Health Insurance, Moral Hazard, and Managed Care

Ching-to Albert Ma
Boston University
Boston, MA 02215
ma@bu.edu
and
Hong Kong University of Science and Technology
Clear Water Bay, Kowloon, Hong Kong

Michael H. Riordan
Graduate School of Business, Columbia University
New York, NY 10027
mhr21@columbia.edu

If an illness is not contractible, then even partially insured consumers demand treatment for it when the benefit is less than the cost, a condition known as moral hazard. Traditional health insurance, which controls moral hazard with copayments (demand management), can result in either a deficient or an excessive provision of treatment relative to ideal insurance. In particular, treatment for a low-probability illness is deficient if illness per se *has little effect on the consumer's marginal utility of income and if the consumer's price elasticity of expected demand for treatment is large relative to the risk-spreading distortion when these are evaluated at a copayment that brings forth the ideal provision of treatment. Managed care, which controls moral hazard with physician incentives, can either increase or decrease treatment delivery relative to traditional insurance, depending on whether demand management results in deficient or excessive treatment.*

1. Introduction

Before the proliferation of managed care, health insurance contracts relied heavily on deductibles, copayments, and other coverage limitations to control the provision of care. These *demand management* instruments cause consumers to limit their demand for health care, and

The authors acknowledge helpful comments from the editor, two referees, David Bradford, David Dranove, Thomas McGuire, and seminar participants at Columbia, Harvard, and Yale Universities, the Federal Reserve Bank of New York, Massachusetts Institute of Technology, Universitat Pompeu Fabra (Barcelona), Universitat Autonoma de Barcelona, and the Universities of New Hampshire, North Carolina, and Toulouse.

Journal of Economics & Management Strategy, Volume 11, Number 1, Spring 2002, 81–107

thereby reduce premiums by restraining health care costs, but may expose consumers to large uninsured risks. In contrast, managed care plans give health care providers explicit financial incentives, such as prospective payments, capitations, and cost-sharing contracts, to limit expensive treatments. Clearly, these *supply management* instruments have become more important.

Economic theory interprets the emergence of institutions as a response to market failures. In the health industry, market failure stems from the difficulty of contracting on illness *ex ante*.[1] Due to this failure, ideal insurance, which shields consumers from the costs of efficiently provided treatment, is infeasible. Second-best insurance pays only part of treatment expenses or limits coverage in other ways. But patients who do not fully face the cost of medical care may demand treatment inefficiently, in the sense that the cost of treatment sometimes exceeds the benefit. This problem is known as *moral hazard*. Demand and supply management policies in health care attempt to control moral hazard while partly insuring consumers against the risks of illness.[2]

In the 1960s, the health economics literature adopted the term moral hazard to describe the difficulty of contracting over health status (Arrow, 1963, 1968; Pauly, 1968; Zeckhauser, 1970), and argued that demand management at best only partly cures the moral hazard problem. The very first lines of Zeckhauser's 1970 article crystallize what is now the conventional wisdom:

> The primary purpose of medical insurance is to spread a risk, the risk of incurring substantial medical expenses. With risk-spreading, individuals will not pay the full amounts of such expenses. Insurance provision will thus introduce a perverse incentive toward overexpenditure if, as is usually the case, (1) the insured has substantial influence over the amount that is spent on their own behalf in any particular medical circumstance, and (2) the level of reimbursement by the insurance plan is a positively associated function of the expenses incurred by its insured.

Given that demand management does not solve the moral hazard problem, it is plausible to interpret supply management (man-

1. See Grossman and Hart (1986) on the difficulties of contracting on complex events.

2. The health economics literature has recognized that demand and supply management policies are complementary responses to market failure in the health industry. Ellis and McGuire (1993) emphasize the potential importance of financial incentives for consumers as well as for providers. Ma and McGuire (1997) use this basic premise to investigate the interaction between optimal insurance and provider payment. Newhouse (1996), in a recent survey, urges more study of supply-side incentives.

aged care) as an additional attempt to do so. On this interpretation, Zeckhauser's statement of the conventional wisdom suggests that moral hazard under demand management alone results in excessive provision of treatment, and that managed care reduces this excess. We demonstrate that this is not necessarily so. Under some conditions, optimal demand management results in deficient treatment, and managed care corrects this deficiency, leading to more treatment.

We construct a model of insurance for a particular treatment, and characterize conditions under which optimal demand management under moral hazard results in an excessive or deficient provision of treatment. By excessive (deficient), we mean that a higher (lower) fraction of individuals receive the treatment compared to the ideal benchmark in the absence of moral hazard. We extend the model to allow supply management (managed care), and prove that this relaxes the moral hazard constraint and, under certain conditions, achieves ideal insurance. In this case, the introduction of managed care increases (decreases) the provision of treatment if treatment under optimal demand management is deficient (excessive), thus alleviating the market failure due to moral hazard.

The provision of treatment can be deficient under optimal demand management if the illness *per se* has little effect on the consumer's marginal utility of income. This possibility depends on the *income effect* of the copayment: the patient's marginal utility of income is higher if the copayment is higher.[3] The intuition is roughly as follows. There are two important conflicting effects on consumer welfare of limiting treatment by charging a copayment higher than the one that elicits the ideal (first-best) provision. First, the consumer directly suffers more income risk, as indicated by a higher marginal utility of income after paying for treatment. *Ceteris paribus*, the cost to the consumer of this departure from optimal risk spreading is greater, the greater is the income effect of the copayment. Second, the consumer enjoys a lower premium because the insurance company has a lower expected cost. This second effect is more pronounced the greater is the income effect of the copayment relative to the marginal utility of income when an ill consumer declines treatment, because the consumer's demand response to the increased copayment is greater and translates into a

3. The health economics literature has paid little attention to the importance of income effects on the demand for treatment. In fact, a standard text in health economics only briefly mentions income effects (Phelps, 1992, pp. 32, 301). There are some notable exceptions. de Meza (1983) considers intertemporal income effects of savings and insurance, but assumes that illness severity is contractible so that optimal insurance involves an indemnity payment when the consumer is sick. Marshall (1976) also allows income effects, focusing on how the optimal insurance gives the consumer an incentive to exercise care to reduce the likelihood of illness.

larger premium reduction. Thus the premium-reduction effect is magnified if illness *per se* does not very much increase the marginal utility of income. The optimal provision of treatment under demand management is less than the ideal amount if the price elasticity of demand for treatment (and hence the premium reduction effect) is large relative to the risk-spreading distortion. While our main results are developed for the case of an infrequent illness, for which these clearly are the most important effects, we also show that the possibility of deficient treatment does not depend crucially on a small illness probability.

The rest of the article is organized as follows. The next section lays out a model of illness, treatment, and insurance. Section 3 studies the first best, when the treatment decision is fully contractible, and Section 4 characterizes the second best, when the provision of treatment is controlled indirectly by demand management. Section 5 compares the two, and establishes conditions under which the second-best provision of treatment with optimal demand management is more or less than the first best with ideal insurance. Section 6 employs these results as benchmarks to study supply management (managed care). Section 7 concludes by summarizing, and describing possible future research. An appendix contains the formal proofs of propositions.

2. The Basic Model

We focus on optimal health insurance for a particular discrete treatment.[4] Implicitly, we hold fixed other health and disability insurance covering the consumer and study insurance for this particular treatment in isolation. To be concrete and to focus our discussion, we interpret the treatment as being suitable for a specific diagnosis.[5] For example, the treatment might be a new class of prescription drugs for high cholesterol (statins) described in a recent *Wall Street Journal* article.[6]

4. This contrasts with a more familiar reduced-form approach in the health care literature, which assumes that a consumer's health is a function of undifferentiated health care expenditures. See, for example, Baumgardner (1991). Our discrete-choice approach perhaps provides a more useful conceptual basis for empirical analyses of micro data that indicate whether a patient received a particular treatment or not. See, for example, Manning and Phelps (1979). Of course, it will be important in future theoretical work to study the bundled coverage of the panoply of treatments typical of health insurance policies.

5. More generally, the treatment in question might be suitable for a variety of diagnoses. For example, the treatment might be an extra day in the hospital due to complications from a surgical procedure. Our model is consistent with this broader interpretation.

6. See "Price Prescription: Powerful Medications for Cholesterol Pose a Paradox for HMO's," by Ron Winslow, *Wall Street Journal*, December 6, 1996, p. A1.

We assume that a consumer becomes ill with probability λ, $0 < \lambda < 1$. For example, λ is the probability that a consumer is diagnosed as having a high cholesterol condition and is potentially a candidate for statins. Many of our results focus on the case where λ is small, which of course is often realistic. For example, according to the aforementioned *Wall Street Journal* article, 13 million people with serious heart problems are candidates for statins. While this is a lot of people, the probability that one of us will become a candidate for statins in the next year is hopefully quite small. It is easy to think of many other expensive treatments and tests that may be beneficial to an individual consumer with a small probability (organ transplants, neural surgeries, etc.). Our model focuses on health insurance coverage for one of these treatments in isolation.

For simplicity, we assume that the cost of diagnosing the illness is zero.[7] Illness varies by its severity, which is described by a random variable ℓ with a cumulative distribution function $F(\ell)$ and a density function $f(\ell)$ with support on the positive real line.[8] As we make clear below, the severity of illness, ℓ, indexes the benefits of treatment, and $F(\ell)$ describes the distribution of benefits across the population of patients. The consumer learns the benefits of treatment (presumably in consultation with a physician) only after becoming ill. We assume that illness is not contractible,[9] and the insurance company can contract neither on the event of illness nor on the realization of ℓ.

We adopt an eclectic approach to modeling the consequences of an untreated illness. Specifically, consumers' preferences are represented by a utility function of the form $U(y - a\ell) - b\ell$. The variable y represents the consumer's expenditure on other goods. The function $U(\cdot)$ is differentiable, increasing, and strictly concave, reflecting the consumer's risk aversion to income fluctuations, and a and b are nonnegative parameters. Nested in this functional form are two special cases of interest. If $b = 0$ and $a > 0$, then illness is completely equivalent to a loss of monetary income; we call this the *monetary loss model*. In this case, we sometimes set $a = 1$ as a normalization, measuring the loss of health resulting from illness in monetary units. The second special case is the *utility loss model* with $a = 0$ and $b > 0$.

7. A fixed and known diagnostic cost will not change our qualitative results, as long as the outcome of the diagnosis is not contractible.

8. The random probability of illness effectively puts a mass point at $\ell = 0$. It is easy to extend the model to allow for bounded severity, and also for a strictly positive lower bound of the support to be strictly positive, as for a Pareto distribution (Ma and Riordan, 2000).

9. If the event of illness were contractible, the optimal insurance contract could involve a monetary transfer payment to patients who become ill. The case is perhaps better understood as disability insurance.

Here, illness affects an additive utility loss and does not directly affect the consumer's marginal utility of income; we sometimes also use the normalization $b = 1$, interpreting ℓ as an expected utility loss from forgoing treatment. This broad interpretation implicitly recognizes the uncertainties surrounding illness and alternative treatments. The general formulation allows for mixtures of monetary losses and (additive) utility losses from illness.

The utility loss model seems more plausible than the monetary loss model, though both are special cases. The monetary loss model implies that the demand for treatment does not depend on income. This seems unrealistic for many discretionary treatments. Manning and Phelps (1979) found significant income effects for specific dental treatments. The RAND Health Insurance Study generally found small but positive income elasticities, and subsequent studies using different methods and data found larger elasticities (Phelps, 1992). In contrast, the utility loss model implies a positive income elasticity of the demand for treatment without restricting its magnitude. The utility loss model does imply that a consumer's attitude toward income risk is independent of illness *per se*. Obviously, this is an empirical question, but seems extremely difficult to answer. We have no *a priori* reason to think that the effect of illness on risk aversion goes one way or another. Therefore, it seems reasonable, at least as a first step, to ignore this effect. For these reasons, we are partial to the utility loss model.

We make several simplifying assumptions about treatment. First, the cost of treatment is fixed at $C > 0$. It will be useful to think of C as a large number, creating a significant demand for insurance, but not prohibitively large.[10] For example, the *Wall Street Journal* article says that statin therapy costs $700 a year for life. While this is an expensive therapy, some consumers' benefits obviously are much higher than $700 a year. Second, if the patient receives treatment, then the losses from illness are eliminated completely.[11] Thus, ℓ indexes the benefits from receiving treatment. These benefits vary across patients, implying that some patients are better candidates for treatment than

10. Obviously, this treatment cost cannot be too large. Otherwise, there would be little demand for insurance in our model. If C is sufficiently small, then there is a demand for insurance to smooth the income fluctuation resulting from illness and treatment. Interestingly, in the monetary loss model, the fluctuation to be smoothed is the monetary equivalent of illness itself. In the utility loss model, the fluctuation to be smoothed is the cost of treatment.

11. As mentioned above, we do not need to assume this in the utility loss model. In this special case, ℓ is the benefit of possibly imperfect treatment.

others.[12] Third, we model the treatment decision as a binary choice. Either the patient receives treatment or not.[13] Consequently, for a fixed copayment, there is a critical value of ℓ such that more severely ill patients demand treatment while less severely ill patients do not, as we show later. Finally, the delivery of treatment is contractible, meaning that the insurance company can prevent fraudulent claims for the reimbursement of nonexistent treatment costs.

3. Ideal insurance

We begin by studying the ideal insurance contract in the absence of moral hazard. In this regime, illness and loss are assumed to be completely contractible; payments and treatment decisions can be contingent on the severity of illness. Under an ideal contract, the consumer pays a fixed premium P, and receives treatment whenever the benefit of treatment, ℓ, is above a fixed threshold L. Moreover, when treatment is withheld, the consumer can be compensated by an indemnity payment, $t(\ell)$, that depends on the severity of the illness. In principle, the consumer receiving treatment can also receive a transfer payment. This possibility, however, plays no role in our model, because treatment eliminates all illness losses. Therefore, under an ideal contract a consumer with income Y enjoys an expected utility given by the formula

$$(1-\lambda)U(Y-P) + \lambda\left\{\int_0^L [U(Y-P-a\ell+t(\ell))-b\ell]f(\ell)d\ell+[1-F(L)]U(Y-P)\right\}. \quad (1)$$

This expected-utility formula is a weighted sum of two components. The first component is the utility of a healthy consumer who has paid the insurance premium.[14] The second component (in curly braces) is the conditional expected utility of an ill consumer, who receives treatment when the benefit is sufficiently great, but otherwise is compensated only by the indemnity payment. The weights are the respective probabilities of illness and health.

12. We emphasize that our interpretation of ℓ is the treatment's potential benefit to patients. One could interpret alternatively that ℓ refers to patients' losses, although we do not think that this is always true. For some illnesses, it may be more natural to allow for multiple treatments when severity varies. See Chernew et al. (2000).

13. A more general, mechanism design approach to optimal insurance would allow the insurance company to randomize the provision of treatment. We ignore this possibility because it is unrealistic.

14. The variable Y can be interpreted as income net of premiums paid for other health insurance. Thus $Y-P$ is net income after the premium insuring for this particular treatment.

The insurance premium must cover the expected costs of treatment and indemnity payments. Thus, the premium constraint for the ideal contract is given by the inequality

$$P \geq \lambda\left(\int_0^L t(\ell)f(\ell)\,d\ell + [1 - F(L)]C\right). \tag{2}$$

The ideal contract maximizes expected utility subject to the premium constraint; it is the solution $(P^*, t^*(\ell), L^*)$ to the problem of maximizing (1) subject to (2). The premium constraint is necessarily binding at an optimum.

Proposition 1: *The ideal insurance contract specifies (a) a treatment threshold that equates the marginal benefits of treatment to the marginal cost, i.e.,*

$$U'(Y - P^*)aL^* + bL^* = U'(Y - P^*)C; \tag{3}$$

(b) indemnity payments that exactly compensate for the monetary losses from an untreated illness, i.e., $t^*(\ell) = a\ell$, *and (c) a premium that exactly covers the expected costs. Finally, fewer patients receive treatment if the cost of treatment is higher, i.e.,* L^* *is strictly increasing in* C.

The characterization of the optimal treatment threshold needs a bit more interpretation. The marginal cost of treatment, C, is in monetary units. To measure the value of C in utility units, it is multiplied by the marginal utility of income, $U'(Y - P^*)$. Equation (3) says that the utility of treatment evaluated at the threshold level of illness, $[U'(Y - P^*)a + b]L^*$, equals the cost of treatment measured in utility, $U'(Y - P^*)C$. In the monetary loss model ($a = 1$, $b = 0$), equation (3) states that treatment will be provided whenever the monetary loss from an untreated illness exceeds the monetary cost of treatment ($\ell \geq C$). In the utility loss model ($a = 0$, $b = 1$), the treatment threshold is set equal to the cost of an untreated illness $[L^* = U'(Y - P)C]$, and there is no indemnity payment for an untreated illness $[t^*(\ell) = 0]$.

4. Demand Management

We now turn to optimal insurance when illness and loss are not contractible. The insurance company contracts with the consumer to reimburse part or all of the cost of treatment, but does not directly observe illness. The treatment decision itself is delegated to the patient, presumably in consultation with a physician.[15] We refer to this scheme of insurance as demand management.

15. Implicitly, we assume that the physician is a perfect agent for consumers. At the end of this section, we briefly discuss competition between self-interested physicians. In Section 5, we examine physician incentives under managed care.

Under a demand-managed insurance contract, the patient pays a fixed premium P up front, and copays D for treatment. Given this contract, the patient demands treatment when the benefits exceed the utility cost of the copayment. Thus treatment is provided when the severity of illness exeeds a threshold value L satisfying the *treatment constraint*

$$U(Y - P - aL) - bL = U(Y - P - D). \tag{4}$$

This constraint states that a patient with illness severity L is indifferent about receiving treatment, and implicitly defines L as a function of D and P. This function determines the expected demand for treatment by consumers who fall ill, which is equal to $1 - F(L)$.

As with ideal insurance, the premium must cover the expected cost of the insurance company. Since treatment is provided only when illness severity exceeds L, the *ex ante* probability of treatment is $\lambda[1 - F(L)]$, and the insurance company breaks even if the premium covers its expected liability[16]:

$$P = \lambda[1 - F(L)](C - D). \tag{5}$$

The breakeven constraint defines a trade-off between the premium and the copayment[17]:

$$\frac{dP}{dD} = -\lambda[1 - F(L)] + \lambda f(L)(C - D)\left(\frac{\partial L}{\partial D} + \frac{\partial L}{\partial P}\frac{dP}{dD}\right).$$

The first term in this expression is the direct effect of shifting more cost to the patient, while the second is the marginal demand reduction from a higher copayment. Solving for the total effect reveals that a higher copayment translates into a lower premium[18]:

$$\frac{dP}{dD} = \frac{\lambda[1 - F(L)][aU'(Y - P - aL) + b] + \lambda f(L)(C - D)U'(Y - P - D)}{-[aU'(Y - P - aL) + b] - \lambda f(L)(C - D)[U'(Y - P - D) - U'(Y - P - aL)]} < 0. \tag{6}$$

16. The premium constraint, $P \geq \lambda[1 - F(L)](C - D)$, must be binding for an optimal insurance contract, because expected utility, given by (7) below, is decreasing in the premium.

17. From the treatment constraint, $\partial L/\partial D = -[U'(Y - P - D)]/[aU'(Y - P - aL) + b] > 0$ and $\partial L/\partial P = -[U'(Y - P - D) - U'(Y - P - aL)]/[aU'(Y - P - aL) + b] > 0$. For a small illness probability, the effect of $(\partial L/\partial P)\ dP/dD$ is of second-order importance.

18. $dP/dD < 0$ follows from the fact that the treatment constraint implies $D \geq aL$.

Consumer expected utility is a probability-weighted average of utility when healthy and when ill:

$$(1-\lambda)U(Y-P) + \lambda\left(\int_0^L [U(Y-P-a\ell)-b\ell]f(\ell)d\ell + [1-F(L)]U(Y-P-D)\right). \quad (7)$$

The event of illness occasions two possible losses. First, a consumer suffers monetary ($a > 0$) and additive utility ($b > 0$) losses from an untreated illness (when $\ell \leq L$). Second, the copayment reduces the disposable income of a patient receiving treatment (when $\ell \geq L$).

Optimal demand management sets a premium and copayment to maximize expected utility, given that consumers determine treatment and the insurance company breaks even. Thus, the optimal contract (P^d, D^d, L^d) maximizes (7) subject to (4) and (5). We call this the *demand management problem*. A first result, traced to Zeckhauser (1970), is that consumers are only partially insured for the cost of treatment. The consumer faces some risk of untreated illness, and copays less than the full cost of treatment. The solution balances the negative consumer welfare effect of a higher copayment against the positive effect of a lower premium.

Proposition 2: *Optimal demand management partially insures consumers against the cost of treatment* ($0 < D^d < C$), *and less severely ill patients decline treatment* ($L^d > 0$). *The optimal copayment balances the expected utility cost of a marginally higher copayment against the corresponding benefits of a lower premium, i.e.,*

$$-\left((1-\lambda)U'(Y-P) + \lambda\int_0^L U'(Y-P-a\ell)f(\ell)\,d\ell + \lambda[1-F(L)]U'(Y-P-D)\right)\frac{dP}{dD} = \lambda[1-F(L)]U'(Y-P-D), \quad (8)$$

where dP/dD is given by (6).

The proposition reveals two marginal effects on consumer welfare of raising the copayment: a direct effect of more income risk, and a premium-reduction effect. At an optimum, there is no need to weigh the marginal effects on the demand for treatment, because these are of second-order importance.[19]

19. Taking the total derivative of (7) with respect to (D, P, L) yields

$$-\left((1-\lambda)U'(Y-P) + \lambda\int_0^L U'(Y-P-a\ell)f(\ell)\,d\ell + \lambda[1-F(L)]U'(Y-P-D)\right)dP$$
$$-\lambda[1-F(L)]U'(Y-P-D)\,dD + \lambda[U(Y-P-aL)-bL-U(Y-P-D)]f(L)\,dL.$$

The treatment constraint (4) implies that the last term vanishes, leaving the two effects identified in the proposition.

The optimal copayment depends on the consumer's risk aversion and on the price elasticity of the expected demand for treatment, both of which depend on the curvature of the utility function. The well-known Arrow-Borch condition for optimal risk-spreading requires that marginal utilities of income be equal in all states of illness and health. However, a copayment will create a distortion from optimal risk-spreading across the states in which the consumer is healthy and the states in which the consumer is treated for illness. The magnitude of the *Arrow-Borch distortion* for a given copayment, $U'(Y-P-D)/U'(Y-P)-1$, depends only on the curvature of the utility function on the domain between $Y-D$ and Y. The consumer's elasticity of expected demand for treatment with respect to the copayment (holding the premium constant),[20] $\{f(L)/[1-F(L)]\}DU'(Y-D)/[aU'(Y-aL)+b]$, also depends on the curvature of the utility function, as well as on the losses from untreated illness and on the severity distribution. These two magnitudes, the Arrow-Borch distortion and the price elasticity of demand, are prominent in determining the optimal cost-sharing ratio $(C-D)/D$ when the probability of illness is small.

Corollary 1: *If the probability of illness is small, then the optimal ratio of cost shares for the insurance company and the patient is approximately equal to the ratio of the Arrow-Borch distortion to the price elasticity of demand for treatment, i.e.,*

$$\left(\frac{f(L^d)}{1-F(L^d)}\frac{D^dU'(Y-D^d)}{aU'(Y-aL^d)+b}\right)\frac{C-D^d}{D^d}=\frac{U'(Y-D^d)}{U'(Y)}-1 \tag{9}$$

in the limit as $\lambda \to 0$.[21]

20. To derive this elasticity, differentiate $\lambda[1-F(L)]$ with respect to D, using (4) and the chain rule, and apply the definition of elasticity.

21. The corollary is proved by substituting (6) into the first-order condition in Proposition 2, dividing through by λ, and taking the limit as $\lambda \to 0$. In the limit, the premium is zero because the consumer never becomes ill ($P^d=0$). However, the premium is positive in the neighborhood of the limit. In general, the consequences of a change in the premium on expected consumer welfare are multifaceted, partly because the marginal utility of income varies across the different states of illness and health. However, some of the premium effects of a small variation of the insurance contract are more important than others, viz., some are proportional to λ (first-order importance) while others are proportional to λ^2 (second-order importance). The small-λ approximation focuses on the first-order effects. The second-order effects that the approximation ignores can be described as follows: any change in the premium (P) affects consumer welfare differently across the different states of illness and health because of the income effects of the copayment (D) and untreated illness (aL). Because the first-order effects are roughly proportional to λ and the second-order effects are roughly proportional to λ^2, there is no particular reason to think that the second-order effects dominate for plausible values of λ.

The corollary illustrates neatly the conflict between providing insurance and controlling moral hazard. On the one hand, if the consumer is highly averse to income risk (a large Arrow-Borch distortion), then the insurance company should bear a high fraction of the treatment cost in order to better insure the patient. On the other hand, if the demand for treatment is sensitive to price (a high price elasticity of demand), then the consumer should face a substantial treatment expense in order to curtail an excessive demand. Optimal cost-sharing balances these two concerns.

We close this section on a technical note. The demand management problem is not in general a concave programming problem. Therefore, the first-order condition for an optimal contract in Proposition 2 is not necessarily sufficient, nor is the solution necessarily well behaved. The following corollary provides an assumption under which this is not an issue.[22]

Corollary 2: *Assume that the following function is strictly increasing in D:*

$$D+\left(\frac{1}{U'(Y)}-\frac{1}{U'(Y-D)}\right)[aU'(Y-aL)+b]\frac{1-F(L)}{f(L)},$$

with L determined by the treatment constraint and $P=0$ *(the hazard-rate assumption). For* λ *sufficiently small, the optimal threshold* L^d *and copayment* D^d *are unique, differentiable, and increasing in the treatment cost C.*

The hazard-rate assumption requires that the inverse hazard rate, $H(L) \equiv [1-F(L)]/f(L)$, not decline too quickly. It is satisfied, for example, if $H(L)$ is constant (exponential distribution) or increasing (e.g., Pareto distribution).

5. Does Moral Hazard Increase Treatment?

We now turn to our main question. How does moral hazard influence treatment? Is the delivery of treatment under demand management more or less than that under the ideal contract? We make two points in this section. First, the provision of treatment can be either deficient ($L^d > L^*$) or excessive. Second, deficient treatment is a robust possibility, and occurs under various plausible conditions. We focus on the case of a small probability of illness, λ.

22. See Proposition 4 in Ma and Riordan (2000) for a formal proof. The assumption is not necessary for the result of the corollary. It is sufficient that the demand management problem have a unique continuous global maximum.

It is convenient to reformulate the demand management problem slightly. The treatment constraint (4) and the binding premium constraint (5) implicitly define the premium P and copayment D as functions of treatment threshold L. Solving for P and D as functions of L, and substituting these into the expression for consumer welfare (7), we obtain an expression for consumer welfare as a function of L alone, which we denote by $Z(L)$. Optimal demand management determines a treatment threshold that maximizes $Z(L)$. Under the hazard-rate assumption (of Corollary 2), $Z(L)$ is a quasiconcave function and therefore achieves a unique local maximum at L^d. We ask whether the derivative $Z'(L^*)$ is positive or negative. If $Z(L)$ is quasiconcave and $Z'(L^*) > 0$, then treatment is provided deficiently under demand management ($L^d > L^*$), and conversely.

Proposition 3: *Under the hazard-rate assumption, if the probability of illness is small, then treatment under optimal demand management is deficient if the cost-share ratio exceeds the ratio of the Arrow-Borch distortion to the price elasticity of demand, evaluated at the copayment for which the consumer demands treatment if and only if the illness severity (ℓ) exceeds the ideal treatment threshold (L^*); i.e., as $\lambda \to 0$, $L^* < L^d$ if and only if*

$$\left(\frac{f(L^*)}{1-F(L^*)}\frac{D^*U'(Y-D^*)}{aU'(Y-aL^*)+b}\right)\frac{C-D^*}{D^*} > \frac{U'(Y-D^*)}{U'(Y)} - 1, \tag{10}$$

where D^* satisfies the treatment constraint (4) evaluated at $P = 0$ and $L = L^*$. The converse also holds.

In the monetary loss model ($b = 0$), the demand price for L^* is equal to the cost of treatment, i.e., $D^* = C$. In this case, inequality (10) obviously fails and we conclude that more treatment is provided under demand management than under the ideal contract. In this special case, an uninsured consumer seeks treatment efficiently (i.e., only when the monetary benefit exceeds the cost: $a\ell > C$), but this cannot be optimal. A risk-averse consumer will always demand some insurance, resulting in an excessive provision of treatment; starting from $D = C$, the consumer is willing to pay a higher premium to achieve a lower copayment, which increases the demand for treatment.[23]

Corollary 3: *In the monetary loss model ($b = 0$), treatment is excessive under demand management ($L^d < L^*$).*

23. The higher premium has a much smaller effect on the demand for treatment, because of the low probability of illness. Actually, though, the following corollary does not require a small illness probability. For the monetary loss model, treatment is always excessive relative to the first best. That is, the conditions in Proposition 3 are unnecessary. See Ma and Riordan (2000).

For the utility loss model ($a = 0$), (3) implies that $b = U'(Y)C/L^*$ as λ tends to 0, because the premium is proportional to λ. In this limiting case, inequality (10) is equivalent to

$$\frac{C - D^*}{C} \frac{U'(Y - D^*)}{U'(Y - D^*) - U'(Y)} > \frac{1 - F(L^*)}{L^* f(L^*)} \tag{11}$$

with D^* satisfying $U(Y) - U(Y - D^*) = CU'(Y)$. Substituting for C using the definition of D^*, we can simplify (11) to

$$\frac{1 - \frac{U'(Y)}{[U(Y) - U(Y - D^*)]/D^*}}{1 - \frac{U'(Y)}{U'(Y - D^*)}} > \frac{1 - F(L^*)}{L^* f(L^*)},$$

whose left-hand side lies between 0 and 1 because $U'(Y - D^*)D^* > U(Y) - U(Y - D^*)$. The elasticity of demand with respect to the disutility ℓ is $\ell f(\ell)/[1 - F(\ell)]$; we refer to this as the *loss elasticity* of demand. If demand is inelastic with respect to the utility loss, then the right-hand side of (11) is greater than 1, in which case (10) is violated and treatment is excessive. The other side of the coin is that moral hazard causes deficient treatment if the loss elasticity is sufficiently large.

Corollary 4: *In the utility loss model ($a = 0$ and $b = 1$), under the hazard-rate assumption, if the illness probability is small, and demand is inelastic with respect to the additive utility loss evaluated at the ideal threshold ($L^* f(L^*)/[1 - F(L^*)] < 1$), then treatment is excessive under demand management ($L^d < L^*$); if the loss elasticity is sufficiently large, then treatment is deficient.*

The corollary can also be interpreted as saying that treatment is deficient in the additive loss model if and only if the price elasticity of demand is sufficiently high. The ordinary price elasticity of demand in the utility loss model $\{L^* f(L^*)/[1 - F(L^*)]\}\, D^* U'(Y - D^*)/CU'(Y)$ is proportional to the loss elasticity. Therefore, holding the utility function constant, an increase in the loss elasticity (which depends only on F) translates directly into an increase in the price elasticity of demand.

An interesting special case of the utility loss model occurs when the utility-of-income function exhibits constant absolute risk aversion, i.e., $U(y) = -\exp(-ry)$, where $r > 0$ is the coefficient of absolute risk aversion. In this case, $D^* = \ln(1 + rC)$, and (11) is equivalent to

$$\frac{rC - \ln(1 + rC)}{rC} \left(\frac{1 + rC}{rC} \right) > \frac{1 - F(L^*)}{L^* f(*)}. \tag{12}$$

The right-hand side of (12) is the elasticity of demand with respect to the disutility of illness, evaluated at L^*. The left-hand side is an

increasing, concave function of rC, ranging from 0.5 as $r \to 0$ to 1.0 as $r \to \infty$. Thus, holding the right-hand side constant,[24] we conclude from this special case that treatment is deficient if the consumer is sufficiently risk-averse and the loss elasticity is sufficiently small. The intuitive explanation for this result is subtle. A higher degree of risk aversion clearly increases the magnitude of the Arrow-Borch distortion when D is raised above D^* to curtail treatment. However, a higher degree of risk aversion also corresponds to a greater income effect of D on the demand for treatment. Thus, if r is higher, then a given increase in D curtails treatment more, resulting in a greater reduction in the premium. The latter effect dominates, making the increase in D attractive to the consumer from an *ex ante* perspective.

The above example and corollaries clearly demonstrate that deficient treatment is a robust possibility. This robustness argument can also be made in a more general way. Toward this end, for a given L^*, normalize the parameters a and b so that

$$b = U'(Y)\left[\frac{C}{L^*} - a\right]. \tag{13}$$

Thus the treatment threshold under the ideal contract is held constant as the parameters a and b vary according to this constraint. The inequality (10) can be expressed alternatively as

$$\frac{C - D^*}{\frac{1}{U'(Y)} - \frac{1}{U'(Y-D^*)}} > \frac{1 - F(L^*)}{f(L^*)}\left(aU'(Y - aL^*) + U'(Y)\left[\frac{C}{L^*} - a\right]\right). \tag{14}$$

The normalization (13) keeps L^* constant at the limit as λ tends to zero. An increase in a increases the right-hand side of (14) because $U'(Y - aL^*) > U'(Y)$. Moreover, since P tends to zero with λ, the treatment constraint implies that D^* increases with a for the same reason. Therefore, the left-hand side of (10) decreases as the value of a increases, and (14) is more likely to hold when a is smaller. Now, fixing values for C, $U(Y)$, $U'(Y)$, and D^*, we can make the left-hand side of (11) arbitrarily large by choosing a utility function with enough curvature on the interval $[Y, Y - D^*]$.[25] This curvature is greater the more risk-averse is the consumer on this interval. Putting together these implications of Proposition 3, we conclude that deficient treatment is a robust possibility over a range of values of a and b.

24. Note that L^* can be held constant as rC varies by adjusting Y suitably.

25. Meanwhile, the right-hand side can be held constant by adjusting a appropriately, and is constant if $a = 0$.

Corollary 5: *Under the hazard-rate assumption and for a small illness probability* λ*, less treatment is delivered under demand management than under the ideal contract* ($L^d > L^*$) *if and only if the parameter a is sufficiently small and the consumer is sufficiently risk-averse.*

The robust possibility of deficient treatment does not hinge on a small illness probability, as mentioned earlier. Let the utility function be logarithmic: $U(Y) = \ln y$, and assume a standard uniform distribution for losses: $f(\ell) = 1$, $F(\ell) = \ell$ with $0 \leq \ell \leq 1$ (thus dispensing with the hazard-rate assumption as well). Let the cost of treatment be expressed as a fraction of income $C = \gamma Y$. Numerical results are displayed in Figure 1. The graphs describe the treatment thresholds L as functions of λ: the lower one corresponds to the treatment threshold under the ideal contract; the higher one, demand management. Figure 1 shows that the threshold under demand management is higher for all values of λ. The graphs correspond to an expensive treatment: the value of γ was set at 0.65. Further numerical comparisons show that the difference between the two thresholds (the second-best threshold minus the first-best threshold) increases with γ. For small values of γ, this difference is uniformly negative, while for high values of γ, it is uniformly positive. We have found that there are two local maxima for the numerical example and that the discontinuity in the second-best L represents a jump between them. Thus, the numerical example also illustrates that our deficient-treatment possibility is robust even when the objective function is not quasiconcave.[26]

In the introduction, we discussed the conventional wisdom that moral hazard causes an excessive provision of treatment. Of course, the conventional wisdom is imprecise, as are most conventional wisdoms. What is the relevant benchmark for determining if treatment is excessive under demand management? We have shown that the conventional wisdom is incorrect if the relevant benchmark is the delivery of treatment under an ideal contract when moral hazard is absent.[27]

We have assumed implicitly that physicians act as perfect agents for the patients. In Ma and Riordan (2000), we extend the utility

26. It also illustrates how moral hazard can result in a rather extreme market failure—no treatment or insurance.

27. Other benchmarks for the conventional wisdom are possible: (I) the amount of treatment that would be provided if consumers lacked insurance or (II) the amount of treatment that is efficient *ex post*, i.e., after consumers have paid the premium. The conventional wisdom according to interpretation (I) is obvious. The conventional wisdom according to interpretation (II) is not necessarily correct. Once consumers have paid the insurance premium P^d, it is socially efficient to provide treatment to

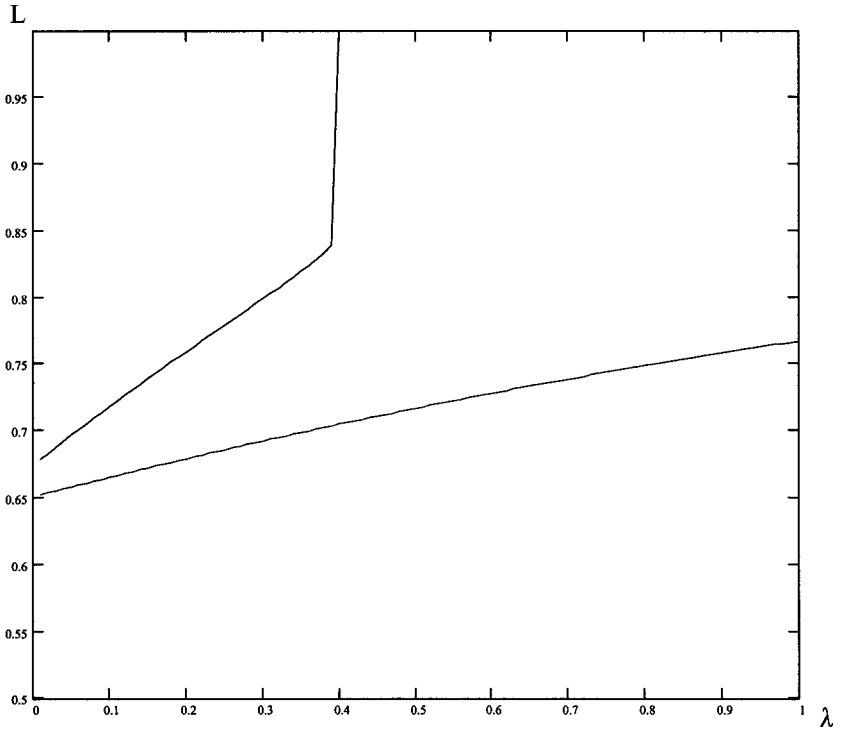

FIGURE 1. OPTIMAL TREATMENT THRESHOLD AS FUNCTION OF PROBABILITY OF ILLNESS

loss model to include physician competition, and consider the other polar case in which the physicians are completely self-interested. In this alternative model, physicians compete for patients by adopting a *practice style* that commits them to a general level of service quality

---

an individual patient whenever the severity of illness, ℓ, exceeds a threshhold L^{*d} satisfying

$$U(Y - P^d) - [U(Y - P^d - aL^{*d}) - bL^{*d}] = U'(Y - P^d)C.$$

If the probability of illness is small, then $L^d > L^*$ implies $L^d > L^{*d}$. The reason is that, if the illness probability is small and P^d is close to zero, then L^{*d} cannot be very much less than L^*. The formal proof is as follows. As $\lambda \to 0$, $P^* \to 0$ and $P^d \to 0$. Therefore, if $L^* < L^{*d}$ in the limit, then

$$bL^{*d} = U'(Y)C - [U(Y) - U(Y - aL^{*d})] \leq U'(Y)[C - aL^{*d}] \leq U'(Y)[C - aL^*] = bL^*,$$

a contradiction.

and a treatment threshold. Surprisingly, optimal demand management when profit-maximizing physicians compete on practice style yields results almost identical to the case when physicians are perfect agents for consumers. Competition forces physicians to act in the interests of their patients, and in equilibrium, physicians adopt a treatment threshold that maximizes a consumer's *ex post* utility.

6. Supply Management and Managed Care

A distinguishing feature of managed care is that health-care providers are given explicit incentives to limit treatments. Contracts with physician groups often feature a "capitated payment" for each patient but leave the physicians responsible for some or all of the treatment costs. This gives physicians an incentive to ration care in order to economize on treatment costs. Such supply management potentially complements and may even replace demand management as a method to control moral hazard. In this section, we introduce a profit-conscious physician into the model and allow a role for managed care. For simplicity, we rely on the utility loss model. We argue that managed care benefits consumers and may expand treatment. Indeed, managed care achieves ideal insurance in some cases.

The technology of illness and treatment is the same as in our basic model. Consumer preferences have a utility loss representation: $U(y) - \ell$. The insurance company contracts with the risk-averse consumer, and also with a risk-neutral physician (or physician group) to diagnose the patient and determine treatment. The physician learns the realization of ℓ from the diagnosis, and subsequently decides on treatment.

We interpret managed care as a capitated payment arrangement in which the physician is credited a total payment of $S + B$ for each diagnosed patient, but is liable for an amount B of the treatment cost. Equivalently, the physician is paid a fixed fee, S, for a diagnosis, and an additional bonus, B, if the diagnosed patient does not receive treatment.[28] If the physician treats a patient with $\ell > L$, then his expected payment is $\lambda[S+F(L)B]$. Without loss of generality, we assume that the treatment cost C is paid directly by the insurance company, i.e., the physician receives a net payment that deducts treatment-cost liability from the capitation payments.

The treatment decision is jointly determined by the physician and patient. To quote Arrow (1963, p. 960):

28. In principle, the contract could specify a payment to the physician even if the consumer does not become ill. However, under the limited liability assumption introduced below, this payment must be zero.

> By certifying to the necessity of given treatment or the lack thereof, the physician acts as a controlling agent on behalf of the insurance companies. Needless to say, it is a far from perfect check; the physicians themselves are not under any control, and it may be convenient for them or pleasing to their patients to prescribe more expensive medication, private nurses, more frequent treatments, and other marginal variations of care.

In this spirit, we interpret the treatment decision as a collective decision that maximizes a weighted sum of the benefits to the physician and the patient.[29] This might be interpreted as a "reduced form" of a bargaining model, with the weights representing the bargaining strengths of the physician and the patient. Implicitly, we assume that bargaining occurs after the consumer becomes ill, but before the patient learns the severity of illness. Thus, the collective treatment decision in the utility loss model establishes a threshold that maximizes the physician's expected treatment payment minus the weighted patient's expected loss from illness, i.e., a value of L that maximizes

$$S + F(L)B - \theta\left[\int_0^L \ell f(\ell)\, d\ell + [1 - F(L)][U(Y - P) - U(Y - P - D)]\right].$$

The magnitude of the weight θ captures the bargaining strength of the patient, i.e., the higher the value of θ, the greater the weight given to the patient's expected loss.

The solution to the collective decision problem balances the physician's bonus payment against the treatment benefits of the marginal patient:

$$B - \theta[L - U(Y - P) + U(Y - P - D)] = 0. \tag{15}$$

This optimality condition is analogous to a treatment constraint in the demand management problem. If $B = 0$, then it is identical. If $B > 0$, so that the physician is at least partially liable for the treatment cost, then it is weaker [$L > U(Y - P) - U(Y - P - D)$] and some patients are rationed, i.e., a patient who wants treatment does not get it.[30]

29. Ellis and McGuire (1990) use a similar assumption.

30. The physician has no incentive to treat the patient and then not report the treatment to the insurer in order to collect the bonus. This is because the physician would have to be responsible for the treatment cost.

The other constraints on contractual relationships of the insurance company are as follows. First, the premium and expected copayment must cover the expected costs to the insurance company:

$$P = \lambda\{S + F(L)B + [1 - F(L)](C - D)\}. \tag{16}$$

Second, the physician must find the relationship profitable:

$$\lambda[S + F(L)B] \geq \overline{U}. \tag{17}$$

The parameter $\overline{U}$ in this individual rationality constraint is the opportunity value of the physician's time and effort in diagnosing and treating the patient. Finally, financial constraints put lower bounds on payments to the physician, i.e.,

$$S \geq \underline{S} \qquad S + B \geq \underline{S}. \tag{18}$$

The second of these limited-liability constraints is generally slack, and will be ignored. The first may or may not be slack. The assumption that S is bounded from below by $\underline{S}$ allows the interpretation that the physician must be guaranteed a minimum compensation for the opportunity cost of diagnosis.[31]

The optimal insurance contract maximizes the *ex ante* expected utility of the patient, given by (7) after setting $a = 0$ and $b = 1$ to conform with the utility loss model, subject to the treatment, individual rationality, and limited-liability constraints, given by (15) to (18). The problem can be simplified further. At an optimum, the premium constraint (16) must bind. Using this and the treatment constraint to eliminate S and B from (17) and (18) leaves a modified individual rationality constraint,

$$P \geq \overline{U} + \lambda[1 - F(L)](C - D), \tag{19}$$

and a modified limited-liability constraint

$$P \geq \lambda\{\underline{S} + \theta F(L)[L - U(Y - P) + U(Y - P - D)] + [1 - F(L)](C - D)\}. \tag{20}$$

This reduces the managed care problem to choosing P, D, and L to maximize (7) subject to the modified individual rationality constraint (19) and the modified limited-liability constraint (20).

31. Alternatively, the limited-liability constraint can be interpreted as capturing risk aversion in a crude way; that is, the physician is risk-neutral with respect to income variation above some critical level, and extremely risk-averse for income variation below this level (Sappington, 1983). What is crucial for our analysis of managed care is that the physician is less risk-averse than the patient with respect to the uncertain cost of treatment, and therefore, better able to absorb this risk. This is natural, because the physician can diversify the treatment-cost risks across a population of patients, and physicians can further pool this risk within a group practice.

The managed care problem has the same structure as the demand management problem analyzed in Sections 3 and 4 (with $\underline{S}$ and $\overline{U}$ normalized to 0), except that the modified limited-liability constraint (20) relaxes the treatment constraint (4). Therefore, since (4) was binding in the demand management problem, managed care must improve *ex ante* consumer welfare. We expect this relaxation of the treatment constraint to move the optimal treatment cutoff in the direction of ideal insurance. Therefore, if optimal demand management results in deficient treatment relative to the ideal contract (e.g., as in Corollary 4), managed care should cause an expansion of treatment. Our next result shows that ideal insurance sometimes can be achieved when supply management is possible.

Consider a relaxed managed care insurance problem in which the modified limited-liability constraint (20) is ignored. That is, consider the maximization of (7) subject only to (19). The only difference between this problem and the maximization problem for the ideal contract in Section 3 is that the premium must also compensate for the physician's opportunity cost of diagnosis. Proposition 1 applies directly, and the solution to the relaxed program specifies the ideal treatment threshold, L^*, and a premium, P^*, that solve $L^* = U'(Y - P^*)C$ and $P^* = \overline{U} + \lambda[1 - F(L^*)]C$. The ideal copayment is zero.[32]

Proposition 4: *The optimal treatment decision, premium, and copayment under managed care are the same as for the ideal contract if and only if the minimum payment for diagnosis is low enough so that the limited-liability constraint does not bind, i.e.,* $\lambda[\underline{S} + \theta F(L^*)L^*] \leq \overline{U}$.

Under what conditions is the limited-liability constraint slack and ideal insurance possible? If copayment is set at zero, then the value of B must be set at θL^* to satisfy the treatment constraint (15) for the implementation of L^*. For the physician's reservation wage constraint to hold, the value of λS must be set to $\overline{U} - \lambda F(L^*)B = \overline{U} - \lambda\theta L^* F(L^*)$. If this value of S is at least $\underline{S}$, then the limited liability is slack. When this is the case, ideal insurance can be achieved while leaving the physician willing to participate. The condition in the proposition is more easily satisfied when the interests of the patient are given little weight in the collective decision, i.e., when θ is small, indicating that the selfish physician has most of the bargaining power, or when C is sufficiently small, indicating that the treatment threshold L^* is low. The former situation suggests ironically that the consumer

32. The formal proof is straightforward and omitted.

may be better off *ex ante* by having less control over the treatment decision *ex post*.

Under the condition in Proposition 4, supply management is sufficient to eliminate the loss due to moral hazard. When the value of θ is high (relative to $\overline{U}$ and L^*), then the condition will not be satisfied and ideal insurance is infeasible. In this case, both demand and supply management are combined in an optimal insurance contract; the physician will be strictly liable for a portion of treatment costs, and the patient will have a strictly positive copayment. This is stated formally in our final result.

Proposition 5: *The physician bonus for withholding treatment and patient copayment are both strictly positive (i.e., $B > 0$ and $D > 0$) if and only if $\lambda[\underline{S} + \theta F(L^*)L^*] > \overline{U}$. In this case, $S = \underline{S}$.*

7. Concluding Remarks

Conventional wisdom presumes an overutilization of health care from the availability of insurance that insulates consumers from the full cost of treatment. We introduce a simple but realistic model of insurance and treatment in order to evaluate this presumption. We show that the conventional wisdom is not generally true. An insurance plan based optimally on demand management (via copayments) can result in deficient treatment relative to ideal insurance. In particular, treatment for a low-probability illness is deficient if illness *per se* has little effect on the consumer's marginal utility of income, and if the consumer's price elasticity of demand for treatment is sufficiently large relative to the Arrow-Borch risk-spreading distortion when these are evaluated at a copayment that brings forth the ideal provision of treatment.

We also consider managed care based on a combination of demand and supply management instruments. Our model of managed care allows an insurer to reward a physician for withholding treatment. This simple form of managed care may be sufficient to achieve ideal insurance. In this situation, the patient makes no copayment because physician incentives are sufficient for an efficient provision of treatment. Thus, managed care corrects the treatment and risk-spreading distortions that arise under pure demand management, and expands or contracts the delivery of treatment accordingly. When ideal insurance is infeasible under managed care, then patient copayments are employed in combination with physician incentives. In this case, managed care alleviates the distortions that arise under pure demand management, but does so incompletely.

For realism and simplicity, we have avoided more general mechanisms for the provision of health insurance. Given the information

structure of our model, a simple demand-managed health insurance contract (premium plus copayment) is not necessarily optimal. The revelation principle states that any feasible insurance contract is equivalent to a direct revelation mechanism specifying a transfer payment and a probability of treatment based on the patient's (or physician's) disclosure of private knowldege about the benefit of treatment. We do not think that contracts providing treatment randomly are realistic (although a cynic might argue that the purpose of utilization review is indeed to randomize the provision of treatment). Thus, we do think that the contracts we have studied are an appropriate normative benchmark given the information structure of our model. Still, a formal treatment of a more general class of insurance contracts under private information may illuminate additional aspects of how moral hazard affects health care treatment decisions. It would also be interesting to consider variable levels of treatment in this more general framework.

Appendix. Proofs of Propositions

Proof of Proposition 1. The first-order conditions for a solution P, $t(\cdot)$, and L are respectively,

$$
\begin{aligned}
&-(1-\lambda)U'(Y-P)-\lambda\int_0^L U'(Y-P-a\ell+t(\ell))f(\ell)\,d\ell\\
&\qquad-\lambda[1-F(L)]U'(Y-P)+\alpha=0,\\
&\lambda U'(Y-P-a\ell+t(\ell))f(\ell)-\lambda\alpha f(\ell)=0,\\
&\lambda[U(Y-P-aL+t(L))-bL]f(L)\\
&\qquad-\lambda f(L)U(Y-P)-\alpha\lambda f(L)[t(L)-C]=0,
\end{aligned}
$$

where $\alpha \geq 0$ is the multiplier. The characterization of the first best follows from solving the first-order conditions.

For the comparative statics of L with respect to C, we totally differentiate (3) and the expression for P^* to obtain

$$
\frac{dL}{dC}=\frac{-U'(Y-P)^2+\lambda[1-F(L)]bLU''(Y-P)}{-[aU'(Y-P)+b]U'(Y-P)-\lambda f(L)(aL-C)bLU''(Y-P)}>0,
$$

where the inequality follows from $aL-C<0$, which in turn follows from (3), and the concavity of U. □

Proof of Proposition 2. Use the constraints (4) and (5) to define P and L as functions of D. Apply the implicit-function theorem, and we obtain the derivative in (6). Now regard the objective function

$$(1-\lambda)U(Y-P)+\lambda\Bigg(\int_0^L [U(Y-P-a\ell)-b\ell]f(\ell)\,d\ell + [1-F(L)]U(Y-P-D)\Bigg)$$

as a function in D, where L and P in the expression are now functions of D. Differentiating with respect to D, applying the chain rule, and using constraint (4), we obtain the derivative of the objective function:

$$-\Bigg((1-\lambda)U'(Y-P)+\lambda\int_0^L U'(Y-P-a\ell)f(\ell)\,d\ell + \lambda[1-F(L)]U'(Y-P-D)\Bigg)\frac{dP}{dD}-\lambda[1-F(L)]U'(Y-P-D). \quad (21)$$

Equation (8) is obtained by setting (21) equal to zero.

It remains to show that $0 < D^d < C$ and $L^d > 0$. But of course, from (4), $0 < D^d < C$ implies $L^d > 0$. So we only need to show that D^d is interior. To show that $D > 0$, we first evaluate (6) at $D = 0$; we get

$$\frac{dP}{dD}=-\lambda-\lambda\frac{f(0)CU'(Y-P)}{aU'(Y-P)+b}, \quad (22)$$

which makes use of the fact that $D = 0$ implies that $L = 0$. Next, after simplification we obtain the value of (21) at $D = 0$:

$$-U'(Y-P)\left[\frac{dP}{dD}+\lambda\right],$$

which is positive by (22). So $D > 0$.

To show that $D < C$, we use the same method. At $D = C$, $P = 0$, and (6) becomes

$$\frac{dP}{dD}=-\lambda[1-F(L)].$$

The value of (21) at $D = C$ is

$$\Bigg((1-\lambda)U'(Y)+\lambda\int_0^L U'(Y-a\ell)f(\ell)\,d\ell + \lambda[1-F(L)]U'(Y-D)\Bigg)\lambda[1-F(L)]-\lambda[1-F(L)]U'(Y-D),$$

which becomes

$$-U'(Y-D)+\Bigg((1-\lambda)U'(Y)+\lambda\int_0^L U'(Y-a\ell)f(\ell)\,d\ell \\ +\lambda[1-F(L)]U'(Y-D)\Bigg)$$

after the common factor $\lambda[1-F(L)]$ has been taken out. Because $U(Y-a\ell)-b\ell > U(Y-D)$ for $\ell < L$, we have $U'(Y-a\ell) < U'(Y-D)$. So the above expression must be negative, and we conclude that $D < C$. □

Proof of Proposition 3. By the discussion preceding it, the proposition is proved if (10) is necessary and sufficient for $Z'(L^*) > 0$ as $\lambda \to 0$. Total differentiation gives

$$Z'(L) = -\lambda[1-F(L)][aU'(Y-P(L)-aL)+b] \\ +\frac{dP}{dL}\Bigg[-(1-\lambda)U'(Y-P(L)) \\ -\lambda\Bigg(\int_0^L U'(Y-P(L)-a\ell)f(\ell)\,d\ell \\ +[1-F(L)]U'(Y-P(L)-aL)\Bigg)\Bigg].$$

The derivative $P'(L)$ is

$$\frac{dP}{dL}=\frac{\lambda[1-F(L)][aU'(Y-P-aL)+b]+\lambda f(L)(C-D)U'(Y-P-D)}{\lambda[1-F(L)][U'(Y-P-D)-U'(Y-P-aL)]-U'(Y-P-D)}, \tag{23}$$

which is obtained from applying the implicit-function theorem to the constraints (4) and (5) to define P and D as functions of L. Using the above, we can show that $P \to 0$ as $\lambda \to 0$, and with (23), $Z'(L)$ has the same sign as

$$(C-D)-\frac{1-F(L)}{f(L)}\left(\frac{1}{U'(Y)}-\frac{1}{U'(Y-D)}\right)[aU'(Y-aL)+b], \tag{24}$$

where

$$bL = U(Y-aL)-U(Y-D). \tag{25}$$

Therefore, we have $Z'(L^*) > 0$ if and only if (10), and conversely. □

Proof of Proposition 5. The first-order conditions for the maximization of (7) with respect to (19) and (20), after simplification, yield

$$U'(Y-P) = (\alpha+\beta)\frac{1-\lambda+\lambda F(L)}{1-\lambda+\lambda F(L)+\lambda\beta\theta F(L)},$$

$$U'(Y-P-D) = (\alpha+\beta)\frac{1-F(L)}{1-F(L)-\beta\theta F(L)},$$

where α and β are respectively the multipliers of (19) and (20). These further simplify to

$$U'(Y-P) = (\alpha+\beta)\left(1+\frac{\lambda\beta\theta F(L)}{1-\lambda+\lambda F(L)}\right)^{-1},$$

$$U'(Y-P-D) = (\alpha+\beta)\left(1-\frac{\beta\theta F(L)}{1-F(L)}\right)^{-1}.$$

Because the first best is infeasible under the condition of the proposition, we have $\beta > 0$. Therefore, $U'(Y-P-D) > U'(Y-P)$, or $D > 0$.

Also, because the first best is infeasible, $S = 0$. Since $\overline{U} > 0$, we have $B > 0$ from (17). □

References

Arrow, K.J., 1963, "The Welfare Economics of Medical Care," *American Economic Review*, 53, 941–973.

——, 1968, "The Economics of Moral Hazard: Further Comment," *American Economic Review*, 58, 537–539.

Baumgardner, J.R., 1991, "The Interaction between Forms of Insurance and Types of Technical Change in Medical Care," *RAND Journal of Economics*, 22, 36–53.

Chernew, M.E., W.E. Encinosa, and R.A. Hirth, 2000, "Optimal Health Insurance: The Case of Observable, Severe Illness," Unpublished Manuscript.

de Meza, David, 1983, "Health Insurance and the Demand for Medical Care," *Journal of Health Economics*, 2, 47–54.

Ellis, R.P. and T.G. McGuire, 1990, "Optimal Payment Systems for Health Services," *Journal of Health Economics*, 9 (4), 375–396.

—— and ——, 1993, "Supply-Side and Demand-Side Cost Sharing in Health Care," *Journal of Economic Perspectives*, 7, 135–152.

Grossman, S.J. and O.D. Hart, 1986, "The Costs and Benefits of Ownership: A Theory of Vertical and Lateral Integration," *Journal of Political Economy*, 94, 691–719.

Ma, C.A. and T.G. McGuire, 1997, "Optimal Health Insurance and Provider Payment," *American Economic Review*, 87, 685–704.

—— and M.H. Riordan, 2000, "Health Insurance, Moral Hazard, and Managed Care," Boston University Discussion Paper.

Manning, W.G., Jr., and C.E. Phelps, 1979, "The Demand for Dental Care," *The Bell Journal of Economics*, 10(2), 503–525.

Marshall, J.M., 1976, "Moral Hazard," *American Economic Review*, 66(5), 880–890.
Newhouse, J.P., 1996, "Reimbursing Health Plans and Health Providers: Selection versus Efficiency in Production," *Journal of Economic Literature*, 34, 1236–1263.
Pauly, M., 1968, "The Economics of Moral Hazard: Comment," *American Economic Review*, 58, 531–537.
Phelps, C.E., 1992, *Health Economics*, Harper Collins: New York.
Sappington, D.E.M., 1983, "Limited Liability Contracts between Principal and Agent," *Journal of Economic Theory*, 29, 1–21.
Zeckhauser, R., 1970, "Medical Insurance: A Case Study of the Tradeoff between Risk Spreading and Appropriate Incentives," *Journal of Economic Theory*, 2, 10–26.