How to Compensate Physicians When Both Patient and Physician Effort are Unobservable

by

Kenneth L. Leonard, Columbia University
Joshua Zivin, School of Public Health, Columbia University

March 2000

Discussion Paper Series No. 9900-1

How To Compensate Physicians When Both Patient and Physician Effort are Unobservable[*]

Kenneth L. Leonard[†] Joshua Zivin[‡]

March 21, 2000

Abstract

In this paper, we construct a joint production model of health with two-sided asymmetric information and ask the question, "How should physicians be compensated?" We demonstrate theoretically that the preferred physician compensation scheme depends on the illness condition. Outcome-contingent payments are better than effort–contingent payments for illnesses in which the efforts of physicians and patients are highly complementary, or in which both types of effort are important to the outcome. Effort-contingent payments are superior when efforts are not highly complementary, or when either physician or patient effort, but not both are important to the outcome. Evidence to support this theory is provided by an empirical analysis of patient choice of health care providers in Africa.

JEL Classification: I1 D8

Keywords: Health Care Organization, Asymmetric Information, Moral Hazard in Teams, Physician Compensation

[*] We wish to thank M Riordan, S Gleid and M Madejewicz for helpful comments.

[†] Department of Economics, Columbia University, Mail Code 3308, 420 W 118th St, New York, NY 10027 (e-mail: KL206@columbia.edu). Research funded by NSF grant #94-22768.

[‡] The Joseph L. Mailman School of Public Health, Division of Health Policy and Management, Columbia University, 600 W 168th St, New York NY 10032

1 Introduction

During the past two decades, academics and policy makers have placed increasing attention on mechanisms to encourage physicians to provide appropriate levels of medical care. Indeed, in the last ten years we have seen managed care exalted as a means to reduce the over-provision of care, and more recently demonized as a mechanism that encourages the under-provision of care. The debate continues, with both sides entrenched and the question remaining, "How should physicians be compensated?"

In this paper, we propose a joint production model of health where both patient and physician behavior affect outcomes. For example, treatment of an asthmatic patient requires the doctor to make the correct diagnosis, prescribe appropriate medications, and describe to the patient any behavioral changes needed to expedite her recovery. Similarly, the patient must take the appropriate medicine and modify her behavior. If the doctor does everything correctly, but the patient leaves the office and immediately begins smoking and does not take her medicine as prescribed, the doctor's effort will be wasted. If the doctor puts forth little effort and misdiagnoses the patient or provides her with inappropriate information, any effort the patient provides will likely be wasted. This concern is greatly exacerbated in an environment where patients cannot evaluate their doctor's activities and doctors cannot evaluate their patient's activities.

This representation of health production in an environment of double-sided asymmetric information is a departure from previous models, which have generally treated the information asymmetry as one-sided (Arrow 1963, Pauly 1980, Ellis and McGuire 1986, Dranove 1988, Reinhardt 1989). It is, however, consistent with commonly raised concerns by physicians about patient compliance and commonly raised concerns by patients about the quality of care they are receiving from their physicians. In this context, the design of contracts that provide each party with the incentives to engage in the appropriate level of effort is very important. With this in mind, we analyze two specific physician compensation schemes: one outcome–contingent and one effort–contingent.

Though double moral hazard has not been discussed in health care, the application of this model has been discussed in other contexts, notably warranties (Cooper and Ross 1985) and general production (Demski and Sappington 1991). Our model of output-contingent contracts is similar to that of Cooper and Ross. The framework of the paper follows the spirit of Weitzman (1975) and Maskin and Riley (1985) in comparing compensation schemes in a 'second-best' framework.

This is not, however, a model constructed to represent all forms of production. The assumptions about information and observability are designed to represent the provision of health care. Effort-contingent payments can be thought of as a traditional fee-for-service compensation scheme, where a third party capable of evaluating the level and appropriateness of care determines the effort-payment schedule. Managed care, with its modified forms of compensation based on physician profiling, utilization review, capitation, and withholds and bonuses tied to utilization procedures can be viewed as a first step toward linking payments and outcomes (Gold et al. 1995). In fact, explicit outcome–contingent contracts for the provision of health care are found outside of the U.S. and are gaining popularity domestically in areas such as reproductive medicine and corrective eye surgery (New York Times 1999, Robertson and Schneyer 1997). The assumptions of the model imply that relevant extensions will be in areas where one of the agents providing effort offers services that can be evaluated by a select group of agents, but not by the population at large. Professional services such as law and accountancy, where trained professionals (who can be evaluated by other professionals) produce a joint product with agents from the general population, come to mind. The purchase of legal services can be made on a per-hour of contingency basis. In the design of law, one can create negligence rules conditioned on precautionary effort, or liability rules where accountability is based on outcomes (see Shavell (1987)).

Our analysis suggests that either compensation regime can achieve the first-best, i.e. full information, solution. In practice, however, these first-best solutions will be difficult to implement. With outcome–contingent contracts it is effectively impossible to give physicians the full incentives to exert effort. Health is a good with extreme valuations and doctors, in general, have a limited liability. Indeed, it is hard to imagine a contract that calls for the execution of a doctor each time a patient dies. Effort-contingent contracts, on the other hand, will generally fail to take proper account of the role of patients in their own health. Optimality requires exertion of physician effort to where its marginal benefit equals marginal cost. If patient and medical effort are complements, the marginal benefit of medical effort will depend on patient effort. Regulators are skilled in determining physician effort, but will have difficulty observing or evaluating patient effort. Their choice of effort thus, cannot always be optimal.

In this world, where outcomes are both the output of interest and observable, economic intuition seems to suggest that outcome–contingent contracts would be superior. However, the best way to compensate physicians depends on the characteristics of the illness being treated. Specifically, when there are large degrees of complementarity between patient and physician effort, compensation should be based on outcomes. When the degree of complementarity is low, compensation should be based on physician effort. In other words, surgery, where short-term success has little to do with patient effort should be compensated based on physician effort, and back pain that relies heavily on the effort of both participants should be based on outcomes. The choice of effort or outcome contingent contracts does not hinge on the importance of unobservable medical effort, but rather on the joint importance of medical and patient effort. This result is worth emphasizing, as the authors are not aware of any academic publication or policy debate that has raised this point.

The theoretical work is then tested using a data set on patient choice of health care providers in Africa. In this setting, patients have no health insurance and can choose which provider to visit for each and every illness condition. Non-governmental health care providers – primarily missions – compensate physicians for their effort, while traditional healers are paid based on outcomes. Using data on the elasticity of health production with respect to unobservable patient effort and unobservable medical effort, our theoretical results are confirmed. Patients with disease conditions that are relatively responsive to patient and practitioner effort are more likely to seek treatment from a traditional healer. When the disease is responsive to either medical or patient effort, though not both simultaneously, patients visit mission centers. The impact of responsiveness of illness conditions on the choice of provider is strong, and patients appear to be willing to incur large costs in order to seek the appropriate provider.

The paper is structured as follows. The next section presents a detailed theoretical model of the joint production of health with two-sided moral hazard. Equilibrium effort levels, utility, and social welfare when payment is effort contingent and when payment is outcome contingent are analyzed. Section three tests our theoretical results using a unique data set from Africa. The final section concludes.

2 A General Model of Health Care

We begin with an individual who has fallen sick from an unknown disease (but a known illness condition, where the illness condition is described by the symptoms of the patient). The given level of health is H. Health intervention might lead to a change in the level of health, ΔH. We simplify the idea of health intervention by assuming that there are only two possible outcomes; the worst outcome $\Delta H = \underline{h}$ and the best outcome $\Delta H = \bar{h}$. These outcomes depend only on the disease condition and not on any characteristics of the patient or the practitioner. We think of $\bar{h}$ as being a full recovery and $\underline{h}$ as being no change in the health status.

The probability of achieving either outcome is determined by two binomial distributions. $\phi^{\star}$ is the 'true diagnosis' distribution and $\phi^{\emptyset}$ is the 'false diagnosis' distribution. We motivate these distributions as follows; if the patient's condition is correctly diagnosed, and the proper treatment regime is prescribed, understood and followed, the patient will have a probability of full recovery of $q^{\star}$. If the diagnosis is incorrect the probability of recovery is $q^{\emptyset}$. The probability of failing to recover is $1 - q^{\star}$ with the 'true diagnosis' and $1 - q^{\emptyset}$ with the 'false diagnosis.' In health, often everything is done as it should be and the patient does not recover. On the other hand, patients frequently recover when nothing has been done for their health (or when incorrect actions have been taken).

Health care is a set of technologies that probabilistically span $\phi^{\star}$ and $\phi^{\emptyset}$. A 'better' technology is one that has a higher probability of choosing the 'correct diagnosis' distribution than another technology. We represent the technology by e $(0 \leq e \leq 1)$ where

$$\Delta H \sim e \cdot \phi^{\star} + (1 - e) \cdot \phi^{\emptyset} \tag{1}$$

The 'best' technology $(e = 1)$ has $q^{\star}$ chance of leading to recovery, and the 'worst' technology $(e = 0)$ leads to a chance of recovery of $q^{\emptyset}$.[1]

The properties of the two binomial distributions are given by the illness condition. The patient cannot choose the distribution under which to seek health care, but she does have some control over the magnitude of health technology (e). e is generally a function of patient effort, patient skill, practitioner effort and practitioner skill. Unobservable efforts imply that the patient does not ever observe e, only whether the outcome was $\bar{h}$ or $\underline{h}$. Since both outcomes are possible with all e the patient can never impute physician effort even if she knows her own level of effort, her own skill and the practitioner skill. Thus, patients can only expect incentive compatible effort which varies according to the means of physician compensation.

[1] We deliberately based this description of ΔH on the Spanning Condition of Grossman and Hart (1983) and the Linear Distribution Function Condition of Hart and Hölmstrom (1987), which will allow us to characterize incentive compatibility constraints as first order conditions or relaxed incentive compatibility constraints.

2.1 The Value of Health

Utility from health can be modeled in a variety of different ways. We follow the basic model of Grossman (1975) and consider health as increasing the hours of time available to consume work and leisure as well as augmenting utility directly. Thus $U = (H, I(H), c(p))$, where H is the health level, $I(H)$ is the income potential at that level of health, p is patient effort and $c(p)$ is the disutility of patient effort. An increase in H leads to an increase in utility through a direct as well as an income effect.

The expected value of health is

$$\begin{aligned} EU &= eq^{\star}\bar{U} + e(1-q^{\star})\underline{U} + (1-e)q^{\emptyset}\bar{U} + (1-e)(1-q^{\emptyset})\underline{U} \\ \bar{U} &= U[\bar{h}, (I(\bar{h}) - C), c(p)] \\ \underline{U} &= U[\underline{h}, (I(\underline{h}) - C), c(p)] \end{aligned} \tag{2}$$

C is the total cost of a visit. We assume a separable utility form such that $U = U'[H, I(H)] - C - c(p)$. Although income and total costs are measured in the same units and need not be separated, we choose this formulation for the following reasons. The income (or earning potential of the patient) and health level for good outcomes is the same whether the patient sought health care or not; it depends on the outcome, not the process. Thus the part of utility inside the utility operator $(U'[H, I(H)])$ depends on the outcome, not on the effort exerted. Costs and disutility have a linear relation to utility. For ease of exposition we write $U'[\bar{h}, I(\bar{h})]$ as $\bar{U}'$ and $U'[\underline{h}, I(\underline{h})]$ as $\underline{U}'$. Thus,

$$EU = \left(e(q^{\star} - q^{\emptyset}) + q^{\emptyset}\right)\bar{U}' + \left(1 - q^{\emptyset} + e(q^{\emptyset} - q^{\star})\right)\underline{U}' - C - c(p) \tag{3}$$

Of interest to the patient is the change in expected utility. We choose as a natural comparison the utility when no health care is sought $(e = 0)$. The change in the expected utility is therefore

$$\Delta EU = e(q^{\star} - q^{\emptyset}) \cdot \left(\bar{U}' - \underline{U}'\right) - C - c(p) \tag{4}$$

At this point we make a number of further simplifying assumptions. First, we assume that $\underline{U}'$ is equal to zero, a simple scaling assumption. Furthermore we assume that utility from health comes from a fixed health affect, $\bar{h} \cdot \bar{w}$ (where $\bar{w}$ is the per unit value of health) and an increased amount of time for leisure or work, $\bar{h} \cdot w$ (where w is the opportunity cost of healthy time.) We cannot separate these two effects and therefore use the combination of effects, $\bar{h} \cdot \omega$ (where $\omega = \bar{w} + w$.) Thus,

$$\Delta EU = e(q^{\star} - q^{\emptyset})\omega\bar{h} - C - c(p)$$

Without loss of generality we define the technology for health production as being a standard production function divided by a 'maximum' level of production for that function, $e = h/\bar{h}$. Thus, where e varies between 0 and 1, h varies between 0 and $\bar{h}$.

$$\Delta EU = (q^{\star} - q^{\emptyset})\omega h - C - c(p) \tag{5}$$

For simplicity we will refer to ΔEU as U.

2.2 The Health Production Technology

The health production technology (h) is viewed as a search for the proper treatment regime. This search is a complex function of a number of different inputs; a production function of health. We assume the following factors are important in the production of health: medical effort, patient effort, medical skill and patient efficiency at transforming health inputs into health. An increase in any of these factors, *ceteris paribus* increases the probability of choosing the 'true diagnosis' distribution. The role of each of these factors will vary according to the illness condition.

The health production technology is represented as a Cobb-Douglas production function.

$$h = \pi p^{\alpha} m^{\beta} \tag{6}$$

where π is the productivity factor, p is the patient effort, α is the elasticity of output with respect to patient effort, m is medical effort and β is the elasticity of output with respect to medical effort. The productivity factor is an increasing function of the skill of the practitioner and the skill of the patient (efficiency of the patient in transforming health inputs into health). We will not specify a functional form for π, but it is increasing in both medical and patient skill. There are decreasing returns to scale in the production of health and therefore we assume that $0 < \alpha < 1$, $0 < \beta < 1$ and $0 < \alpha + \beta < 1$. For simplicity of notation we will refer to the product of productivity, the value of health and the difference in probability with full effort and with no effort ($\pi\omega(q^{\star} - q^{\emptyset})$) as A. This variable can be thought of conceptually as the value of obtaining health care, a measure that embodies the benefits from being healthy and the ability of the practitioner (relative to letting the disease run its natural course) to provide that health. We assume that disutility of effort is a linear function of the effort, and normalize the coefficient for patient effort to one, with a coefficient of D for practitioner effort.[2]

2.3 Production with Full Information

As a basis of comparison for the cases with asymmetric information, we will first analyze the utility maximization for the case with full information. This case corresponds to a world where both the practitioner and the patient observe the other's

[2] This is more general than it would seem at first. Consider the standard model of Cobb-Douglas with exponential disutilities (Bhattacharyya and Lafontaine 1995) where social welfare would be represented by $Ap^{\alpha}m^{\beta} - \frac{\delta p^{k}}{k} - \frac{\gamma m^{l}}{l}$. Since patient and medical effort and their responsivenesses have no intuitive units we can arbitrarily define a new variable as follows: $p = p'^{\frac{1}{k}}; m = m'^{\frac{1}{l}}; \alpha = k\alpha'; \beta = l\beta'$ with the result : $S = Ap'^{\beta'}m'^{\alpha'} - \frac{\delta}{k}p' - \frac{\gamma}{l}m'$. Defining $A = A'\frac{\delta}{k}$ and $D = \frac{\gamma}{l}\frac{k}{\delta}$ we obtain $S = \left(Ap'^{\beta'}m'^{\alpha'} - p' - Dm'\right)\frac{\delta}{k}$ which is functionally the same as the equation we will work with.

effort and there is no coordination problem. Social welfare is the utility from health net of effort costs.

$$Ap^{\alpha}m^{\beta} - p - Dm. \tag{7}$$

Maximizing welfare with respect to p an m we obtain:

$$\alpha Ap^{\alpha-1}m^{\beta} - 1 = 0 \tag{8a}$$

$$\beta Ap^{\alpha}m^{\beta-1} - D = 0 \tag{8b}$$

These conditions simply state that the marginal productivity of each input equals the marginal cost of that input. Together, these first order conditions allow us to define an optimal level of patient and practitioner effort that are simply a function of the value of health care, the marginal productivities of effort, and the costs of effort.

$$p^{\star}_{\text{FI}} = \alpha \left(A\alpha^{\alpha}\left(\beta/D\right)^{\beta}\right)^{\frac{1}{1-\alpha-\beta}} \tag{9a}$$

$$m^{\star}_{\text{FI}} = \left(\beta/D(A\alpha^{\alpha}\left(\beta/D\right)^{\beta}\right)^{\frac{1}{1-\alpha-\beta}} \tag{9b}$$

The subscript FI denotes the full information solution. These expressions for optimal effort levels can then be employed to determine social welfare and practitioner and patient utility.[3]

$$U_{\text{FI}} = (1-\alpha)\left(A\alpha^{\alpha}\left(\beta/D\right)^{\beta}\right)^{\frac{1}{1-\alpha-\beta}} \tag{10a}$$

$$W_{\text{FI}} = (1-\alpha-\beta)\left(A\alpha^{\alpha}\left(\beta/D\right)^{\beta}\right)^{\frac{1}{1-\alpha-\beta}} \tag{10b}$$

It should be clear that social welfare under any regime with informational asymmetries can at best be equivalent to social welfare with perfect information. In the

[3] We assume that patients retain the full value of their health, minus the disutility of their effort and a fixed fee (which we drop for notational simplicity). This derivation of utility makes the most sense in the health context (where fees are generally fixed). Social welfare more accurately reflects the surplus created in a general context.

regimes that follow, we will concentrate our analysis on both patient utility and social welfare. The reasons for the additional focus on patient utility are twofold. First, patient welfare is often the impetus behind the design, or redesign, of health care institutions and regulations. Second, our empirical analysis will examine patient choice of practitioner, a comparison in which patient utility and not social welfare drives behavior.

2.4 Joint Production with Dual Unobservable Effort

Here patients cannot observe practitioner's effort and vice versa, i.e. a world with joint production and double-sided asymmetric information. We introduce a social planner who can implement contracts and whose goal is to maximize welfare. The social planner can observe medical effort and outcomes, but not patient effort. One can think of the social planner as a regulator with the resources and the skills to observe and evaluate the activities of practitioners. Given these abilities, the social planner can design an effort–contingent payment system. Alternatively, the social planner can design contracts that reward both the patient and the practitioner according to outcomes. We will show that in a second-best world, where effort–contingent contracts cannot incorporate behavioral responses of patients or where physicians and/or patients do not retain the full value created by health outcomes, neither payment scheme is uniformly superior. Social welfare is maximized through outcome–contingent contracts for some illnesses and through effort–contingent contracts for others.

2.5 Effort—Contingent Contracts

The social planner, since he can observe medical effort, can choose the level of medical effort. However, he cannot observe patient effort. When medical and patient effort are complements (as they are with our choice of production function) the optimal level of medical effort depends on the level of patient effort and therefore the social planner will not be able to force the practitioner to exert the socially optimal level of effort. We model the lack of information as a social planner who sets m assuming patient effort is invariant to medical effort. The social planner does know that patient effort is important in health, but does not know (or cannot model) how patient effort reacts to medical effort. The social planner maximizes welfare assuming $p = \bar{p}$.

$$\max_{m} A\bar{p}^{\alpha}m^{\beta} - \bar{p} - Dm \tag{11}$$

The first order condition for the social planner's maximization problem is:

$$\beta A\bar{p}^{\alpha}m^{\beta-1} - D = 0 \tag{12}$$

The marginal productivity of medical effort, evaluated at the social planner's estimate of patient effort, $\bar{p}$, is equal to the marginal cost of medical effort. However, the patient does not stay idle. The patient responds to practitioner effort through her choice of effort. The reaction function of the patient is determined by the first order condition of the maximization problem above with respect to p.

$$\alpha Ap^{\alpha-1}m^{\beta} - 1 = 0 \tag{13}$$

This expression is identical to equation (8a) in the full information context. Patient effort will be provided to a point where the marginal productivity of that effort, taking into consideration the level of medical effort, is equal to the marginal cost of patient effort. We can combine these first order equations to determine optimal expressions for patient and practitioner effort, patient utility and social welfare (equations (14a

-14d).)

$$p_{\mathrm{E}}^{\star} = (A\alpha^{1-\beta}(\beta/D)^{\beta}\bar{p}^{\alpha\beta})^{\frac{1}{(1-\beta)(1-\alpha)}} \tag{14a}$$

$$m_{\mathrm{E}}^{\star} = (A(\beta/D)\bar{p}^{\alpha})^{\frac{1}{1-\beta}} \tag{14b}$$

$$U_{\mathrm{E}} = U_{\mathrm{FI}}(\frac{\bar{p}}{p_{\mathrm{FI}}^{\star}})^{\frac{\alpha\beta}{(1-\alpha)(1-\beta)}} \tag{14c}$$

$$W_{\mathrm{E}} = \left(1 - \alpha - \beta(\frac{\bar{p}}{p_{\mathrm{FI}}^{\star}})^{\frac{\alpha(1-\alpha-\beta)}{(1-\alpha)(1-\beta)}}\right)\left(A\alpha^{\alpha}\left(\beta/D\right)^{\beta}\right)^{\frac{1}{1-\alpha-\beta}}(\frac{\bar{p}}{p_{\mathrm{FI}}^{\star}})^{\frac{\alpha\beta}{(1-\alpha)(1-\beta)}} \tag{14d}$$

The subscript E denotes the effort–contingent solution. U_{FI} is the full information utility and $p_{\mathrm{FI}}^{\star}$ is the level of patient effort under the full information solution. When $\bar{p} = p_{\mathrm{FI}}^{\star}$ the full information solution will obtain. When $\bar{p} < p_{\mathrm{FI}}^{\star}$, practitioner effort decreases at the expense of patient utility. The practitioner is not working hard enough. When $\bar{p} < p_{\mathrm{FI}}^{\star}$, practitioner effort increases to the benefit of patient utility. The practitioner is working too hard.

Social welfare under effort–contingent contracts is equal to full information social welfare when $\bar{p} = p_{\mathrm{FI}}^{\star}$. When $\bar{p} > p_{\mathrm{FI}}^{\star}$ or when $\bar{p} < p_{\mathrm{FI}}^{\star}$ welfare under the effort based contracts is strictly less than welfare under full information. Patient utility can be greater than under full information because the patient does not have to compensate the practitioner for working too hard. However, welfare can never be greater than full information welfare.

2.6 Outcome-Contingent Payments

In this case, the medical practitioner and the patient receive payment as a function of output. We call the share to the patient s_p and the share to the practitioner s_m. The social planner seeks to maximize welfare choosing the levels of the shares. The optimal solution will be the case where each share is equal to 1, where both patient and practitioner face the full incentives. In general the shares will not be equal to one. As we have stated, it would be difficult to set the practitioner share to 1 since this would imply that he would experience the same disutility from cancer, for example, that the patient experiences. The patient share may also be bounded away from one due to risk sharing arrangements, such as disability, life or health insurance. The social planner will set these shares at their highest possible levels, but they will not generally be one. Given the share, the patient seeks to maximize her utility, which is now represented as

$$s_p A p^\alpha m^\beta - p \tag{15}$$

Similarly, the practitioner will maximize his utility, which can be represented as

$$s_m A p^\alpha m^\beta - Dm \tag{16}$$

The two agents play a Nash game and equilibrium is found where each player's choice of effort is equal to the other players expectation. The social planner plays no role beyond choosing the shares and implementing the terms of the contract. He does not need to observe m and cannot observe p. Equations (17a - 17d) are the equilibrium levels of patient effort, medical effort, patient utility and social welfare for outcome–contingent contracts.

$$p^\star_{\mathrm{O}} = s_p\alpha\left(A\alpha^\alpha(\beta/D)^\beta\right)^{\frac{1}{1-\alpha-\beta}}(s_p^\alpha s_m^\beta)^{\frac{1}{1-\alpha-\beta}} \tag{17a}$$

$$m^\star_{\mathrm{O}} = s_m(\beta/D)\left(A\alpha^\alpha(\beta/D)^\beta\right)^{\frac{1}{1-\alpha-\beta}}(s_p^\alpha s_m^\beta)^{\frac{1}{1-\alpha-\beta}} \tag{17b}$$

$$U_{\mathrm{O}} = U_{\mathrm{FI}} s_p (s_p^\alpha s_m^\beta)^{\frac{1}{1-\alpha-\beta}} \tag{17c}$$

$$W_{\mathrm{O}} = (1 - s_p\alpha - s_m\beta)\left(A\alpha^\alpha(\beta/D)^\beta\right)^{\frac{1}{1-\alpha-\beta}}(s_p^\alpha s_m^\beta)^{\frac{1}{1-\alpha-\beta}} \tag{17d}$$

The subscript O denotes the outcome–contingent solution. Note that $s_p(s_p^\alpha s_m^\beta)^{\frac{1}{1-\alpha-\beta}}$ becomes 1 when s_m and s_p are both equal to 1; each participant receives the full rewards for their effort. In this case the full information solution obtains. When either $s_m < 1$ or $s_p < 1$ or both then $s_p(s_p^\alpha s_m^\beta)^{\frac{1}{1-\alpha-\beta}}$ is less than one. Either the patient or the practitioner, or both, do not face the full incentives to provide effort, so they each under–provide it. Patient utility and social welfare under outcome–

contingent contracts are inferior to patient utility and social welfare under the full information solution.

2.7 Effort– vs. Outcome–Contingent Payments

Now we are ready to compare patient utility across regimes. The difference in utilities can be represented as follows:

$$U_{\mathrm{O}} - U_{\mathrm{E}} = U_{\mathrm{FI}} \left(s_p (s_p^{\alpha} s_m^{\beta})^{\frac{1}{1-\alpha-\beta}} - (\frac{\bar{p}}{p_{\mathrm{FI}}^{\star}})^{\frac{\alpha\beta}{(1-\alpha)(1-\beta)}} \right) \tag{18}$$

We cannot sign this expression without knowledge of $\bar{p}$. Clearly if $\bar{p}$ is very small then $U_{\mathrm{O}} > U_{\mathrm{E}}$. On the other hand, if $\bar{p}$ is very large then $U_{\mathrm{O}} < U_{\mathrm{E}}$. We do not know $\bar{p}$ a priori, but if $\bar{p}$ is fixed we can determine the conditions under which U_{O} is most likely to be greater than U_{E} and when U_{O} is least likely to be greater than U_{E}. Define $\hat{p}$ as the value for $\bar{p}$ when patient utility is equivalent in both regimes. Thus, by construction, when $\bar{p} > \hat{p}$, the expression above is negative and patient utility is larger when physician compensation is effort–contingent. When $\bar{p} < \hat{p}$, the opposite is true. $\bar{p}$ is fixed and therefore U_{O} is more likely to be greater (less) than U_{E} when $\hat{p}$ is larger (smaller). The magnitude of $\hat{p}$ depends on the nature of the disease condition, specifically the elasticity of health production with respect to patient and practitioner effort. Therefore, regime performance can be characterized through an analysis of changes in $\hat{p}$ with respect to α and β.[4]

[4] Note that a given illness condition is defined by α and β and therefore α and β do not change. Changes in α and β reflect comparisons between illness conditions.

$$\hat{p} = p^\star_{\mathrm{O}} \left(s_p(\frac{s_m}{s_p})^\beta \right)^{\frac{1}{\alpha\beta}} \tag{19a}$$

$$\frac{\partial \hat{p}}{\partial \alpha} = \frac{\hat{p}}{1-\alpha-\beta} \left(\ln p^\star_{\mathrm{O}} + \frac{1-\beta}{\alpha} - \frac{(1-\alpha-\beta)\ln\left(s_p(\frac{s_m}{s_p})^\beta\right)}{\alpha^2\beta} \right) \tag{19b}$$

$$\frac{\partial \hat{p}}{\partial \beta} = \frac{\hat{p}}{1-\alpha-\beta} \left(\ln m^\star_{\mathrm{O}} - \frac{(1-\alpha-\beta)\ln s_p}{\beta^2\alpha} + 1 \right) \tag{19c}$$

$$\begin{aligned} \frac{\partial^2 \hat{p}}{\partial\alpha\partial\beta} &= \frac{\hat{p}}{(1-\alpha-\beta)^2} \\ &\left(\begin{array}{c} \left(\ln p^\star_{\mathrm{O}} - \frac{(1-\alpha-\beta)\ln\left(s_p(\frac{s_m}{s_p})^\beta\right)}{\alpha^2\beta} + 1 \right) \left(\ln m^\star_{\mathrm{O}} - \frac{(1-\alpha-\beta)\ln s_p}{\beta^2\alpha} + 1 \right) \\ + \frac{1-\beta}{\alpha}(\ln m^\star_{\mathrm{O}} + 1) + \ln p^\star_{\mathrm{O}} + 1 \end{array} \right) \end{aligned} \tag{19d}$$

$s_p(\frac{s_m}{s_p})^\beta$ and s_p are always less than one. If $p^\star_{\mathrm{O}}$ and $m^\star_{\mathrm{O}}$ are greater than one, all three derivatives above are positive. Inputs with values greater than one is the standard Cobb-Douglas assumption, but takes on special meaning in this context.[5] In this model, the level of inputs supplied is endogenous, so we cannot assume that patient effort and medical effort are greater than one, but must examine the conditions necessary for this result to obtain. Ensuring that $p^\star_{\mathrm{O}}$ and $m^\star_{\mathrm{O}}$ are greater than one simply requires that seeking health care is valuable relative to the costs of effort.[6] If this were not the case, one would imagine that the health care market for this disease would not arise. For example, patients do not generally seek medical care for a bruised elbow because the benefit to jointly producing health with a physician is not worth the effort. Therefore, if health care is worth seeking, $\frac{\partial \hat{p}}{\partial \alpha}$, $\frac{\partial \hat{p}}{\partial \beta}$, and $\frac{\partial^2 \hat{p}}{\partial\beta\partial\alpha}$ are all positive.

The signs of the $\hat{p}$ derivatives imply that utility in the outcome–contingent regime

[5] This assumption is standard because when the inputs are less than one, increases in the productivity of an input yields lower levels of output. This peculiar property occurs because fractions raised to a higher power produce smaller numbers.

[6] A must be 'large' compared to both 1 and D. Since A has no directly measurable units, but is meant to capture value, 'large' means that the value of health care exceeds the effort costs. When A is 'large' increasing the elasticity of outcomes with respect to either effort increases the utility of the patient. In other words, when medical effort (for example) is more productive, patient utility is improved.

is most likely to exceed utility in the effort–contingent regime when α and β are both large. In other words, outcome–contingent payment schemes are best for disease conditions when both physician and patient effort are productive. The intuition is straightforward. When both productivities are high, a feedback mechanism is necessary so that one agent's effort encourages provision by the other. This feedback is achieved by conditioning payments on outcomes, which are, of course, a result of joint effort. When physician effort is productive, but patient effort is not, payment on physician effort is sufficient. When patient effort is productive, but physician effort is not, the compensation scheme of the practitioner is unimportant when patients face the full incentives, which they do under effort–contingent contracts but do not under outcome–contingent contracts.[7]

Proposition 1 *In a second-best world, the physician compensation scheme preferred by patients depends on the illness condition. Outcome-contingent payments are better than effort–contingent payments for illnesses where the marginal productivities of both patient and physician effort are high. Effort-contingent payments are better than outcome–contingent payments for illnesses where the marginal productivity of medical or patient effort is high, but not both.*

The proof is above. Second-best, in this and all subsequent references, implies $s_p < 1$ or $s_m < 1$ and $\bar{p}$ constant. The welfare implications are similar, though not as straight-forward to illuminate.

$$W_{\mathrm{O}} - W_{\mathrm{E}} = \frac{W_{\mathrm{FI}}}{1-\alpha-\beta}$$
$$\left((1 - s_p\alpha - s_m\beta)\left(s_p^{\alpha} s_m^{\beta}\right)^{\frac{1}{1-\alpha-\beta}} - \left(1-\alpha-\beta\left(\frac{\bar{p}}{p^{\star}_{\mathrm{FI}}}\right)^{\frac{\alpha(1-\alpha-\beta)}{(1-\alpha)(1-\beta)}}\right)\left(\frac{\bar{p}}{p^{\star}_{\mathrm{FI}}}\right)^{\frac{\alpha\beta}{(1-\alpha)(1-\beta)}} \right) \tag{20}$$

When health care is valuable ($m^{\star}_{\mathrm{FI}} > 1$ and $p^{\star}_{\mathrm{FI}} > 1$) both full information welfare (W_{FI}) and patient effort ($p^{\star}_{\mathrm{FI}}$) are increasing in both α and β. When $\bar{p} = p^{\star}_{\mathrm{FI}}$ (the

[7] Note that the above holds true when both $s_p < 1$ and $s_m < 1$. If $s_p = 1$ then outcome–contingent contracts will be superior to effort–contingent ones in this case.

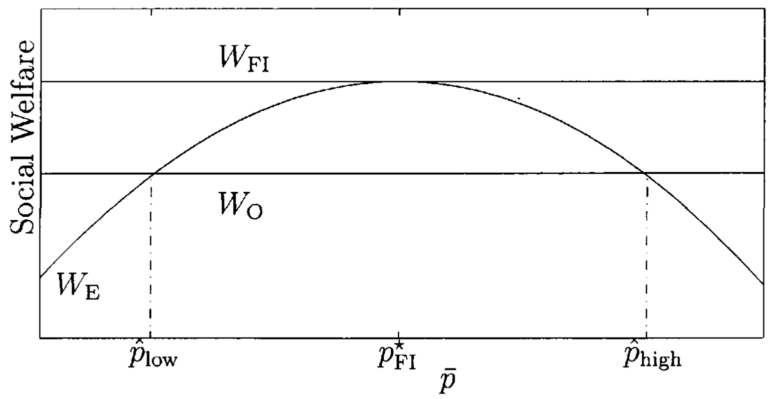

Figure 1: Full Information, Outcome–Contingent and Effort–Contingent Welfare as $\hat{p}$ Changes

social planner happens to have guessed patient effort correctly), welfare under effort–contingent contracts is equal to welfare in the full information case ($W_{\mathrm{E}} = W_{\mathrm{FI}}$). Welfare under outcome–contingent contracts (W_{O}), on the other hand, is never equal to full information welfare (W_{FI}) if $s_m < 1$ or $s_p < 1$. Thus, when $\bar{p} = p^{\star}_{\mathrm{FI}}$, welfare under outcome–contingent contracts is less than welfare under effort–contingent contracts ($W_{\mathrm{O}} < W_{\mathrm{E}}$).

Figure 1 represents full information, outcome–contingent and effort–contingent welfare as $\bar{p}$ changes. Neither W_{FI} nor W_{O} change with $\bar{p}$. As discussed earlier, outcome–contingent welfare is always less than welfare under full information. Effort–contingent welfare is equal to full information welfare at the point where $\bar{p} = p^{\star}_{\mathrm{FI}}$. However, welfare under the effort contingent contract is strictly less than full information welfare ($W_{\mathrm{E}} < W_{\mathrm{FI}}$) at all points where $\bar{p} \neq p^{\star}_{\mathrm{FI}}$, both $\bar{p} < p^{\star}_{\mathrm{FI}}$ and $\bar{p} > p^{\star}_{\mathrm{FI}}$. Thus outcome–contingent welfare (W_{O}) can potentially be greater than effort–contingent welfare (W_{E}) both when $\bar{p}$ is small and when $\bar{p}$ is large. For small values of $\bar{p}$, physicians provide insufficient effort which yields low levels of health production and, in turn, low levels of social welfare. When $\bar{p}$ is very large, health production levels are high but at the expense of physician effort, again yielding low levels of social welfare. These points where the regime producing greater social welfare switches are represented on Figure 1 as $\hat{p}_{\text{LOW}}$ and $\hat{p}_{\text{HIGH}}$. For this reason there is not a unique counterpart to $\hat{p}$ as there was in the case of patient utility; there will, in general, be two $\hat{p}$'s.

If we examine $W_{\text{O}} - W_{\text{E}}$ at the point $\hat{p}$ (where $U_{\text{O}} = U_{\text{E}}$) we obtain the following

$$W_{\text{O}} - W_{\text{E}}\,(U_{\text{O}} = U_{\text{E}}) = \frac{W_{\text{FI}}}{1-\alpha-\beta}(s_p^{\alpha} s_m^{\beta})^{\frac{1}{1-\alpha-\beta}} \left(1 - s_m\beta - s_p + \beta s_m s_p^{\frac{1}{\beta}}\right) \qquad (21)$$

$1 - s_m\beta - s_p + \beta s_m s_p^{\frac{1}{\beta}}$ is always greater than 0 when $s_p < 1$.[8] Thus, expression 21 is always positive.

Proposition 2 *In a second-best world, when outcome–contingent patient utility is equal to effort–contingent patient utility, outcome–contingent welfare is superior to effort–contingent welfare.*

The proof is above. Furthermore,

Proposition 3 *In a second-best world, when outcome–contingent patient utility is greater than effort–contingent patient utility, outcome–contingent welfare is always superior to effort–contingent welfare.*

For proof see the appendix.

Figure 2 is a graphical representation of the relationship between utilities and welfares in the outcome–contingent and effort–contingent worlds. Figure 2 is a two dimensional representation of all possible α and β pairs. Any illness condition is represented by a point on the graph. Shown are two isoquants; the iso-utility line which represents all illness conditions for which the patient utility under outcome- and effort–contingent contracts are equal (for a given A, D, s_m, s_p and $\bar{p}$) and the iso-welfare line representing all points where welfare is equal across the two regimes.

[8] When $s_p = 1$, $1 - s_m\beta - s_p + s_p^{\frac{1}{\beta}}\beta s_m = 0$, and for all s_p $\frac{\partial\left(1-s_m\beta-s_p+s_p^{\frac{1}{\beta}}\beta s_m\right)}{\partial s_p} = s_m s_p^{\frac{1-\beta}{\beta}} - 1$, which is always negative. In the limit, as s_p approaches 1, $1 - s_m\beta - s_p + s_p^{\frac{1}{\beta}}\beta s_m$ approaches 0 from above, and is therefore always greater than 0.

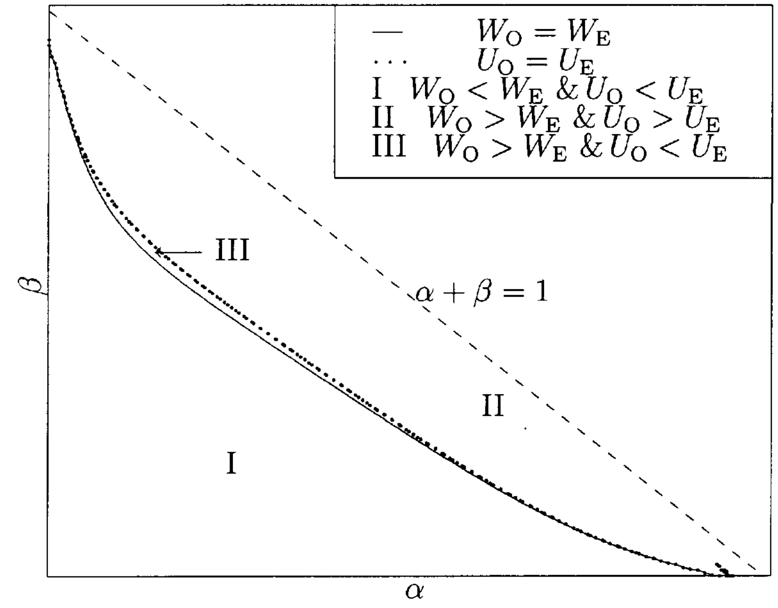

Figure 2: Iso-Utility and Iso-Welfare lines with Nash and Blind Social Planner Contracts

The fact that these two lines are close to each other in $\alpha\beta$ space is a robust result and does not depend on our choice of A, D, s_m, s_p and $\bar{p}$. For any illness condition in region II both utility and welfare with outcome–contingent contracts are superior to that with effort–contingent contracts. In region I, both utility and welfare are superior with effort–contingent contracts. Region III represents the area where welfare is superior with outcome–contingent contracts, but utility is superior with effort–contingent contracts. We do not know whether any specific area of this figure is populated with an actual illness condition (there might not exist any illness condition for which $\alpha = 0.95$ and $\beta = 0.04$, for example), but we can see that conditions for which both α and β are large simultaneously are more likely to populate the area in which outcome–contingent welfare and utility are superior to effort–contingent welfare and utility.

We have been able to derive unambiguous results for patient utility and show that social welfare closely follows these results. Whenever health care is worth seeking, illness conditions with a high α and β are likely to lead to higher social welfare and patient utility if they are treated under an outcome–contingent regime. When both α and β are high, conditions exhibit a high degree of effort complementarity, where complementarity implies that both efforts are necessary for the treatment of the illness condition. On the other hand, for illness conditions in which either α or β is large, but not both, social welfare and patient utility are higher under effort–contingent regimes. Here efforts do not exhibit high degrees of complementarity. One effort or the other is necessary, but not high levels of both.

3 Empirical Evidence

We submit that it is surprising that we can test this theory at all. We will not be able to obtain measures of social welfare, as is almost always the case with empirical work. Even patient utility depends crucially on the patient's valuation of her own health at both outcomes, information that is difficult to collect reliably. However if patients have expectations of the outcome (and the value of this outcome) we should be able to observe some sort of revelation of preferences if patients choose practitioners for each illness condition. In other words, if one contract is "better" than another in some cases, we should observe that patients are more likely to seek that contract in those cases than otherwise. In the South-West province of Cameroun, and indeed in most of Africa, patients have a wide variety of choices available to them and we observe that patients exercise this choice. The same individual is likely to visit a wide variety of practitioners over her lifetime and patients claim that they evaluate the condition from which they suffer before they choose a practitioner to visit. Of particular interest to this paper is that two of the practitioners available use the contracts described in the theoretical section of this paper. We will briefly discuss some important features of the data.

Traditional healers in Africa in general, and specifically in Cameroun, are paid only if the patient is cured (after a fixed fee). Thus, traditional healers offer an outcome–contingent contract. The value of outcome is shared between healer and patient according to a sharing rule, such that $s_p + s_m = 1$.[9] Implicit or explicit contracts are negotiated between patients and healers before the healer diagnoses the patient. Although patients often pay healers very little, when they are cured payments can be substantial. Healers feel no obligation to accept every patient though they refuse patients infrequently. As part of the data collection effort described below extensive interviews were conducted with a few healers. These interviews are discussed in Leonard (1998).

The second type of provider of interest to this work is the church–operated health system (hereafter referred to as the mission). Missions are seen as the "high quality" providers in this area. Patients pay a fixed fee to the mission and practitioners are monitored and compensated by their employers. If their effort levels are deemed to be within protocols, they are financially rewarded. If they are deemed below accepted standards, they are punished. Practitioner compensation at this institution is effort contingent. There are two types of mission health facilities available to patients in our data set, clinics and hospitals. Clinics and hospitals fall under the same hierarchical management within churches and we therefore assume that the contracts offered at these two locations is similar. Providers at clinics and hospitals have different levels of skill.

To test our theory we use data on individual choices of practitioner collected in Mbonge Sub-Division, in the South-West province of Cameroun in 1994. Forty villages were randomly chosen and twenty randomly selected households from each village were interviewed. Data were collected on all members of the household. 4,489 individuals were thus polled, and 681 illness episodes were reported within the month previous to the survey. Of primary interest to this work was the first location visited in the search for care and 252 of these episodes resulted in first visits to either traditional healers, mission clinics or mission hospitals. The other major source of health care is

[9] This is an example of a balanced budget contract. In the absence of a third party who can credibly inject or remove output, all contracts between two parties must be balanced: payments from one party must equal the payments received by the other party. This has a very important effect on contracts with dual unobservable effort and implies that the full information solution can never be obtained (Hölmstrom 1982).

the government health system (289 visits) with drug peddlers, pharmacists, neighbors, private hospitals, private clinics and parastatal hospitals rounding out the sample.[10]

Despite its wealth (relative to other areas of Cameroun and Africa) and the importance of commerce, roads in this area are terrible. There is only one all weather road, and many of the villages surveyed are far from roads with any transportation infrastructure. Nevertheless, we observe significant bypassing of facilities. To compensate for vastly different road conditions we use the transport cost per kilometer to normalize all distances to the distance your fare would take you on the main (paved) road in the sample area. Nearly 80% of all visits were to a provider who was not the closest provider, suggesting a strong revealed preference for the care that is available there.

We suggest that it is information that patients possess about the illness from which they suffer that drives them to incur significant cost in the search of care. The survey polled respondents on the characteristics of the episode from which they suffered: all of the symptoms they experienced; the self–declared severity of the disease; the number of days sick before seeking care; and the number of those days in which the patient was bedridden. With these characteristics of the disease plus the age and sex of the individual and information about endemic diseases in the area (but not information on the choice of provider or the diagnosis), two doctors and one nurse[11] (all experienced in rural tropical medicine) independently scored all the cases using the following definitions:

Responsiveness of the condition to Medical Effort The degree to which outcome depends on the effort of the practitioner. This is our estimate of β.

Responsiveness of the condition to Patient Effort The degree to which outcome depends on the effort of the patient. This is our estimate of α.

[10] All of the regressions reported below were also run with the government as a third type of institution from which patients could choose and none of the coefficients on the choice between traditional healers and missions were significantly affected.

[11] We are indebted to Dr. Hailemariam, Dr. Djomand and Ms. Pouani for their assistance in this endeavor.

Responsiveness of the condition to skill Patients can choose between three levels of skill and capacity: untrained or informally trained providers (corresponding to traditional healers), providers at clinics and providers at hospitals. This variable represents three data points for each illness condition. This is our estimate of π, which is a major component of A.

Outcome Range What is the possibility for a very bad health outcome given the disease from which the patient suffers? This is an estimate of $(q^{\star} - q^{\emptyset})$, an important element of A.

In addition to these three sets of scores, we created scores for each case using basic medical references (Griffith 1985, Strickland, ed 1984, Werner 1977).[12]

The codings are correlated, albeit not perfectly. The key informational content in these codings is ordinal ranking of disease properties. Creating an average score from the four codings necessarily treats them as cardinal rankings and unnecessarily discards valuable information. Taking the average, therefore makes little sense and rather than merge these four data points, we treat them as four different data sets and test each one individually.

3.1 Estimation

Patients choose providers on the basis of the expected utility at that provider, minus fixed costs and travel costs. Expected utility will be affected by the contract under which medical and patient effort are delivered as well as the skill of the provider in question. The fixed costs are constant and are therefore not a source of variation, but travel costs differ significantly. We know the distance to the nearest mission clinic and hospital for each individual but we do not know the distance to the nearest traditional healer. We know that there are many healers and that they are widely dispersed and therefore assume that travel costs to traditional healers are zero.

[12] The data for this survey as well as correlation tables of the scores discussed above are available online at the Inter-university Consortium for Political and Social Research at the University of Michigan (Study # 1138).

Individuals choose between two types of providers and three locations. Types (indexed by k) are traditional (TH) – outcome–contingent payment – and missions (M) – effort–contingent payment. The locations (indexed by j) are traditional (TH), mission clinic (MC), and mission hospital (MH). Thus k=TH if j=TH and k=M if j=MC or MH. Coefficients are obtained by maximizing the following log likelihood with respect to η, γ and ρ.

$$\log L = \sum_{i=1}^{n} \sum_{j \in \{\text{TH,MC,MH}\}} \delta_{ij} \log P_{ij} \tag{22a}$$

$$P_{ij} = \frac{\exp(\eta_j' x_i + \gamma' y_{ij} + \rho_k' z_i)}{\sum_{m \in \{\text{TH,MC,MH}\}} \exp(\eta_m' x_i + \gamma' y_{im} + \rho_k' z_i)} \tag{22b}$$

$\delta_{ij} = 1$ if the i'th individual visits provider j and 0 otherwise.

x is a vector of characteristics of the individual. There is only one vector per individual, but there are three sets of coefficients, representing the three locations between which a patient can choose.[13] x includes individual income, household wealth[14], years of schooling and a dummy variable for whether or not the patient is an adult. Thus, for example, any patient has only one level of income, but income has a potentially different effect at each of the three providers. The characteristics of individuals are included to control for the possibility that the observed bypassing is done by only a select group of individuals and is not a function of the illness condition: a hypothesis the results reject. y is a vector of information about the locations visited. The data varies across providers but the coefficient does not.[15] y includes the travel cost to each provider and the skill of the provider for the illness condition reported. Thus, while each provider potentially has a different travel cost the effect of travel cost is the same at each provider; for this variable two providers each 100 kms from the patient

[13] This is the standard multinomial logit framework.

[14] To get a measure of household wealth we estimated total household income and regressed this on observable characteristics of the household (employment type, construction of primary residence, ownership of consumer durables, etc.) and used the predicted household income as a measure of household wealth.

[15] This corresponds to the McFadden Conditional Logit.

are treated as the same. z is a vector of information about the illness condition and is therefore only one vector of information with two sets of coefficients representing traditional healers and missions. z includes the elasticity of the given condition to patient effort (α), the elasticity with respect to medical effort (β), the product of the two ($\alpha \cdot \beta$) and the outcome range for the given condition. Each illness condition has only one set of characteristics but these characteristics can have different effects at a traditional healer than at a mission.[16] Note that in order to solve the model we normalize γ_{TH} and ρ_{TH} to zero. The entire regression is just a specific case of the more general conditional logit model (Maddala 1983, pp 44) and therefore has the required properties for obtaining a solution.

3.2 Results

In running the regressions that follow, after controlling for other important variables, we are looking for the following patterns. We expect that patient utility at traditional healers is higher than at missions when effort complementarity is high; when α and β are both large. We expect that patient utility is higher at missions when effort complementarity is low; when α or β are large but not both simultaneously. Thus we have included the product of α and β. When $\alpha \cdot \beta$ is large, the probability of a visit to a mission should decrease. When $\alpha \cdot \beta$ is small and when α or β is large the probability of a visit to a mission should increase. Thus, when the visit to the mission is the visit we are trying to explain, the coefficient for α and β should be greater than zero and the coefficient for $\alpha \cdot \beta$ should be less than zero.

Table 1 displays the results of the logit regression on the four data sets[17]. The coefficients for the individual characteristics are not reported, but were part of the regressions and have therefore been controlled for. The coefficients for the first four variables can be read as the effect of the variable on the likelihood of a visit to a mission

[16] Adding the additional terms $\rho_k' z_k$ has the same effect as restricting some of the coefficients in the η vector to be equal to each other.

[17] We differentiate between scorings by individuals with medical expertise from the coding by reference to medical texts. However, we choose not to identify individuals coders with their scores.

Table 1: Conditional Logit Coefficients

Dependent variable is the choice of provider

data set	Medical Ref		Individual 1		Individual 2		Individual 3	
variable	coef	std err	coef	std err	coef	std err	coef	std err
α	0.359‡	0.121	0.378‡	0.156	0.310‡	0.130	0.494	0.443
β	0.301‡	0.140	0.263†	0.130	0.554‡	0.164	-0.148	0.099
$\alpha \cdot \beta$	-0.103‡	0.024	-0.085‡	0.029	-0.081‡	0.025	0.095	0.056
outcome range	0.376‡	0.132	0.264	0.134	-0.037	0.075	-0.049	0.107
travel cost	-0.763‡	0.104	-0.684‡	0.099	-0.719‡	0.099	-0.713‡	0.100
provider skill	0.168	0.122	-0.185	0.130	0.015	0.048	-0.044	0.061
individual characteristics (income, household wealth, schooling and adult) controlled for but not reported								
log likelihood	-213.02		-224.15		-227.17		-222.99	

‡significant at 99% for one-sided test

†significant at 97.5% for one-sided test

provider (either a clinic or hospital) over a traditional healer. Thus, increasing α or β increases the probability of a visit to a mission provider in the first three data sets. However, increasing the size of the product reduces the likelihood of a visit in those first three data sets. Thus when β is large, increasing α decreases the probability of a visit to a mission, increasing the probability of a visit to a traditional healer. None of the illness condition characteristics are significant on the 3rd individual coding.

The coefficients on the last two variables (the γ vector) can be read as the effect of the variable on the likelihood of a visit to each provider. Thus when the cost of travel to any given provider increases the likelihood of a visit to that provider falls. Skill does not appear to have any significant affect on the probability of a visit after controlling for α, β and the outcome range.

Table 2 reports the marginal impact of the variables on the probability of a visit to any given provider for the first data set (coding by medical references). The entries can be read as follows. Increasing the outcome range by 1% leads to a decrease of 0.2% in the probability of a visit to a traditional healer, an increase of 0.1% in the probability of a visit to a mission clinic and a 0.1% increase in the probability of a visit to a mission hospital. The elasticities with respect to α and β reported in the table combine the direct and interaction effects. The effect of an increase in α or β from

Table 2: Elasticities of Probabilities with respect to Characteristics

	Change in percentage probability of visit from a 1% change in variable from its mean		
variable	Traditional Healer	Mission Clinic	Mission Hospital
outcome range	-0.205	0.146	0.059
travel to MC	0.101	-0.218	0.117
travel to MH	0.119	0.348	-0.467
α at low β	-0.038	0.027	0.011
α at $\bar{\beta}$	0.083	-0.060	-0.024
α at high β	0.128	-0.092	-0.036
β at low α	-0.063	0.045	0.018
β at $\bar{\alpha}$	0.038	-0.027	-0.011
β at high α	0.115	-0.082	-0.032

Low indicates 20th percentile and high indicates 80th percentile

their mean values depends on the magnitude of the other elasticity. When β is low, increasing α decreases the probability of choosing an outcome–contingent contract (the traditional healer), but when β is large increasing α increases this probability. The same pattern holds for β with respect to α.

Patterns of patient choices between contracts display exactly the characteristics predicted by a model of two–sided asymmetric information. Outcome-contingent contracts are preferred when α and β are both large. Effort-contingent contracts are preferred when α alone is large or when β alone is large. These results are significant in three out of the four data sets that we collected. They are robust to empirical specifications and offer strong support that patient utility is affected by the contract available at any given provider.

4 Conclusions

The existence of double–sided asymmetric information, where physicians cannot observe patient behavior and patients cannot observe physician behavior, is an important problem in health care that has received little attention in the economics literature. This paper develops a model that attends to these informational concerns and compares the relative performance of two physician compensation strategies: one where compensation is effort contingent, and one where compensation is outcome contingent. These two strategies can be viewed as stylized representations of existing payment schemes that include such things as capitation, salary, physician profiling, and traditional fee-for-service compensation.

Our analytic results indicate the need for compensation regimes that vary according to disease conditions. There is no one-size-fits-all solution. When patient effort and physician effort are highly complementary, physician compensation should be outcome contingent. When the two efforts are not very complementary, effort–contingent payments are better. In so far as effort complementarity is specific to medical specialty areas, we should compensate each specialty accordingly. It is important to recognize that these results can be generalized to a wide range of service industries that are often characterized by production in teams with two-sided information asymmetries. These industries range from legal services to electronics and automobile repair and the framework developed here may provide insight into the scope and limitations of such things as warranties and money-back guarantees, which can be thought of as a means of transforming effort–contingent payments to outcome–contingent ones.

Evidence to support this theory is provided by an empirical analysis of patient choice of health care providers in Africa. The analysis provides strong evidence for the principal theoretical result. Patients with disease conditions that are relatively responsive to patient and practitioner effort are more likely to seek treatment from a traditional healer who is paid based on outcomes. When the disease is not particularly responsive to one of the two types of effort, patients visit effort-compensated physicians at mission health care providers. Elasticity measures with respect to effort complementarity are large and on the same scale as the significant travel costs facing patients in this area. Contracts matter crucially in this context, and offer reason to believe that they matter in health care markets outside of Africa as well.

The framework developed here has been kept relatively basic for analytic simplicity. Given the stochasticity of health production, and in turn patient and physician utility, future research should extend this analysis to the case where agents are risk averse. In this case, the differential ability of agents to bear risk should also play a vital role in determining appropriate compensation schemes. Additional research should also examine the role of monitoring and the legal system in the social planner's derivation of $\bar{p}$, the social planner's expectations of patient effort. A deeper understanding of the institutions and norms that govern health care provision will enhance our ability to overcome these informational market failures.

References

Arrow, Kenneth J., "Uncertainty and the Welfare Economics of Medical Care," *American Economic Review*, 1963, *53* (5), 941–973.

Bhattacharyya, Sugato and Francine Lafontaine, "Double-sided moral hazard and the nature of share contracts," *Rand Journal of Economics*, 1995, *26* (4), 761–781.

Cooper, Russell and Thomas W. Ross, "product warranties and double moral hazard," *Rand Jounal of Economics*, 1985, *16* (1), 103–113.

Demski, Joel and David Sappington, "resolving double moral hazard problems with buyout agreements," *Rand Jounal of Economics*, 1991, *22* (2), 232–240.

Dranove, David, "Demand Inducement and the Physician/Patient Relationship," *Economic Inquiry*, 1988, *26* (2), 281–98.

Ellis, Randall P. and Thomas G. McGuire, "Provider Behavior under Prospective Reimbursement: Cost Sharing and Supply," *Journal of Health Economics*, 1986, *5* (2), 129–51.

Gold, M. et al., "A National Survery of the Arrangements Managed Care Plans Make with Physicians," *The New England Journal of Medicine*, 1995, *333* (25), 1678–38.

Griffith, H. Winter, *The Complete Guide to Symptoms, Illness and Surgery*, Los Angeles, CA: The Body Press, 1985.

Grossman, Michael, "On the Concept of Health Capital and the Demand for Health," *Journal of Political Economy*, 1975, *80*, 223–255.

Grossman, Sanford and Oliver Hart, "An Analysis of the Principal Agent Problem," *Econometrica*, 1983, *51* (1), 7–45.

Hart, Oliver and Bengt Hölmstrom, "The Theory of Contracts," in Truman F. Bewley, ed., *Advances in Economic Theory: Fifth World Congress*, number 12. In 'Econometric Society monographs.', Cambidge: Cambridge University Press, 1987.

Hölmstrom, Bengt, "Moral Hazard in Teams," *Bell Journal of Economics*, 1982, *13*, 324–40.

Leonard, Kenneth L., "African Traditional Healers: Incentives and Skills in Health Care Delivery," Discussion Paper Series 9798-13, Columbia University 1998.

Maddala, G.S., *Limited-Dependent and Qualitative Variables in Econometrics*, Cambridge, UK: Cambridge University Press, 1983.

Maskin, Eric and John Riley, "input vs. output incentive schemes," *Journal of Public Economics*, 1985, *28*, 1–23.

New York Times, "Vital signs: fertility; Help for Vasectomy Reversals that Fail," *The New York Times*, 1999.

Pauly, M., *Doctors and Their Workshops*, Chicago: University of Chicago Press, 1980.

Reinhardt, U., "Economists in Health Care: Saviors, or Elephants in a Porcelain Shop?," *American Economic Review*, 1989, *79* (2), 37–42.

Robertson, John A. and Theodore J. Schneyer, "Professional Self-Regulation and Shared-Risk Programs for In Vitro Fertilization," *Journal of Law Medicine and Ethics*, 1997, *25* (4), 283–291.

Shavell, S., *Economic Analysis of Accident Law*, Cambridge, MA: Harvard University Press, 1987.

Strickland, G. Thomas, ed., *Hunters' Tropical Medicine*, 6 ed., Philadelphia: W.B. Saunders Co., 1984.

Weitzman, Martin L., "Prices and Quantities," *Review of Economic Studies*, 1975, *41* (4), 477–491.

Werner, David, *Where There is No Doctor; A village health care handbook*, Palo Alto, CA: The Hesperian Foundation, 1977.

A Proof of Proposition 3

Recall that the difference welfare under the two regimes is:

$$W_{\mathrm{O}} - W_{\mathrm{E}} = \frac{W_{\mathrm{FI}}}{1-\alpha-\beta} \left(\begin{array}{c} (1 - s_p\alpha - s_m\beta)\left(s_m^{\beta} s_p^{\alpha}\right)^{\frac{1}{1-\alpha-\beta}} - \\ \left(1-\alpha-\beta\left(\frac{\bar{p}}{p^{\star}_{\mathrm{FI}}}\right)^{\frac{\alpha(1-\alpha-\beta)}{(1-\alpha)(1-\beta)}}\right)\left(\frac{\bar{p}}{p^{\star}_{\mathrm{FI}}}\right)^{\frac{\alpha\beta}{(1-\alpha)(1-\beta)}} \end{array} \right) \tag{20}$$

The difference is positive if

$$(1 - s_p\alpha - s_m\beta)\left(s_m^{\beta} s_p^{\alpha}\right)^{\frac{1}{1-\alpha-\beta}} - \left(1-\alpha-\beta\left(\frac{\bar{p}}{p^{\star}_{\mathrm{FI}}}\right)^{\frac{\alpha(1-\alpha-\beta)}{(1-\alpha)(1-\beta)}}\right)\left(\frac{\bar{p}}{p^{\star}_{\mathrm{FI}}}\right)^{\frac{\alpha\beta}{(1-\alpha)(1-\beta)}} > 0 \tag{23}$$

Recall that $\hat{p}$ is the value of $\bar{p}$ for which the utility under the two regimes is equal. We introduce a notation for $\bar{p}$, $\bar{p} = \hat{p}t$. When t is equal to one therefore, $\bar{p} = \hat{p}$ and we have the solution outlined in equation (21). When t is less than 1, $\bar{p} < \hat{p}$ and we are in the territory described in figure 2 as region II, where the utility with outcome–contingent contract is greater than the utility with effort–contingent contracts. Thus to prove proposition 3, we need to show that when t is less than one, equation (23) always holds.

We start with the fact that when $t = 1$ equation (23) is positive by proposition 2. Let g denote the expression in equation (23). Taking the derivative of g with respect to t we obtain

$$\frac{\partial g}{\partial t} = \left(s_p^{\frac{1-\beta}{\beta}} t^{\alpha\frac{1-\alpha-\beta}{(1-\beta)(1-\alpha)}} s_m - 1 \right) s_p t^{\frac{1-\alpha-\beta}{(1-\beta)(1-\alpha)}} \alpha \frac{\beta}{1-\beta} \tag{24}$$

Since either s_p or s_m is always less than or equal to one, with one strictly less than one, $\frac{\partial g}{\partial t} < 0$ whenever t is less than one — it is increasing as t falls toward 1. If g is decreasing in t when t is less than one, and g is positive when t is equal to one, then g must be positive whenever t is less than one. Thus the difference between welfare with outcome–contingent contracts and welfare with effort–contingent contracts is always positive when $\bar{p} < \hat{p}$, or when $U_{\mathrm{O}} > U_{\mathrm{E}}$. QED.

1999-2000 Discussion Paper Series

Department of Economics
Columbia University
1022 International Affairs Bldg.
420 West 118th Street
New York, N.Y., 10027

The following papers are published in the 1999-00 Columbia University Discussion Paper series. The series run on an academic year: from November 1st to October 31st.

All papers from **9798 to 9900** will be accessible on-line at the following website:
http://www.SSRN.Com/

Some papers from **1995 to 1997** are available and may be downloaded from the following website:
http://www.columbia.edu/dlc/wp/econ/index.html.

Copy Requests

Current Papers (backdated to 9798 series):

Please go to the SSRN website (see above) to access discussion papers.
Hardcopies are not available.

Past Discussion Papers:

Past discussion papers (1996-97 and prior papers) are available for purchase in *U.S. dollars only*, at the cost of :

U.S. per paper $4.00 / Per series: $140.00
Canada per paper $5.00 / Per series: $150.00
Overseas per paper $7.00 / Per series: $185.00

To order any of the *past series*, as above stated, please write to the Discussion Paper Coordinator at the above address, along with a check for the appropriate amount, made payable to: Department of Economics, Columbia University. *Please be sure to indicate the discussion paper number or the particular series in your written request.* Orders cannot be process without payment, and they cannot be taken over the phone, or by fax or email.

3/22/00

1999-00 Discussion Papers
Economics Department, Columbia University

Series no.	Title	Authors
9900-01	How to Compensate Physicians When Both Patient and Physician Effort are Unobservable	Leonard, K. Zivin, J.