Optimal Health Insurance and Provider Payment

By CHING-TO ALBERT MA AND THOMAS G. MCGUIRE*

We derive optimal insurance for patients and payment method for physicians when neither the input decided by the patient (quantity of treatment) nor the input decided by the physician (effort) are contractible. The equilibrium in this third-best regime may sometimes be second best, in which both the physician input and the report of treatment are verifiable. Otherwise, truthful reporting forces a third best, characterized by provider "prospective payment" and suboptimal effort, while consumers' demand becomes excessive. We also analyze how "professional ethics" alters the equilibrium. Finally, collusive reporting mechanisms imply more stringent constraints, while competition among physicians relaxes them. (*JEL* I10, B80)

Health care markets are changing rapidly as doctors and hospitals integrate with insurance companies, either through mergers or by complex contracts, to sell health services to consumers (Stephen M. Shortell et al., 1994). These "organized delivery systems" tie together insurers, providers, and consumers with elaborate contracts. Simple "fee-for-service" contracts, in which providers are reimbursed for their full costs of providing treatment while consumers are free to choose their providers, are becoming rare. A typical "health plan," as distinguished from an ordinary insurance policy, imposes significant risk of health care costs for a population on doctors and hospitals, as well as restricts consumers' choices of providers.[1] From the point of view of the consumer, choice of a health plan has become a choice of insurance coverage *together* with a set of providers paid according to certain terms.

Economic theory often interprets the emergence of social institutions as a response to market failures. Indeed, economists have recognized for a long time that a significant market is missing in the health sector: insurance policies based on health outcomes. Kenneth J. Arrow (1963) observed that an efficient (first-best) health insurance policy would specify payment contingent on the individual's state of health. For example, an individual who suffered a sudden health problem would be paid a specified amount by the insurance company; afterwards, the individual could make his own decision to purchase health care. A state-contingent payment scheme protects the individual from the financial risk of illness *ex ante* and retains incentives for the patient to consume health care efficiently *ex post*. Nevertheless, insurance

* Department of Economics, Boston University, 270 Bay State Road, Boston, MA 02215. An earlier version of this paper was circulated under the title "Noncontractible Inputs and Health Insurance-Payment Systems." Research support has been received from Grant K05-MH00832 from the National Institute of Mental Health. We thank Richard Arnott, Randall Ellis, Richard Frank, Esther Gal-Or, Martin Gaynor, Michael Manove, Joseph Newhouse, Michael Riordan, David Salkever, Richard Zeckhauser, and numerous seminar participants at various universities and conferences for their helpful comments. Advice by two referees is gratefully acknowledged. The authors alone are responsible for the analysis and conclusions.

[1] One recent survey reports that nonfee-for-service contracts make up 16 percent of total income of office-based physicians, a percentage projected to grow annually at 40 percent for the next several years (Michael Quint, 1995). According to the Physician Payment Review Commission (1995), about 17 percent of the U.S. population was enrolled in health maintenance organizations (HMOs) in 1994, a number that is projected to grow to one-third by the year 2000 (Office of Technology Assessment, 1994). In 1994, 65 percent of insured workers in firms having 200 or more employees had some restrictions on their choice of physician (Physician Payment Review Commission, 1995 p. 185). In the leading-edge state of California, nearly 90 percent of public employees are enrolled in HMOs.

policies contingent on health status are nonexistent because health status is too costly to verify. As a result, the market for health-status-based insurance is missing.[2]

The lack of a health-status-based insurance market further rules out any policy that commits the patient to certain bundles of treatment contingent upon health status. The economics literature emanating from Arrow's observation derives the optimal policy under the assumption that insurance coverage may only be based on the patient's choice of treatment quantity. Mark V. Pauly (1968) and Richard Zeckhauser (1970) were among the first researchers to characterize the trade-off between risk sharing and moral hazard: the optimal copayment that the patient must pay for each unit of treatment exposes him to some risk *ex ante,* but also partly remedies his incentives to consume an excessive amount of health care *ex post.* Indeed, the basic premise from this literature—that health care services with a greater demand response to insurance should require higher patient copayment—has laid a foundation for an economic policy toward health insurance and has spawned a great deal of empirical research (see Joseph P. Newhouse and the Insurance Experiment Group, 1993).

While this literature certainly has generated deep insight, its single focus on the insurer-patient relationship cannot account for the elaborate contracts among insurers, patients, and providers that are so common today. Contracts between payers and physicians now include many new features. Some of these are clearly designed to deal with moral hazard. Thus, "capitation contracts" give a physician or a group of physicians an annual per-person payment, and make physicians responsible for all physician costs and sometimes costs of hospitalization as well (Joan B. Trauner and Julie S. Chestnut, 1996). Other contract features, however, are clearly not about controlling costs. Harold S. Luft (1996), for example, reports that in some physician contracts up to 30 percent of income could be based on bonuses related to measures of "quality" of care.[3] The potential conflicts in using physician payment and other contractual relations to pursue both the goal of limiting cost and satisfying consumers have been recognized in the medical sociology literature since the dawn of prepaid health care employing salaried physicians (Eliot Friedson, 1993). To date there is no economic theory coming to grips with the changes in physician contracting taking place.

In our view, a theory to interpret these complex contractual arrangements must recognize that there is more going on than that captured in the moral-hazard-versus-risk paradigm. This observation leads to our central point: a model for the health market must consider the interaction among insurers, physicians, and patients, to derive simultaneously optimal insurance contracts to consumers and payment contracts to providers. Furthermore, we believe that two additional kinds of missing markets, or contractibility problems, are very important.

First, the quantity of treatment is actually not contractible, even *ex post;* the market for insurance and payment policies based on actual quantity is missing. In practice, insurers contract on the basis of *reports* of treatment submitted by providers, not treatment itself; an insurance "claim," as the name suggests, is only a report. Such report-based contracts are used because verifying quantity of treatment is costly. Distinguishing between a report and actual treatment immediately reveals an incentive problem. If a patient bears some cost sharing, for example by paying coinsurance, he has an incentive to ask the provider to underreport treatment. A physician paid by a capitation contract with a financial penalty assessed for each unit of service delivered has a similar incentive. If these insurance- and provider-payment contracts are paired in a health plan,

[2] It must be noted that this incomplete-market framework takes health status as an exogenous state of nature. Thus, it ignores the moral hazard problem when consumers can influence their health status by their own actions.

[3] Vertical integration of hospitals, physicians, and payers is another significant trend. See James C. Robinson and Lawrence P. Casalino (1996). One piece of evidence that these contracting changes are important is that physician incomes seem to be affected: the American Medical Association reports that real median physician income fell in 1994 (Physician Payment Review Commission, 1996).

the patient and the provider will agree to report a smaller amount of treatment. The key implication is that truthful reporting translates to restrictions on insurance-payment policies, which, as we show, may interfere with the implementation of second-best, efficient allocations. Although false reporting of patient diagnosis on hospital claims or false reporting of procedures conducted have been noted as a "fraud" issue (Grace M. Carter et al., 1991), our discovery that incentives to report truthfully impose restrictions on payment parameters probably is novel.

A second contractibility problem is arguably more important and fundamental. Some elements of treatment are never reported at all. Insurance coverage and provider payment are based on reports of measures such as number of "visits" or "days" in a hospital, or accounting "costs," which only partially indicate the resources devoted to treatment. A physician or other health care provider must be relied upon to prescribe the clinical content of the services connected with a "visit" or a "day," and to invest (costly) effort into making these services productive in terms of the patient's health. Thus, the physician almost always supplies her own input into the production of health care for the patient. This input, which we call "effort" in our model, is simply not contractible; the market for insurance and payment policies based on the physician's effort is missing. Because of this market failure, an insurance-payment system must motivate the physician to make an appropriate investment in her "hidden action."[4]

In this paper, we analyze the interaction between the two kinds of missing markets or contractibility problems just described. Our theory can be used to interpret modern complex contractual arrangements among insurer, patient, and physician as social responses to these missing markets, supplying incentives to remedy inefficiency. We emphasize that our theory focuses on the interaction between different kinds of market failures. Indeed, the second kind of market failure—the noncontractibility of physician effort—leads to an additional distortion in resource allocation if the restrictions placed on the payment system by the first—the requirement that patients and physicians truthfully report treatment quantity—are binding. The noncontractibility of physician effort alone may be overcome if the insurance-payment system is unconstrained, or if treatment quantity is verifiable *ex post*. Nevertheless, because of the need to induce truthful reporting due to the noncontractibility of treatment quantity, a third-best allocation may result; under a broad class of conditions, this third-best allocation involves an undersupply of physician effort.

Section I describes our basic model of the strategic interactions between a physician and a patient. Health care is produced by a patient-controlled input (number of visits) and a doctor-controlled input (effort), neither of which is directly contractible. Given the insurance and payment parameters, the physician chooses her effort level, and upon observing the effort, the patient decides on the quantity to purchase. Thus, the basic strategic interaction can be summarized as a *demand response*—the physician may influence the patient's demand reaction by investing in effort. After the production of health care, the patient and physician must make a claim to the insurer to fulfill the insurance and payment contract. We use a simple reporting game for the claims reporting process, capturing the idea that truthful reports in the claim must be individually rational; an alternative form that allows (implicit) collusion is considered in Section V. At the close of Section I we derive the constraints on insurance and payment systems that are consistent with truthful reports in the claims process.

In Section II we characterize the set of effort and quantity pairs that can be implemented by insurance and payment parameters, when physician-patient interaction as well as the requirement for truthful reporting are taken into account. We demonstrate that the requirement for truthful reporting limits the feasible range of payment parameters and, hence, the set of implementable efforts. In Section III, we derive the optimal insurance and payment systems. There, we prove that even though the

[4] The hidden action terminology is Arrow's (1986), where a distinction between contracting problems caused by hidden action or moral hazard, and hidden information or adverse selection, is made.

inputs are not contractible (or there are markets missing), the second best may still be attainable. Indeed, whenever the second best is feasible, its implementation does not impose extra costs even when physician-patient interaction and truthful reporting requirements must be fulfilled; the second best must arise in equilibrium as a result. But when the truth-telling constraint on the payment system binds, the second best fails to be implementable, and physician-patient interaction and truthful reporting result in a distortion.

Section IV considers a class of models in which the second best must be infeasible. In this case, equilibrium effort must be lower than the second best, and the equilibrium provider payment must be purely prospective—the physician receives a fixed fee for each patient, but must be responsible for the patient's treatment costs. Then we generalize the physician's simple profit-maximizing behavior, and endow her with a concern for the health benefits received by the patient. We show how this altruism or "ethical" behavior, as we call it, may be exploited to improve upon the third-best allocation. Welfare can improve because the ethical behavioral constraint can be used to get around the limitations of the payment system due to the truth-telling requirement.

In Section V, we study two extensions. First, we discuss the more severe constraint imposed on insurance and payment parameters if physicians and patients act as a coalition. There, the parties are allowed to write "side-contracts" when they consider making reports after treatments to fulfill the insurance-payment contract. The requirement for truthful reporting is shown to constrain *both* payment and insurance parameters, and the second best is not an equilibrium. Second, we consider competition among physicians as a design choice of a health plan. We find that competition may be a useful instrument when insurance-payment contracts fail to implement the second best. Generally, competition among physicians can be used as an incentive to induce costly efforts. Finally, Section VI contains some discussion and concluding remarks.

I. The Model

Our model describes an agency relationship between a physician and a patient. When a patient becomes sick, he may recover some of the loss due to sickness by undergoing medical treatment or services, which formally are modelled by a production process. The production of health for the patient requires two inputs, quantity of treatment and effort.

By quantity of treatment we mean health services as they are conventionally defined, such as the number of physician office visits, or the length of stay at a hospital. The quantity of treatment may be measured and verified *ex post*. By effort we mean other inputs contributed by the physician which increase the intensity or quality of treatment but are difficult to measure and verify. Following Gerald Wedig et al. (1989), we interpret effort as any costly activity that affects the patient's valuation of the services he receives, including dimensions of convenience, comfort, communication about medical conditions, as well as some narrowly defined "clinical" quality of care. For example, this may be the physician's task of matching a patient and his health problem to a specific therapy, monitoring a patient's progress and either recommending changes in treatment or termination of care, or learning about a problem that is unfamiliar to her. The most concrete way to think about effort in our context is simply in terms of physician "time" per visit or encounter. Physicians are paid according to "procedures," not time. More physician time leads to a higher quality but more costly visit. There is good evidence that physicians have considerable discretion about the time they spend per procedure.[5] The physician's decision about time or effort is likely to be related to the incentives in a payment system. One internist told us that when he conducts a routine examination of an elderly Medicare patient, he prepares a letter summarizing the findings for patients who pay a fee supplementing the Medicare rate; those patients relying only on the Medicare rate do not get the letter.

[5] William C. Hsiao et al. (1988) studied physician time and the subjective difficulty of medical procedures. The standard deviation of the time physicians spent per procedure was about half the mean in most cases, indicating that even in the carefully standardized vignettes prepared in this research, physicians can exercise substantial discretion in the time spent on tasks.

As in any model of optimal insurance, we assume that health improvement itself (what might be called "outcome") is noncontractible, and that an insurance contract cannot specify the quantity of treatment *ex ante*. Moreover, we assume that an insurance-payment contract is not based on the actual quantity of treatment. Instead, we assume that the contract can only be based on the treatment quantity *reported* by the patient and the physician: after the patient receives treatment, the physician and the patient file a claim with the insurer, and the patient's copayment and the physician's reimbursement are based on the claim information.[6] (The process of making a claim is described shortly.) The physician incurs disutility from supplying effort, which is assumed to be observable to the patient, but nonverifiable to the insurer. Thus, doctors can only be paid on the basis of the reported quantity of treatment, not effort or patient benefit.

The extensive form of our game consists of five stages.[7] In stage 1, the insurer chooses the elements of the insurance and payment systems. In stage 2, "Nature" determines whether the patient is ill with probability p. If the patient is healthy, the game ends. Otherwise, the patient seeks health care from the physician. Then in stage 3, the doctor chooses her effort, ε. In stage 4, after observing the doctor's choice of effort, the patient chooses the quantity of treatment, τ. Finally, in stage 5, the patient and the doctor play a reporting subgame; subsequently, the financial terms of the insurance-payment contract between the patient, the physician, and the insurer will be settled.

It is instructive to compare this model to the one analyzed by Zeckhauser (1970). His model is the same as the one presented here if in stage 5 in our game the patient and the physician must report the true treatment quantity, and if in stage 3 either the physician makes no choice of effort into treatment or her effort is contractible. Thus, our model generalizes Zeckhauser's.

We now define the reporting subgame. The physician first suggests to the patient a report of treatment quantity τ', not necessarily equal to the actual quantity τ. If the patient agrees, then τ' is reported in the claim. If he disagrees, then τ is reported. Implicitly, we assume that medical records cannot be falsified; alternation of medical records is deterred by fraud penalty from potential audits or possible malpractice suits. Ordinarily, however, the insurer uses only the information in the claim filed by the physician, not medical records, to collect the patient's copayments and to reimburse the doctor.

In our reporting subgame, either the patient or the physician can reveal the medical records to the insurer if he or she so chooses: the doctor always can enforce a truthful report by suggesting $\tau' = \tau$; the patient can reveal the true τ by disagreeing with a nontruthful suggestion. If the doctor misrepresents treatment quantity and the patient agrees, then the patient's copayment and the physician's reimbursement will be based on the false report. This subgame captures the idea that misreporting quantity information to the insurer is possible if, and only if, it is in the self-interest of both the physician and the patient; in other words, collusion is possible only if it is individually rational. This is a minimal requirement, one that imposes the least restriction on an insurance-payment system. Later, in Section V, we consider another subgame that captures the idea that collusion is motivated by the joint interest of the physician-patient coalition.

Besides the reporting subgame, the other difference between our model and Zeckhauser's is the physician's effort decision. In our extensive form, the physician's effort decision is made in stage 3, *before* the patient chooses how many units of quantity to purchase from the physician. Providing effort is costly for the physician, but effort is not verifiable, and cannot be rewarded by the insurance-payment

[6] In adopting the assumption that the insurance contract is based entirely on reported information of quantity, we do not mean that this information may never be made verifiable. The interpretation is that insurance-payment contracts that require credible verification of each piece of information related to a patient's course of treatment is very costly to enforce. Our assumption then allows us to study those contracts that are written to save the costs of verification.

[7] We use Greek letters for endogenous variables, small Latin letters for parameters, and large Latin letters for functions.

contract directly. Nevertheless, our sequential structure allows the creation of incentives by the insurance-payment contract through the patient's reaction. Indeed, since effort is important to the patient, different levels of it will lead to different quantity decisions by the patient. By changing her effort, therefore, the physician will induce different patient demands. This fact, together with the ability of the payment contract to reimburse the physician more or less than the (marginal) cost of quantity, creates incentives for the physician to provide costly effort.

This methodology of *demand response* can be given a number of interpretations. First, often physicians have long-term relationships with their patients, as with the case of family practitioners, many dentists, and those doctors who treat chronic illnesses. Then it is reasonable to assume that a patient actually observes the physician's quality of care before he decides the total quantity of treatment, as we do in the formal model.

In other situations (such as acute, emergency, and specialty services), repeated physician-patient interactions may not occur, but a physician's quality of care may still influence the demand for her services. We can interpret our model as a reduced form or a stationary state of a more general, dynamic model with many patients. Suppose that consumers may get some information from friends or other physicians about the quality of care that a physician has provided to other patients.[8] By changing her effort and quality, the physician changes the information available to her potential patients. Given his information about the physician's past quality, a patient may rationally and correctly believe that he will receive the same quality, because the physician is interested in maintaining her reputation. Thus, the physician's choice of quality alters the information available to her potential patients, and creates a demand response.[9]

We should point out that the alternative assumptions—the physician effort decision is made either simultaneously with, or after the patient's quantity decision—are unpalatable: in both cases, neither the patient's quantity choice nor the payment contract can provide any incentive for the physician to undertake costly actions.[10]

We now define the remaining elements of the game in detail, beginning with the patient's and the doctor's preferences. The patient's utility depends on health, the benefits to health care treatment received, and income available for spending on other goods after any insurance premiums and his share of costs for treatment are paid. Initial income is w. The patient is ill with probability p. When ill, the patient is subject to a negative health shock with a monetary equivalent of s. Health can be (partially) recovered according to a strictly concave function $F(\tau, \varepsilon)$ representing the monetary equivalent of benefits to treatment, where τ denotes quantity of treatment, and ε a physician's input into the production of health benefits.[11] We assume F is increasing in τ and ε, and bounded between 0 and s. The variables τ and ε are bounded below, with their lower bounds set at zero.

The patient's copayment per unit of treatment is β; this copayment is paid to the in-

[8] This information need not be the quality itself, but can be anything related to it, such as satisfaction levels, outcomes, etc.

[9] Through repeated interactions, more powerful contracts may be used to induce stronger physician quality incentives than those considered formally in this paper. This, however, does not undermine the focus of this paper. First, repeated interaction may only generate nonverifiable information. Contracts of the form we consider here must still be used. Second, any verifiable information generated by repeated interaction will be used with incentive contracts. Thus, we expect that our qualitative results here will continue to hold.

[10] Another alternative assumption is that a patient may learn about a physician's quality through treatments, and decides on the total quantity as this information accumulates. We believe that this is consistent with our second interpretation of demand response (in which patients and physicians do not interact repeatedly, but physicians invest in quality to increase or maintain their reputations).

[11] In adopting this formulation of the health production function, we build on Michael Grossman's (1972) pure investment model. We assume that the consumer derives utility from goods, other than health, that can be purchased. Medical care is bought because the health improvement that it generates can be translated into additional income to be spent on other goods and services. The model abstracts from time allocation issues by assuming implicitly that a consumer's own time and treatment τ are used in fixed proportions to produce health.

surer. For convenience, it is assumed that the patient can always obtain treatment at constant unit cost c *ex post*. Hence, under an insurance policy, copayment β will be less than the marginal cost of treatment: $0 \leq \beta \leq c$. The patient must also pay the insurer an insurance premium $\alpha \geq 0$ before his health status is known. Income available for other goods is $w - \alpha$ when the individual is healthy and does not require treatment. When the patient becomes ill, he seeks medical care. When the physician supplies an effort ε and the patient chooses a quantity τ, he obtains the benefit $F(\tau, \varepsilon)$. If the treatment quantity τ^r is reported to the insurer, then his income becomes $w - \alpha - \beta\tau^r$. Hence, using the strictly concave function $U(\cdot)$ to represent the risk averse patient's preferences, we write his expected utility as:

$$(1)\quad EU = pU(w - \alpha - \beta\tau^r - s + F(\tau, \varepsilon)) + (1 - p)U(w - \alpha).$$

The physician is risk neutral with respect to money,[12] has a utility function separable in money and effort, and a reservation utility level normalized at zero. It is assumed that the cost of treatment, c per unit, is borne by the physician. The physician can be paid in two components. First, for each patient, she receives a fixed fee ρ, which can be regarded as a prospective payment. Second, for each unit of treatment reported to the insurer, she receives a reimbursement $\delta + c$. Hence, the variable δ is the margin over variable cost and can be positive or negative. A negative δ means the doctor is paid less than cost at the margin. Following Randall P. Ellis and McGuire (1986), a payment system with $\delta < 0$ will be referred to as containing supply-side cost sharing. A fully prospective payment system sets $\delta = -c$, the doctor receiving her total income from the prospective component ρ.

If actual quantity of treatment is τ and the reported quantity of treatment τ^r, the physician's total revenue becomes $\rho + (\delta + c)\tau^r - c\tau$. The physician must also bear the cost of effort, ε, designated $G(\varepsilon)$. We assume $G(0) = 0$, $G' > 0$, and $G'' > 0$. Because effort is noncontractible, the physician receives no direct payment for it. The physician's utility can be written as:

$$(2)\quad V = p[\rho + (\delta + c)\tau^r - c\tau - G(\varepsilon)].$$

The insurer maximizes the patient's expected utility subject to a balanced budget. This implies that the premium paid by the patient must equal the expected value of the insurer's payment to the physician:

$$(3)\quad \alpha = p[\rho + (\delta + c - \beta)\tau^r].$$

We complete this section by deriving the constraint imposed on the payment system if in a (subgame-perfect) equilibrium the true treatment quantity is reported. First, since $\beta \geq 0$, the patient pays a positive amount for each unit of reported quantity. Because he always can reveal the true quantity, he rejects a report that is bigger than the true quantity. Conversely, he will accept a report that is lower than the actual quantity. Second, when $\delta + c \geq 0$, the physician receives a positive payment per unit of reported treatment. In this case, she will never underreport. Nevertheless, she cannot overreport quantity, since the patient will reject her suggestion. Hence, when $\delta + c \geq 0$, truthful reporting must be the equilibrium in the subgame in stage 5. Conversely, consider $\delta + c < 0$. In this case, the physician pays a positive amount to the insurer for every unit of reported quantity, so she has an incentive to report $\tau^r < \tau$. Moreover, such a reported quantity will be accepted by the patient, who prefers to reduce his copayment. Thus, whenever $\delta + c < 0$, the equilibrium report will be zero.[13] In summary, truthful reports will be the equilibrium in stage 5 if, and only if, $\delta + c \geq 0$. In our framework, a payment system that relies on unverified claim information cannot make the physician be

[12] The physician is not exposed to any health risk the patient faces; the assumption that she is risk neutral is unimportant.

[13] In this case, the doctor only receives a prospective payment ρ in equilibrium, and must always perform the least costly effort. In the analysis to follow, we will concentrate on insurance-payment systems that implement costly efforts.

responsible for more than the cost of treatment.[14] From now on we impose the constraint $\delta + c \geq 0$.

II. Patient Choice of Quantity and Doctor Choice of Effort

As we have just shown, equilibrium truthful reporting occurs in stage 5 if, and only if, the physician's payment per unit of treatment is nonnegative. We now proceed to find the subgame-perfect equilibrium of this game, and we begin with stage 4. So suppose that the insurer has chosen $(\alpha, \beta, \rho, \delta)$ in stage 1 (with $\delta + c \geq 0$), and that in stage 2 Nature has determined that the patient is ill. Further, suppose that in stage 3, the doctor has chosen her effort level ε. In stage 4, the patient chooses τ to maximize $U(w - \alpha - \beta\tau - s + F(\tau, \varepsilon))$. His best response—the optimal choice of τ—is given by the following necessary and sufficient first-order condition,[15] obtained by maximizing $U(\cdot)$ with respect to τ: $U'(\cdot)(-\beta + F_\tau(\tau, \varepsilon)) = 0$, or

$$\beta = F_\tau(\tau, \varepsilon). \tag{4}$$

In this standard description of demand behavior, the patient chooses τ to set the marginal benefit of treatment equal to the out-of-pocket cost per unit β. Because both the health shock and health care production are expressed in monetary equivalents, the patient's insurance premium α does not affect his demand for treatment.

To the physician, the function (4) is the patient's reaction[16] against her choice of ε given the copayment β. The doctor's effort ε is her instrument to influence the patient's demand for treatment, on which her reimbursement will be based. In fact, from (4), we obtain

$$\frac{d\tau}{d\varepsilon} = -\frac{F_{\tau\varepsilon}}{F_{\tau\tau}}, \tag{5}$$

which measures the increase of the patient's choice of treatment quantity (his reaction) per unit increase of physician effort. The sign of the slope of this reaction function is the same as the sign of $F_{\tau\varepsilon}$. If effort and treatment quantity are *substitutes,* meaning that $F_{\tau\varepsilon} < 0$, then to induce the patient to demand a higher τ, a lower ε will be necessary. Alternatively, if effort and treatment are *complements,* meaning that $F_{\tau\varepsilon} > 0$, then a higher τ can only be induced by a higher ε.

The classification of substitutes and complements between effort and quantity is important; in the sequel, it will be shown that distortions of equilibrium allocations may arise under the case of substitutes. In general, the sign of the cross partial derivative of F may change according to both τ and ε. For example, effort and quantity may be complements at low effort levels ($F_{\tau\varepsilon}(\tau, \varepsilon) > 0$), and substitutes at high levels ($F_{\tau\varepsilon}(\tau, \varepsilon) < 0$), or vice versa.

We now consider the doctor's decision rule in stage 3. Here, anticipating the patient's treatment quantity reaction in stage 4, the doctor chooses ε to maximize her revenue less costs (including the cost of ε). Thus, in any subgame-perfect equilibrium, the doctor's choice of ε and the patient's subsequent choice of τ are given the solution of the following Program A: for $0 \leq \beta \leq c$ and $\delta \geq -c$, choose ε and τ to maximize

$$\rho + \delta\tau - G(\varepsilon) \tag{6}$$

subject to Program A

$$\beta = F_\tau(\tau, \varepsilon).$$

[14] Notice that by symmetry, when $\beta < 0$ and $\delta + c > 0$, the physician and the patient both have incentives to overreport treatment quantity. Thus, if $\beta < 0$, then only $\delta + c \leq 0$ is consistent with truthful reporting. Nevertheless, an optimal insurance-payment system ever setting a *negative* copayment for the patient seems an anomaly, and throughout the paper we assume that the copayment is always positive.

[15] A subscript of a function signifies a partial derivative with respect to that variable; a double subscript denotes a second-order derivative.

[16] It is also easy to verify that when β increases, the reaction function (4) must shift towards the origin. Reaction functions do not cross: that is, $\beta_1 \neq \beta_2$ implies that $\beta_1 - F_\tau(\tau, \varepsilon) = 0$ and $\beta_2 - F_\tau(\tau, \varepsilon) = 0$ do not possess a solution. Clearly, then, from (4) $d\beta = F_{\tau\varepsilon}d\varepsilon + F_{\tau\tau}d\tau$. For $d\beta > 0$, if $F_{\tau\varepsilon} < 0$, then it cannot be true that $d\varepsilon > 0$ and $d\tau > 0$. Thus, (4) must shift to the left as β is increased.

Observe that the doctor's prospective payment, $\rho \geq 0$, does not affect her supply of effort. Hence, only β and δ appear as parameters in Program A. Our goal now is to characterize the set of subgame-perfect equilibrium (τ, ε) pairs in stage 3. First, we define

$$\Omega = \{ (\tau, \varepsilon) : \text{there exist } (\beta, \delta), \text{ with } 0 \leq \beta \leq c \text{ and } \delta \geq -c, \text{ for which } (\tau, \varepsilon) \text{ solves Program A given } (\beta, \delta) \}.$$

The set Ω contains all those (τ, ε) pairs that can arise as subgame-perfect equilibria given some combination of copayment and reimbursement margin parameters (β and δ). The set Ω will be called the *implementable set.*

From the constraint of Program A, one obtains τ as a function of ε, with its derivative given by (5). Hence, the first-order derivative of the objective function with respect to ε becomes

$$(7) \qquad \delta \frac{d\tau}{d\varepsilon} - G'(\varepsilon) = -\frac{F_{\tau\varepsilon}(\tau, \varepsilon)}{F_{\tau\tau}(\tau, \varepsilon)} \delta - G'(\varepsilon).$$

This derivative is negative if τ and ε are substitutes ($F_{\tau\varepsilon} < 0$) and $\delta \geq 0$, or if τ and ε are complements ($F_{\tau\varepsilon} > 0$) and $\delta \leq 0$. Indeed, from (6), if $\delta > 0$, the doctor's utility increases only if τ increases. When $F_{\tau\varepsilon} < 0$, an increase in τ can only result from a decrease in ε. The doctor will therefore reduce effort as much as she can if $\delta > 0$. It follows that in the case of substitutes, the implementation of costly effort must require the doctor to bear some of the costs of treatments. A similar argument establishes that when effort and treatment are complements, the implementation of costly effort must require that the doctor be reimbursed more than the costs of treatment.[17] To summarize, we state:

[17] Incentive problems associated with the discrepancy between the reimbursement rate and the marginal cost of treatment have been appreciated also by Pauly (1991).

PROPOSITION 1: *Consider a treatment quantity-physician effort pair (τ, ε) belonging to the implementable set Ω. Suppose at (τ, ε), ε and τ are substitutes; that is, $F_{\tau\varepsilon} < 0$. Then ε is above its minimum ($\varepsilon > 0$) if, and only if, the physician payment system includes supply-side cost sharing ($\delta < 0$). Alternatively, suppose at (τ, ε), ε and τ are complements; that is, $F_{\tau\varepsilon} > 0$. Then ε is above its minimum ($\varepsilon > 0$) if, and only if, the physician payment system does not include supply-side cost sharing ($\delta > 0$).*

The characterization of the implementable set Ω in Proposition 1 demonstrates the fundamental incentive problem in the physician-patient interaction. To implement a pair of effort and treatment quantity, the copayment and reimbursement margin must be chosen carefully to ensure that in fact the physician and the patient have the incentives to select the effort and treatment in equilibrium. Therefore, the range of implementable effort and treatment quantity, the "size" of Ω, is directly related to the range of variation of the copayment and reimbursement margin parameters, β and δ. The range of β is limited by insurance considerations, and common to both regimes of verifiable effort and nonverifiable effort. The range of δ is only relevant when effort is nonverifiable, the focus of our model, and limited by the truth-telling constraint: the physician cannot be made responsible for more than the actual cost of treatment quantity, $\delta \geq -c$, if the actual quantity is to be reported for reimbursement.

If effort and quantity are complements, Proposition 1 requires that the implementation reimbursement margin be positive. Thus, the constraint $\delta \geq -c$ is not relevant for the pair to be implementable. But if they are substitutes, according to Proposition 1, the margin must be negative. Then the constraint $\delta \geq -c$ puts a limit on the penalty the payment system can impose. As we next show, the higher the effort level to be implemented under substitutes, the more negative δ would have to be set. Thus, when effort and treatment quantity are substitutes, very high effort levels may not belong to the implementable set because the constraint $\delta \geq -c$ binds.

Consider the case of substitutes, and assume that the first-order conditions for Program A are necessary and sufficient for its solution. Then it can be shown that the comparative statics with respect to δ are given by:

$$(8) \qquad \frac{\partial \tau}{\partial \delta} = -\frac{F_{\tau\varepsilon}(\tau, \varepsilon)^2}{H} > 0$$

$$\frac{\partial \varepsilon}{\partial \delta} = \frac{F_{\tau\varepsilon}(\tau, \varepsilon) F_{\tau\tau}(\tau, \varepsilon)}{H} < 0,$$

where $H < 0$ is the bordered Hessian.[18] We confirm that under substitutes, implementation of higher effort levels requires setting increasingly negative values for δ.

Figure 1 illustrates the implementable set under substitutes. Here, the patient's reaction functions ($F_\tau(\tau, \varepsilon) = \beta = 0$ and $F_\tau(\tau, \varepsilon) = \beta = c$) are negatively sloped. The doctor's (concave) indifference curves are also shown for three sets of values of ρ and δ, with slopes $G'(\varepsilon)/\delta$. An equilibrium is a tangency of an indifference curve with a reaction function, such as (ε', τ'). As the parameters δ and β vary over their ranges, different pairs of τ and ε will become equilibrium in stage 3, and these are represented by the shaded area. The southeast boundary of the implementable set reflects the constraint $\delta \geq -c$: any effort level to the east of this boundary is infeasible.

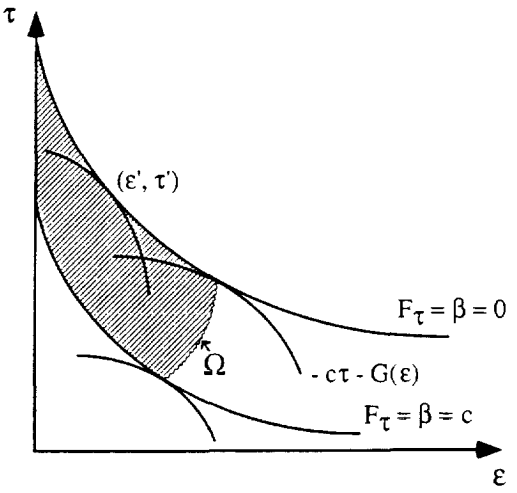

FIGURE 1. THE SET OF IMPLEMENTABLE EFFORT AND TREATMENT LEVELS

III. Optimal Insurance-Payment Systems

In this section, we characterize the optimal insurance-payment system. That is, we analyze the equilibrium in stage 1, or the choice of the insurance-payment parameters that maximize the patient's expected utility [see (1)], given that the premium for the patient is actuarially fair [see (3)], that the physician's expected utility [see (2)] is at least her reservation level (normalized at zero), and that the choices of ε and τ are given by an equilibrium of the subgame in stage 3 defined by the payment parameters.

Clearly, the optimal insurance-payment system will only pay the physician her reservation utility level. Combining the physician's binding reservation utility constraint and the insurance breakeven or budget constraint (3), we obtain

$$(9) \qquad p[c\tau + G(\varepsilon)] = \alpha + p\beta\tau,$$

the left-hand side of which is the total expected cost of treatment and effort; the right-hand side, the total premium and expected copayment from the patient. Conversely, for any α, β, ε, and τ that satisfy (9), and for an arbitrary δ, one can set $\rho = G(\varepsilon) - \delta\tau$ so that both (3) is satisfied and (2) is equal to the doctor's reservation utility (zero). So without loss of generality, we can eliminate the choice ρ, and replace (3) and (2) by (9). Hence, it is sufficient to consider those insurance-payment parameters that maximize (1) subject to (9), and that given the parameters, the doctor's choice of ε and the patient's choice of τ form an equilibrium in the subgame in stage 3. Formally, the equilibrium insurance-payment parameters, as well as the effort and treatment quantity, will be given by the solu-

[18] The expression for H is $F_{\tau\varepsilon}[\delta F_{\tau\tau\varepsilon} + F_{\tau\tau\tau} G'(\varepsilon)] - F_{\tau\tau}[\delta F_{\tau\varepsilon\varepsilon} + F_{\tau\tau\varepsilon} G'(\varepsilon) + F_{\tau\tau} G''(\varepsilon)]$. Its negative value follows from the second-order necessary condition.

tion to the following program: choose α, β, δ, ε, and τ to maximize

$$pU(w - \alpha - \beta\tau - s + F(\tau, \varepsilon)) + (1 - p)U(w - \alpha)$$

subject to (9) and $(\varepsilon, \tau) \in \Omega$. It is important to observe that the payment margin parameter δ necessary for the implementation of ε and τ only appears in the definition of Ω, not in the other constraint (9) or the objective function. Because the prospective payment element ρ can be adjusted to extract the physician's rent for any given value of δ, the use of δ to implement a quantity-effort pair does not result in extra social costs.

Now, consider a benchmark regime in which the doctor's input ε is contractible. Then a payment policy can specify a payment to the doctor contingent on her performance of a particular effort level; moreover, she can be relied upon to reveal truthfully the patient's treatment quantity in the insurance claim, since her welfare is independent of the patient's quantity choice. This regime corresponds to the Zeckhauser model. Here, the quantity of treatment is again given by (4), and the budget constraint (9) still applies. However, the doctor makes no decision about her effort level and the policy instrument δ can be deleted. In this Zeckhauser, second-best regime (second best in the sense that treatment quantity is not contractible *ex ante*), the optimal insurance-payment system is given by the solution of the following program: choose α, β, ε, and τ to maximize

$$pU(w - \alpha - \beta\tau - s + F(\tau, \varepsilon)) + (1 - p)U(w - \alpha)$$

subject to (9) and (4).

Clearly, this program is a relaxed version of the previous one for the regime in which physician effort is noncontractible: the constraint (4) is implied by the condition that $(\varepsilon, \tau) \in \Omega$, but not conversely (see Program A). Let α^{SB}, β^{SB}, ε^{SB}, τ^{SB} denote the solution to this relaxed program, or the second best; let EU^{SB} be the patient's second-best expected utility. The unrelaxed program above thus represents a "third best"; in addition to the patient treatment quantity being *ex ante* noncontractible (as in the second best), the doctor's effort level also is noncontractible. Furthermore, the physician's payment parameter, δ, is restricted to be at least $-c$, in order to guarantee truthful claim reports about treatment quantity. Let α^{TB}, β^{TB}, δ^{TB}, ε^{TB}, τ^{TB} denote the solution to the unrelaxed program, and let EU^{TB} be the patient's expected utility at this solution.

The next proposition relates the second best and the third best. It states that the second best is achievable whenever $(\varepsilon^{SB}, \tau^{SB})$ belongs to the implementable set Ω.

PROPOSITION 2: *Suppose there exists* δ^*, *with* $\delta^* \geq -c$, *such that* $(\varepsilon^{SB}, \tau^{SB})$ *belongs to the implementable set* Ω. *Then the second best is achieved in the third-best regime. That is,* $\alpha^{TB} = \alpha^{SB}$, $\beta^{TB} = \beta^{SB}$, $\varepsilon^{TB} = \varepsilon^{SB}$, $\tau^{TB} = \tau^{SB}$, *and* $EU^{TB} = EU^{SB}$.

PROOF:

Because EU^{SB} is the patient's expected utility from the solution of a more relaxed program, $EU^{SB} \geq EU^{TB}$. Under the hypothesis of the proposition, the set of variables α^{SB}, β^{SB}, δ^*, ε^{SB}, τ^{SB} is a feasible policy. But under this policy, $EU^{TB} = EU^{SB}$. Thus, the patient's expected utility when the physician input is noncontractible remains second best. This also implies that the solution to the unrelaxed program can be obtained by solving the relaxed program.

Proposition 2 says that if the doctor's payment δ can be adjusted to align her effort incentive with that in the second best, then the noncontractibility of effort is inconsequential. The design of optimal insurance can proceed as if the doctor's effort was at the second-best level. Then the implementation of the second-best effort can be achieved through an appropriate choice of δ.

From Proposition 2, the second best is not achieved when the bound on δ due to the truth-telling requirement, namely $\delta \geq -c$, binds. In this case, the patient's equilibrium expected utility will be strictly inferior to the second best. From Proposition 1, we know that when $F_{\tau\varepsilon} > 0$, so that treatment quantity and effort are complements, the implementation of any

positive ε requires a positive δ, which does not conflict with the constraint that δ must be greater than $-c$. When τ and ε are substitutes, however, the implementation of positive effort levels requires supply-side cost sharing, or setting δ less than zero. For some cases, the implementation of the second best merely requires setting δ at a level strictly above $-c$. In the following section, we investigate situations in which the constraint $\delta \geq -c$ binds, a genuine third-best regime. We also will define professional ethics formally, and identify circumstances in which the professional ethics constraint improves the third best.

IV. A Third-Best Equilibrium and Ethical Behavior

Having established the limits of the physician payment system to achieve efficiency when physicians are motivated strictly by economic self-interest, we turn our attention to the role of professional ethics. Many models of physician behavior include some role for "altruistic" or "ethical" behavior, although there is no consensus about how this should be done. (See Martin Gaynor [1994] for a recent review.) Arrow (1963) suggested that physicians may behave in the interest of their patients, that is, act ethically, in a kind of social exchange for the professional autonomy patients and society grant them; in addition, physicians may take pride in their work, and enjoy providing health care. Here, we incorporate a physician's concern for patients in a simple way by assuming that a physician must provide health benefits above a certain threshold (given a health shock). Formally, the ethics constraint is given by the requirement that

$$F(\tau, \varepsilon) \geq \bar{F}, \tag{10}$$

where $\bar{F}$ is a constant. Under the ethics constraint, in stage 3, the physician must choose an effort level, which produces a health benefit of $\bar{F}$ when combined with the patient's demand response. Clearly, this constraint need not be binding always. For example, when β is sufficiently low, then for some δ, the patient may already receive a benefit more than $\bar{F}$ in the equilibrium in stage 3 of the game. Our interest is the identification and analysis of those equilibria in which (10) binds.

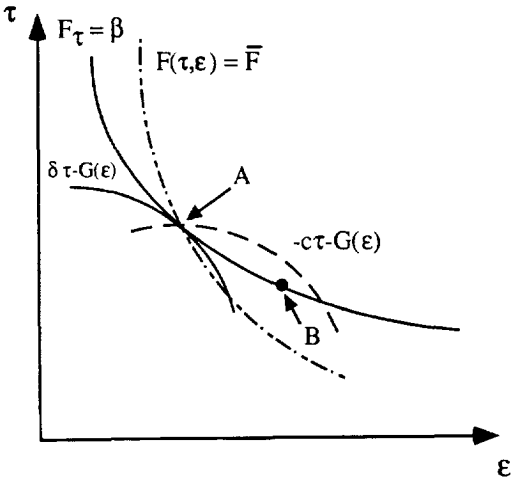

FIGURE 2. A THIRD-BEST EQUILIBRIUM ALLOCATION

First, we characterize a class of situations in which the second best cannot be attained; the next proposition describes a condition under which the constraint $\delta \geq -c$ must bind and prevents the second best from belonging to the implementable set Ω.

PROPOSITION 3: *Suppose* $F_{\tau\varepsilon} < 0$, *and that for all* ε *and* τ

$$-\frac{F_{\tau\varepsilon}(\tau, \varepsilon)}{F_{\tau\tau}(\tau, \varepsilon)} > -\frac{F_{\varepsilon}(\tau, \varepsilon)}{F_{\tau}(\tau, \varepsilon)}.$$

In equilibrium, $\delta = -c$, *and the patient's expected utility is less than* EU^{SB}.

The proofs of this and the following proposition are in the Appendix. Figure 2 illustrates Proposition 3. A point like A cannot be second best. At point A, the indifference curve ($\delta\tau - G(\varepsilon)$) is tangent to the reaction function ($F_\tau = \beta$). But by hypothesis, the isobenefit line ($F(\tau, \varepsilon) = \bar{F}$) is steeper than the reaction function.[19] The isosocial cost

[19] The assumption that the the isobenefit line is steeper than the reaction function is reasonable. Consider what must be true if it is not the case. Then if ε is reduced, the

line $(-c\tau - G(\varepsilon))$ is indicated by the dashed line. By considering a point like B, with the same β but a lower δ, it is possible to reduce τ compared to that in A (reducing risk to the consumer), reduce total social cost, reduce premium α (although this cannot be shown in the figure), and increase health benefits. Thus, all components of expected utility improve with a decrease in δ. Hence, the tangency point A cannot be an equilibrium if $\delta > -c$. The only possible equilibrium allocations are those on the southeast boundary of Ω where $\delta = -c$. An example of a health production function F that satisfies the hypothesis of the proposition is $F(\tau, \varepsilon) = \ln(\tau + \varepsilon) + f(\varepsilon)$, where f is an increasing and concave function, and where ln denotes the natural logarithm.[20]

Returning to the ethics constraint, we recall that in a subgame-perfect equilibrium, the doctor's choice of ε and the patient's choice of τ will be given by the solution of Program A with the additional constraint (10). Clearly, if the ethics constraint does not bind, the equilibrium will be the solution to the unmodified Program A. When the ethics constraint does bind, then (10) together with (4) will determine the equilibrium τ and ε.

In Figure 3, an ethics constraint adds those points on the dotted isobenefit line F^* but not in Ω to the set of possible equilibria. All implementable points must be on a reaction function for $0 \leq \beta \leq 1$. The ethics constraint may be useful for improving resource allocation, because, without it, the equilibrium effort level is too low. (In Figure 3, the physician reaction function and isobenefit line designated by Δ are used in the proof of Proposition 4 in the Appendix.)

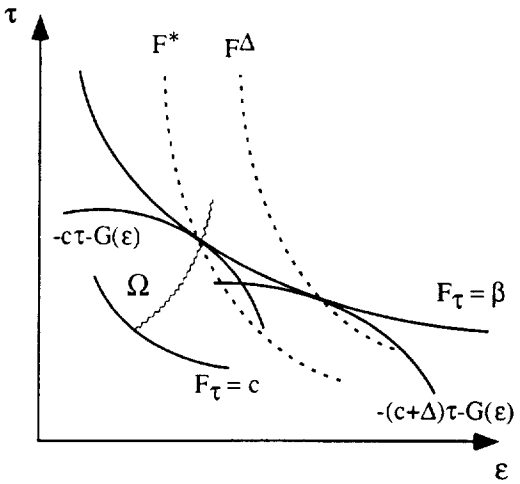

FIGURE 3. ETHICS AND AN IMPROVEMENT ON THE THIRD BEST

Consider first an equilibrium in the absence of an ethics constraint: $\alpha^*, \beta^*, \delta^*, \varepsilon^*, \tau^*$ that solve the program for the optimal payment system in Section III. Also, denote by EU^* the patient's expected utility from the equilibrium allocation. Given the equilibrium, suppose we set the professional ethics constraint at $F(\tau, \varepsilon) \geq F(\tau^*, \varepsilon^*) \equiv F^*$. One interpretation is that physicians are "accustomed" to providing the level of health care F^*, and this level has become to be accepted as the professional norm. We investigate how the replacement of the physician's profit-maximizing choice of ε by her choice of ε satisfying $F(\tau, \varepsilon) = F(\tau^*, \varepsilon^*) \equiv F^*$ affects the equilibrium insurance-payment system.

Since F^* is defined with respect to the equilibrium when the ethics constraint is absent, the patient's equilibrium expected utility when the ethics constraint is present will be at least EU^*. The next proposition shows that when the isobenefit line $(F(\tau, \varepsilon) = \bar{F})$ is steeper than the reaction function $(F_\tau(\tau, \varepsilon) = \beta)$, the patient's equilibrium expected utility will increase through an increase of β from β^*. By imposing a higher coinsurance rate on the patient, a smaller amount of treatment will result. To maintain the health care level at F^*, the physician must increase her effort. Although the cost of effort will have to be paid for, the

increase in τ must compensate for more than the reduction in ε, in the sense that a higher level of health benefit will be produced. In other words, when she reduces effort, the doctor *improves* the patient's health by inducing him to choose significantly more quantity. Convexity assumptions do not eliminate this perverse possibility, but we regard it as unlikely.

[20] This function is bounded by s if suitable bounds are imposed on τ and ε.

result is a net gain in the patient's expected utility.

PROPOSITION 4: *Suppose* $F_{\tau\varepsilon} < 0$, *and that for all* ε *and* τ

$$-\frac{F_{\tau\varepsilon}(\tau, \varepsilon)}{F_{\tau\tau}(\tau, \varepsilon)} > -\frac{F_{\varepsilon}(\tau, \varepsilon)}{F_{\tau}(\tau, \varepsilon)}.$$

Under the ethics constraint $F(\tau, \varepsilon) \geq F(\tau^*, \varepsilon^*) \equiv F^*$, *the patient's equilibrium expected utility must be higher than the equilibrium expected utility without the ethics constraint* (EU^*), *and* β *must increase above* β^*.

In circumstances in which the (Zeckhauser) second-best equilibrium is unattainable because of bounds on feasible forms of doctor payment, an ethics constraint of the form stipulated here can be exploited by the insurer to improve the patient's expected utility. In comparison to the third-best equilibrium with $\delta = -c$, the insurer can enlist the ethics constraint by lowering treatment demanded by the patient by raising β, forcing the doctor to raise her effort to satisfy the ethics constraint. Under the hypothesis of Proposition 4, the patient's benefit will have to increase, and he attains a higher expected utility as a result.

V. Extensions

In this section, we consider two extensions of our model. First, we study an alternative form of physician-patient reporting (sub)game. Second, we consider the effect and desirability of competition among physicians.

In the main model in Section I, we adopt a very simple and natural reporting subgame. The reader may recall that the central reason for introducing this subgame is that reimbursements often are based on reported quantities, creating an asymmetry of information on the actual quantities between physicians and patients on the one hand and insurance companies on the other. If the physicians and patients as a group can exploit the insurers by misrepresenting this information, explicit incentives must be introduced to mitigate the potentially detrimental effects.

Our reporting subgame in Section I describes a simple reporting mechanism for the patient and physician. The construction there is based on the idea that misrepresenting information is feasible if, and only if, each party finds it in his or her own private interest to collude. This implicitly assumes that there is no outside enforcement mechanism that the parties can appeal to when they coordinate to make reports; in other words, successful misrepresentation must be individually rational. Although this appears to be an accurate assumption in many circumstances (and a minimal requirement for any payment system based on reported information), sometimes one observes cooperation between health care providers and clients as if the parties can bind themselves to making false reports. For example, a physician may waive a copayment in exchange for an understanding that she bills for a higher quantity of service than was actually provided. The collusive transaction—the copayment or deductible waiver in exchange for cooperation—mimics a kind of (implicit) contract between the doctor and the patient, although such contracts are clearly illegal.[21]

It is thus of some interest to study an alternative reporting subgame in which "side-contracts" are allowed; that is, information misrepresentation results from a *group* incentive. Clearly, the assumptions of costless side-contracting and the absence of it are both extreme, but their simplicity implies tractability, and they may serve as benchmarks for broad classes of other reporting subgames.[22]

Under side-contracting the reporting subgame is described below. First, the patient and physician jointly decide on a report of treatment quantity τ' and a transfer from the physician to the patient (the side-contract).

[21] If *ex post* the provider has reneged on her promise and seeks to collect copayment from the patient, the patient would have to pay. The patient's threat of reporting the fraud to the authority probably is an incredible threat, since he, himself, is subject to severe penalty.

[22] A long-term relationship such as the one between a physician and a patient may discourage or encourage collusion. On the one hand, physician and patient may fear that collusion in the long run may trigger reactions by payers, making them worse off. Nevertheless, the trust between physician and patient may encourage them to act in their "group" interest, inducing them to collude.

Second, the quantity τ^r is reported in the claim, the copayment β is paid, and reimbursement $\delta + c$ received, by the patient-physician coalition. The equilibrium of this reporting subgame with a side-contract is straightforward to derive. First, the transfer settles the division of surplus between the patient and the physician. Second, by the side-contract, for each unit of treatment quantity the patient-physician coalition reports, it pays copayment β but receives reimbursement $\delta + c$. Effectively, the coalition can trade $\delta + c$ for β! Hence, whenever $\delta + c > \beta$, the report will be the feasible maximum. Likewise, when $\delta + c < \beta$, the report will be feasible minimum. Truthful reporting obtains if, and only if, $\delta + c = \beta$: side-contracting imposes a more stringent constraint on the insurance-payment parameters.

To ensure truthful reporting in a side-contracting regime, an insurance-payment policy must set patient copayment and physician reimbursement per unit of treatment to be identical. As a result, the net reimbursement to the physician from the insurance company is a prospective payment. Proposition 1 continues to apply. That is, when τ and ε are complements, the margin on physician reimbursement per unit of τ must be positive if costly effort is to be implemented; when they are substitutes, the margin must be negative. The equilibrium allocation under side-contracting can be obtained by solving the program for the third best (shown before Proposition 2) with the additional constraint that $\delta + c = \beta$. Formally, the payment margin δ in Program A must be replaced by $\beta - c$. Because of this new restriction on the payment margin, the implementation of costly effort (whether under effort-quantity complements or substitutes) must be at the expense of further distortions in optimal risk sharing, and the second-best (Zeckhauser) allocation will not be implementable generally. We will not derive formally the distortion due to the deviation from the second best, but conjecture that equilibrium effort can be too low or too high, depending on the parameters of the model.

We now return to the case where side-contracts are infeasible, and turn our attention to the effects of physician competition on our model. Our methodology regards insurance and payment systems as outcomes of a design process. This can be applied to competition, too. That is, the structure of the contract between the payer and physicians may influence or explicitly specify the form of competition between physicians—more on this below. We must, however, first clarify the potential role of physician competition within our framework.

The simplest way to illustrate the effect of physician competition is to imagine for the time being that treatment quantity is verifiable and hence contractible. In this circumstance, an insurance contract may specify the quantity of treatment when the consumer seeks help from a physician; only the physician's input, her effort, is noncontractible. Under competition, a physician must choose an effort to compete with other physicians for patients; a higher effort enhances patient utility, increasing a doctor's demand. Indeed, a physician will raise effort to increase her demand if, and only if, she expects to make a profit from treating a patient. Because the marginal return from attracting a patient must equal the marginal disutility of effort, raising the prospective and margin parameters implements higher efforts.[23] When physicians must compete for patients, the payment parameters may be used as an instrument for the implementation of costly efforts.

We now reinstate the assumption that both treatment quantity and physician effort are noncontractible. Suppose physicians are allowed to compete for patients. Given the insurance-payment parameters, and the physician's effort, each consumer decides whether to seek help from the physician, and the treatment quantity if he does. Anticipating these reactions from the consumers, a physician picks a profit-maximizing effort level, which determines her demand and each of her patient's treatment quantity. Thus, in contrast to our model in earlier sections (where physician competition is not considered), a set of insurance-payment parameters implements not

[23] Ma (1994) considered the cost and quality incentives of various payment systems. In that model, each physician faces an *upward* sloping demand curve, which is a function of the physician's quality of care.

only the physician's effort and her patient's treatment quantity, but her demand as well.

To present a simple but formal analysis of physician competition, we first extend our model by including in it a population of consumers with total mass normalized to one, each of these consumers having identical utility functions but different out-of-pocket costs for using different physicians (in addition to the copayment). This introduces a "horizontal" product differentiation dimension in the physician-provider market. Let θ_i represent this uninsurable out-of-pocket cost for obtaining service from physician i, and the (distribution) function $\Phi_i(x)$ denote the proportion of consumers who must incur a cost of at most x when they use physician i.

Suppose the insurer has contracted with a set of physicians. For our purpose, it is sufficient to study the strategic choice of a representative physician, say physician i, who is allowed to compete with others. Fixing the insurance-payment parameters and the efforts of all other physicians, we look at the representative physician's incentive to compete for patients by her effort decision.

Now suppose that a consumer who has signed the insurance contract obtains an expected utility of $\bar{U}$ if he does not use physician i.[24] For a given set of insurance-payment parameters and the physician's effort level, the consumer's optimal choice of treatment quantity τ is given by (4). Clearly, the consumer's utility *ex post* (optimized with respect to τ) is increasing in physician effort. Those consumers with values of θ_i below a threshold level will use physician i for service; this threshold $\bar{\theta}_i$ is defined by:

$$(11) \qquad U(w - \alpha - \beta\tau - s + F(\tau, \varepsilon) - \bar{\theta}_i) = \bar{U}.$$

Solving this equation, we express the threshold as an increasing function of ε: $\bar{\theta}_i(\varepsilon)$. Thus, setting her effort at ε, physician i gets $\Phi(\bar{\theta}_i(\varepsilon))$ of all consumers. To simplify notation, denote $\Phi(\bar{\theta}_i(\varepsilon))$ by $D(\varepsilon)$. The increasing function $D(\varepsilon)$ represents the number of consumers who will choose physician i for service when they become sick, or simply the demand function facing the representative physician.[25]

We can now study the incentives for setting effort due to competition. Given the payment parameters, and her patient's *ex post* choice of treatment quantity in (4), the representative physician's profit can be written as:

$$(12) \qquad D(\varepsilon)[\rho + \delta\tau - G(\varepsilon)],$$

where τ is again given by (4). Consider the first-order derivative of (12) with respect to ε:

$$(13) \qquad D'(\varepsilon)[\rho + \delta\tau - G(\varepsilon)] + D(\varepsilon)\left[\delta \frac{d\tau}{d\varepsilon} - G'(\varepsilon)\right].$$

Comparing this with (7), we see that physician competition provides an extra incentive for costly effort: if the physician expects to earn a positive profit from providing service to each patient ($\rho + \delta\tau - G(\varepsilon) > 0$), then the first set of terms of (13) represents the marginal return of effort through its effect of increasing demand.

Let us continue to consider the implementation of a fixed level of effort ε^*; this level can be identified with an appropriately defined second best. We will regard competition as an instrument for the insurer when payment contracts to physicians are drawn up. Two simple and stylized regimes are considered. First, physicians who sign payment contracts with the insurer are free to compete against other participating physicians for patients (who have established insurance contracts with the insurer). Second, physicians are allocated a preassigned set of patients; treating patients

[24] This "reservation" expected utility $\bar{U}$ will be determined endogenously if the equilibrium of the competition game between physicians is constructed. We are not directly concerned with this construction, but with illustrating the role of competition as an incentive mechanism for costly efforts.

[25] The $D(\varepsilon)$ function could be the result of other differences among consumers. For example, consumers may have different reservation utilities of seeing an alternative physician. We have captured consumers' differences by θ for simplicity.

outside her assignment will not lead to payments from the insurer.

Suppose now that treatment and effort are complements: $d\tau/d\varepsilon > 0$. Then, it follows from Proposition 2 that even when $\rho + \delta\tau - G(\varepsilon) = 0$, the value of δ can be so chosen that ε^* can be an equilibrium choice by the physician. Thus, allowing physicians to compete for patients does not improve the allocations; under complements, the implementation of the second best is already achieved by payment parameter δ without the help of competition.[26]

Next, suppose that treatment and effort are substitutes: $d\tau/d\varepsilon < 0$. We have seen from Propositions 2 and 3 that the implementable set may be restricted by $\delta \geq -c$. Suppose that without competition (or alternatively, by setting $\rho + \delta\tau - G(\varepsilon) = 0$), the effort level ε^* is not implementable. Then physician competition together with an appropriate choice of ρ and δ may enable ε^* to be implemented.

Typically, the failure of the implementation of ε^* without competition is due to the lower limit of δ: because of the truth-telling constraint, a physician cannot be made responsible for more than the cost of treatment ($\delta \geq -c$). When ε and τ are substitutes, lowering δ reduces the physician's reward for visits, and therefore increases the incentive to supply effort. Thus, without competition, even when δ is set equal to $-c$, the effort level is still less than ε^*. Competition provides an additional incentive for the physician to increase effort; by raising the value of ρ, the marginal return of effort to attract patients is increased, leading to a higher effort level. More precisely, pick ρ^* such that

$$D'(\varepsilon^*)[\rho^* - c\tau^* - G(\varepsilon^*)] + D(\varepsilon^*)\left[-c\frac{d\tau}{d\varepsilon} - G'(\varepsilon^*)\right] = 0,$$

then ε^* will be implemented. In summary, competition strictly expands the set of implementation efforts.

[26] Allowing physicians to compete for patients (and allowing $\rho + \delta\tau - G(\varepsilon) > 0$) implies that the value of δ will have to be readjusted for the implementation of ε^*.

Notice that the implementation of ε^* by the above method lets the physician earn a significant amount of profit *ex post*. This excess profit *ex post* likely imposes a social cost for implementing ε^*. The payer may need to set up a participation fee *ex ante* for physicians who join the "network" in order to make up for the costs of paying physicians excess profits *ex post*. This "entry" fee allows a physician to compete for patients who have purchased insurance contracts with the payer, and it can be so chosen that a participating physician earns zero economic profits overall.

VI. Discussion

The extant literature on optimal health insurance regards the *ex ante* noncontractibility of quantity of treatment to be the underlying market failure. The choice of level of insurance coverage balances the risk-spreading gains from more insurance against the welfare loss from exacerbating patient moral hazard. It seems clear, however, that the delivery of health care often involves joint decisions by a doctor and a patient.[27] The questions of the optimal insurance system and optimal provider payment system should be answered in a unified model explicitly including doctor and patient interaction. Thus, the issue of insurance and payment design is more complex and involves more than one market failure.

Extending the earlier optimal insurance literature, we consider a number of issues that are both practically and analytically important. First, we assume that insurance-payment contracts are based on reported information, and discover that truthful reporting imposes constraints on the contract parameters. Second, we model physician-patient interaction by a form of demand response: the patient chooses the amount of an input after observing the physician's input (effort). We find that the implementation of desirable input combinations may be possible, but often this may be

[27] Ellis and McGuire (1993), Newhouse (1996), and others have pointed out that provider contracts may limit health care consumption and are alternatives to patient coinsurance for dealing with patient demand-side moral hazard.

prevented by the truthful reporting constraint. Third, we show that when physicians behave ethically, in that they would achieve a minimum level of benefit for a patient, the restriction of the truthful reporting constraint can be relaxed. Fourth, we extend the analysis to consider the effect of competition on insurance-payment contracts, and find that competition as a policy design may alleviate the restriction due to the truthful reporting requirement.

Our work suggests several directions for research. Our paper mainly studied the optimal design of health plans by insurance and payment contracts. When extending it to consider competition as an element of plan design, we assumed that competition had an unrestricted structure; a physician was allowed to accept any patients who chose to enroll. Study of market structure within a model of the effects of choice and competition would be worthwhile. Indeed, our unified insurance-payment framework provides a foundation for the study of policies at an industry level, such as the implementation of "managed competition," and the design of competition between conventional insurance plans, health maintenance organizations, and preferred provider organizations.

We have concentrated on contractual incentives to induce physician effort or quality. As we have observed at the beginning, existing insurance and provider contracts far exceed the complexity of those we have studied in this paper. For example, managed care, utilization reviews, service authorization, second-opinion requirements, gate-keeping, auditing, quality assurances, etc., are commonly observed in the health care industry. Our model must be extended significantly if these arrangements are to be studied carefully. Nevertheless, our framework may provide a foundation from which advances can be made.

We incorporate doctor's "ethical" behavior into a model of treatment determination and optimal payment and insurance, but in a simple way. Other approaches to modelling ethics or altruism are certainly worth pursuing. Also, our perfect information assumption may be relaxed. Introduction of asymmetric information between the doctor and patient is a natural next step. For example, the physician's effort may be interpreted as an *information structure*. In the course of diagnosis, the physician receives a private, informative signal about a patient's illness. Whether the physician has an incentive to reveal this information truthfully and whether she can be motivated to exert costly effort to obtain a more informative signal (in the sense of Blackwell) clearly would depend on the reimbursement mechanism. Our preliminary findings indicate that there is a conflict between motivating costly efforts for informative signals and the truthful revelation of those received signals, resulting in distortions in equilibrium allocations.

APPENDIX

PROOF OF PROPOSITION 3:

Consider a feasible policy α', β', δ', ε', and τ', and suppose that $\delta' > -c$. Let us now consider another feasible policy α, β, δ, ε, and τ, where $-c \leq \delta < \delta'$, and $\beta = \beta'$. We show that the patient's expected utility is higher under this alternative policy.

First, observe that since $\delta < \delta'$, from the comparative statics of τ on δ, (8), we know that $\tau < \tau'$, and that $\varepsilon > \varepsilon'$. Second, we argue that $\alpha < \alpha'$. Consider the first-order derivative (with respect to δ) of the left-hand-side expression of (9):

$$c\frac{\partial \tau}{\partial \delta} + G'(\varepsilon')\frac{\partial \varepsilon}{\partial \delta}$$

$$= \frac{F_{\tau\varepsilon}}{H}\left[-cF_{\tau\varepsilon} + G'(\varepsilon')F_{\tau\tau}\right] > 0,$$

where the equality follows from substituting (8) for the partial derivatives with respect to δ, and the inequality from the fact that $\delta' > -c$, and from the first-order condition for ε in Program A by setting (7) to zero: $\delta' F_{\tau\varepsilon} + G'(\varepsilon')F_{\tau\tau} = 0$. Hence, the left-hand side of (9) falls when δ' is decreased to δ. Now since $\beta \leq c$, it must follow that $\alpha < \alpha'$ as well.

Third we prove that $F(\tau, \varepsilon) > F(\tau', \varepsilon')$. By hypothesis, we have

$$-\frac{F_{\tau\varepsilon}(\tau, \varepsilon)}{F_{\tau\tau}(\tau, \varepsilon)} > -\frac{F_{\varepsilon}(\tau, \varepsilon)}{F_{\tau}(\tau, \varepsilon)}.$$

From (4), we have $d\beta = F_{\tau\tau}d\tau + F_{\tau\varepsilon}d\varepsilon$. Furthermore, if $d\varepsilon > 0$, we have

$$(14) \quad F_{\tau\tau}d\tau + F_{\tau\varepsilon}d\varepsilon > \frac{F_{\tau\tau}}{F_\tau}[F_\tau d\tau + F_\varepsilon d\varepsilon].$$

Since β remains constant, we have

$$\lim_{\delta \to \delta'} F_{\tau\tau}(\tau - \tau') + F_{\tau\varepsilon}(\varepsilon - \varepsilon') = 0,$$

where the derivatives are evaluated at (τ, ε). Hence, from (14) and the facts that $\varepsilon > \varepsilon'$ and $F_{\tau\tau} < 0$, we know that

$$\lim_{\delta \to \delta'} F_\tau(\tau - \tau') + F_\varepsilon(\varepsilon - \varepsilon') = 0,$$

again with the derivatives evaluated at (τ, ε). So for $\delta > \delta'$ and sufficiently close to δ', we have $F(\tau, \varepsilon) > F(\tau', \varepsilon')$. In summary, we have shown that the expected utility has increased.

Thus, comparing the two feasible policies, we know that both the total premium and the patient's *ex post* copayment have decreased, while the benefit has increased. Hence, $\delta' < -c$ cannot be an equilibrium. It follows immediately that the patient's equilibrium expected utility must be less than EU^{SB}.

PROOF OF PROPOSITION 4:

From the hypothesis of the proposition and from Proposition 3, $\delta^* = -c$. Thus, the constraint $\delta \geq -c$ binds. Suppose that the constraint $\delta \geq -c$ can be relaxed to $\delta = -c - \Delta$, $\Delta > 0$. Because the constraint $\delta \geq -c$ binds in the third best, for a sufficiently small Δ, this relaxation, together with a feasible policy $\beta = \beta^*$, α^Δ, ε^Δ, and τ^Δ, must yield an equilibrium with a higher expected utility for the patient, say EU^Δ, than EU^*. Moreover, by the same method in the proof of Proposition 3, we know that with $\delta = -c - \Delta$ and $\beta = \beta^*$, the ε^Δ and τ^Δ belonging to Ω must yield $F(\tau^\Delta, \varepsilon^\Delta) > F(\tau^*, \varepsilon^*) \equiv F^*$. That is, because the isobenefit line $(F(\tau, \varepsilon) = F(\tau^*, \varepsilon^*))$ is steeper than the reaction function $(F_\tau(\tau, \varepsilon) = \beta)$, the point (τ^*, ε^*) (implemented by $\delta = -c$ and β^*) must yield a lower level of benefit compared to $(\tau^\Delta, \varepsilon^\Delta)$ (implemented by $\delta = -c - \Delta$ and β^*).

The same argument is shown in Figure 3. Here, the original equilibrium (without the ethics constraint) is the tangency point between the reaction function $F_\tau = \beta$ and the indifference curve $-c - G(\varepsilon)$. If the constraint $\delta \geq -c$ could be relaxed, then setting $\delta = -c - \Delta$ would implement a point outside Ω, and would increase the patient's expected utility. By hypothesis the isobenefit line F^* is steeper than the reaction function; hence, $\delta = -c - \Delta$ will induce a higher benefit, F^Δ.

We now demonstrate that the patient's expected utility must be at least EU^Δ under the ethics constraint $F(\tau, \varepsilon) \geq F^*$. Consider strengthening the ethics constraint to $F(\tau, \varepsilon) \geq F(\tau^\Delta, \varepsilon^\Delta) > F^*$. Under this stronger ethics constraint, the policy $\beta = \beta^*$, $\tau = \tau^\Delta$, $\varepsilon = \varepsilon^\Delta$, and α satisfying (9) yields an expected utility EU^Δ for the patient. Thus, the patient's equilibrium expected utility must be at least EU^Δ. Now, when the more stringent ethics constraint $F(\tau, \varepsilon) \geq F(\tau^\Delta, \varepsilon^\Delta) > F^*$ is restored to the original $F(\tau, \varepsilon) \geq F^*$, the patient's equilibrium expected utility cannot fall. Thus, under $F(\tau, \varepsilon) \geq F^*$, the patient's expected utility must be at least $EU^\Delta > EU^*$.

It remains to show that the equilibrium coinsurance rate must have increased from β^*. First, observe that any ε and τ belonging to Ω and satisfying the ethics constraint $F(\tau, \varepsilon) \geq F^*$ can only yield an expected utility at most EU^* to the patient. Thus, to achieve an expected utility strictly above EU^*, the equilibrium allocation in the regime with ethics must be outside the implementable set. Second, the ethics constraint must bind, so that the equilibrium allocation must be on the isobenefit line $F(\tau, \varepsilon) = F^*$ and outside the implementable set. By the hypothesis of the proposition, the isobenefit line is steeper than the reaction function. Therefore, the equilibrium allocation must lie on a reaction function corresponding to a higher coinsurance rate.

REFERENCES

Arrow, Kenneth J. "Uncertainty and the Welfare Economics of Medical Care." *American Economic Review*, December 1963, *53*(5), pp. 941–69.

_____. "Agency and the Market," in K. J. Arrow and M. D. Intrilligator eds., *Handbook of mathematical economics, volume*

III. Amsterdam: North-Holland, 1986, pp. 1183–95.

Carter, Grace M.; Newhouse, Joseph P. and Relles, Daniel. "Has DRG Creep Crept Up?" RAND Publication R-4098-HCFA/ProPAC, Santa Monica, CA, 1991.

Ellis, Randall P. and McGuire, Thomas G. "Provider Response to Prospective Payment: Cost Sharing and Supply." *Journal of Health Economics*, June 1986, *5*(2), pp. 129–51.

_____. "Supply-Side and Demand-Side Cost Sharing in Health Care." *Journal of Economic Perspectives*, Fall 1993, *7*(4), pp. 135–51.

Friedson, Elliot. "Prepaid Group Practice and the New 'Demanding Patient'." *Health and Society*, Fall 1993, *51*(4), pp. 473–88.

Gaynor, Martin. "Issues in the Industrial Organization of the Market for Physician Services." *Journal of Economics & Management Strategy*, Spring 1994, *3*(1), pp. 211–55.

Grossman, Michael. "On the Concept of Health Capital and the Demand for Health." *Journal of Political Economy*, March–April 1972, *80*(2), pp. 223–55.

Hsiao, William C. et al. "A National Study of Resource Based Relative Values for Physician Services." Final Report, Harvard School of Public Health, September 22, 1988.

Luft, Harold S. "Modifying Managed Competition to Address Cost and Quality." *Health Affairs*, Spring 1996, *15*(1), pp. 23–38.

Ma, Ching-to Albert. "Health Care Payment Systems: Cost and Quality Incentives." *Journal of Economics & Management Strategy*, Spring 1994, *3*(1), pp. 93–112.

Newhouse, Joseph P. "Reimbursing Health Plans and Health Providers: Efficiency in Production versus Selection." *Journal of Economic Literature*, September 1996, *34*(3), pp. 1236–63.

Newhouse, Joseph P. and the Insurance Experiment Group. *Free for all? Lessons from the RAND health insurance experiment*. Cambridge, MA: Harvard University Press, 1993.

Office of Technology Assessment, U.S. Congress. *Understanding estimates of national health expenditures under health reform*. Washington, DC: U.S. Government Printing Office, May 1994.

Pauly, Mark V. "The Economics of Moral Hazard." *American Economic Review*, June 1968, *58*(3), pp. 531–37.

_____. "Fee Schedules and Utilization," in H. E. Frech III, ed., *Regulating doctors' fees: Competition, benefits, and controls under Medicare*. Washington, DC: AEI Press, 1991, pp. 288–305.

Physician Payment Review Commission. *Annual report to Congress*. Washington, DC: American Medical Association, 1995.

_____. *Annual report to Congress*. Washington, DC: American Medical Association, 1996.

Quint, Michael. "Health Plans Force Changes in the Way Doctors are Paid." *New York Times*, February 9, 1995, p. A1.

Robinson, James C. and Casalino, Lawrence P. "Vertical Integration and Organizational Networks in Health Care." *Health Affairs*, Spring 1996, *15*(1), pp. 7–22.

Shortell, Stephen M.; Gilles, Robin R. and Anderson, David A. "The New World of Managed Care: Creating Organized Delivery Systems." *Health Affairs*, Winter 1994, *13*(5), pp. 46–64.

Trauner, Joan B. and Chestnut, Julie S. "Medical Groups in California: Managing Care Under Capitation." *Health Affairs*, Spring 1996, *15*(1), pp. 159–70.

Wedig, Gerald; Mitchell, Janet B. and Jerry Cromwell. "Can Optimal Physician Behavior Be Obtained Using Price Controls?" *Journal of Health Politics, Policy and Law*, Fall 1989, *14*(3), pp. 601–20.

Zeckhauser, Richard. "Medical Insurance: A Case Study of the Tradeoff between Risk Spreading and Appropriate Incentives." *Journal of Economic Theory*, March 1970, *2*(1), pp. 10–26.