

# The Hidden Evidence of Defensive Medicine

Daniel Montanera\*

Ph.D. Candidate, Department of Economics  
University of Western Ontario, London, ON, dmontan2@uwo.ca

Despite physician surveys indicating the widespread practice of defensive medicine, empirical investigations into the subject have produced inconsistent findings. This lack of clear evidence has confounded efforts to formulate, implement, and evaluate various tort reforms designed to lower costs and improve access to medical care. This paper presents an innovative model of the interaction between patients, physicians, and health insurers that provides a unified framework within which the existing empirical findings can be understood. The model predicts that the practice of defensive medicine would manifest itself as non-monotonic relationships between malpractice liability costs and both the cost and quality of health care. Both relationships are positive at low levels of malpractice pressure, and become negative at higher levels. These findings are consistent with tendencies in the empirical literature for broad data sets to produce weak or nonexistent relationships, and highly specific data sets to produce clear relationships that conflict with one another. In addition to an explanation for conflicts in the empirical literature, the model provides testable predictions allowing policymakers to better anticipate the effects of tort reform.

*Key words:* health care expenditure, medical malpractice, defensive medicine

*History:* Job market paper, October 2011.

---

## 1. Introduction

An important question among researchers and policymakers is whether or not the fear of medical malpractice liability induces physicians to change the way they practice medicine. This behaviour is commonly known as defensive medicine, and surveys of physicians on the subject regularly report its widespread practice (Paik et al. 2011). Positive defensive medicine is the over-utilization of medical services in an effort to forestall claims of negligence, while negative defensive medicine is the avoidance of risky patients or procedures believed likely to result in malpractice claims. Such behavior illustrates the adverse relationships that could exist between the liability costs born by physicians and both the cost and quality of the health care enjoyed by consumers. Empirical investigations designed to uncover and measure these relationships have produced inconsistent findings, thus clear evidence of whether or not defensive medicine is practiced on a significant

\*The author wishes to acknowledge the patience and invaluable insight of his thesis advisor, Dr. Al Slivinski, as well as the suggestions from committee members, Dr. James B. Davies and Dr. Greg Zaric. Thoughtful and constructive comments were provided by the attendees of both the 2nd Annual Midwest Health Economics Conference as well as the 10th Annual Meeting of the Canadian Health Economics Study Group. Finally, the author directs much appreciation toward classmate David Fieldhouse for his assistance with numerical estimation methods.

scale remains elusive. In several countries, the United States in particular, this lack of empirical verification has confounded efforts to formulate, implement, and evaluate tort reforms intended to make high-quality medical care more affordable to consumers.

This paper presents an innovative model of the interaction between patients, physicians, and health insurers that provides a unified framework within which the existing empirical findings can be understood. Ultimately, the model is a decision problem on the part of an insurer involving whether and how to structure their contracts with physicians as malpractice pressure increases, with the goal of surviving in a competitive market for health insurance. The model predicts that rising malpractice pressure induces insurers to provide higher quality health insurance at higher prices to consumers as long as the level of malpractice pressure is relatively low. As pressure reaches higher levels, however, insurers reverse this behaviour and provide cheaper and lower quality insurance policies if malpractice pressure continues to rise. This creates a non-monotonicity in the relationships between malpractice pressure and both health care quality and expenditure, which can leave evidence of defensive medicine hidden from empirical methods designed to investigate for monotonic relationships.

These predictions are broadly consistent with the literature, and offer an explanation for conflicting results in past empirical studies. Data focusing on specialties facing relatively low malpractice pressure tend to report positive relationships while those with high malpractice pressure reveal negative relationships. Studies including more than one region or specialty are likely to mix the two cases, causing the opposite relationships to cancel each other and produce weak or inconclusive relationships over the entire sample. In addition to this explanation for conflicts in the empirical literature, the model predicts that accounting for measures of access to care when estimating the effects of changes in the malpractice environment would allow policymakers to better anticipate the effects of tort reform.

## 2. Empirical Puzzle

Despite clear intuition behind the idea of defensive medicine, a number of studies have discovered weak or nonexistent relationships between medical malpractice liability costs and both health care quality and expenditure. Baicker & Chandra (2005) found little evidence of defensive medicine based on 9-year long differences in malpractice liability and utilization rates of certain procedures at the state level. In a more recent analysis (Baicker et al 2007), this time using 8-year long differences at the state level, the authors found that a 10% increase in either average malpractice premiums or average malpractice payments per physician yield a 0.6% to 1% increase in Medicare spending per beneficiary, and a 1% to 1.3% increase in Medicare spending per beneficiary on physicians' services. Hellinger & Encinosa (2006) utilize a popular measure of malpractice pressure: the existence of certain legislative reforms designed to lower malpractice payments. While it is a coarser measure than average malpractice payments or premiums, it is arguably more policy-relevant given tools available to legislators. The authors found that the states that introduced legislation capping awards for damages due to malpractice experienced between 3.25% and 3.4% lower per-capita health care expenditure than states without caps. These results are typical of those studies using state-level data. The coefficients reported in the four studies suggest limited scope for effective reductions in health care expenditure through tort reform on a national scale.

It is notable that similar methods produced an insignificant relationship between malpractice pressure and health care spending in Baicker & Chandra (2005), but a statistically significant positive relationship in Baicker et al (2007). The 2007 study focussed on a more restricted population of patients; those on Medicare. This is part of a general trend in the literature, and the first piece of the empirical puzzle examined here: studies that narrow their focus to specific geographic regions, medical specialties, or demographics tend to uncover more statistically and economically significant relationships than broader studies. The most oft-cited of the narrower variety of study is by Kessler & McClellan (1996). It found that certain tort reforms were able to lower hospital expenditures on elderly Medicare patients suffering heart attack and heart disease, without significantly affecting the quality of care. The authors interpret these findings as evidence that defensive medicine is

practiced on a significant scale. The authors later published a study on a similar population breaking down several key mechanisms through which liability costs might affect the cost and quality of care (Kessler & McClellan 2002). They found that changes to some components of liability costs had a greater effect on health care expenditures than others. For example, reductions in the time and trouble to physicians of drawn-out court defenses had a greater impact than the frequency of claims and the severity of awards. This may indicate that uninsurable components of liability costs produce more defensive medicine than insurable ones.

While Kessler & McClellan's findings may suggest how malpractice pressure affects the cost and quality of health care among specialties treating elderly patients with heart problems, the other piece of the empirical puzzle is that specific studies utilizing data from narrowly defined populations tend to conflict with one another. In a study examining county-level data mostly from California and New York, Lakdawalla & Seabury (2009) found that even though lower malpractice pressure can result in sizable reductions in health care expenditure, it also results in lower quality health care. Other authors have tried to duplicate Kessler & McClellan's results while either broadening the patient population or focussing on a different population altogether. The most thorough examination is Sloan & Shadle (2009) where, using a different timeframe and a more general definition of expenditure, the authors found that tort reforms generally had no significant effect on expenditure or quality of care enjoyed by Medicare patients suffering from heart attack, breast cancer, diabetes, or stroke; either jointly or separately. This has led policymakers to question whether Kessler & McClellan's findings are generalizable to other populations (Congressional Budget Office 2004).

Studies examining obstetrics have found the effects of changes in malpractice pressure on the cost and quality of care to be the opposite of those from the previously mentioned studies. In Currie & MacLeod (2008), caps on noneconomic damages increased the incidence of delivery by C-section, considered to be a more expensive alternative to natural childbirth in non-emergency cases. These findings imply that, by increasing the incidence of the more expensive alternative, this particular tort reform increases expenditure in the obstetrics specialty, the opposite of the result found in Kessler & McClellan (1996). Another widely cited paper, Dubay et al (2001), found

that higher malpractice premiums were associated with a greater incidence of late prenatal care, particularly among poorer socioeconomic groups. Where the timing of care is a vital component of health care quality, these findings indicate tort reforms would deliver improvements in quality, which goes against the findings from both Kessler & McClellan (1996) and Lakdawalla & Seabury (2009). Given the findings from studies using data from obstetrics, tort reforms could produce the opposite of their expected effect in certain medical specialties.

### 3. Theoretical Literature

Overall, the evidence on the relationships between malpractice pressure and both the cost and quality of consumer health care is unclear. This suggests that there may be room for economic theory to make a contribution. There is an established theoretical literature investigating medical malpractice penalties and their effects on physician behaviour, including Leger (2000), Zeiler (2004), Arlen & MacLeod (2005), Olbrich (2008), and Kpelitse (2010). These studies tend to utilize variations of principal-agent problems with the goal of determining an “optimal” level of malpractice pressure. While certainly addressing an important issue with broad policy implications, these studies nonetheless investigate different questions than are posed in the empirical literature, and are thus of limited use in explaining the mixed empirical results.

An important distinction between this study and much of the existing theoretical literature lies in the role of incentive compatibility in contracts between insurers and physicians. In studies utilizing the principle-agent framework, physicians tend to have information over the health status of a given patient that is unavailable to insurers, thus requiring incentive compatibility constraints on any contract guaranteeing the appropriate treatment for a patient of a particular health status. The model presented here contains no information asymmetries, as it focuses on interactions that occur before physicians gain private information over their patients’ health statuses. Two such interactions involve the contracts formed both between insurers and consumers regarding the terms of health insurance, and between insurers and physicians determining methods and levels of payment,

as well as the size and scope of health care networks. The contracts in this model require incentive compatibility for a different reason, namely the degree of independence with which physicians manage their individual practices. Besides the inherent difficulties in micro-managing individual physician behaviour, insurers maintain an “arm’s length” approach in their health networks due to a tendency to hire physicians as independent contractors rather than employees (Arlen & MacLeod 2005). While this limits insurers’ exposure to malpractice liability, it also makes physicians more responsible for making decisions. For this reason, the model presented here allows insurers to influence physician behaviour only through incentives rather than commands.

Further, in most other models, a single insurer (principal) typically induces a single physician (agent) to provide a certain quality of treatment to a single patient. This framework is useful for examining health care quality on the intensive margin, but implicitly assumes the extensive margin - the number of patients treated by each physician - to be determined exogenously. This is unlikely to be problematic where all malpractice liability costs are insurable, since holding malpractice insurance would bring the marginal liability cost of treating a patient to zero. It could be a strong assumption, however, if some liability costs are uninsurable, and thus increasing in the number of patients treated. Examples of uninsurable liability costs faced by physicians include reputational damage, the prospect of suspension, the time and trouble of court proceedings, and the chance that the physician would lose coverage due to a malpractice insurer bankruptcy or exit from the market. If there are significant uninsurable liability costs, physicians could choose to limit exposure by reducing patient rolls or the size of their practice. Both Kessler & McClellan (2002) and Currie & MacLeod (2008) elude to differences between insurable and uninsurable liability cost in their impact on physician behaviour, and so the model presented here endogenizes the extensive margin of health care quality in addition to the intensive margin.

#### **4. The Model**

There is a continuum of identical consumers of measure 1, each endowed with income  $m$ . Consumers have the utility function  $U(H, y)$ , where  $H$  is health status and  $y$  is consumption, with positive

first derivatives, negative second derivatives, and a positive cross-partial derivative. Each consumer expects to become ill with probability  $q$  and remain healthy with probability  $1 - q$ . A healthy consumer enjoys a health status of  $H_1$  while an ill consumer enjoys  $H_2$ , where  $H_1$  is greater than  $H_2$ . Consumers are willing to purchase health insurance as long as the expected utility ( $EU$ ) of owning insurance is at least as large as the expected utility of going uninsured. The parameters  $m$ ,  $q$ ,  $H_1$ ,  $H_2$ , and the function  $U(\cdot)$  are also common knowledge. Since consumers are identical, a strict preference toward health insurance for a single consumer results in the entire measure of consumers purchasing health insurance, with exactly  $q$  insured consumers expected to become ill and seek treatment.

If any consumer  $i$  with insurance falls ill, he is able to obtain treatment ( $t_i$ ) from a physician and will recover with probability  $1 - \rho(t_i)$ , but suffer an adverse outcome with probability  $\rho(t_i)$ , a function that is common knowledge. Ill consumers without insurance have zero probability of recovery. The function  $\rho(t_i)$  is positive, decreasing, strictly convex, continuous, and differentiable with at least the first, second, and third derivatives being finite for any  $t_i$ . Assume that  $\rho(t_i)$  is equal to 1 when  $t_i = 0$  and approaches zero as  $t_i$  approaches infinity. Note that these assumptions imply that the derivatives of  $\rho(t_i)$  alternate in sign (the first derivative being negative) and all approach zero as  $t$  increases. Since  $\rho(t_i)$  is always decreasing, the model does not allow treatment to become harmful to the patient at any level. For this reason, large increases in treatment are better considered as gold-plating; costly inputs that deliver marginal improvements in care, rather than the potentially harmful continued administration of any one particular medical treatment.

There is a continuum of identical physicians of measure  $D$ , which is common knowledge. There are two margins of physician behaviour of interest in this paper: the number of patients that each physician chooses to treat, as well as the amount of treatment the physician devotes to each patient. Due to the significant cost of entry into the medical profession, the measure of physicians is determined exogenously rather than by a clearing condition in the physician labour market. On the other hand, a fixed  $D$  prohibits this model from addressing questions of physician mobility or exit in response to changing malpractice pressure. This is another interesting margin of behaviour

worth further study (See Kessler, Sage, & Becker 2005; Mello et al. 2005; and Klick & Stratman 2007).

Physicians are risk-neutral income maximizers. Each one receives a revenue  $w$  from an insurer for each policyholder treated, as well as a stock of resources ( $s$ ) for use in the treatment of policyholders. Rather than any one particular explicit method of payment (prospective payment, fee-for service, etc.), the pair  $\{w, s\}$  should be considered as an implicit contract between insurer and physician. Any explicit contract should leave a physician with a sense of the revenue they can expect to realize from taking on a patient, as well as the resources at their disposal for their patients' treatment. This allows the model to abstract away from the finer details of any one explicit contract and make the incentive compatibility constraints in the insurer's problem more tractable. Individual physicians are small relative to a health insurer, and therefore cannot behave strategically, taking both  $w$  and  $s$  as given.

Each patient  $i$  that a physician treats brings an expected liability cost of  $g(t_i, P)$ . Patient  $i$ 's expected liability cost is a function of the amount of resources he receives from the physician ( $t_i$ ), and a parameter serving as a measure of malpractice pressure ( $P$ ), which is common knowledge. Assume that  $g(t_i, P)$  has the form:

$$g(t_i, P) = \rho(t_i) \cdot P$$

This form implies that the physician expects the liability cost of treating a patient to be proportional to the risk of that patient suffering an adverse health outcome, and that this proportion is fixed at the level of malpractice pressure prevalent in the physician's jurisdiction. This form doesn't allow for a negligence rule of liability, since such a rule would require liability costs to equal zero once a physician reached some legal standard of care. There is evidence, however, that types one and two errors on the part of courts in malpractice cases make physicians capable of reducing

exposure to malpractice claims, but not able to eliminate it completely (Danzon 1985), which is consistent with the form assumed here. Essentially, if large amounts of resources are devoted to the treatment of a patient, then the patient is less likely to have grounds to claim negligence, and therefore are not likely to bring liability costs upon their physician. Each physician chooses the size of his patient roll ( $n$ ) and the amount of treatment to provide to each patient ( $t_i$ ) as a share of available resources ( $s$ ). Due to the properties of  $\rho(t_i)$ , expected liability costs are at their lowest for any number of patients when all patients receive an equal amount of treatment. This means that  $t_i = \frac{s}{n}$  for all  $i$  in equilibrium. Given values of  $w$  and  $s$ , each physician solves the problem:

$$\max_{n \geq 0} \left\{ wn - n \cdot g\left(\frac{s}{n}, P\right) \right\}$$

Let

$$\begin{aligned} n^* &= n^*(w, s, P) = \arg \max_{n \geq 0} \left\{ wn - n \cdot g\left(\frac{s}{n}, P\right) \right\} \\ \tilde{n} &= \tilde{n}(w, s, P) = \min \left\{ n^*, \frac{q}{D} \right\} \end{aligned}$$

$n^*(w, s, P)$  is the number of patients the physician would like to treat given any contract  $\{w, s\}$  and level of malpractice pressure  $P$ . Since the expected number of ill consumers is  $q$ , the maximum number of patients any of the  $D$  identically-behaving physicians could treat is  $\frac{q}{D}$ . Therefore, the number of patients the physician would *actually* treat given any contract  $\{w, s\}$  and level of malpractice pressure  $P$  is  $\tilde{n}(w, s, P)$ . The physician's total liability costs are a convex function in  $n$ , so  $n^*$  and  $\tilde{n}$  are unique.

The final decision maker is a managed care organization (MCO). Consistent with the margins of physician behaviour of interest here, the assumption of a single firm avoids issues of physician

mobility between competing MCOs. This firm offers consumers a health insurance policy delivering a probability of recovery equal to  $Q$  (hereafter referred to as “quality”) at a policy price of  $\tau$ . The successful treatment of an ill consumer is the result of two events. First, the consumer must obtain a place on a physician’s patient roll. If ill consumers are assigned randomly to patient rolls, the probability of this happening is equal to the total number of places available ( $Dn$ ) divided by the total number of ill consumers seeking placement ( $q$ ). This probability serves as the level of “access” in the health care system. Second, conditional on success with the first event, the patient receives treatment and recovers with probability  $1 - \rho(t)$ . Quality,  $Q(n, t) = \frac{Dn}{q}(1 - \rho(t))$ , is thus the product of these two probabilities.

A perfectly competitive market for health insurance determines the objective of the MCO, as well as its constraints. The MCO must offer an insurance policy that maximizes expected consumer utility subject to a zero profit constraint. Any other bundle, whether one that brought non-zero profits to the MCO or one delivering a lower  $EU$  to consumers without violating the zero profit constraint, would induce firm entry or exit and thus cannot hold in equilibrium. In order to be credible, the quality of the insurance policy must be incentive compatible with physicians’ behaviour under the MCO’s contract  $\{w, s\}$ , meaning  $n = \tilde{n}$  in the insurer’s problem. The firm acquires resources at a marginal cost of  $c$ , which is assumed to be constant. The MCO’s problem is thus:

$$\max_{w, s, \tau} \left\{ EU(\tilde{Q}, \tau) = \text{prob}(H_1) \cdot U(H_1, m - \tau) + \text{prob}(H_2) \cdot U(H_2, m - \tau) \right\}$$

$$\text{subject to } \tau = Dw\tilde{n} + Dcs$$

$$\text{where } \text{prob}(H_1) = (1 - q) + q\tilde{Q}$$

$$\text{prob}(H_2) = q - q\tilde{Q}$$

$$\tilde{Q} = Q\left(\tilde{n}, \frac{s}{\tilde{n}}\right) = \frac{D\tilde{n}}{q} \left[ 1 - \rho\left(\frac{s}{\tilde{n}}\right) \right]$$

Notice that, since the insurer's choice set contains  $\{w, s, \tau\} = \{0, 0, 0\}$ , a policy  $\{\tilde{Q}, \tau\} = \{0, 0\}$  can always be delivered that is equivalent to no insurance. Thus, the solution to the MCO's problem will always be such that consumers weakly prefer purchasing health insurance over going uninsured.

This environment can be analyzed in the form of a decision problem from the perspective of the MCO. The firm maximizes expected consumer utility by offering consumers a health insurance policy at a certain price and of a certain quality. In order to obtain such quality, the MCO must set a contract to procure physicians' services, while also accounting for how the structure of that contract affects physicians' treatment decisions. Given any set of parameters  $\{q, m, D, c, H_1, H_2, P\}$ , the solution to this problem is defined as a contract between the MCO and consumers  $\{\tau^*\}$ , a contract between the MCO and physicians  $\{w^*, s^*\}$ , and a choice of patient roll size given any contract with the MCO  $\{n^*(w, s, P)\}$  such that:

$$\begin{aligned} 1. \quad & n^*(w, s, P) = \arg \max_{n \geq 0} \left\{ wn - n \cdot g\left(\frac{s}{n}, P\right) \right\} \\ 2. \quad & \{w^*, s^*, \tau^*\} \in \arg \max_{w, s, \tau} \left\{ EU(\tilde{Q}, \tau) \mid \tau = Dw\tilde{n} + Dcs \right\} \end{aligned}$$

where  $\tilde{n}^* = \min \left\{ n^*(w^*, s^*, P), \frac{q}{D} \right\}$  is the number of patients each physician treats in equilibrium. Also, let  $t^* = \frac{s^*}{\tilde{n}^*}$  and  $\tilde{Q}^* = \frac{D\tilde{n}^*}{q} \left[ 1 - \rho\left(\frac{s^*}{\tilde{n}^*}\right) \right]$  be equilibrium values of treatment and system quality respectively.

## 5. Equilibrium: Physician

As this paper is focused on the changes in equilibrium outcomes induced by changes in malpractice liability costs, the effects of changes in the parameters  $q, m, D, c, H_1,$  and  $H_2$  are not investigated here, and are largely suppressed in the notation.

Solving backward, the first problem to be considered is the physician's. The solution to the physician's problem is the patient roll that satisfies:

$$n^*(w, s, P) \text{ such that } \frac{w}{P} = \rho\left(\frac{s}{n}\right) - \left(\frac{s}{n}\right) \cdot \rho'\left(\frac{s}{n}\right)$$

where  $\rho'(\cdot)$  is the first derivative of  $\rho(\cdot)$ . Due to the properties of  $\rho(t)$  there will be a unique, non-negative, and finite  $n^*$  as long as  $0 \leq w < P$ . Also, since  $\frac{s}{n} = t$  and  $n$  only enters the right side through  $t$ , the physician's problem also characterizes a unique and non-negative level of treatment  $t(w, P) = \frac{s}{n^*}$  that is independent of  $s$  and determined entirely by  $w$  and  $P$ . The characterization of  $n^*$  yields the physician's responses to changes in the contract with the MCO and the level of malpractice pressure:

$$\begin{aligned} \frac{\partial n^*}{\partial P} &= -\left(\frac{w}{P}\right) \left(\frac{n^3}{s^2}\right) \left[ P \cdot \left(\frac{\partial^2 \rho}{\partial t^2}\right) \right]^{-1} &< 0, \\ \frac{\partial t}{\partial P} &= \left(\frac{w}{P}\right) \left(\frac{n}{s}\right) \left[ P \cdot \left(\frac{\partial^2 \rho}{\partial t^2}\right) \right]^{-1} &> 0, \\ \frac{\partial n^*}{\partial w} &= \left(\frac{n^3}{s^2}\right) \left[ P \cdot \left(\frac{\partial^2 \rho}{\partial t^2}\right) \right]^{-1} &> 0, \\ \frac{\partial t}{\partial w} &= -\left(\frac{n}{s}\right) \left[ P \cdot \left(\frac{\partial^2 \rho}{\partial t^2}\right) \right]^{-1} &< 0, \\ \frac{\partial n^*}{\partial s} &= \left(\frac{n}{s}\right) &> 0 \end{aligned}$$

These together illustrate two points about physician behaviour under this model. First, physicians practice both positive and negative defensive medicine in the face of rising malpractice pressure. Given any contract with the MCO, rising pressure makes the marginal patient too risky to treat, causing the physician to remove some patients from his patient roll (negative) and also to increase the amount of treatment provided to each of the remaining patients (positive). Second, physicians respond to financial incentives in their contract with the MCO. A greater stock of resources lowers the risk of treating the marginal patient, and induces the physician to increase

his patient roll size. Increases in compensation-per-patient increases the revenue from treating the marginal patient, leading the physician to take on more patients and accept greater expected liability costs. This behaviour is accounted for when the MCO chooses its contracts.

## 6. Equilibrium: MCO

From the physician's problem, for a given value of  $P$ , the revenue per patient necessary for the MCO to induce a physician to choose to treat a given roll size ( $\bar{n}$ ), as a function of resources, is given by  $\omega(s; \bar{n}, P)$ . Characteristics of this function are retrieved from the physician's problem:

$$\begin{aligned}\omega(s; \bar{n}, P) &= P \left[ \rho(t) - t \left( \frac{\partial \rho}{\partial t} \right) \right] > 0 \\ \frac{\partial \omega}{\partial s} &= -Pt \left( \frac{\partial^2 \rho}{\partial t^2} \right) \left( \frac{1}{\bar{n}} \right) < 0 \\ \frac{\partial^2 \omega}{\partial s^2} &= -P \left[ \frac{\partial^2 \rho}{\partial t^2} + t \left( \frac{\partial^3 \rho}{\partial t^3} \right) \right] \left( \frac{1}{\bar{n}} \right)^2\end{aligned}$$

Where  $t = \frac{s}{\bar{n}}$ . The function  $\omega(\cdot)$  is always decreasing in  $s$  and must have at least one inflection point. This is because  $\left( \frac{\partial^2 \rho}{\partial t^2} \right)_{s=0} > 0$  and  $t \left( \frac{\partial^3 \rho}{\partial t^3} \right)_{s=0} = 0$ , causing  $\omega(\cdot)$  to be concave at  $s = 0$ . It must eventually become and remain convex, however, since  $\omega(\cdot)$  is always decreasing in  $s$  and must remain positive.

The existence of an inflection point is counter-intuitive, as the convexity in  $\rho(t)$  implies that there are diminishing returns to treatment. That being the case, one would expect the impact of a marginal increase in the stock of resources available to physicians on  $\omega(\cdot)$  to decrease monotonically as more resources are provided. The reason for the existence of the inflection point is that  $\omega(\cdot)$  approaches  $P$  as  $s$  approaches zero. From the solution to the physician's problem, if  $w = P$ , the physician would want to see an infinite number of patients since the large payment would fully compensate for the liability cost of taking on a new patient, even as the probability of an adverse outcome approaches 1. In such a situation, small changes in  $w$  produce large changes in

$n$ , and so maintaining  $\bar{n}$  requires only small changes in  $\omega(\cdot)$  at levels of  $s$  close to 0. An inflection point in  $\omega(\cdot)$  should therefore be expected at low levels of  $s$ , and thus  $t$ , when  $\omega(\cdot)$  is nearly as large as  $P$ . Without reason to expect more than one, Assumption 1 is made to rule out cases of multiple unconnected inflection points:

**Assumption 1 (A1):** The function  $\rho(t)$  is such that  $\exists! \underline{t} > 0$  where:

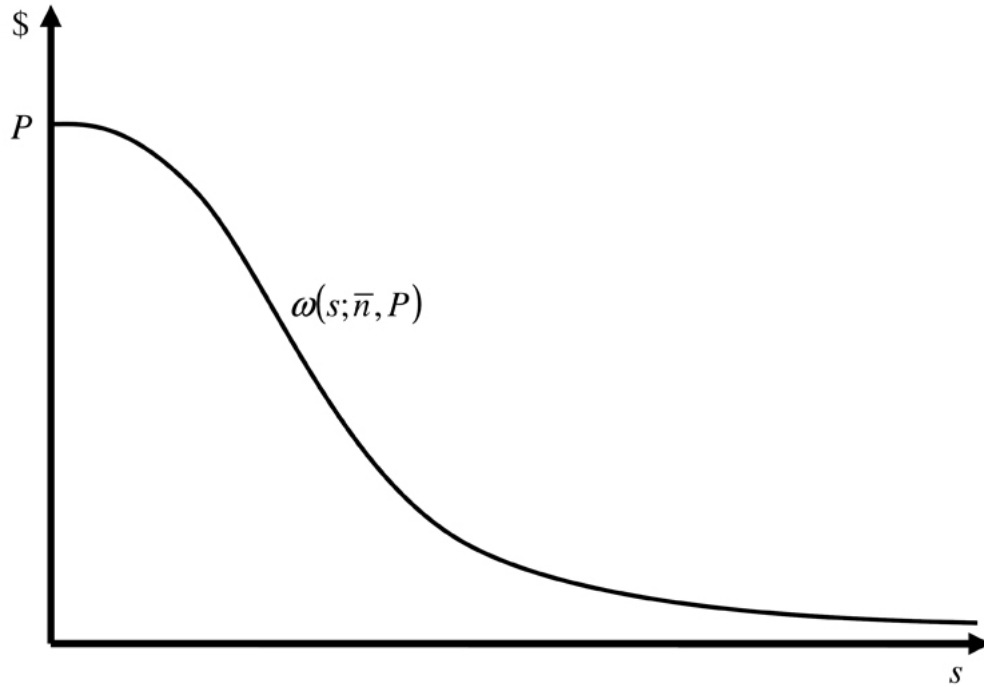
$$\frac{\partial^2 \rho}{\partial t^2} + t \left( \frac{\partial^3 \rho}{\partial t^3} \right) \begin{cases} \geq 0 \quad \forall t \in [0, \underline{t}]; \\ < 0 \quad \textit{otherwise}. \end{cases}$$

Let  $\theta(s; \bar{n}, c, P) = D\omega(s; \bar{n}, P)\bar{n} + Dcs$ . The function  $\theta(\cdot)$  represents the minimum policy price that the insurer can charge in order to induce a representative physician to treat  $\bar{n}$  patients, given that the physician is provided with  $s$  resources. As resources are removed, the total cost of inducing each of physicians to keep treating  $\bar{n}$  patients approaches  $D\bar{n}P$ . This cost would be high in environments where malpractice pressure is high. In such instances, it is entirely likely that the policy price necessary to get  $\bar{n}$  patients treated would be lower when physicians are provided with a positive amount of resources than when they are provided with no resources at all. For this to be case, treatment must be effective at lowering the probability of an adverse outcome, and the marginal cost of resources ( $c$ ) must be low relative to the prevailing level of malpractice pressure. It is to restrict the analysis to these cases that Assumption 2 is made:

**Assumption 2 (A2):** The function  $\rho(t)$  and parameters  $\{c, P\}$  are such that:

$$\exists \bar{t} > 0 \textit{ where } c\bar{t} = P[1 - \rho(\bar{t}) + \bar{t}\rho'(\bar{t})]$$

The condition in A2 is necessary and sufficient for  $\theta(s; \bar{n}, c, P) = \theta(0; \bar{n}, c, P)$  at  $\bar{t}$ . Notice that the right side is actually equal to  $\omega(0; \bar{n}, P) - \omega(\bar{s}; \bar{n}, P)$  where  $\frac{\bar{s}}{\bar{n}} = \bar{t}$ . Intuitively, for some level of



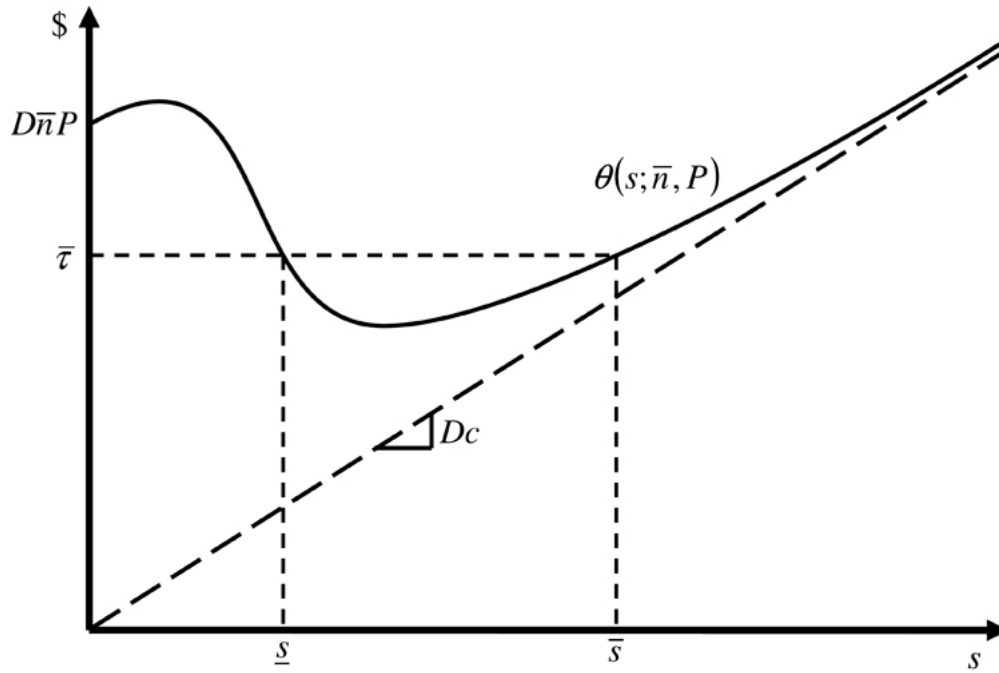
**Figure 1** The revenue-per-patient necessary to induce a physician to treat  $\bar{n}$  patients, as a function of resources provided. By Assumption 1, it consists of a weakly concave segment followed by a strictly convex one.

treatment  $\bar{t}$ , the cost of providing a patient with that treatment ( $c\bar{t}$ ) is balanced by the reduction in the payment to a physician ( $\omega(0; \bar{n}, P) - \omega(\bar{s}; \bar{n}, P)$ ) that is required to get the physician to see that patient when the physician is provided with  $\bar{s} = \bar{n}\bar{t}$  in resources instead of zero. Essentially, it means that the least expensive way to get doctors to treat  $\bar{n}$  patients is not to provide them with zero resources for use in treatment.

**Lemma 1:** If A1 and A2 hold, then in any equilibrium,  $t^*$  must be such that:

$$\frac{\partial^2 \rho}{\partial t^2} + t \left( \frac{\partial^3 \rho}{\partial t^3} \right) < 0$$

**Proof:** Where A1 holds, the function  $\omega(\cdot)$  consists of a weakly concave segment followed by a strictly convex segment, as shown in Figure 1. The function  $\theta(\cdot)$  can be derived by scaling  $\omega(\cdot)$  by  $D\bar{n}$  and shifting it up by a vertical distance of  $Dcs$ , as shown in Figure 2. The condition determining the concavity of  $\theta(\cdot)$  is identical to that of  $\omega(\cdot)$ , and so  $\theta(\cdot)$  also consists of a weakly



**Figure 2** The amount of spending per patient necessary to induce physicians to treat  $\bar{n}$  patients, as a function of resources provided. By Assumption 2, the global minimum cannot occur at  $s=0$ .

concave segment followed by a strictly convex segment. Where A2 holds,  $\theta(\cdot)$  has an interior global minimum. Also note that this global minimum must be in the segment where  $\omega(\cdot)$  and  $\theta(\cdot)$  are convex. Any amount of resources in the domain of the concave segment ( $\underline{s}$ ) is dominated by another amount of resources ( $\bar{s}$ ), just as  $\theta(\underline{s}) = \theta(\bar{s}) = \bar{\tau}$  in Figure 2. This is because the two points induce the same patient roll size and entail the same policy price, but the greater amount of resources results in more treatment and higher quality at  $\bar{s}$ . Thus any equilibrium must be in the convex segment of  $\theta(\cdot)$ , where  $\frac{\partial^2 \rho}{\partial t^2} + t \left( \frac{\partial^3 \rho}{\partial t^3} \right) < 0$ .

Solving the MCO's problem requires working with the function  $\tilde{n}(w, s, P)$ . However, this function is not well-behaved, so the insurer's problem is modified using the function  $n^*(w, s, P)$  instead. This change requires the subsequent step of verifying whether or not the solution to the MCO's modified problem is feasible (ie. satisfies  $n^*(w, s, P) = \tilde{n}$ ) and if not, the optimal way for the MCO to restructure its contracts. As shown in the proof of Proposition 1, it turns out to be optimal that the restructured contract be such that  $n^* = \tilde{n} = \frac{q}{D}$ , and the solution to the insurer's problem can be determined in this case as well. It is convenient define equilibria in terms of whether or not a restructuring is necessary, so let:

$$\left\{ \hat{w}(\tau, P), \hat{s}(\tau, P) \right\} = \arg \max_{w, s} \left\{ Q\left(\tilde{n}, \frac{s}{\tilde{n}}\right) \mid \tau \geq Dw\tilde{n} + Dcs \right\}$$

$$\bar{\tau}(P) = \min \left\{ \tau \mid \tilde{n}(\hat{w}, \hat{s}, P) = \frac{q}{D} \right\}$$

The insurance policy price  $\bar{\tau}(P)$  represents the lowest price which, if allocated optimally between physician payments and resources, would be sufficient to provide access for the expected number of ill policyholders, given the level of malpractice pressure. Define a “full-access equilibrium” as any equilibrium  $\{\tau^*, w^*, s^*, n^*(w, s, P)\}$  such that  $\tilde{n}^* = \frac{q}{D}$  and  $\tau^* > \bar{\tau}(P)$ . Also define a “limited-access equilibrium” as any equilibrium such that  $\tau^* \leq \bar{\tau}(P)$ . The first order conditions in the MCO’s modified problem are:

$$q \cdot \Delta U \cdot \frac{\partial Q^*}{\partial w} = WMU_y \cdot \frac{\partial \tau}{\partial w} \quad (1)$$

$$q \cdot \Delta U \cdot \frac{\partial Q^*}{\partial s} = WMU_y \cdot \frac{\partial \tau}{\partial s} \quad (2)$$

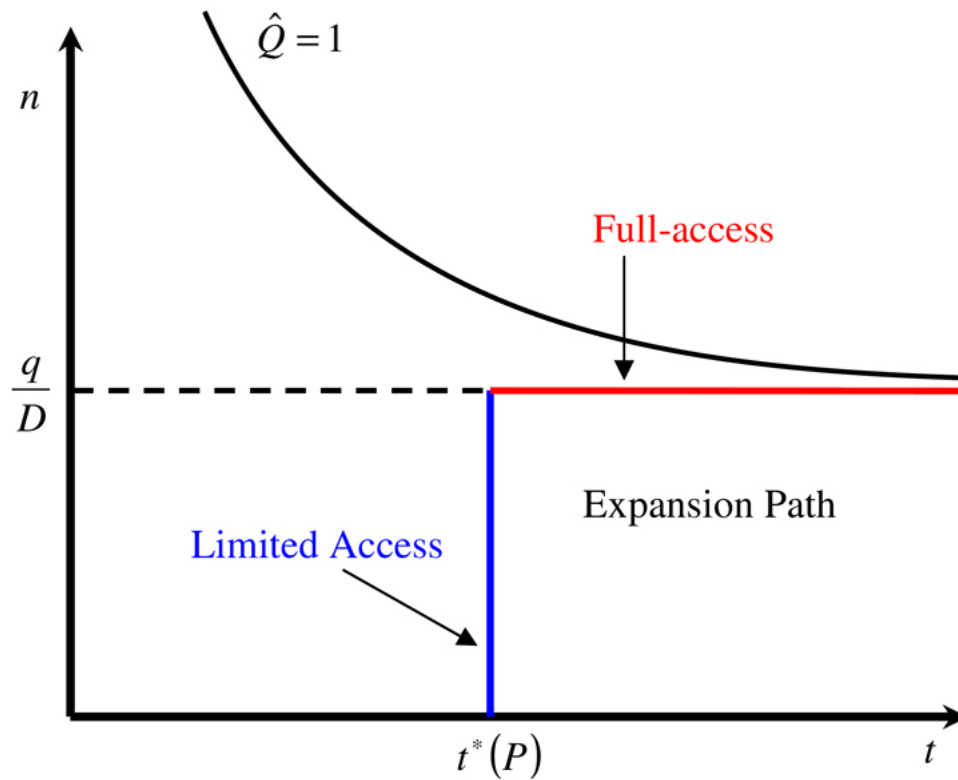
$$\tau = Dw n^* + Dcs$$

where,

$$\Delta U = U(H_1, m - \tau) - U(H_2, m - \tau) > 0, \quad \text{and}$$

$$WMU_y = (1 - q) \cdot \frac{\partial U(H_1, y)}{\partial y} + q \cdot \frac{\partial U(H_2, y)}{\partial y} + qQ^* \cdot \frac{\partial \Delta U}{\partial y} > 0$$

The term  $WMU_y$  stands for “weighted marginal utility of consumption”, and represents the marginal value of a unit of consumption that a policyholder faces ex ante. The  $\Delta U$  term represents the difference in ex post utility between those consumers who did not become ill or recovered from illness, and those who did not receive treatment or suffered an adverse outcome from treatment.



**Figure 3** Expansion path showing the optimal allocation of insurer revenues in procuring health care inputs  $t$  and  $n$  as health insurance policies become more expensive.

**Proposition 1:** If A1 and A2 hold, then there exists a unique solution to the MCO's problem that is either a full-access or a limited-access equilibrium.

**Proof:** In appendix.

Part of the insurer's problem is to decide, given some level of malpractice pressure, how to divide up its revenues from selling insurance policies ( $\tau$ ) in order to procure inputs ( $t$  and  $n$ ) to health care quality ( $Q$ ). This problem is shown in Figure 3. The proof to Proposition 1 shows that there is a unique level of treatment  $t^*(P)$  that solves the insurer's modified problem, and that this level of treatment is independent of the amount of funds that the insurer has available for procurement. This means that, if the insurer were to charge higher policy prices, the optimal way to use the higher revenues is to hold the level of treatment constant at  $t^*$  and use the additional funds to induce physicians to treat more patients. Therefore, as  $\tau$  increases, the expansion path is initially vertical in  $\{t, n\}$  space.

This is true until  $\tau = \bar{\tau}(P)$ , which is the lowest amount of spending where  $n^* = \frac{q}{D}$ . The expansion path cannot continue along the vertical path from the modified problem, because the values of  $n$  in these allocations would be greater than  $\frac{q}{D}$ , and thus infeasible. In such a case, the insurer would need to choose the best alternative contract that brings about a feasible  $n$ . Since there is a unique solution to the insurer's first-order conditions,  $n^* > \frac{q}{D}$  implies that the choice  $n = \frac{q}{D}$  dominates all  $n < \frac{q}{D}$ , and so the alternative contract must be structured such that  $n = \frac{q}{D}$ . Thus, if  $\tau$  increases beyond  $\bar{\tau}(P)$ , the additional funds are used to increase  $t$  instead of  $n$ . This gives rise to the horizontal section of the expansion path in Figure 3. Any value of  $\tau$  results in an allocation somewhere along this expansion path, and so the unique equilibrium value must fall in either the full- or limited-access segments.

## 7. Full-Access Equilibrium

By definition, every ill policyholder receives treatment in a full-access equilibrium, so it must be that  $\tilde{n} = \frac{q}{D}$ . Analysis can therefore be confined to the values of  $w$  and  $s$  that induce a choice of  $n^*(w, s, P) \geq \frac{q}{D}$  in the physician's problem. Since any  $w$  and  $s$  such that  $n^*(w, s, P) > \frac{q}{D}$  would result in  $\frac{\partial \tilde{Q}}{\partial w} = 0$  and  $\frac{\partial \tau}{\partial w} > 0$ , which would violate (1), analysis can further be confined to those  $w$  and  $s$  inducing  $n^*(w, s, P) = \frac{q}{D}$ . From the previous section, setting  $w = \omega(s; \frac{q}{D}, P)$  is the unique value of  $w$  that would bring about such a choice for any value of  $s$ . Substituting  $w = \omega(s; \frac{q}{D}, P)$ ,  $\tilde{n} = \frac{q}{D}$ , and  $t = \frac{Ds}{q}$  into the MCO's problem and maximizing with respect to  $s$  yields the first order conditions:

$$\begin{aligned} \Delta U \cdot \left( -\frac{\partial \rho}{\partial t} \right) &= WMU_y \cdot \left[ c - Pt \left( \frac{\partial^2 \rho}{\partial t^2} \right) \right] \\ \tau &= q \cdot \omega \left( s; \frac{q}{D}, P \right) + Dcs \end{aligned} \quad (3)$$

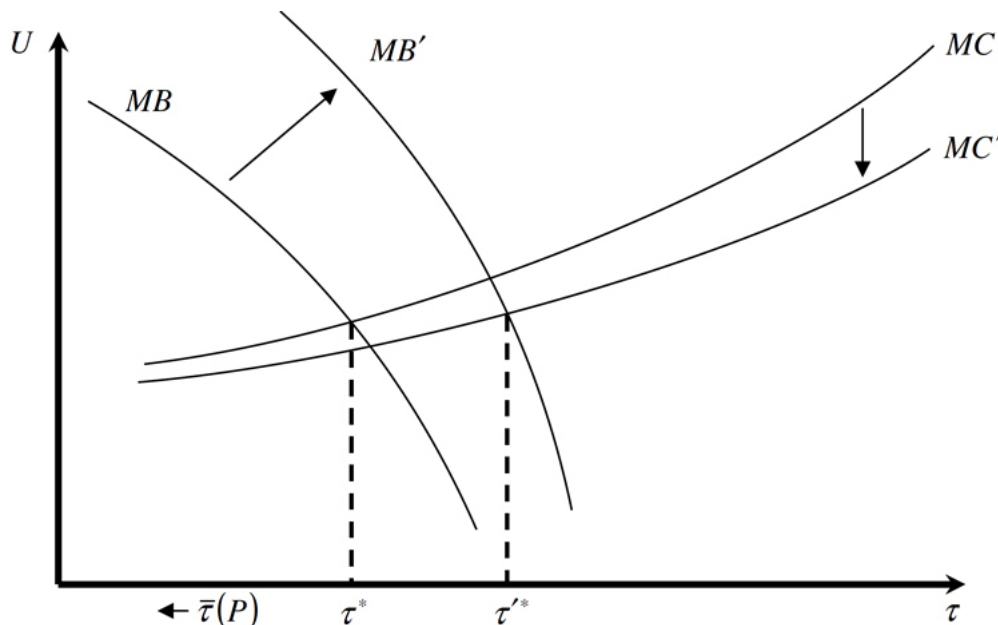


Figure 4 The effect of an increase in malpractice pressure on equilibrium health care spending in a full-access equilibrium.

**Proposition 2:** If A1 and A2 hold, then  $\frac{\partial \tau^*}{\partial P} > 0$ , and  $\frac{\partial \bar{n}^*}{\partial P} = 0$  in any full-access equilibrium.

**Proof:** In appendix.

This proposition states that, starting in any full-access equilibrium, the MCO would respond to rising malpractice pressure by increasing prices in order to maintain the level of access to physicians enjoyed by their customers. Since the set of consumers is of measure 1, and all consumers purchase health insurance in equilibrium, the policy price is equal to total health care spending. Therefore, total health care spending is also increasing in malpractice pressure in any full-access equilibrium.

The intuition for these comparative statics can be found in Figure 4. The marginal benefit to consumers of an increase in the policy price (and thus increased funds for the procurement of resources and physicians' services) is  $MB = q \cdot \Delta U \frac{\partial \bar{Q}}{\partial \tau}$ , which represents the value of a marginally higher probability of ending up with  $H_1$  instead of  $H_2$ .  $MB$  is monotonically decreasing in  $\tau$  for all  $\tau > \bar{\tau}(P)$ . The marginal cost is  $MC = WMU_y$ , which is increasing in  $\tau$  and represents the marginal value of forgone consumption.

An increase in  $P$  has two effects. First, given any amount of resources, the MCO must provide physicians with a greater per-patient payment in order to maintain  $n^* = \frac{q}{D}$ . This means that, given

a level of spending, an increase in  $P$  leaves less funds left over for the procurement of resources, causing ill consumers to receive less treatment and enjoy a lower-quality insurance policy. Given the assumptions on  $U(H, y)$ , the marginal utility of consumption is lower for consumers enjoying  $H_2$  instead of  $H_1$ . Since holding a lower-quality insurance policy increases the likelihood of ending up with  $H_2$ , a higher  $P$  causes consumption to be marginally less valuable ex ante to consumers. This in turn causes  $MC$  to shift down at every  $\tau$ . Second, since there are diminishing returns to treatment, and treatment is lower at a given  $\tau$  and a higher  $P$ , a marginal increase in spending (and thus resources) delivers a greater marginal increase in quality. Thus, an increase in  $P$  causes  $MB$  to shift up at every  $\tau$ . Taken together, these two effects cause the new equilibrium to occur at an unambiguously higher level of spending. Since spending in the initial full-access equilibrium is greater than  $\bar{\tau}(P)$ , full access is preserved in the new equilibrium with the marginally higher  $P$ .

**Proposition 3:** If A1) and A2) are satisfied, then  $\frac{\partial t^*}{\partial P} > 0$  and  $\frac{\partial \bar{Q}^*}{\partial P} > 0$  in a full-access equilibrium if and only if:

$$\epsilon_{J,\tau} < -\frac{\epsilon_{\omega,t}}{\epsilon_{\tau,t}} \quad (4)$$

where  $\epsilon_{x,y}$  is the percent change in  $x$  due to a one-percent change in  $y$  and  $J = \frac{WMU_y}{\Delta U}$ .

**Proof:** In appendix.

Proposition 3 describes the impact of increasing malpractice pressure on health care quality in a full-access equilibrium. From Proposition 2, the MCO raises the the policy price in response to a marginal increase in malpractice pressure in any full-access equilibrium. Since all consumers purchase health insurance in equilibrium, and the MCO makes zero profits, the entire increase in revenues must be divided between increased per-patient payments and resources in the new

equilibrium contract between the MCO and physicians. Condition (4) is necessary and sufficient to determine whether or not the new equilibrium contract provides physicians with more resources than were available under the old contract. Also from Proposition 2, the number of patients that each physician would treat in equilibrium would remain unchanged. Therefore, where (4) holds, the new equilibrium contract would allow physicians to divide a greater amount of resources among the same number of patients, leading to a greater amount of treatment, a lower chance of an adverse outcome, and greater health care quality.

There is reason to expect that Condition 4 holds in all full-access equilibria with reasonable parameter values and functional forms. From (3), where both A1 and A2 hold and (4) holds at least initially,  $Pt \frac{\partial^2 \rho}{\partial t^2}$  must approach  $c$  as  $P$  increases, driving  $\epsilon_{\tau,t}$  to zero and the right side of (4) to positive infinity. This occurs as treatment levels increase to the point of significant gold-plating, where  $\rho(t)$  is relatively flat. If the Inada conditions hold, then the left side of (4) only approaches infinity if the amount consumers spend on health care approaches their total income, driving consumption to zero. Given that the United States, the highest spender on health care both per-capita and as a percentage of GDP, spent less than 15 percent of GDP on health care as recently as 2002 (Anderson et al. 2005), the amount of spending necessary to violate (4) seems unrealistic, particularly if cheaper limited-access insurance policies could be provided. Taking Propositions 1 and 2 together, jurisdictions or medical specialties in full-access equilibria can expect rising malpractice pressure to cause increases in the cost of health insurance and total health care spending. Regarding quality, the effect is analytically ambiguous, but it is likely that equilibrium health care quality would increase with malpractice pressure as long as the level of health care spending is not extremely high.

## 8. Limited-Access Equilibrium

An important finding in the proof of Proposition 1 is that the treatment  $t$  solving the modified insurer's problem is independent of the level of spending, as shown for  $t^*(P)$  in Figure 3. Since,

for a given  $P$ , the level of treatment is uniquely determined in the physician's problem by  $w$ , then  $w^*(P)$  such that  $t(w^*(P), P) = t^*(P)$  is also independent of  $\tau$ . Since modified and unmodified solutions are equivalent in limited-access equilibria, both  $t^*(P)$  and  $w^*(P)$  from the physician's modified problem arise in a limited-access equilibrium for a given level of  $P$ .

**Proposition 4:** If A1 and A2 hold, then  $\frac{\partial t^*}{\partial P} > 0$  in any limited-access equilibrium.

**Proof:** In appendix.

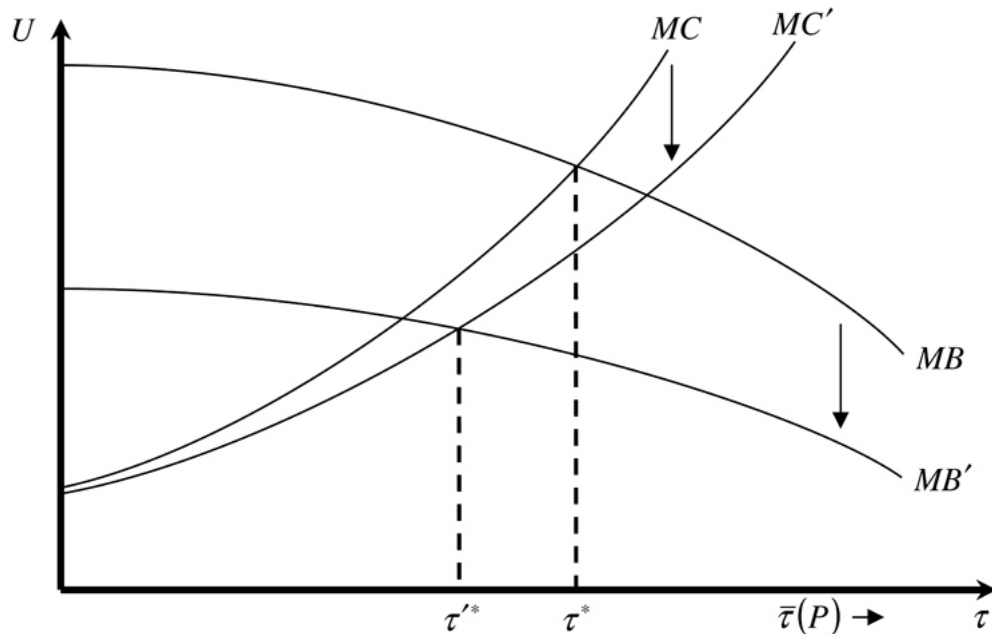
This result is derived from the way  $w^*$  changes in response to rising malpractice pressure. As shown in the proof, even if  $\frac{\partial w^*}{\partial P}$  is positive,  $w^*$  cannot increase in equilibrium faster than  $P$  increases. From the solution to the physician's problem, this means that the left side must decrease with  $P$  in equilibrium, inducing physicians to choose higher values of  $t$  in order to balance the equation. Furthermore, since  $(1 - \rho)$  is solely a function of  $t$ , then the equilibrium probability of patient recovering from illness, conditional on gaining access to a physician, would increase in  $P$  as well.

**Proposition 5:** If A1 and A2 hold, then  $\frac{\partial \tau^*}{\partial P} < 0$  in a limited-access equilibrium if and only if:

$$\tau < \frac{\Delta U}{\frac{\partial \Delta U}{\partial y}} \quad (5)$$

**Proof:** In appendix.

By reason of the concavity of  $\Delta U$  in  $y$  alone, the lowest upper bound on the levels of spending for which Condition (5) can hold is greater than half of total consumer income. Similar to (4), (5) can be expected to hold for all realistic parameter values. The reason that a qualifying condition



**Figure 5** The effect of an increase in malpractice pressure on equilibrium health care spending in a limited-access equilibrium satisfying (5).

is required here is that increases in malpractice pressure have different implications for the marginal benefit and cost of health care spending in the two types of equilibrium. In comparing Figures 4 and 5, a rise in malpractice pressure has the same qualitative effect on  $MC$  in the two equilibrium types, but the opposite effect on  $MB$ . Since  $\tau^*$  is greater than  $\bar{\tau}(P)$  in the full-access case, marginal increases in spending are devoted solely to procuring more resources to improve treatment. In the limited-access case, where  $\tau^*$  is less than  $\bar{\tau}(P)$ , marginal spending increases must be split between increasing treatment and increasing access. An increase in malpractice pressure makes it more expensive for the MCO to induce physicians to treat a given number of patients, which makes access more expensive. This increase in the marginal cost of access, along with no change in the marginal cost of resources, means that a marginal increase in spending is able to provide a relatively smaller increase in quality when  $P$  is high. Essentially, malpractice pressure makes each unit of health care spending less effective in producing quality, which causes the downward shift in  $MB$ . Where (5) holds, the shift in  $MB$  dominates the shift in  $MC$ , leading to a lower equilibrium policy price.

**Corollary 5a:** In any limited-access equilibrium,  $\frac{\partial \tau^*}{\partial P} < 0 \Rightarrow \frac{\partial \tilde{n}^*}{\partial P} < 0$  and  $\frac{\partial \tilde{Q}^*}{\partial P} < 0$ .

**Corollary 5b:** If condition 5 holds in a limited-access equilibrium at  $P'$  then it holds for all  $P \in (P', \infty)$ .

The implication in Corollary 5a is fairly straightforward. Once  $\{w, t\} = \{w^*, t^*\}$ , the only endogenous variable left to affect health care quality  $\tilde{Q}$  is  $\tau$ . Holding other parameters constant,  $\frac{\partial \tilde{Q}^*}{\partial P} = \frac{\partial \tilde{Q}}{\partial \tau} \frac{\partial \tau^*}{\partial P} + \frac{\partial \tilde{Q}}{\partial P} \Big|_{\tau^*}$ . Since quality is increasing in spending for a given level of malpractice pressure, and decreasing in malpractice pressure for a given level of spending,  $\frac{\partial \tau^*}{\partial P} < 0$  makes the right side unambiguously negative. Essentially, rising malpractice pressure makes quality more expensive to provide, so if the funds available for spending on quality decrease then the amount of quality produced must also decrease. Since  $t^*$  increases with  $P$  by Proposition 4, the only way that  $\tilde{Q}^*$  can decrease in  $P$  is if  $\tilde{n}^*$  is also decreasing in equilibrium.

If spending is decreasing in malpractice pressure, then an increase in  $P$  causes the left side of (5) to decrease and the right side to increase, proving corollary 5b. Taken together, the propositions and corollaries in this section imply there is a threshold level of malpractice pressure, after which any further increases in malpractice pressure would cause decreases in health care spending. Since  $\tilde{n}^*$  decreases in  $P$  in limited-access equilibria, this further implies that any full-access equilibria must occur before this threshold, at lower levels of malpractice pressure. Also, despite each patient receiving better treatment, the reduction in ill consumers' access to medical care causes the overall quality of the health care system to decrease in limited-access equilibria.

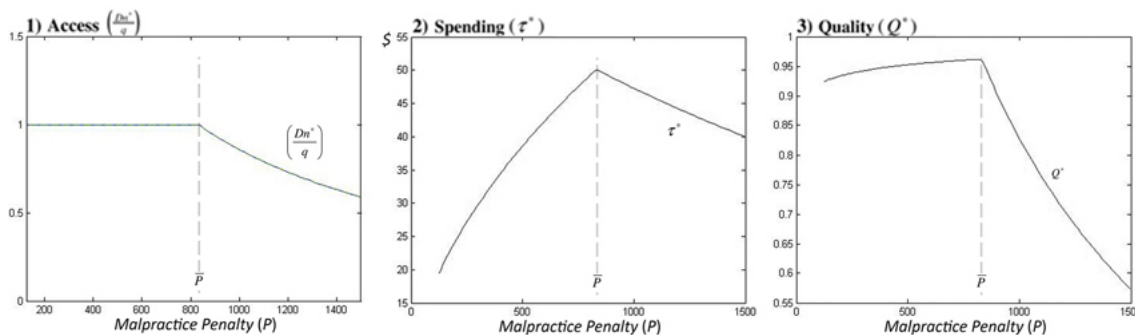
Notice in Figure 6 that the effects of malpractice pressure on spending and quality in the limited-access equilibrium are the opposite of those in the full-access equilibrium. Intuitively, there are two inputs into the quality of health insurance: access and treatment. Where malpractice pressure is low, physicians do not require much incentive to treat a large number of patients. When physicians are so amenable to treating patients, access is cheap relative to treatment, making it optimal to provide as much access as possible. There is an upper bound on the amount of access consumers

Where it occurs	Equilibrium Type	
	Full-Access Low $P$ (likely)	Limited-Access High $P$
Access $\left(\frac{\partial \tilde{n}^*}{\partial P}\right)$	0	—
Treatment $\left(\frac{\partial r^*}{\partial P}\right)$	+	+
Price of Insurance/ Total Expenditure $\left(\frac{\partial \tau^*}{\partial P}\right)$	+	—
System Quality $\left(\frac{\partial \tilde{Q}^*}{\partial P}\right)$	+	—

**Figure 6** Summary of key analytical results from the two possible types of equilibrium.

can be provided, that being “full access”. While full access is not costly at low levels of malpractice pressure, it becomes more costly as pressure increases. Therefore, at low levels of malpractice pressure, the price of insurance and total health care spending are increasing in pressure in order to maintain full access. As long as (4) holds, it is optimal to use enough of the increased revenue from higher policy prices to procure more resources for the physicians than they previously had, which causes quality to increase in malpractice pressure as well.

As the cost of full access increases, eventually consumers’ willingness to pay becomes exhausted. Rather than continue to maintain full-access, the MCO instead offers a lower-priced insurance policy with more limited access to physicians. Should malpractice pressure continue to increase, each level of access becomes more costly. Since the marginal cost of resources remains constant, the MCO substitutes away from access toward treatment in the production of health care quality, causing access to decrease and treatment to increase in malpractice pressure. Finally, since there are decreasing returns to treatment, this substitution makes each dollar spent on health insurance less effective at producing quality. Therefore, consumers would prefer to substitute away from spending on health insurance and instead spend more income on consumption. These two substitution effects cause both spending and quality to decrease in malpractice pressure once pressure is high enough to make limited access preferable to full access. Altogether, given the different comparative statics in the two types of equilibrium, and the values of malpractice pressure



**Figure 7** Numerical example illustrating non-monotonic relationships between malpractice pressure and 1) the level of access, 2) the price of health insurance, and 3) the quality of the health care system.

for which we can expect each type of equilibrium to occur, the model predicts non-monotonic relationships between malpractice pressure and both health care quality and expenditure.

### 9. Discussion and Conclusion

A numerical example is useful to illustrate the predictions of the model. A Cobb-Douglas utility function and the functional form  $\rho(t) = \frac{1}{1+\alpha t}$  where  $\alpha > 0$  produce the relationships in Figure 7. The non-monotonic relationships predicted by the model are clearly apparent. This pattern suggests an explanation for the findings of past empirical studies, namely Baicker & Chandra (2005), Hellinger & Encinosa (2006), Congressional Budget Office (2006), and Baicker et al (2007). Essentially, studies using broad data sets spanning multiple jurisdictions and/or medical specialties are likely to include observations from both types of equilibrium. Their empirical methods uncover the best monotonic relationship that fits observations from a fundamentally non-monotonic relationship. Such a regression is likely to return coefficients that are weaker than those that would emerge from a data set with observations from a single equilibrium type.

Studies utilizing data on a more narrowly defined population, either a single medical specialty or small number of jurisdictions, are more likely to contain observations from a single equilibrium type. If that were the case, then studies from environments with relatively low malpractice pressure should report increases in health care spending and quality as pressure rises, while those from high-pressure environments should report decreases instead. The difference in findings between specific studies with differing malpractice climates are generally consistent with the predictions of

the model. Kessler & McClellan (1996) and (2002) find a positive relationship regarding spending in data covering heart-attack and heart disease patients on Medicare. Elderly patients are relatively unlikely sources for malpractice claims, and cardiology is not among the most risky of medical specialties (Kessler & McClellan 1996). Lakdawalla & Seabury (2009) draw many of their observations from California, which has experienced some of the lowest malpractice insurance rate increases (General Accounting Office 2003) and been at the forefront of crafting tort reform concerning medical malpractice (Brenner & Smith 2004). It seems plausible, therefore, that physicians serving the populations studied in the works of Kessler & McClellan and Lakdawalla & Seabury practice under below-average malpractice pressure, and thus possibly in full-access equilibrium.

On the other hand, studies reporting negative relationships use data on infant health and obstetric care (Dubay et al. 2001, Currie & MacLeod 2008). This specialty is one of the riskiest for the frequency and severity of malpractice lawsuits. The cost of malpractice insurance for obstetricians increased 180 percent between 1977 and 1984, versus 109 percent for lower risk specialties (Danzon, Pauly, & Kington 1990), and the average malpractice premium for obstetricians was more than double that for all physicians by 1992 (Dubay et al. 2001). If the high level of malpractice pressure has caused the obstetrics specialty to tend toward limited-access equilibria, with comparative statics being the opposite of full-access equilibria, then it is unsurprising that the empirical studies would conflict with each other. Overall, the monotonic relationships found (or not found) in past empirical studies are consistent with the non-monotonic relationships predicted by the model presented here.

There are several implications of these findings to public policy. First, tort reform is not a “silver bullet” policy capable of raising health care quality while also lowering the cost of care. Even if a given reform was successful in changing the prevailing level of malpractice pressure, quality and spending would move together. Policymakers, therefore, face a tradeoff and must decide whether quality improvements or cost reductions would be of greater benefit to the affected population. Second, given a clear goal, it is necessary to determine which of the two types of equilibrium currently holds in a given jurisdiction or medical specialty in order to accurately

predict the qualitative effects of tort reforms. This is because tort reforms have opposite effects in the two equilibrium types, which suggests a more targeted approach to tort reform would be more successful than sweeping changes. Finally, the qualitative effect of tort reforms can be determined by the current level of access that ill consumers enjoy with their physicians. Excessive waiting times, a high incidence of late treatment, or other significant difficulty in securing a physician's services can be considered examples of poor access. Where these are at feasibly low levels, tort reforms lowering malpractice pressure should lower insurance policy prices and total health care expenditure, while causing some reductions in health care quality. Where they are unnaturally high, the same reforms would have the opposite effect. There is a need for data that includes measures of access alongside measures of malpractice pressure, health care spending, and quality in order to separate observations from different equilibrium types. Failing to do so puts the accuracy of reported results on the impact of defensive medicine and the effect of tort reforms in doubt.

## References

- Anderson, J.F., P.S. Hussey, B.K. Frogner, H.R. Waters. 2005. Health Spending in the United States and the Rest of the Industrialized World. *Health Affairs* **24**(4) 903–914.
- Arlen, J., and W.B. MacLeod. 2005. Torts, Expertise, and Authority: Liability of Physicians and Managed Care Organizations. *The RAND J. of Economics* **36**(3) 494–519.
- Baicker, K., and A. Chandra. 2005. The Consequences of the Growth of Health Insurance Premiums. *AEA Papers and Proceedings* **95**(2) 214–218.
- Baicker, K., E.S. Fisher, and A. Chandra. 2007. Malpractice Liability Costs and the Practice of Defensive Medicine in the Medicare Program. *Health Affairs* **26**(3) 841–852.
- Brenner, R.J., and J.J. Smith. 2004. The Malpractice Liability Crisis. *J. of the American College of Radiology* **1**(1) 18–22.
- Congressional Budget Office. 2004. Limiting Tort Liability for Medical Malpractice. Economic and Budget Issue Brief, Congressional Budget Office.

- Congressional Budget Office. 2006. Medical Malpractice Tort Limits and Health Care Spending. Background Paper, Congressional Budget Office.
- Currie, J., and W.B. MacLeod. First Do No Harm? Tort Reforms and Birth Outcomes. *The Quarterly J. of Economics* May, 2008. 795–830.
- Danzon, P.M. 1985. Liability and Liability Insurance for Medical Malpractice. *The J. of Health Economics* **4** 309–331.
- Danzon, P.M., M.V. Pauly, and R.S. Kington. 1990. The Effects of Malpractice Litigation on Physicians' Fees and Incomes. *AEA Papers and Proceedings* **80**(2) 122–127.
- Dubay, L., R. Kaestner, and T. Waidmann. 2001. Medical malpractice liability and its effect on prenatal care utilization and infant health. *J. of Health Economics* **20** 591–611.
- General Accounting Office. 2003. Medical Malpractice Insurance: Multiple Factors Have Contributed to Increased Premium Rates. GAO-03-702 Report to Congressional Requesters, General Accounting Office.
- Hellinger, F.J. and W.E. Encinosa. 2006. The Impact of State Laws Limiting Malpractice Damage Awards on Health Care Expenditures. *American J. of Public Health* **96**(8) 6–12.
- Kessler, D., and M. McClellan. 1996. Do Doctors Practice Defensive Medicine?. *The Quarterly J. of Economics* **111**(2) 353–390.
- Kessler, D.P., and M.B. McClellan. 2002. How liability law affects medical productivity. *J. of Health Economics* **21** 931–955.
- Kessler, D.P., W.M. Sage, and D.J. Becker. 2005. Impact of Malpractice Reforms on the Supply of Physician Services. *J. of the American Medical Association* **293**(21) 2618–2625.
- Klick, J., and T. Stratmann. 2007. Medical Malpractice Reform and Physicians in High-Risk Specialties. *J. of Legal Studies* **36**(2) S121–S142.
- Kpelitse, K.A. 2010. Physician Payments and Medical Malpractice Mechanisms. Working Paper.
- Lakdawalla, D.A., and S.A. Seabury. 2009. The Welfare Effects of Medical Malpractice Liability. NBER Working Paper.
- Leger, P.T. 2000. Quality control mechanisms under capitation payment for medical services. *Canadian J. of Economics* **33**(2) 564–586.

- Mello, M.M., D.M. Studdert, C.M. DesRoches, J. Peugh, K. Zapert, T.A. Brennan, and W.M. Sage. 2005. Effects of a Malpractice Crisis on Specialist Supply and Patient Access to Care. *Annals of Surgery* **242**(5) 621–628.
- Olbrich, A. 2008. Heterogeneous physicians, lawsuit costs, and the negligence rule. *International Review of Law and Economics* **28** 78–88.
- Paik, M., B.S. Black, D.A. Hyman, and C.M. Silver. 2011. Will Tort Reform Bend the Cost Curve? Evidence from Texas. Working Paper.
- Sloan, F.A. and J.H. Shadle. 2009. Is there empirical evidence for Defensive Medicine? A reassessment. *J. of Health Economics* **28** 481–491.
- Zeiler, K. 2004. Medical Malpractice and Contract Disclosure: An Equilibrium Model of the Effects of Legal Rules on Behaviour in Health Care Markets. *American Law & Economics Association Annual Meetings* **60**. 1–75.

## 10. Appendix

**Proposition 1:** If A1 and A2 hold, then there exists a unique solution to the MCO's problem that is either a full-access or a limited-access equilibrium. **Proof of Proposition 1:**

Combining FOCs 1 and 2:

$$\frac{\left(\frac{\partial \tilde{Q}}{\partial w}\right)}{\left(\frac{\partial \tilde{Q}}{\partial s}\right)} = \frac{\left(\frac{\partial \tau}{\partial w}\right)}{\left(\frac{\partial \tau}{\partial s}\right)}$$

and

$$\frac{w}{P} = \rho - t \frac{\partial \rho}{\partial t}$$

FOCs become:

$$\frac{\left(\frac{\partial \tilde{Q}}{\partial w}\right)}{\left(\frac{\partial \tilde{Q}}{\partial s}\right)} = \frac{\left(\frac{\partial \tau}{\partial w}\right)}{\left(\frac{\partial \tau}{\partial s}\right)}$$

...

$$\frac{1 - \rho + t \frac{\partial \rho}{\partial t}}{1 - \rho} = \frac{tP \frac{\partial^2 \rho}{\partial t^2} + \frac{w}{t}}{c + \frac{w}{t}} \quad (6)$$

...

$$\frac{1 - \rho}{w + ct} = \frac{-\frac{\partial \rho}{\partial t}}{c - Pt \frac{\partial^2 \rho}{\partial t^2}} \quad (7)$$

Condition (7) is the same tangency condition necessary to solve the problem:

$$\max_{n,t} \left\{ Q(n,t) = \left(\frac{D}{q}\right) n(1-\rho) \right\}$$

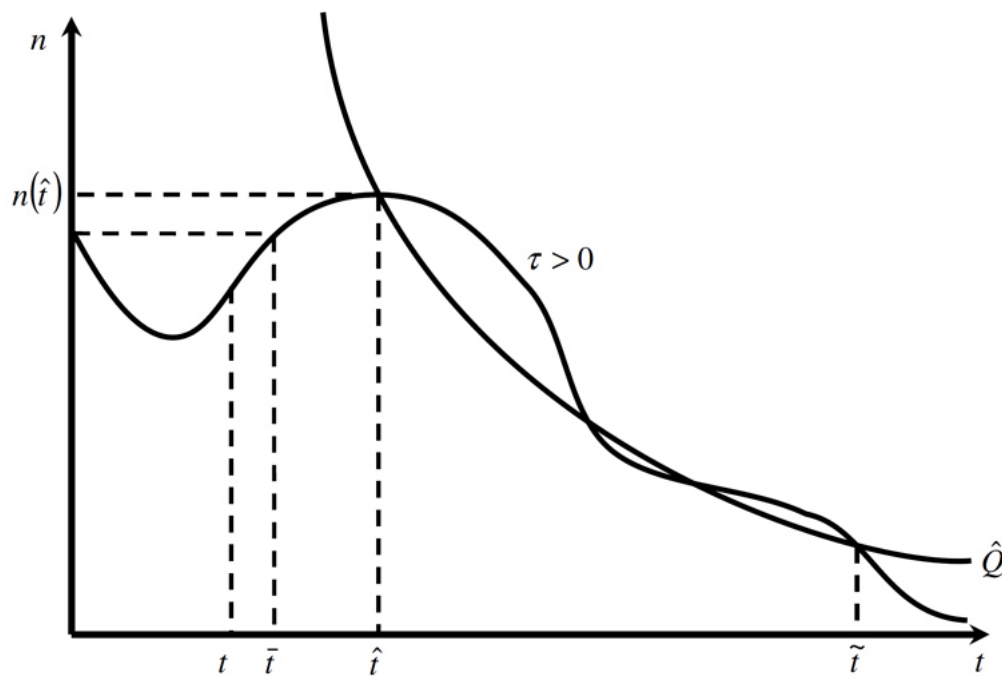


Figure 8 Existence of a point satisfying (7) in the range of  $t$  where A1 and A2 hold.

$$\text{subject to } \tau = Dn(\omega(t; P) + ct)$$

$$\text{where } \omega(t; P) = P\left(\rho - t \frac{\partial \rho}{\partial t}\right)$$

The isoquants in this problem are convex, but isocosts are not. It is thus necessary to investigate whether there is a  $t$  in the choice set that can satisfy condition (7) when A1 and A2 hold, and if it does exist, whether or not it is unique.

Existence of tangency point

Not only is it necessary to show that a  $t$  satisfying (7) exists, but it must exist given that A1 and A2 hold, that is where  $\frac{\partial^2 \rho}{\partial t^2} + t \frac{\partial^3 \rho}{\partial t^3} < 0$  in order for that  $t$  to be an argmax.

An illustration of all points described can be found in Figure 8. By A2, on the same isocost curve, there exists an interior level of treatment  $\bar{t}$  such that  $n$  is as high as where  $t = 0$ . Let  $\hat{Q}$  represent

the level of quality described by the isoquant that passes through the maximum  $n$  achievable on a given isocost curve, defined as  $n(\hat{t})$  where  $\hat{t}$  is the level of treatment at which the maximum  $n$  is reached. Note that  $\hat{Q} > 0$ ,  $\hat{t} > 0$ , and by A1),  $\frac{\partial^2 \rho}{\partial t^2} + t \frac{\partial^3 \rho}{\partial t^3} < 0$  holds for all  $t \geq \hat{t}$ . At  $\{\hat{t}, n(\hat{t})\}$ , the slope of the isocost curve is zero while the slope of the isoquant is negative. Also, since  $\hat{Q} > 0$  and for a given  $\tau > 0$ ,  $\lim_{t \rightarrow 0} Q = 0$ , there is a level of treatment  $\tilde{t}$  where the isoquant and isocost intersect and the isoquant remains above the isocost for all  $t > \tilde{t}$ . Thus, there is a negative difference between isocost and isoquant for all  $t \notin [\hat{t}, \tilde{t}]$ , at least a subset of  $[\hat{t}, \tilde{t}]$  such that there is a positive distance, and zero distance at  $\hat{t}$  and  $\tilde{t}$ . Note that all  $t \notin [\hat{t}, \tilde{t}]$  are irrelevant alternatives, and thus the set of relevant choices is compact.

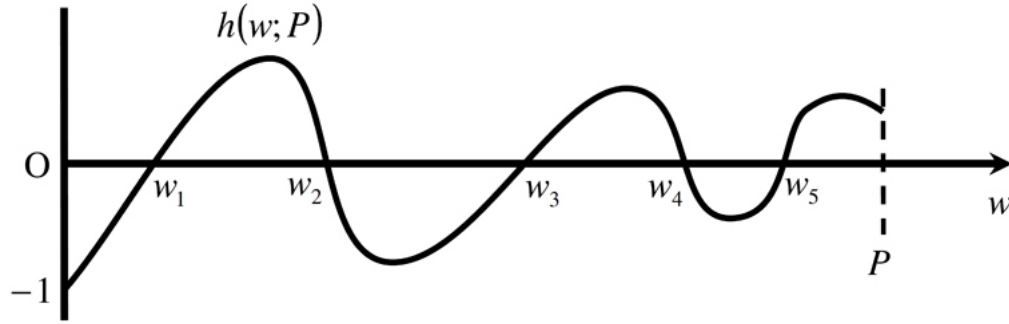
At  $\hat{t}$ , the values of  $n$  in the isoquant describing  $\hat{Q}$  and the isocost describing  $\tau$  are equal, while the slope of the isocost is greater than that of the isoquant. This implies that, at  $\hat{t}$ , the left side of (7) is less than the right side. At  $\tilde{t}$ , the values of  $n$  in the isoquant describing  $\hat{Q}$  and the isocost describing  $\tau$  are equal, while the slope of the isocost is less than that of the isoquant. This implies that, at  $\tilde{t}$ , the left side of (7) is greater than the right side. Since both isoquant and isocost are continuous, by intermediate value theorem there exists at least one point  $t^* \in [\hat{t}, \tilde{t}]$  such that (7) holds. Since (7) is solely a function of  $t$ , tangency will hold at  $t^*$  for any  $n$ . Therefore, there always exists a point of tangency in the region where A1 and A2 hold.

### Unique maximum

By subbing in the solution to the physician's problem, (7) can be rearranged into:

$$\frac{w}{P} = \frac{c - PA(t)}{c - PB(t)} \quad (8)$$

where  $A(t) = (1 - \rho) t \frac{\partial^2 \rho}{\partial t^2} > 0$  and  $B(t) = \frac{\partial \rho}{\partial t} < 0$ . Note from the physician's problem that for a given  $P$ , there is a unique finite  $t$  for any  $w \in (0, P)$ , where  $t$  is monotonically decreasing in  $w$ . Let:



**Figure 9** Illustration of uniqueness result. Even though (7) holds at  $w_1$  to  $w_5$ , only  $w_1$  can hold in equilibrium where A1 and A2 hold.

$$h(w; P) = \frac{w}{P} - \frac{c - PA(t)}{c - PB(t)}$$

For an illustration, see Figure 9. Conditions (7) and (8) are satisfied where  $h(w; P) = 0$ . From the physician's problem,  $\lim_{w \rightarrow 0} t = \infty$  and  $\lim_{w \rightarrow P} t = 0$ . This means that  $\lim_{w \rightarrow 0} h(w; P) = -1$  and  $\lim_{w \rightarrow P} h(w; P) > 0$ . Since  $\rho$  is continuous, by intermediate value theorem there exists at least one  $w$  such that  $h(w; P) = 0$ . The first derivative of  $h(\cdot)$  is:

$$\frac{\partial h(\cdot)}{\partial w} = \frac{1}{P} + \frac{P \frac{\partial t}{\partial w} \left[ \frac{\partial A}{\partial t} (c - PB) - \frac{\partial B}{\partial t} (c - PA) \right]}{(c - PB)^2}$$

where the sign of  $\frac{\partial h(\cdot)}{\partial w}$  depends on the condition:

$$\begin{aligned} \frac{\partial h(\cdot)}{\partial w} &\begin{matrix} \geq \\ < \end{matrix} 0 \\ \left( \frac{\rho}{1-\rho} \right) \left( \frac{\partial^2 \rho}{\partial t^2} \right) + \frac{A(c - PB)}{P(1-\rho)^2} - \left( \frac{\frac{\partial^2 \rho}{\partial t^2}}{1-\rho} \right) \cdot h(\cdot) &\begin{matrix} \geq \\ < \end{matrix} \frac{\partial^2 \rho}{\partial t^2} + t \frac{\partial^3 \rho}{\partial t^3} \end{aligned}$$

Where the left side is greater than zero whenever  $h(w; P) = 0$ . Let  $\{w_1, w_2, \dots\}$  be values of  $w$  such that  $h(w; P) = 0$ , and let  $\{t_1, t_2, \dots\} = \{t(w_1), t(w_2), \dots\}$ . Let  $w_i$  be increasing in  $i$ , which

implies that  $t_i$  is decreasing in  $i$ . If any  $t_i$  is such that  $\frac{\partial h(\cdot)}{\partial w} \leq 0$ , then  $\frac{\partial^2 \rho}{\partial t^2} + t \frac{\partial^3 \rho}{\partial t^3} \geq 0$  at  $t_i$  and, by A1, at all  $t_j$  such that  $j > i$ . Since we've already shown that a point of tangency exists where  $\frac{\partial^2 \rho}{\partial t^2} + t \frac{\partial^3 \rho}{\partial t^3} < 0$ , and that  $h(w; P)$  is continuous, it must be true that that  $t_1$  is the only element of the set  $\{t_1, t_2, \dots\}$  that can hold in equilibrium where A1 and A2 hold. Therefore, the level of treatment  $t^*(c, P)$  and revenue-per-patient  $w^*(c, P)$  satisfying (1) and (2) are unique and depend only on  $c$  and  $P$ . The  $c$  is suppressed in the remainder of this proof.

Multiplying both sides of FOC (1) by  $-P \left(\frac{s}{n^2}\right) \left(\frac{\partial^2 \rho}{\partial t^2}\right)$  and adding each side to each side of FOC (2) yields the necessary condition:

$$\begin{aligned} \Delta U \cdot \left(-\frac{\partial \rho}{\partial t}\right) &= WMU_y \cdot \left[c - Pt \left(\frac{\partial^2 \rho}{\partial t^2}\right)\right] \\ \frac{\left(-\frac{\partial \rho}{\partial t}\right)}{c - Pt \left(\frac{\partial^2 \rho}{\partial t^2}\right)} &= \frac{WMU_y}{\Delta U} \end{aligned}$$

Substituting  $w^*(P)$  and  $t^*(P)$  into this condition leaves the left side unchanging in  $s$  while the right side is monotonically increasing in  $s$  through its effect on  $n$ . There is thus a unique  $s^*$  solving this necessary condition whenever  $\tau^* > 0$ , and therefore a unique  $n^*(w^*, s^*, P)$  for any given  $P$  solving the insurer's first-order conditions. If this  $n^*(w^*, s^*, P)$  is less than  $\frac{q}{D}$ , then the unique limited-access equilibrium is  $\{w^*, s^*\}$ . If it is greater than  $\frac{q}{D}$ , then the  $n$  solving the insurer's first-order conditions is infeasible and setting  $n = \frac{q}{D}$  on the boundary dominates any  $n < \frac{q}{D}$  due to the uniqueness of  $n^*(w^*, s^*, P)$ . This would give rise to the full-access equilibrium, where  $t = \frac{Ds}{q}$ . FOC (3) is the same as the necessary condition above, with the exception that the left side is also decreasing in  $s$  over all values that can hold in equilibrium. This means that there is a unique  $s^*$  in the full-access equilibrium, and that the full-access equilibrium  $\{\omega(s^*; \frac{q}{D}, P), s^*\}$  is also unique. QED.

**Proposition 2:** If A1 and A2 hold, then  $\frac{\partial \tau^*}{\partial P} > 0$ , and  $\frac{\partial \bar{n}^*}{\partial P} = 0$  in any full-access equilibrium.

**Proof of Proposition 2:**

A1 and A2 ensure that any equilibrium  $s^*$  must occur to the right of the global minimum of  $\tau(s; \bar{n}, P)$ , where  $\tau(s; \bar{n}, P)$  is monotonically increasing in  $s$ . The function can therefore be inverted, the inverse  $\sigma(\tau; \bar{n}, P)$  being resources as a function of spending. Using the inverse, the insurer's problem can be rewritten with  $\tau$  as the only choice variable:

$$\max_{\tau} \quad \left\{ (1-q)U(H_1, m - \tau) + qU(H_2, m - \tau) + q\tilde{Q}(\tau, P)\Delta U \right\}$$

$$\text{where} \quad \tilde{Q}(\tau, P) = 1 - \rho \left( \frac{D \cdot \sigma(\tau; \frac{q}{D}, P)}{q} \right)$$

This problem yields the first-order condition:

$$q \left( \frac{\partial Q}{\partial \tau} \right) \Delta U = WMU_y$$

and the comparative static:

$$\frac{d\tau}{dP} = \frac{\left( \frac{\partial WMU_y}{\partial P} \right) - q\Delta U \left( \frac{\partial^2 Q}{\partial \tau \partial P} \right)}{q \left( \frac{\partial \Delta U}{\partial \tau} \right) \left( \frac{\partial Q}{\partial \tau} \right) + q\Delta U \left( \frac{\partial^2 Q}{\partial \tau^2} \right) - \left( \frac{\partial WMU_y}{\partial \tau} \right)}$$

Since  $WMU_y > 0$ ,  $\frac{\partial WMU_y}{\partial P} < 0$ ,  $\frac{\partial \Delta U}{\partial \tau} < 0$ ,  $\frac{\partial Q}{\partial \tau} > 0$ ,  $\frac{\partial WMU_y}{\partial \tau} > 0$ , the sign of  $\frac{d\tau}{dP}$  depends on the signs of  $\frac{\partial^2 Q}{\partial \tau \partial P}$  and  $\frac{\partial^2 Q}{\partial \tau^2}$ .

$$\frac{\partial^2 Q}{\partial \tau \partial P} = - \left( \frac{D}{q} \right) \left[ \frac{\partial^2 \rho}{\partial t^2} \left( \frac{D}{q} \right) \frac{\partial \sigma}{\partial P} \frac{\partial \sigma}{\partial \tau} + \frac{\partial \rho}{\partial t} \left( \frac{\partial^2 \sigma}{\partial \tau \partial P} \right) \right]$$

where,

$$\frac{\partial^2 \sigma}{\partial \tau \partial P} = D^{-1} \left[ c - Pt \frac{\partial^2 \rho}{\partial t^2} \right]^{-2} \left[ t \frac{\partial^2 \rho}{\partial t^2} + P \left( \frac{D}{q} \right) \frac{\partial \sigma}{\partial P} \left[ \frac{\partial^2 \rho}{\partial t^2} + t \frac{\partial^3 \rho}{\partial t^3} \right] \right]$$

Since A1 and A2 imply that  $\frac{\partial^2 \rho}{\partial t^2} + t \frac{\partial^3 \rho}{\partial t^3} < 0$  in any equilibrium,  $\frac{\partial^2 \sigma}{\partial \tau \partial P}$  is positive, and therefore  $\frac{\partial^2 Q}{\partial \tau \partial P}$  is positive as well.

$$\frac{\partial^2 Q}{\partial \tau^2} = - \left( \frac{D}{q} \right) \left[ \frac{\partial^2 \rho}{\partial t^2} \left( \frac{D}{q} \right) \left( \frac{\partial \sigma}{\partial \tau} \right)^2 + \frac{\partial \rho}{\partial t} \left( \frac{\partial^2 \sigma}{\partial \tau^2} \right) \right]$$

where,

$$\frac{\partial^2 \sigma}{\partial \tau^2} = P \left( \frac{D^2}{q} \right) \left( \frac{\partial s}{\partial \tau} \right)^3 \left[ \frac{\partial^2 \rho}{\partial t^2} + t \frac{\partial^3 \rho}{\partial t^3} \right]$$

Once again, A1) and A2) make  $\frac{\partial^2 \rho}{\partial t^2} + t \frac{\partial^3 \rho}{\partial t^3} < 0$  in any equilibrium. This makes  $\frac{\partial^2 \sigma}{\partial \tau^2}$  negative and thus  $\frac{\partial^2 Q}{\partial \tau^2}$  negative. Along with the previously mentioned signed expressions,  $\frac{\partial^2 Q}{\partial \tau \partial P} > 0$  and  $\frac{\partial^2 Q}{\partial \tau^2} < 0$  make  $\frac{d\tau^*}{dP}$  unambiguously positive.

Since full-access equilibria are defined such that  $\tau^*(P) > \bar{\tau}(P)$ , the marginal change in spending produced by a marginal change in malpractice pressure  $P$  to  $P'$  still leaves  $\tau^*(P') > \bar{\tau}(P')$ . Therefore, a marginal change in  $P$  does not precipitate a departure from the full-access equilibrium, and  $\tilde{n}$  remains unchanged. Thus,  $\frac{\partial \tilde{n}^*}{\partial P} = 0$ . QED.

**Proposition 3:** If A1) and A2) are satisfied, then  $\frac{\partial t^*}{\partial P} > 0$  and  $\frac{\partial \tilde{Q}^*}{\partial P} > 0$  in a full-access equilibrium if and only if:

$$\epsilon_{h,\tau} < -\frac{\epsilon_{\omega,t}}{\epsilon_{\tau,t}} \quad (9)$$

where  $\epsilon_{x,y}$  is the percent change in  $x$  due to a one-percent change in  $y$  and  $h = \frac{WMU_y}{\Delta U}$ . **Proof of Proposition 3:**

In the full-access equilibrium, the FOC with respect to  $s$  is:

$$q \left( \frac{\partial Q}{\partial s} \right) \Delta U = WMU_y \left( \frac{\partial \tau}{\partial s} \right)$$

Substituting for the first derivatives and rearranging yields:

$$\begin{aligned} \frac{(-\frac{\partial \rho}{\partial t})}{c - Pt \left( \frac{\partial^2 \rho}{\partial t^2} \right)} &= \frac{WMU_y}{\Delta U} \\ H \left( \frac{Ds}{q}, P \right) &= J \left( \frac{Ds}{q}, \tau \right) \end{aligned}$$

Since  $\tau = \tau(s; \frac{q}{D}, P)$  in the full access equilibrium, this is an implicit function describing equilibrium resources ( $s^*$ ). By the Implicit Function Theorem:

$$\frac{ds^*}{dP} = \frac{- \left( \frac{\partial H}{\partial P} - \frac{\partial J}{\partial \tau} \frac{\partial \tau}{\partial P} \right) \Big|_s}{\frac{\partial H}{\partial t} \frac{\partial t}{\partial s} - \frac{\partial J}{\partial t} \frac{\partial t}{\partial s} - \frac{\partial J}{\partial \tau} \frac{\partial \tau}{\partial s}}$$

Where A1 and A2 hold, the denominator is unambiguously negative. Therefore,  $\frac{ds^*}{dP}$  will be positive if and only if  $\left( \frac{\partial H}{\partial P} - \frac{\partial J}{\partial \tau} \frac{\partial \tau}{\partial P} \right) \Big|_s$  is positive for a given amount of resources. The expression for  $\left( \frac{\partial H}{\partial P} - \frac{\partial J}{\partial \tau} \frac{\partial \tau}{\partial P} \right) \Big|_s$  can be rearranged into:

$$\left( \frac{\frac{\partial \Delta U}{\partial y}}{\Delta U} \right) WMU_y + \frac{\partial WMU_y}{\partial \tau} < \left[ \frac{t \cdot \Delta U \left( \frac{\partial^2 \rho}{\partial t^2} \right) \left( -\frac{\partial \rho}{\partial t} \right)}{q \left( \rho - t \frac{\partial \rho}{\partial t} \right) \left( c - Pt \frac{\partial^2 \rho}{\partial t^2} \right)^2} \right]$$

Since:

$$\begin{aligned} \frac{\partial \omega}{\partial t} &= -Pt \frac{\partial^2 \rho}{\partial t^2} \\ \frac{\partial \tau}{\partial t} &= q \left[ c - Pt \frac{\partial^2 \rho}{\partial t^2} \right] \\ \omega &= P \left[ \rho - t \frac{\partial \rho}{\partial t} \right] \\ \frac{\left( -\frac{\partial \rho}{\partial t} \right)}{c - Pt \left( \frac{\partial^2 \rho}{\partial t^2} \right)} &= \frac{WMU_y}{\Delta U} \end{aligned}$$

The expression for  $\left( \frac{\partial H}{\partial P} > \frac{\partial J}{\partial \tau} \frac{\partial \tau}{\partial P} \right) \Big|_s$  can further be rearranged into:

$$\epsilon_{J,\tau} < -\frac{\epsilon_{\omega,t}}{\epsilon_{\tau,t}}$$

where  $\epsilon_{x,y}$  is the percent change in  $x$  due to a one-percent change in  $y$  and  $J = \frac{WMU_y}{\Delta U}$ . This expression implies  $\frac{ds^*}{dP} > 0$ , and since  $\tilde{n}^*$  is constant in the full-access equilibrium,  $\frac{ds^*}{dP} > 0$  implies that  $\frac{\partial t^*}{\partial P} > 0$  and  $\frac{\partial \tilde{Q}^*}{\partial P} > 0$ . QED.

**Proposition 4:** If A1 and A2 hold, then  $\frac{\partial t^*}{\partial P} > 0$  in any limited-access equilibrium. **Proof of Proposition 4:**

From the physician's problem, there is a unique level of treatment for any combination of  $w$  and  $P$ , defined as  $t(w, P)$ . From proposition 1, there is a unique equilibrium  $w$  ( $w^*$ ) and thus a unique

equilibrium level of treatment for any value of  $P$ , or  $t^*(P) = t(w^*(P), P)$ . The effect of a change in  $P$  on  $t^*(P)$  is thus:

$$\frac{\partial t^*}{\partial P} = \frac{\partial t}{\partial w} \frac{\partial w^*}{\partial P} + \left( \frac{\partial t}{\partial P} \right) \Big|_w$$

Using comparative statics from the physician's problem, this can be rearranged into:

$$\frac{\partial t^*}{\partial P} = \frac{\partial t}{\partial w} \left( \frac{\partial w^*}{\partial P} - \frac{w}{P} \right)$$

If  $\frac{\partial w^*}{\partial P}$  is negative, then  $\frac{\partial t^*}{\partial P}$  is unambiguously positive. If  $\frac{\partial w^*}{\partial P}$  is positive, then  $\frac{\partial t^*}{\partial P}$  will be positive as long as  $\frac{w}{P} > \frac{\partial w^*}{\partial P}$ . Using (8) where  $t = t^*(P)$ , The first derivative of  $w^*(P)$  can be written as:

$$\frac{\partial w^*}{\partial P} = \frac{(c - PA)^2 - P^2 A(A - B) - \left(\frac{w}{P}\right) P^2 G(t) \frac{\partial t}{\partial w}}{(c - PB)^2 - P^2 G(t) \frac{\partial t}{\partial w}}$$

where  $G(t) = (c - PA) \frac{\partial B}{\partial t} - (c - PB) \frac{\partial A}{\partial t}$ . The denominator will be positive as long as:

$$\frac{\partial^2 \rho}{\partial t^2} + t \frac{\partial^3 \rho}{\partial t^3} < \left( \frac{\rho}{1 - \rho} \right) \left( \frac{\partial^2 \rho}{\partial t^2} \right) + \frac{A(c - PB)}{P(1 - \rho)^2}$$

The right side is positive, so A1 and A2 ensure that this inequality holds. The denominator is therefore positive. Equilibrium treatment is increasing in  $P$  if and only if:

$$\frac{w}{P} > \frac{\partial w^*}{\partial P}$$

...

$$B < A$$

Since  $B$  is negative and  $A$  is positive, this condition always holds, so  $\frac{\partial t^*}{\partial P} > 0$ . QED.

**Proposition 5:** If A1 and A2 hold, then  $\frac{\partial \tau^*}{\partial P} < 0$  in a limited-access equilibrium if and only if:

$$\tau < \frac{\Delta U}{\frac{\partial \Delta U}{\partial y}} \quad (10)$$

### Proof of Proposition 5:

Since the equilibrium values  $\{w^*, t^*\} = \{w^*(c, P), t^*(c, P)\}$  depend only on  $c$  and  $P$ , resources can be expressed as a function of  $\tau$  and  $P$  alone. Let this function be  $s = \psi(\tau, P)$ :

$$\begin{aligned} \tau &= Dw^*n^*[w^*, \psi(\tau, P), P] + Dc\psi(\tau, P) \\ \Rightarrow \frac{\partial \psi}{\partial \tau} &= \frac{t}{D(w+ct)} > 0, \\ \frac{\partial \psi}{\partial P} &= \left(\frac{nt}{w+ct}\right) \left[ \left(\frac{w+ct}{t} - \left(c - t \frac{\partial^2 \rho}{\partial t^2}\right)\right) \frac{\partial t^*}{\partial P} - \frac{w}{P} \right] \end{aligned}$$

Since (7) must hold in equilibrium,  $\frac{\partial \psi}{\partial P}$  can be rearranged into:

$$\frac{\partial \psi}{\partial P} = \left(\frac{n}{1-\rho}\right) \left[ \left(1 - \frac{w}{P}\right) \frac{\partial t^*}{\partial P} - t \left(\frac{1-\rho}{w+ct}\right) \frac{w}{P} \right]$$

Similar to Proposition 2, the insurer's problem can be written in terms of  $w^*$  and  $\psi(\tau, P)$ , with  $\tau$  being the choice variable:

$$\max_{\tau} \quad \left\{ (1-q)U(H_1, m-\tau) + qU(H_2, m-\tau) + qQ(\tau, P)\Delta U \right\}$$

$$\text{where} \quad Q(\tau, P) = \left( \frac{Dn^*(w^*, \psi, P)}{q} \right) (1 - \rho(t^*))$$

Solving yields the same first order condition as in Proposition 2, but where  $\frac{\partial Q}{\partial \tau} = \frac{1}{q} \left( \frac{1-\rho}{w+ct} \right) > 0$  and  $\frac{\partial^2 Q}{\partial \tau^2} = 0$ . The FOC can thus be simplified into:

$$\frac{1-\rho}{w+ct} = \frac{WMU_y}{\Delta U}$$

Applying the implicit function theorem yields the same comparative static as in Proposition 2, with the exception that  $\frac{\partial^2 Q}{\partial \tau^2} = 0$ , the signs of the terms in the denominator are the same as in Proposition 1, making the denominator unambiguously negative. Therefore, the sign of  $\frac{\partial \tau^*}{\partial P}$  will be determined by the sign of the numerator. Breaking down each component of the numerator. Using  $\tilde{Q}(\tau, P) = \left( \frac{Dn^*(w^*, \psi, P)}{q} \right) (1 - \rho(t^*))$  and (7), we find  $\frac{\partial WMU_y}{\partial P} = -q \frac{\partial \Delta U}{\partial y} \frac{Dn}{q} \left( \frac{1-\rho}{w+ct} \right) \frac{w}{P}$ , and  $\frac{\partial^2 Q}{\partial \tau \partial P} = -\left( \frac{1}{q} \right) \left[ \frac{(1-\rho) \frac{w}{P}}{(w+ct)^2} \right]$ . Substituting the simplified forms of  $\frac{\partial WMU_y}{\partial P}$  and  $\frac{\partial^2 Q}{\partial \tau \partial P}$  into the numerator of  $\frac{\partial \tau^*}{\partial P}$ :

$$\frac{d\tau}{dP} < 0$$

$$\Leftrightarrow \left( \frac{\partial WMU_y}{\partial P} \right) > q\Delta U \left( \frac{\partial^2 Q}{\partial \tau \partial P} \right)$$

$$\dots$$

$$\tau < \frac{\Delta U}{\frac{\partial \Delta U}{\partial y}}$$

Therefore, if and only if  $\tau < \frac{\Delta U}{\frac{\partial \Delta U}{\partial y}}$ , then  $\frac{d\tau^*}{dP} < 0$ . QED.