

NBER WORKING PAPER SERIES

OUTCOME VERSUS SERVICE BASED
PAYMENTS IN HEALTH CARE:
LESSONS FROM AFRICAN TRADITIONAL HEALERS

Kenneth Leonard
Joshua Graff Zivin

Working Paper 9797
<http://www.nber.org/papers/w9797>

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
June 2003

We gratefully acknowledge the support of a seed grant from the Institute for Social and Economic Research and Policy (ISERP) at Columbia University. We wish to thank M Riordan, S Glied and M Madajewicz for helpful comments. The views expressed herein are those of the authors and not necessarily those of the National Bureau of Economic Research.

©2003 by Kenneth Leonard and Joshua Graff Zivin. All rights reserved. Short sections of text not to exceed two paragraphs, may be quoted without explicit permission provided that full credit including © notice, is given to the source.

Outcome Versus Service Based Payments in Health Care:
Lessons from African Traditional Healers
Kenneth Leonard and Joshua Graff Zivin
NBER Working Paper No. 9797
June 2003
JEL No. I1, D8

ABSTRACT

We compare the more common physician compensation method of fee-for-service to the less common payment-for-outcomes method. This paper combines an investigation of the theoretical properties of both of these payment regimes with a unique data set from rural Cameroon in which patients can choose between outcome and service based payments. We show that consideration of the role of patient effort in the production of health leads to important differences in the performance of these contracts. Theory and empirical evidence show that when illnesses require (or are responsive to) large amounts of both patient and practitioner effort, outcome based payment schemes are superior to effort based schemes. The traditional healer — a practitioner who offers health services on an outcome-contingent basis — is advanced as an important example of how patient effort can be better understood and tapped in health care.

Kenneth Leonard
Department of Economics
Columbia University
Mail Code 3308
420 W 118th St.
New York, NY 10027
kl206@columbia.edu

Joshua Graff Zivin
Department of Health Policy and Management
Columbia University
600 West 168th St. #608
New York, NY 10032
and NBER
jz126@columbia.edu

In the vast majority of health care delivery systems around the globe, physician compensation is input based—physicians are compensated based on the effort or services that they provide to patients. Intuition, however, suggests that outcome based compensation should be preferred. Paying directly for the item of value (outcomes) makes more sense than paying for inputs which have no direct value, and when physician behavior is difficult to properly monitor or evaluate, it should be easier to align incentives with outcome based payments. The absence of such compensation schemes is generally based on practical concerns about the verifiability of outcomes or the potential manipulation of such a system by physicians to the disadvantage of patients (refusing difficult patients for example.) Even in situations where these concerns can be overcome, such as the African setting that forms the basis of our empirical analysis, another more fundamental concern about the nature of health production poses challenges to the presumed supremacy of outcome-contingent compensation schemes. Physician effort is not the only determinant of health outcomes; patient effort also matters. Healthiness is generally created by the joint effort of both patients and physicians and this has important implications for the implementation and effectiveness of both outcome and effort based compensation.

In this paper, we introduce a basic model of joint production of health where both patients and physicians provide unobservable effort that affects outcomes following Hölmstrom (1982).¹ We model two types of payment schemes: outcome-contingent and effort-contingent. For both forms of contracts there exists a third party who can implement contracts and will seek to maximize social welfare (the joint utility of patient and practitioner). This third party can be thought of as the employer or regulator of the practitioner. In the outcome-contingent scheme, physician and patient each earn utility from the outcome of treatment. In the effort-contingent contract the patient earns utility from outcomes and pays for ser-

¹Unobservable effort in physician behavior is referred to as imperfect agency in the health economics literature and moral hazard in the general economics literature. For background on views of imperfect agency in health care see Gaynor (1994, p. 222) and McGuire (2000, p. 499). In the model used in this paper, we intend imperfect agency to cover unobservable diagnostic quality or effort. Thus the term *hidden effort* is a more precise term and terms such as quality, shirking, or slacking-off mirror the concerns of this paper.

vices delivered. The third party, who can observe physician effort, forces the physician to provide a particular level of effort. The structure of information plays an important role in this model, as in all models of asymmetric information. We assume the following:

- Both physicians and patients can observe outcomes but patients cannot evaluate medical effort and physicians cannot observe patient effort.
- The third party (regulator) can observe outcomes and medical effort but cannot observe or infer patient effort.

Under these assumptions, we show that both contracts achieve the full information result—the result that would be obtained if every action were observable—and are therefore interchangeable. Administrative simplicity or any other factor can determine choice of contract form. However, we assume that in the real world there are at least two additional informational restrictions: First, that the regulator cannot use payments that transfer the full benefits or costs of health outcomes to physicians (no scheme can make a practitioner care as much about outcomes as the patient). Second, that the regulator cannot adequately model the behavioral response of patients as a function of physician behavior, and therefore sets medical effort assuming patient effort is either unimportant on the margin or unresponsive to medical effort.

Under these conditions, 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 (if the outcome is verifiable.) 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.

There is, to the best of our knowledge, only one health delivery institution that delivers a wide spectrum of health care services and is paid on the basis of outcomes: the African traditional healer. The reason traditional healers are able to use this contract is that the institution of traditional medicine allows for verifiable outcomes for all illnesses; patients believe that traditional healers are the agents of higher powers and that these higher powers can verify all outcomes (Leonard, 2003). This allows us to test our findings on a data set from rural Cameroon in which patients choose between different types of health care providers; one compensated through effort-contingent contracts and the other through outcome-contingent ones. We show that patients choose practitioners according to exactly the criterion outlined in the theory section; they prefer traditional healers for illness conditions where the elasticities of outcomes with respect to both medical and patient effort are high. These basic patterns are consistent with those found in other African settings (Mwabu, 1986).

In examining the practices of traditional healers, one of the features most salient to this paper is the focus they place on the effort of patients. Healers spend almost as much effort molding patient activities as diagnosing and curing patients. We suggest this is not a coincidence, but rather a feature of the contracts they offer. We do not suggest that traditional healers can or should be emulated, but rather that if outcome-contingent contracts (of any form) are to be used, they should be used primarily where both patients and practitioners provide cooperative effort towards a cure.

The paper is structured as follows. The next section introduces a basic theoretical model of the joint production of health with two-sided moral hazard. The full derivation of the model is presented in Appendix A. Equilibrium effort levels, utility, and social welfare when payment is effort contingent and when payment is outcome contingent are analyzed. Section two tests our theoretical results using a unique data set from Africa. The third section discusses alternative methods of modeling patient-practitioner interaction and the final section concludes.

1 A Model of Health Care

We model the net expected value from seeking health care as a function of the opportunity cost of healthy time (ω) and the expected increase in health (h), net the cash costs of seeking care (C) and the disutility of the patient's own effort in producing her own health (p).

$$\Delta EU = \omega h - C - c(p) \tag{1}$$

h is the expected value of H .² One obvious way to motivate this simple specification is by a health capital model where investments in health increase the amount of time available to patients for work and leisure (Grossman, 1975). In this view ω is the value of an additional hour created, i.e. the wage or the opportunity cost of leisure. All subsequent analysis defines patient and social welfare in terms of the net utility of expected health outcomes, and for simplicity we will refer to ΔEU as U .

The expected value of increased health, h , is a 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 the patient being cured. The role of each of these factors will vary according to the illness condition.

We hypothesize that for an illness and treatment regime medical and patient effort are complements. When more of either effort is provided the marginal impact of the other will increase. It is also equivalent to state that when more of one effort is provided the cost of the other effort decreases. Either patient effort makes medical effort more effective or it makes it less difficult for the practitioner to provide.

We do not suggest that medical and patient effort are global complements. It is important

²The astute reader will note that this could contain some undesirable assumptions about the nature of risk aversion. Appendix A.1 demonstrates that the model employed here can, with some basic assumptions about the distribution of health outcomes, be derived from a model that is consistent with risk averse patients.

to recognize the difference between *ex ante* and *ex post* substitutability. An illness may be treatable by two different technologies (an injury may be treated by surgery or physical therapy) and each technology uses very different levels of medical and patient effort. A patient effort–intensive technology may substitute for a medical effort–intensive technology, but this is *ex ante* substitutability. Once the technology has been chosen, medical and patient effort are complements: increased medical effort enhances the impact of the patient effort under physical therapy and increased patient effort increases the impact of medical effort under surgery. We assume that the *ex ante* choice of technology depends on factors outside the concern of this paper (such as relative costs of inputs). In developing countries (the concern of our empirical section) patient effort–intensive technology prevails.³

Thus, h is modeled as a Cobb-Douglas production function which assumes complementarity.

$$h = \pi p^\alpha m^\beta \tag{2}$$

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 (π) and the value of health (ω) as A , a technology scale. A can be thought of as the value of obtaining health care, a measure that embodies the benefits (relative to letting the disease run its natural course) from being

³Van der Geest and Sarkodie (1998) provide an eye-opening description of life as a patient in a typical African hospital, including the extensive reliance on family to provide important nursing functions.

healthy and the ability of the practitioner to provide that health. Thus utility is,

$$U = Ap^\alpha m^\beta - p - C \quad (3)$$

and practitioner utility is a function of effort ($-m$) and a transfer T , which may or may not be equal to the payment of the patient.

$$Y = T - m \quad (4)$$

1.1 Production with Full Information

As a basis of comparison for the cases with asymmetric information, we will first analyze the problem for the case with full information. This case corresponds to a world where both the practitioner and the patient observe the other's effort and there is no coordination problem. The full information solution is interesting because it represents the best possible outcome. Social welfare is the sum of patient utility and practitioner income, which is simply patient utility from health net of effort costs (all transfers balance out).

$$W = Ap^\alpha m^\beta - p - m \quad (5)$$

Maximizing welfare with respect to p and m we obtain the two first order conditions. Together, these first order conditions allow us to define an optimal level of patient and practitioner effort that are a function of the value of health care, the marginal productivities of effort, and the costs of effort. We represent these optimal levels as p_{FI}^* and m_{FI}^* , where the subscript FI denotes the full information solution. These expressions for optimal effort levels can then be employed to determine social welfare (W_{FI}) and practitioner and patient utility (U_{FI}). Under the assumption that health care is valuable (A is large enough), p_{FI}^* , m_{FI}^* , W_{FI} and U_{FI} are all increasing in the elasticity of outcomes with respect to medical and patient effort (α and β) as well as the technology, A (See Appendix A.2.)

1.2 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. Under the effort-contingent contract the third party (employer or regulator) chooses physician effort to maximize social welfare. The regulator cannot observe patient effort, but he uses some approximation which we represent as \tilde{p} , the regulator's guess of patient effort. Under outcome-contingent contracts the only parameter of choice is the share of the outcome for the physician (s_m) and the patient (s_p). This share could be set by the regulator, by bargaining between physician and patient or by some other convention. Under both contracts, the full information solution can be achieved.

Effort-Contingent Contracts The regulator, since he can observe medical effort, can choose the level of medical effort. However, he cannot observe patient effort and, due to the stochasticity of health outcomes, he cannot infer it by looking at outcomes. Thus, the regulator maximizes welfare assuming patient effort is \tilde{p} .

$$\max_m A\tilde{p}^\alpha m^\beta - \tilde{p} - m \tag{6}$$

The regulator chooses medical effort such that the marginal productivity of medical effort, evaluated at the regulator's estimate of patient effort, \tilde{p} , is equal to the marginal cost of medical effort.

The patient responds to practitioner effort through her choice of effort, choosing her effort so as to maximize her utility. Using patient and practitioner optimal effort we obtain equilibrium patient utility and social welfare, which we express as a function of full information

utility and welfare.

$$U_E = \left(\frac{\tilde{p}}{p_{\text{FI}}^*}\right)^{\frac{\alpha\beta}{(1-\alpha)(1-\beta)}} U_{\text{FI}} \quad (7a)$$

$$W_E = \frac{\left(1 - \alpha - \beta\left(\frac{\tilde{p}}{p_{\text{FI}}^*}\right)^{\frac{\alpha(1-\alpha-\beta)}{(1-\alpha)(1-\beta)}}\right)}{1 - \alpha - \beta} \left(\frac{\tilde{p}}{p_{\text{FI}}^*}\right)^{\frac{\alpha\beta}{(1-\alpha)(1-\beta)}} W_{\text{FI}} \quad (7b)$$

The subscript E denotes the effort–contingent solution. U_{FI} is the full information utility and p_{FI}^* is the level of patient effort that would have been provided under the full information solution. When $\tilde{p} = p_{\text{FI}}^*$ the full information solution will obtain. Thus, if the regulator can model patient behavior accurately he will achieve the full information (optimal) solution. When $\tilde{p} < p_{\text{FI}}^*$ (the regulator assumes patients do less than the actually do), practitioner effort decreases at the expense of patient utility. The practitioner is not working hard enough. When $\tilde{p} > p_{\text{FI}}^*$ (the regulator assumes patients do more than they actually do), 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 $\tilde{p} = p_{\text{FI}}^*$. When $\tilde{p} > p_{\text{FI}}^*$ or when $\tilde{p} < p_{\text{FI}}^*$ welfare under the effort based contracts is strictly less than welfare under full information. Patient utility, however, can be greater than under full information because the patient does not have to compensate the practitioner for working too hard.

Outcome–Contingent Payments In this case, practitioner and 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 . Given the shares, we obtain

$$U = s_p A p^\alpha m^\beta - p \quad (8)$$

$$Y = s_m A p^\alpha m^\beta - m \quad (9)$$

We assume a Nash solution and equilibrium is found where each player's choice of effort is equal to the other player's expectation. The regulator 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 . We represent utility and social welfare as functions of full information utility and social welfare:

$$U_O = s_p (s_p^\alpha s_m^\beta)^{\frac{1}{1-\alpha-\beta}} U_{\text{FI}} \quad (10a)$$

$$W_O = \frac{(1 - s_p\alpha - s_m\beta)}{1 - \alpha - \beta} (s_p^\alpha s_m^\beta)^{\frac{1}{1-\alpha-\beta}} W_{\text{FI}} \quad (10b)$$

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; when each participant receives the full rewards for their effort 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. If either the patient or the practitioner (or both) do not face the full incentives to provide effort they will each under-provide it. In such a case patient utility and social welfare under outcome-contingent contracts are inferior to patient utility and social welfare under the full information solution.⁴

1.3 The Constrained Regulator

If $\tilde{p} = p_{\text{FI}}^*$ and $s_p = s_m = 1$ both contracts achieve full information and there is no difference between them. In a second-best world where these conditions do not hold, neither payment scheme is uniformly superior. Welfare is maximized through outcome-contingent contracts for some illnesses and through effort-contingent contracts for others. To form hypotheses on the relative advantages of these contracts we develop a more explicit model of these restrictions.

⁴This is an extension of the well-established finding that under dual effort when the budget is balanced, the full information solution cannot be obtained (Hölmstrom, 1982).

Modeling of Patient Effort We assume that the regulator’s view of patient effort, \tilde{p} is invariant with respect to illness conditions. This corresponds immediately to three possible views on the part of the regulator. First, he may see patient effort as being fixed; patients provide the same amount of effort for every illness condition. Second he may not recognize the role of patient effort in health care, maximizing welfare of the form $\hat{A}m^\beta - m$ (with $\hat{A} = A\tilde{p}^\alpha$); the regulator sees the benefit of patient effort as being an illness specific technology shift, not the input of a rational participant. Or third, he may believe that patients provide effort, perhaps even rational amounts of effort, but that the level of patient effort should not change the level of medical effort provided. The solution to the regulator’s problem is identical for all three views of patient effort.

This view of regulators matches the empirical setting in which we will test the theory. Regulators completely ignore the possibility that patient effort is important, and certainly do not see patient effort as a variable input that they can manipulate to their advantage. In general, though most health professionals would readily admit that patient effort matters, little attention is paid to whether or not patient effort should affect the level of medical effort. Thus, whether or not regulators are *de jure* constrained in their views of patient effort they are certainly *de facto* constrained.

In our model of complementary efforts in team production, patient effort is useful to the practitioner. It increases the effectiveness of his own effort. This will be particularly important for illnesses that require extensive joint effort. The regulator must not only recognize that patient effort increases the probability of a cure, he must also recognize that patient effort increases the usefulness of medical effort, and that increased medical effort will induce the patient to do more. There is a virtuous circle that the regulator must recognize *ex ante*.

Less-than-complete incentives We model the regulator’s inability to force physicians and patients to face the full impact of outcomes as a contract in which $s_p < 1$ and/or $s_m < 1$.

It is difficult to provide the practitioner with the full incentives to exert medical effort since this would imply that he would experience the same disutility from failed cancer treatment, for example, as the patient. The patient may also not face the full incentives to exert effort due to risk sharing arrangements, such as disability, life or health insurance, that try to reduce the disutility of adverse health outcomes. A special case of reduced incentives is that where the shares are forced to sum to one $s_p + s_m = 1$.⁵ This will be the case for the traditional healer, who is our example of an outcome-contingent contract used in the empirical section.

When the shares are less than one, each participant has a marginally reduced incentive to provide effort. It does not mean that they provide no effort, but only that they provide less than the optimal level of effort.

Anticipated impact of these constraints Under these assumptions we can anticipate that both the outcome- and effort-contingent contract will fall short of the full information solution. However, they will do so in different ways. Specifically, the outcome-contingent contract will always be inferior to the full information setting but still contains a reduced element of the virtuous circle. Even though each party has less incentive to exert effort than they would if they faced the full share of outcome, they still recognize the beneficial impact of the other player's effort. This will not be true with the effort-contingent contract as we have modeled it. Thus, the outcome-contingent contract is likely to do better for those illness conditions where joint effort is important in the cure. We formalize this intuition in the following section.

1.4 Effort- vs. Outcome-Contingent Payments

Outcomes are different under these two possible contracts in the second-best⁶ world. We examine patient utility and total welfare. The difference in patient utility across regimes is:

⁵This is an example of the balanced budget constraint discussed in Hölmstrom (1982).

⁶Second-best, in this and all subsequent references, implies $s_p < 1$ or $s_m < 1$ and \tilde{p} constant.

$$U_O - U_E = U_{FI} \left(s_p (s_p^\alpha s_m^\beta)^{\frac{1}{1-\alpha-\beta}} - \left(\frac{\tilde{p}}{p_{FI}^*} \right)^{\frac{\alpha\beta}{(1-\alpha)(1-\beta)}} \right) \quad (11)$$

The sign of this expression depends on the value of \tilde{p} . If \tilde{p} is very small then utility is greater under the outcome-contingent contract ($U_O > U_E$). On the other hand, if \tilde{p} is very large then the opposite is true ($U_O < U_E$). We do not know \tilde{p} a priori, but if \tilde{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 . Whatever the level of patient effort assumed by the practitioner, we can determine the conditions under which one contract is most likely to be superior to the other.

Define \hat{p} as the value of \tilde{p} such that patient utility is equivalent in both regimes: $U_O - U_E|_{(\tilde{p}=\hat{p})} = 0$. By construction, when $\tilde{p} > \hat{p}$, Equation 11 is negative and patient utility is larger when physician compensation is effort-contingent. When $\tilde{p} < \hat{p}$, the opposite is true. Although \hat{p} varies with α and β , \tilde{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 .⁷

We show in the appendix that $\frac{\partial \hat{p}}{\partial \alpha}$, $\frac{\partial \hat{p}}{\partial \beta}$ and $\frac{\partial^2 \hat{p}}{\partial \alpha \partial \beta}$ are all positive. These signs of the \hat{p} derivatives imply that utility in the outcome-contingent regime 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

⁷Note 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.

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.⁸ Thus we have our first proposition.

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 welfare implications are similar, though not as straight-forward to illuminate. We can derive the following proposition

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

The proof is contained in Appendix 21. Both welfare and patient utility are more likely to be greater under outcome–contingent contracts when α and β are both high. When both α and β are high, illness 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.

Proposition 2 is important because it suggests that when we observe patient utility greater under the outcome–contingent contract we can conclude that welfare is also greater under the

⁸Note 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.

outcome-contingent contract. In the empirical analysis that follows our analysis is of utility, not welfare. We choose this focus because our empirical strategy depends on observing patient choice, which is determined by utility not welfare. However, the implications of the model (in terms of the potential superiority of outcome-contingent contracts) will hold for welfare as well.

2 Empirical Evidence

This paper makes strong predictions about factors that impact a patient's choice of contract when they suffer from illnesses that have observable and verifiable outcomes. When practitioners are associated with certain types of contracts these same factors (among others) should impact a patient's choice of practitioner. In particular, we expect that the degree of complementarity of medical and patient effort should be an important determinant of the choice of practitioner when one practitioner offers only the outcome-contingent contract and the other offers only the effort-contingent contract.

Patients in rural Africa face precisely this choice of practitioners. Patients can choose between traditional healers, who offer health care on an outcome-contingent basis, and modern medicine where health care is delivered in a fee-for-service environment. As these fee-for-service practitioners are part of organizations that monitor and enforce the provision of effort, the fee-for-service model is an effort-contingent contract. In most parts of Africa, the highest quality facilities are operated by nongovernmental organizations (NGOs), almost always religious orders (missions). Importantly, patients choose providers based on the illness from which they are suffering. When patients have health insurance and/or a primary care physician they might choose a physician and then visit that physician for almost any medical condition (if they choose to seek any care). However, in rural Africa, there is no health insurance and patients must carry all their records (if they have any) with them to every visit. Though a patient might choose to visit the same physician because of proximity,

experience or reputation, they are free to choose between all possible practitioners for each and every illness. It is very common all over Africa to observe the same patients choosing different providers when they suffer from different illnesses. This is important because we contend that the choice of practitioner is, in part, determined by the illness from which people suffer.

Traditional Healers Traditional healers in Africa are paid (after a fixed fee) only if the patient is cured: an outcome-contingent contract. The value of the outcome is shared between healer and patient according to a sharing rule, such that $s_p + s_m = 1$. 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. Leonard (2003) discusses interviews with traditional healers and the anthropological literature and notes that healers talk at length about the importance of patient effort and their understanding of patient effort in their cures. Many healers use modern medicines as well as herbal medicines and therefore the differential access to technology (compared to modern practitioners), though great, is not as great as casual observation would suggest. In addition, traditional healers—because they are seen as being the agents of higher powers and are respected in the community (and sometimes feared)—are able to behave as if they could verify outcomes. Thus patients are choosing between traditional healers and a series of modern providers (some of whom provide high powered incentives in an effort-contingent contract) for every illness condition.

NGO health care providers Both government and mission facilities use an effort-contingent contract. The government, however, uses very low-powered incentives and does not represent a compelling example. On the other hand, missions facilities are well-run, and medical personnel are frequently supervised. Patients pay a fixed fee to the mission and practitioners are monitored and compensated by their employers. Monitoring typically

involves examination of records kept on patients as well as actual observation and further training. The patients' symptoms and complaints are part of all records and therefore it is possible to verify whether procedures and records follow protocols developed for each set of complaints. If a particular record or collection of records is determined to be in violation of standards, the practitioner is punished in accordance with the gravity of the deviation.⁹ In the previous section we showed that optimal effort varies with α and β , and each illness condition represents a unique α β pair, therefore an illness condition protocol is equivalent to an effort-contingent contract in which medical effort is optimally determined by α and β . Practitioner compensation at this institution is effort-contingent.

Evidence of illness based choice of providers Mwabu (1986) analyzed patient choice of health providers in rural Kenya, where the contract used by traditional healers was similar to the one described above.¹⁰ Table 1 shows the relationship of chief complaints to the first practitioner visited in this study and clusters of chief complaint/practitioner matches generated by cluster analysis. The cluster analysis suggests that visits to providers are being determined (at least in part) by illness conditions (or chief complaints). Cluster three corresponds to illnesses that lead to visits primarily to traditional healers and mission clinics. Clearly, asthma, body pain and joint pain are illnesses that require effort from both the doctor and the patient as well as a degree of cooperation between them, providing some suggestive evidence for our theoretical hypothesis.

The empirical results summarized in Table 1, however, do not control for patient or household characteristics, or travel costs. Moreover, chief complaint is essentially the diagnosis, information that the patient learned *after* visiting a provider. In a model of patient choice, careful attention must be paid to information patients have *before* they choose a provider. In the analysis that follows, we employ a dataset from rural Cameroon that allows us to

⁹In practice, facilities with stronger incentives use discretionary bonuses, and the threat of termination to encourage the provision of effort. Mliga (2000) reports that, in Tanzania, where he studied 4 different health care provision systems, those organizations that had the power to use these forms of incentives provided significantly superior quality of care, as judged by other clinicians.

¹⁰Personal communication with the author.

overcome each of these limitations.

Table 1: Distribution of first visits by illnesses and illness clusters across providers (Rural Kenya)

Cluster	Chief complaint	Govt clinic	Missn clinic	Priv clinic	Govt hosp.	Phmcy or shop	Trad. healer	Self	None
1	Ear	25.0	12.0	12.0		12.5		25.0	25.5
1	Eye	40.0	26.7		6.7	6.7		6.7	13.3
1	Cough	36.4	27.3		2.3	22.7		3.5	13.8
1	Vomiting	40.0	20.0			20.0			20.0
1	Backache	36.1	27.8	8.3		11.1	11.1		5.5
1	Abdomen	41.5	16.9	4.6	1.5	24.6	3.1	6.2	1.5
1	Rib pain	30.0	20.0	10.0	10.0	30.0			
1	Diarrhea	25.0	35.0			20.0	5.0	15.0	
	Mean	34.3	23.3	4.4	2.6	18.5	2.4	7.0	8.3
2	Wounds	52.6	15.8				10.5	15.8	5.2
2	Fainting	66.7	33.3						
	Mean	59.7	24.6				5.3	7.9	2.6
3	Asthma	20.0	40.0				40.0		
3	Bodypain	23.5	17.7	11.8		5.9	29.4	5.9	5.9
3	Joint pain	20.0	20.0		6.7	6.7	20.0	13.3	6.7
3	Other	28.6	14.3	14.3	14.3		28.6		
	Mean	23.0	23.0	6.6	5.3	3.0	29.5	4.8	3.2
4	Malaria		50.0		37.7			7.1	7.1
4	Leprosy		60.0		20.0		20.0		
	Mean		55.5		28.9			3.5	3.5
5	Swelling	20.0	60.0			20.0			
5	Heart	10.0	40.0	10.0	20.0	20.0			
	Mean	15.0	50.0	5.0	10.0	20.0			
6	Headache	21.9	9.5	2.9	0.9	47.6	3.8	4.8	8.6
6	Fever	17.7	8.8			58.8		5.9	8.8
	Mean	19.8	9.2	1.5	0.5	53.2	1.9	5.4	8.7
7	Tuberculosis		66.7			33.3			

All entries are percentages. Blanks represent 0.0%. Source: Mwabu (1986)

2.1 Mbonge sub-division, South West Province Cameroon

To test our theory, we use data on individual choices of practitioner collected in Mbonge Sub-Division, in the South-West province of Cameroon 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 traditional healers, mission clinics or mission hospitals. The other major source of health care is the government health system (289 visits) with drug peddlers, pharmacists, neighbors, private hospitals, private clinics and parastatal hospitals rounding out the sample.¹¹ Mission clinics and hospitals are both under the same organization and are monitored using the same technology. The major difference between clinics and hospitals is one of skills not incentives.

Despite its wealth (relative to other areas of Cameroon 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 regular traffic. Nevertheless, we observe significant bypassing of facilities. Nearly 80% of all visits to modern providers were to a provider who was not the closest provider, suggesting a strong revealed preference for the care that is available at facilities visited. However, it is not the case that patients are visiting a few types of providers, but rather that patients are sometimes visiting one provider and other times visiting another. We suggest that it is information that patients possess about the illness from which they suffer that drives them to exercise choice and incur significant cost in the search for 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), all illness episodes were scored using the definitions below. The scores were created by reference to medical texts which contained

¹¹All 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.

information on diagnostic tests necessary for collections of symptoms, as well as information about severity, possible outcomes and the possibility for patient effort to impact outcomes. These scores were validated by scores created by two doctors and a nurse experienced in tropical medicine (Leonard, 2003).

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 .

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 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 important element of A .

Summary statistics for all of these and other variables used in this analysis are available in Leonard (2003).

2.2 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.

Individuals choose between two types of providers and three locations. Types (indexed by k) are traditional (TH) and missions (M). 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 .

$$\log L = \sum_i \sum_{j \in \{TH, MC, MH\}} \delta_{ij} \log P_{ij}$$

$$P_{ij} = \frac{\exp(\eta'_j x_i + \gamma' y_{ij} + \rho'_k z_i)}{\sum_{m \in \{TH, MC, MH\}} \exp(\eta'_m x_i + \gamma' y_{im} + \rho'_k z_i)} \quad (12)$$

$\delta_{ij} = 1$ 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.¹² x includes the age, gender, education level and income of the patient as well as the same variables for the caretaker in the case where the patient is a child or invalid. x also includes a constant and estimated household wealth¹³, 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.

y is a vector of information about the locations visited. The data varies across providers but the coefficient does not.¹⁴ y includes the estimated 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 are treated as the same.

¹²This is the standard multinomial logit framework.

¹³To 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.

¹⁴This corresponds to the McFadden Conditional Logit.

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.¹⁵ 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.

Thus • mission hospitals, mission clinics and traditional healers are potentially different with respect to individual characteristics, • missions (both clinics and hospitals) are different from traditional healers in their comparative advantage for different illness conditions and • all three are different distances from patients and have different skills (for each illness condition).

2.3 Results

In the regression that follows, after controlling for other important variables, we are looking for the following patterns. We expect that patient utility at traditional healers is more likely to be 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 , as well as (in separate regressions) the residual of the product regressed on both α and β ($\widehat{\alpha \cdot \beta}$). This residual is uncorrelated with both α and β and therefore represents a ‘complementarity’ effect. 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 we are trying to explain the visit to the mission the coefficient

¹⁵Adding 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.

for α and β should be greater than zero and the coefficient for $\alpha \cdot \beta$ should be less than zero.

Table 2: Conditional Logit of Choice of Practitioner (contract type) on illness condition characteristics

	Model A	Model B	Model C	Model D
Restricted multinomial variables (γ): Mission Facilities				
α	0.587 (0.267)‡	-0.155 (0.105)	0.69 (0.268)‡	-0.122 (0.099)
β	0.416 (0.221)†	-0.153 (0.133)	0.418 (0.217)†	-0.205 (0.126)†
$\alpha \cdot \beta$	-0.142 (0.047)‡		-0.155 (0.047)‡	
$\widehat{\alpha \cdot \beta}$		-0.142 (0.047)‡		-0.155 (0.047)‡
outcome range	0.38 (0.138)‡	0.38 (0.138)‡	0.413 (0.130)‡	0.413 (0.130)‡
Restricted multinomial variables (γ): Government Facilities				
α			0.376 (0.248)	-0.087 (0.092)
β			0.042 (0.201)	-0.313 (0.121)‡
$\alpha \cdot \beta$			-0.088 (0.042)‡	
$\widehat{\alpha \cdot \beta}$				-0.088 (0.042)‡
outcome range			0.499 (0.127)‡	0.499 (0.127)‡
Conditional variables (ρ)				
travel cost	-0.379 (0.161)‡	-0.379 (0.161)‡	-0.742 (0.095)‡	-0.742 (0.095)‡
provider skill	0.236 (0.121)†	0.236 (0.121)†	0.234 (0.081)‡	0.234 (0.081)‡
General Multinomial variables (η)				
constant	included	included	included	included
individual chars	included	included	included	included
caregiver chars	included	included	included	included
observations	252	252	533	533
log-likelihood	201.80	201.80	699.86	699.86

Dependent variable is the choice of provider. Default choice is traditional healer. Standard errors in parentheses. Positive coefficient for γ represents increased probability of choosing mission (either clinic or hospital) over a traditional healer. Positive coefficient for ρ represents increased probability of choosing providers with a greater value for that variable. (Negative coefficient for travel implies patients prefer providers who are closer (smaller travel cost) after controlling for other factors).

‡significant at 97.5% for one-sided test †significant at 95% for one-sided test

Table 2 displays the results of the four logit regressions. In Model A and B, we use the data restricted to the choice between traditional healer and both types of mission facilities. In Models C and D, we add the data on government facilities. In Model A and C we use the standard definition for joint effort ($\alpha \cdot \beta$) and in Models B and D we use the residual definition ($\widehat{\alpha \cdot \beta}$). Note that the only difference between A and B and between C and D is in the coefficients (and standard errors) for α , β .

In all models the impact of the product of efforts is negative and significant. When

patients suffer from an illness that requires large amounts of effort on the part of both patients and practitioners they are less likely to visit a mission facility (more likely to visit a traditional healer). In addition, the impact of outcome range, travel costs and skill is constant across models. Patients prefer mission facilities (and government facilities) when the possibility of a bad health outcome is higher, they prefer facilities that are closer and they prefer practitioners with a greater skill for the illness from which they suffer.

The impact of α and β directly varies with the specification. Since these variables interplay with their product it is easier to see the behavior implied by this specification in a table of elasticities.

Table 3: Elasticities of Probabilities with respect to Characteristics

variable	Change in percentage probability of visit from a 1% change in variable from its mean		
	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

Table 3 reports the marginal impact of the variables on the probability of a visit to any given provider. The table corresponds to the data in Models A and B. The entries can be read as follows. Increasing the outcome range by 1% leads to a decrease of 0.21% in the probability of a visit to a traditional healer, an increase of 0.15% in the probability of a visit to a mission clinic and a 0.06% 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 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 . These elasticities are significant, but also large. They are on the same scale as the impact of travel costs, something we know to be very important in the search for medical care.

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 offer strong support to the hypothesis that patient utility is affected by the contract available at any given provider.

3 Modeling Patient Practitioner Interaction

The institution of traditional medicine is not generalizable, but we suggest that the approach to patient effort might be. Although traditional healers are severely handicapped by their health technology they continue to offer attractive services to patients because they understand the importance of the interaction between patients and providers. This strength is clear in interviews with healers, but it is also apparent in the data we have introduced.

The model we have described and the informational restrictions that we have put forward fit the data we have analyzed. However, in order to draw conclusions about the relative attractiveness of outcome- versus effort-contingent contracts we need to discuss the generality of our assumptions. Are traditional healers better at patient-practitioner interaction because they are a culturally based institution with strong roots in the community, or because they use an outcome-contingent contract? Can a regulator develop a model of patient behavior that eliminates the inefficiency in the effort-contingent contract?

Health care and advances in health care technology appear increasingly 'sterile', in that the patient is viewed as less and less of an actor in her own health. The fact that doctors do not make house calls is driven by other factors, but it prevents doctors from seeing patients in

their own environment. Not only do traditional healers make house calls, but they frequently interview family members about the condition. In addition, healers adopt the valuations of patients. A healer would not say “I cured her, but she continues to complain.” In his practice well-being and healthiness are inseparable.

The NGO providers that enter the empirical analysis appear to be using an unnecessarily restricted form of monitoring that does not properly take into account patient effort. We know this is the view of health care regulators in this area, but we do not know if this is a necessary view. Although the empirical evidence is compelling, the theory we have advanced suggests that the regulator could easily adopt a superior technology. In order to know whether such a technology would in fact improve outcomes we need to have a more complete understanding of the interaction between patients and practitioners. Even if the regulator assumes a better technology he will always be distant from the actual consultation and will never be able to observe patient effort. Is there something about the relationship that requires the presence of a medical practitioner who is compensated on the basis of outcomes?

We do not have the data to answer this question, but theory offers an interesting insight. The Nash solution that we have assumed in the outcome-contingent problem can be achieved through many mechanisms, some of which take the form of sequential announcements of intentions before any effort is exerted. In this tatonnement process, the players discuss their intentions until they reach a point where neither wishes to change their action; the Nash equilibrium. It is easy to imagine that the bargaining that takes place between a patient and a traditional healer as capturing the benefits of such a process. With a regulator enforcing effort, this communication would have to take place between the regulator and the patient, or the regulator would have to force the practitioner to engage in this communication. This might be difficult to do. Our model does not capture these complexities¹⁶ and our empirical

¹⁶In our model there is only one round of communication; the regulator makes a crude guess (\bar{p}) and the patient reacts to this guess. We can advance the model so that the regulator reacts to the patient's reaction to his guess. The influence of the first crude guess is reduced, but not eliminated and our results still stand. Continuing the cycle—the patient reacts to the regulator's reaction to the patient's reaction to the crude

analysis does not contain the ‘experiment’ that would allow us to make comments about this. As we observe few modern providers under an outcome–contingent contract, we cannot know whether the informational assumptions of our model are truly binding.

In addition, the regulator who seeks to maximize social welfare does not necessarily earn or obtain direct utility from social welfare. If the regulator has mistakenly modeled the patient reaction function, he is unlikely to be presented with any evidence of his error. On the other hand, a physician paid on the basis of outcomes could quickly develop an accurate understanding of patients both because of physical proximity to them, and because his payment increases when he gets it right. Paying a practitioner on the basis of outcomes rather than effort will, at the very least, force practitioners to consider the role of patient effort.

4 Conclusions

This paper develops a model of dual hidden effort and compares the relative performance of two physician compensation strategies: one where compensation is effort contingent, and one where compensation is outcome contingent. Although it is clear that either contract could achieve the first best solution if there were no additional information restrictions, under relatively general restrictions we can show that each contract is likely to be superior for a range of illnesses. In particular, outcome–based contracts are most likely to be successful when both the patient and the practitioner play important complementary roles in the cure of the illness, and effort–contingent contracts are likely to be successful when either effort is necessary, but not both.

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

guess and so on—the influence of the initial guess gets smaller and smaller and the solution approaches the full information solution. The set of illness conditions for which outcome–contingent contracts are superior to effort–contingent contracts gets smaller and smaller until it is eliminated. However, our basic intuition remains.

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.

The ability to verify all outcomes in health care is not a transferable technology. Since outcome-contingent contracts cannot be implemented for non-verifiable outcomes, this limits the set of illnesses for which such a contract can be implemented. However, for that set of outcomes that are verifiable, outcome-contingent contracts appear most attractive when medical and patient effort are both very useful in the cure.

Extensions The most immediate use for outcome-contingent contracts is to pay doctors or organizations on the basis of population average outcomes. For public health interventions, doctors could be paid in an effort-contingent manner (for example, time spent on vaccination campaign) or an outcome-contingent manner (for example, reducing cholera outbreaks or malaria prevalence). In this case the outcomes represent average or public health targets, rather than individual observations. This will raise some concerns about who does the measuring, but these are not insurmountable. Our model suggests two concerns: First, if patients contribute to the achievement of public health goals (as they would in the cholera and malaria example) the regulator should use an outcome rather than input compensated scheme. In the narrow confines of our model, the outcome contingent contract could lead to superior outcomes because the effort-contingent contract would fail to take proper account of patient contributions. In the broader spirit of the model, the outcome-contingent contract would force the practitioner to think about the relationship between his effort and that of the population he serves and to engage in a more cooperative endeavor.

Second, for any type of outcome but particularly for outcomes where patient effort mat-

ters, results will be improved by insuring the maximum possible share of outcome is gained by the patient (s_p is as close as possible to 1). This will increase their incentives to cooperate. Functionally this means that if patients themselves have to compensate the practitioner for meeting a mark, their share will fall and their participation will be diminished. It would be better to implement a scheme in which local populations pay to a fund according to the expected outcome of any project and the fund then compensates the physician according to the actual outcome—injecting funds for better than expected outcomes, and withholding funds for less than expected outcomes. If these marginal contributions had to be made by the population served, it would decrease their incentives to provide proper effort.

References

- Gaynor, M.**, “Issues in the Industrial Organization of the Market for Physician Services,” *Journal of Economic Management and Strategy*, 1994, 3 (1), 211–255.
- 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*, Cambridge: Cambridge University Press, 1987.
- Hölmstrom, Bengt**, “Moral Hazard in Teams,” *Bell Journal of Economics*, 1982, 13, 324–40.
- Leonard, Kenneth L.**, “African Traditional Healers and Outcome-Contingent Contracts in Health Care,” *Journal of Development Economics*, 2003, 71 (1), 1–22. (available on-line).
- Maddala, G.S.**, *Limited-Dependent and Qualitative Variables in Econometrics*, Cambridge, UK: Cambridge University Press, 1983.
- McGuire, T.**, “Physician Agency,” in A.J. Culyer and J.P. Newhouse, eds., *Handbook of Health Economics*, Amsterdam: Elsevier Science Publishers, 2000, pp. 461–536.
- Mirlees, J.**, “The Theory of Moral Hazard and Unobservable Behavior – Part I,” Mimeo, Nuffield College, Oxford 1975.
- Mliga, Gilbert R.**, “Decentralization and the Quality of Health Care,” in David K. Leonard, ed., *Africa’s Changing Markets for Human and Animal Health Services*, London: Macmillan, 2000, chapter 8.
- Mwabu, Germano M.**, “Health Care Decisions at the Household Level: Results of a Rural Health Survey in Kenya,” *Social Science and Medicine*, 1986, 22 (3), 315–19.
- Van der Geest, Sjaak and Samuel Sarkodie**, “The Fake Patient: A Research Experiment in a Ghanaian Hospital,” *Social Science and Medicine*, 1998, 47 (9), 1373–1381.

A Mathematical Appendices

A.1 A Model of Health Care

In this section we develop a model of health care from basic principals. The form derived is the same as the one used in the paper, but the assumptions necessary to put aside concerns about relaxed incentive compatibility constraints¹⁷ and risk aversion are spelled out in detail.

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. ϕ^* is the ‘true diagnosis’ distribution and ϕ^\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^* . If the diagnosis is incorrect the probability of recovery is q^\emptyset . The probability of failing to recover is $1 - q^*$ 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 ϕ^* and ϕ^\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

$$\Delta H \sim e \cdot \phi^* + (1 - e) \cdot \phi^\emptyset \tag{13}$$

The ‘best’ technology ($e = 1$) has q^* chance of leading to recovery, and the ‘worst’ technology ($e = 0$) leads to a chance of recovery of q^\emptyset .¹⁸

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

¹⁷Mirlees (1975) as cited in Hart and Hölmstrom (1987)) shows that the first order conditions do not describe globally optimal actions for distributions such as $H = h + \theta$ or $H = h \cdot \theta$ when θ is any of the standard candidates for random distributions. Thus in order to obtain some theoretical results the choice of functional form for H is crucial.

¹⁸We 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.

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.

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

$$\begin{aligned} EU &= eq^* \bar{U} + e(1 - q^*) \underline{U} + (1 - e)q^\theta \bar{U} + (1 - e)(1 - q^\theta) \underline{U} \\ \bar{U} &= U[\bar{h}, (I(\bar{h}) - C), c(p)] \\ \underline{U} &= U[\underline{h}, (I(\underline{h}) - C), c(p)] \end{aligned} \quad (14)$$

C is the total cost of a visit. We assume a separable utility form such that $U = V[H, I(H)] - C - c(p)$. 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 ($V[H, I(H)]$) depends on the outcome, not on the effort exerted. Costs and disutility have a linear relation to utility. For ease of exposition we write $V[\bar{h}, I(\bar{h})]$ as \bar{V} and $V[\underline{h}, I(\underline{h})]$ as \underline{V} . For any given patient and illness condition there are only two possible V ; \bar{V} and \underline{V} . Different health technologies represent different probabilities of each event occurring but do not change the value of the event. Note that because stochastic health outcomes are measured in utility (not dollars), all forms of risk aversion can be accommodated in this model. Further, so long as treatment and patient effort costs are small, i.e. they do not change wealth so much that they change the marginal utility of *health*, the linear separability assumption is a reasonable one. In this representation, ‘risk aversion’ is represented by a high \bar{V} , not by changing marginal utility of *income*.

With the new notation we get,

$$EU = (e(q^* - q^\theta) + q^\theta) \bar{V} + (1 - q^\theta + e(q^\theta - q^*)) \underline{V} - C - c(p) \quad (15)$$

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

$$\Delta EU = e(q^* - q^\theta) \cdot (\bar{V} - \underline{V}) - C - c(p) \quad (16)$$

At this point we make a number of further simplifying assumptions. First, we assume that \underline{V} is equal to zero, a simple scaling assumption. Furthermore, we assume that utility from health is of the form $\bar{h} \cdot \omega$ where ω represents the combination of the opportunity cost of healthy time and a per unit value of health. Thus,

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

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} .

$$\Delta EU = (q^* - q^0)\omega h - C - c(p) \quad (17)$$

For ease of exposition, in the body of the paper, we move the expression $(q^* - q^0)$ into h , and refer to the net expected value of health as

$$\Delta EU = \omega h - C - c(p) \quad (1)$$

By using the spanning condition we have created a random distribution of health outcomes in which efforts cannot be inferred from outcomes and incentive compatibility constraints can be represented by first order conditions. In addition, by allowing for only two outcomes and assuming costs and disutilities are small relative to health valuations, we have a final specification which appears to be a utility over expected outcomes but is in fact an expected utility formulation. Thus our choice of utility for the model of the paper, though restrictive, is not unrealistic.

A.2 Production with full information, under outcome- and effort-contingent contracts

This section shows the derivation of the results discussed in the text. We begin with Equation 5 from the text. Maximizing welfare with respect to p and m and solving the system of equations we obtain the representations of patient and medical effort under full information:

$$p_{\text{FI}}^* = \alpha (A\alpha^\alpha \beta^\beta)^{\frac{1}{1-\alpha-\beta}} \quad (18a)$$

$$m_{\text{FI}}^* = (\beta/D(A\alpha^\alpha \beta^\beta))^{\frac{1}{1-\alpha-\beta}} \quad (18b)$$

These expressions for optimal effort levels can then be employed to determine social welfare and practitioner and patient utility.¹⁹

$$U_{\text{FI}} = (1 - \alpha) (A\alpha^\alpha \beta^\beta)^{\frac{1}{1-\alpha-\beta}} \quad (19a)$$

$$W_{\text{FI}} = (1 - \alpha - \beta) (A\alpha^\alpha \beta^\beta)^{\frac{1}{1-\alpha-\beta}} \quad (19b)$$

¹⁹We 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.

Under effort–contingent contracts we obtain

$$p_E^* = (A\alpha^{1-\beta}\beta^\beta\tilde{p}^{\alpha\beta})^{\frac{1}{(1-\beta)(1-\alpha)}} \quad (20a)$$

$$m_E^* = (A\beta\tilde{p}^\alpha)^{\frac{1}{1-\beta}} \quad (20b)$$

$$U_E = U_{FI}\left(\frac{\tilde{p}}{p_{FI}^*}\right)^{\frac{\alpha\beta}{(1-\alpha)(1-\beta)}} \quad (20c)$$

$$W_E = \left(1 - \alpha - \beta\left(\frac{\tilde{p}}{p_{FI}^*}\right)^{\frac{\alpha(1-\alpha-\beta)}{(1-\alpha)(1-\beta)}}\right) (A\alpha^\alpha\beta^\beta)^{\frac{1}{1-\alpha-\beta}} \left(\frac{\tilde{p}}{p_{FI}^*}\right)^{\frac{\alpha\beta}{(1-\alpha)(1-\beta)}} \quad (20d)$$

Under outcome–contingent contracts we obtain

$$p_O^* = s_p\alpha (A\alpha^\alpha\beta^\beta)^{\frac{1}{1-\alpha-\beta}} (s_p^\alpha s_m^\beta)^{\frac{1}{1-\alpha-\beta}} \quad (21a)$$

$$m_O^* = s_m\beta (A\alpha^\alpha\beta^\beta)^{\frac{1}{1-\alpha-\beta}} (s_p^\alpha s_m^\beta)^{\frac{1}{1-\alpha-\beta}} \quad (21b)$$

$$U_O = U_{FI}s_p(s_p^\alpha s_m^\beta)^{\frac{1}{1-\alpha-\beta}} \quad (21c)$$

$$W_O = (1 - s_p\alpha - s_m\beta) (A\alpha^\alpha\beta^\beta)^{\frac{1}{1-\alpha-\beta}} (s_p^\alpha s_m^\beta)^{\frac{1}{1-\alpha-\beta}} \quad (21d)$$

Equation 22a and Equation 22b show how \hat{p} changes with α and β respectively and Equation 22c is the cross partial of \hat{p} with respect to α and β .

$$\frac{\partial \hat{p}}{\partial \alpha} = \frac{\hat{p}}{1 - \alpha - \beta} \left(\ln p_O^* + \frac{1 - \beta}{\alpha} - \frac{(1 - \alpha - \beta) \ln \left(s_p \left(\frac{s_m}{s_p} \right)^\beta \right)}{\alpha^2 \beta} \right) \quad (22a)$$

$$\frac{\partial \hat{p}}{\partial \beta} = \frac{\hat{p}}{1 - \alpha - \beta} \left(\ln m_O^* - \frac{(1 - \alpha - \beta) \ln s_p}{\beta^2 \alpha} + 1 \right) \quad (22b)$$

$$\begin{aligned} \frac{\partial^2 \hat{p}}{\partial \alpha \partial \beta} &= \frac{\hat{p}}{(1 - \alpha - \beta)^2} \\ &\left(\left(\ln p_O^* - \frac{(1 - \alpha - \beta) \ln \left(s_p \left(\frac{s_m}{s_p} \right)^\beta \right)}{\alpha^2 \beta} + 1 \right) \left(\ln m_O^* - \frac{(1 - \alpha - \beta) \ln s_p}{\beta^2 \alpha} + 1 \right) \right. \\ &\quad \left. + \frac{1 - \beta}{\alpha} (\ln m_O^* + 1) + \ln p_O^* + 1 \right) \end{aligned} \quad (22c)$$

$s_p \left(\frac{s_m}{s_p} \right)^\beta$ and s_p are always less than one. If p_O^* and m_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.²⁰ 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_O^* and m_O^* are greater than one simply requires that seeking health care is valuable relative to the costs of effort.²¹ If this were not the case, one would imagine

²⁰This 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.

²¹ A must be ‘large’ compared to both 1 and D . Since A has no directly measurable units, but is meant

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, $\frac{\partial \hat{p}}{\partial \alpha}$ (Equation 22a), $\frac{\partial \hat{p}}{\partial \beta}$ (Equation 22b), and $\frac{\partial^2 \hat{p}}{\partial \beta \partial \alpha}$ (Equation 22c) are all positive.

Proof of Proposition 2 The proof of proposition 2 is outlined in two parts. In the first part the proposition is established in the case the outcome-contingent utility is equal to effort-contingent utility and in the second that outcome-contingent utility is greater the effort-contingent utility. The difference in welfare is:

$$W_O - W_E = \frac{W_{FI}}{1 - \alpha - \beta} \left((1 - s_p \alpha - s_m \beta) (s_p^\alpha s_m^\beta)^{\frac{1}{1-\alpha-\beta}} - \left(1 - \alpha - \beta \left(\frac{\tilde{p}}{p_{FI}^*} \right)^{\frac{\alpha(1-\alpha-\beta)}{(1-\alpha)(1-\beta)}} \right) \left(\frac{\tilde{p}}{p_{FI}^*} \right)^{\frac{\alpha\beta}{(1-\alpha)(1-\beta)}} \right) \quad (23)$$

If we examine $W_O - W_E$ at the point \hat{p} (where $U_O = U_E$) we obtain the following

$$W_O - W_E (U_O = U_E) = \frac{W_{FI}}{1 - \alpha - \beta} (s_p^\alpha s_m^\beta)^{\frac{1}{1-\alpha-\beta}} \left(1 - s_m \beta - s_p + \beta s_m s_p^{\frac{1}{\beta}} \right) \quad (24a)$$

$1 - s_m \beta - s_p + \beta s_m s_p^{\frac{1}{\beta}}$ is always greater than 0 when $s_p < 1$.²² Thus, Equation 24a is always positive.

The difference in welfare is positive if:

$$(1 - s_p \alpha - s_m \beta) (s_m^\beta s_p^\alpha)^{\frac{1}{1-\alpha-\beta}} - \left(1 - \alpha - \beta \left(\frac{\tilde{p}}{p_{FI}^*} \right)^{\frac{\alpha(1-\alpha-\beta)}{(1-\alpha)(1-\beta)}} \right) \left(\frac{\tilde{p}}{p_{FI}^*} \right)^{\frac{\alpha\beta}{(1-\alpha)(1-\beta)}} > 0 \quad (25)$$

Recall that \hat{p} is the value of \tilde{p} for which the utility under the two regimes is equal. We introduce a notation for \tilde{p} , $\tilde{p} = \hat{p}t$. When t is equal to one therefore, $\tilde{p} = \hat{p}$ and we have the solution outlined in equation (24a). When t is less than 1, $\tilde{p} < \hat{p}$ 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 (25) always holds.

We start with the fact that when $t = 1$ equation (25) is positive by proposition 2. Let g denote the expression in equation (25). Taking the derivative of g with respect to t we

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.

²²When $s_p = 1$, $1 - s_m \beta - s_p + s_p^{\frac{1}{\beta}} \beta s_m = 0$, and for all s_p $\frac{\partial(1 - s_m \beta - s_p + s_p^{\frac{1}{\beta}} \beta s_m)}{\partial s_p} = s_m s_p^{\frac{1-\beta}{\beta}} - 1$, 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.

obtain

$$\frac{\partial g}{\partial t} = \left(s_p^{\frac{1-\beta}{\beta}} t^{\alpha \frac{1-\alpha-\beta}{(1-\beta)(1-\alpha)}} s_m - 1 \right) s_p t^{\frac{1-\alpha-\beta}{(1-\beta)(1-\alpha)}} \alpha \frac{\beta}{1-\beta} \quad (26)$$

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 $\tilde{p} < \hat{p}$, or when $U_O > U_E$. QED.