [p_j(b+q_i) U'(y-z) - e(y-z, -p_j(b+q_i))].$$

As $U'' < 0$, $e(s, t) < 0$ if $t \neq 0$. By Proposition 2, $-x_l + q_l = -d < 0$. Hence the only negative term in (18) is $p_j e(y-z, -x_l)$ whereas the positive terms are certainly bounded from below by $p_j^2 b U'(y)$. Since $e(y-z, -x_l) \rightarrow 0$ as $x_l \rightarrow 0$, (18) is positive for x_l sufficiently low.

Second, if a loss probability p_i is sufficiently high, the optimal coverage $q_i = 0$: Suppose that $q_i > 0$ is optimal. Then, as in the previous paragraph, write $\pi = z + p_j(b+q_i)$ and expected utility of the optimal policy as

$$p_0 U(y-z-p_j(b+q_i)) + p_l U(y-z-p_j(b+q_i) - x_l+q_l) \quad (19)$$

$$+ \sum_{i \neq j} p_i U(y-\pi-x_i+q_i) \\ < p_0 U(y-z-p_j(b+q_i)) + p_l U(y-z-p_j(b+q_i) - x_l+q_i) \\ + \sum_{i \neq j} p_i U(y-z-x_i+q_i).$$

The expected utility for a modified policy with $q_i = 0$ and premium of only z is

$$p_0 U(y-z) + p_l U(y-z-x_l) + \sum_{i \neq j} p_i U(y-z-x_i+q_i). \quad (20)$$

It needs to be shown that if p_i is sufficiently high, (20) exceeds the expression following the inequality sign in (19). Subtracting the latter from the former,

$$p_0 [U(y-z) - U(y-z-p_j(b+q_i))] + p_l [U(y-z-x_l) \\ - U(y-z-p_j(b+q_i) - x_l+q_l)].$$

When $p_j = 1$ (in which case $p_0 = 0$), (21) equals $U(y-z-x_l) - U(y-z-b-x_l) > 0$, so that a continuity argument implies that (21) is positive for p_j sufficiently close to 1.²⁸

²⁸ The variable z is implicitly determined by the probabilities and is not necessarily

The discussion of the problem of the administrative costs of insurance as presented here should be compared with Arrow's well-known result about deductibles. This result is that the optimal insurance policy must be the usual kind of deductible if the premium paid is based only on the expected payments which are made to the insured under the terms of the policy. If the administrative costs are just the fixed costs of taking out a policy, then Arrow's assumption holds (if $\bar{q}$ is expected payments, then $\pi(\bar{q}) = a + \bar{q}$), but Proposition 1 does not contradict his result, since he did not rule out a deductible of zero—full coverage. On the other hand, if there are claims-processing costs, the premium is not a function merely of expected payments to the insured. It also depends on the probability that positive payment will be made. Therefore, Proposition 2 does not contradict Arrow's result either.

continuous in p_i and p_j . However, as $z = \pi - p_i(q_i + b)$ where π is assumed to be the optimal premium, z is certainly bounded from above by $(1 - p_i) \max_{1 \leq j \leq J} x_j$ (by Proposition 2, at the optimum $q_i \leq x_i$ or $\pi \leq \sum_{j=1}^J p_j x_j$) and from below by zero. Thus $z \rightarrow 0$ as $p_i \rightarrow 1$, so that t_i is clear that (21) is positive for p_i close to 1.

MEDICAL MALPRACTICE AND MEDICAL COSTS

Bruce C. N. Greenwald and Marnie W. Mueller

Do doctors and hospitals pass on the increased cost of their malpractice insurance? Do doctors move away from states where they think they are likely to be sued for malpractice? Are millions of dollars of scarce medical resources being used by doctors and hospitals for defensive medicine? Finally, have the suspicions generated by increased malpractice claims destroyed the therapeutically beneficial trust between doctors and their patients? Recent requests for legislative relief from the rising costs of malpractice suits and of insurance premiums have been justified by belief that the answer to each of these questions is affirmative. The purpose of this paper is to bring the weight of empirical evidence from cross-sectional data by state from 1970 to bear on these questions and to determine the likely magnitude of these effects.

The first section of the paper examines the effect of malpractice suits against doctors on doctors' services, while the second section examines the effect of malpractice suits against doctors and hospitals on hospital services. These sections are organized in parallel fashion. Each begins by constructing measures of the complexity of medical procedures appropriate to the area in question. These measures form the basis for analyzing the extent of the practice of defensive medicine and provide evidence of the influence of insurance coverage and income on medical procedures and through them on costs. Similar models of the market for doctors' services and of the market for hospital services are developed next. These models are then used to determine the final incidence of increases in malpractice insurance premiums through their direct impact on doctors' fees and hospital prices, on the demand for doctors' services and for hospital services, and on the location decisions of doctors. The third section concludes the paper with estimates of the additions to medical costs caused by the phenomenon of malpractice suits.

Malpractice Suits and Doctors' Services

Doctors who believe themselves threatened by the prospect of a malpractice suit will take defensive actions. Most obviously, they will buy