

Contributions to Economic Analysis & Policy

Volume 2, Issue 1

2003

Article 9

A Theory of Utilization Review

David Dranove*

Kathryn E. Spier†

*Kellogg School of Management, Northwestern University, d-dranove@kellogg.northwestern.edu

†Kellogg School of Management, Northwestern University and NBER, k-spier@kellogg.northwestern.edu

Copyright ©2003 by the authors. All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording, or otherwise, without the prior written permission of the publisher, bepress. *Contributions to Economic Analysis & Policy* is one of *The B.E. Journals in Economic Analysis & Policy*, produced by The Berkeley Electronic Press (bepress). <http://www.bepress.com/bejeap>.

A Theory of Utilization Review*

David Dranove and Kathryn E. Spier

Abstract

Through utilization review (UR), managed care organizations (MCOs) monitor and alter physician treatment decisions. We show that the value of UR depends on physician incentives. Not surprisingly, when physicians have incentives to significantly overtreat patients, UR can improve social welfare by eliminating unnecessary utilization. More surprisingly, UR can also improve welfare when physicians have incentives to significantly undertreat patients. In this case, UR filters out the least valuable cases, encouraging physicians to recommend more treatments. We also show that the effectiveness of UR depends on MCO precommitment to a treatment approval threshold. Ex ante optimal precommitment can make it appear that the MCO is inappropriately withholding care ex post.

*d-dranove@kellogg.northwestern.edu and k-spier@kellogg.northwestern.edu. We gratefully acknowledge the helpful suggestions we received from Albert Ma, the editors, and two referees.

Introduction

Ever since Arrow's (1963) seminal paper, economists studying health care markets have recognized the centrality of agency relationships between patients and physicians. Patients rely upon their physicians to diagnose their ailments and recommend treatments. However, as Arrow suggests, problems with information and financial incentives may lead physicians to make decisions that are not in the best interest of their patients.¹ Managed care organizations (MCOs) have arisen, in part, to mitigate these conflicts. Through a variety of practices known collectively as utilization review (UR), MCOs use information about the appropriateness of medical treatment to monitor and alter treatment decisions. For example, an MCO may require a physician to obtain authorization prior to a hospital admission or a medical procedure. Based on specific clinical information it obtains from the physician, the MCO may approve or deny the requested treatment. In some cases, it may request additional information.

UR is usually regarded as a mechanism to reduce utilization. It was first introduced at a time when fee-for-service reimbursement was the norm, and physicians had financial incentives to overtreat. In recent years, fee-for-service reimbursement has been replaced by capitation, in which physicians receive a lump sum payment per patient and bear part of the expense for the services they deliver. Although capitation gives physicians financial incentives to undertreat patients, UR remains firmly intact.

Although UR is a part of virtually every MCO plan and has received considerable scrutiny from policy makers and the media, it has received relatively little attention from economists. In particular, there is to our knowledge no detailed theoretical framework for evaluating the merits and drawbacks of UR, or understanding how the terms of UR may evolve along with other changes in the health care marketplace.

This paper begins to fill that void. We model the medical decision making process as one in which the physician and the UR agency possess independent private information about the value of a treatment. The physician makes a treatment recommendation that the UR agency may approve or deny. Both the physician and the UR agency act strategically, and account for each other's private information as best as they can when making decisions.

We show that the social value of UR depends crucially upon the nature of physician incentives. UR is of greatest value when physicians' financial incentives lead them to excessively overtreat *or* undertreat patients relative to the social optimum. At one extreme, UR is of great value when there are strong fee-for-service incentives. As one would expect in this case, UR prevents the provision of costly but low value services. At the other extreme, UR is of great value when capitation leads physicians to aggressively undertreat their patients. In this case, UR encourages physicians to be less conservative in their recommendations; physicians anticipate correctly that the UR agency will use their information to weed out the least valuable cases. The latter finding challenges the conventional wisdom that UR and capitation are substitutes whereby the desirability of one is diminished when the other is introduced.² In this study, we show under some conditions, capitation and UR are actually complements.³

In addition to this finding, we identify strategies that UR agencies should adopt to maximize social welfare, and how physicians are likely to respond. Not surprisingly, we find

¹ Either physicians ignore the impact of their decisions on costs, which affects premiums, or they abuse their private information, such as occurs in inducement models (Dranove, 1988).

² For example of the conventional wisdom, see Wagner and Wagner, (1999).

³ In their study of capitation and copayments, Pauly and Ramsey (1999) have shown that these two cost containment strategies can complement one another, so that their simultaneous use enhances social welfare.

that physicians will be more liberal in their treatment recommendations whenever the UR agency becomes more conservative in approving them. On the other hand, if changes in financial incentives cause physicians to become more conservative, UR agencies may find it optimal to commit to being more conservative in response. We also show that both the effectiveness of UR and the optimal UR strategy depend on whether the UR agency can precommit to its strategy.

These results shed important light on the consequences and desirability of utilization review. In the past three years, many MCOs have weakened UR enforcement, shifting from regulatory control to an informational role. Some analysts interpret this as a sign that MCOs are caving in to public pressure. The interpretation afforded by our model is that the financial incentives that MCOs give to their physicians have led them to adopt recommendation strategies that are closer to the social optimal; hence, the value of restrictive UR may have fallen.⁴ Our model also demonstrates why UR agencies find it optimal to precommit to conservative acceptance rules, even though they appear to be *ex post* irrational. UR agencies may receive a lot of criticism for their tough stance, but this may, in fact, be socially desirable.

While not concerned with UR *per se*, Dranove's (1988) model of supplier-induced demand is related to ours.⁵ In that model, physicians recommend unnecessary treatments to their patients, who consent because they lack adequate information about the benefits. In Dranove (1988), treatment occurs if and only if both the physician and patient approve. In the present model, treatment occurs if and only if the physician and UR agency approve. The present model differs from Dranove (1988) in several ways. In Dranove (1988), the physician and patient receive random draws about the same value; in the present model the physician obtains a draw on the treatment's value, while the UR agency obtains a draw on the probability of success. This alteration allows us to derive more definitive results. In Dranove (1988), the physician is assumed to maximize income under fee-for-service reimbursement; in the present model the physician can have a range of financial incentives. This allows us to evaluate UR under different reimbursement mechanisms. In Dranove (1988), individual patients have formed accurate expectations about physician practice styles. In the present paper, UR agencies form these expectations. Economies of scale in information gathering suggest that the latter assumption is more reasonable.

There is a modest body of empirical research about the effects of UR on the practice of medicine. Wickizer (1992) summarizes a series of older studies he conducted with various colleagues, reporting that "hospital inpatient UR can be effective, reducing hospital admissions by approximately 10-15 percent."⁶ More recent studies generally support this claim. For example, Wickizer (1992) finds that one private insurer's UR program reduces hospital admissions by 12 percent. On the other hand, Wickizer and Lessler (1998) find that a different insurer's UR plan failed to limit hospital admissions, but did reduce lengths of stay, mostly for mental health patients. Robinson et al. (1995) compare utilization of 20 categories of medical procedures among New York City union members and their families undergoing UR and a

⁴ United Healthcare uses a variety of financial incentives to compensate physicians, including capitation. As long as capitation has not caused physicians to become too conservative, these incentives may have limited the value of UR.

⁵ Dranove (1993) sketches a model of UR that introduces the idea that physicians and the UR agency may each have private information. However, physicians and the UR agency neither optimize nor act strategically; nor are there any normative results.

⁶ See Wickizer (1992), p. 104.

comparable group receiving "sham" review (the physicians did not know it, but all requests were automatically reviewed.) They find a nearly 10 percent reduction in utilization from UR.

Pauly and Ramsey's (1999) study of MCO practices is related to ours because they show that seeming substitutes, in their case capitation and copayments, can actually complement one another. They show that capitation combined with copayments can lead to more effective control over moral hazard when consumers have different severities of illness and different price sensitivities within severity class. In our study, the complementarity between UR and capitation results from strategic interactions between the UR agency and physicians, and does not depend on multiple dimensions of consumer willingness to pay.

The next section provides some institutional background on UR. Section 3 presents the basic model and derives some preliminary results. Section 4 presents social welfare comparisons. Further results and extensions are presented in Section 5. Section 6 offers concluding remarks. All proofs are in the appendix.

Background on UR

The basic premise of UR is that there is a wealth of information available to enhance medical decision-making. However, in the absence of UR, physicians may not use this information efficiently. Physicians may find it too costly to read and assess the available information, or they may lack the proper financial incentives to alter their practice patterns in response to the information. By taking advantage of scale economies in information assessment and threatening to withhold payments if physicians fail to follow UR recommendations, MCOs may be able to improve efficiency and even boost quality. As Wolff and Schlesinger (1998) argue, UR can rationalize medical utilization by "reducing the variance of clinical procedures that conflict with professional norms." UR critics counter that it threatens physician decision making autonomy, and that UR agencies often deny treatment requests as a way of reducing costs and boosting MCO profits, at the expense of quality. In addition, UR imposes an administrative burden that might actually drive up the cost of care.⁷

UR is not new. For decades, hospital staffs have reviewed medical records to identify inappropriate medical decisions and take steps to prevent similar mistakes from reoccurring. In the 1970s, many traditional fee-for-service indemnity insurers began requiring patients to obtain second opinions prior to surgery. In 1983, the Health Care Finance Administration established Peer Review Organizations (PROs) in every state, and charged them with reviewing Medicare hospitalizations for appropriateness, effectively making nearly half of all hospitalizations eligible for UR. As MCOs came to dominate the market in the 1990s, UR retained its central role. Virtually all MCOs, including HMOs that pay physicians on a capitated basis, use UR.

Nowadays, UR includes preauthorization of hospitalizations and surgeries, ongoing review of costly inpatient stays, and review of post-discharge placement into nursing homes, home care, or other forms of treatment. Our model pertains specifically to preauthorization review, in which the UR agency has the opportunity to deny payments prior to the treatment being rendered. This is the most controversial element of UR.

While many insurers perform their own UR, many others outsource it to companies such as Interqual. Their UR methodology for hospitalizations and surgeries suggests the ways in which UR agencies use information to rationalize clinical practice. Consider a physician treating a patient whose MCO has contracted with Interqual to perform UR. Prior to

⁷ Fees paid to UR service agencies add about one percent to insurance premia, and compliance costs may be even higher.

hospitalizing or operating on that patient, the physician must provide certain clinical information to Interqual. By examining this information and reviewing its standards of care, Interqual determines whether the requested intervention is necessary, and whether the intervention can be performed in a less costly setting. If Interqual does not approve the initial request, the physician can appeal the decision, perhaps by providing more extensive clinical information. Ultimately, if the physician performs a service that was not approved by Interqual, then the MCO can refuse payment.

The Basic Model

The Essential Elements

We develop a model that captures the essential elements of UR while remaining tractable. These elements are as follows:

- 1) A patient seeks treatment of uncertain value.
- 2) The physician has some private information about the value of the treatment
- 3) An independent UR agency also has private information about the value of treatment.
- 4) This information is noncontractible.
- 5) Although the private information is continuous, as a practical matter, the physician must make a discrete treatment recommendation and the UR agency must make a discrete approve/disallow decision.⁸
- 6) While in practice the physician usually provides some details about the diagnosis, etc., it is impossible for him to precisely reveal his private information. As a result, the discrete treatment recommendation is informative.
- 7) If a recommended treatment is approved, the physician provides the treatment. The physician's resulting financial reward depends on the nature of the reimbursement scheme.

It is plausible to assume that the UR agency is a "Stackelberg" leader in the decision process. That is, the UR agency selects its decision rule prior to physicians making their recommendation decisions. This makes sense because the UR agency interacts with thousands of physicians, using a decision rule that it develops over considerable time and then codifies for implementation. Thus, the UR agency is likely to adopt and stick with approval criteria, which are reasonably well-known by providers. Providers observe these criteria and choose their recommendation strategy accordingly. Once providers select a recommendation strategy, the UR agency may regret its choice of approval criteria; in general, a Stackelberg leader is off of its reaction curve. Later on, we consider the equilibrium when the UR agency and physicians choose their strategies simultaneously.

In addition to modeling the elements of UR, we must also make assumptions about the objectives of the UR agency and physicians. UR policies are chosen by MCOs that compete for the business of employers offering health benefits to their employees. Most metropolitan areas have several MCOs, suggesting that the MCO market is reasonably competitive. It follows that

⁸ There is a large theoretical literature on information transmission between informed agents and uninformed principals who base their decisions on the information received. In the "cheap talk" game of Crawford and Sobel (1982), communication is imperfect and pooling outcomes emerge in equilibrium. Here, we have exogenously assumed the communication structure for the physician is binary: he either recommends treatment or remains silent. Wolinsky (2002) considers a more general environment with multiple experts and models the value of commitment and communication.

to be successful MCOs must meet the objectives of employers. In an afterward to a set of articles on employee benefits, Pauly (2001) finds indirect evidence that when selecting MCOs, employers act as good agents for their employees. Assuming that employees seek to balance costs and quality, we think it is fairly reasonable to assume that UR policies are chosen to maximize social welfare.

Some critics of UR may suggest that market forces have failed to force UR agencies to act as perfect agents, citing anecdotal cases where UR agencies disallowed treatments whose benefits apparently exceeded their costs. As we show, the UR agency may appear *ex post* to be excessively restrictive even when they do act as perfect agents. Thus, we do not take such anecdotes as proof that they are imperfect agents. Even so, we acknowledge that market forces may fail to ensure perfect UR agency. By assuming perfect UR agency, we keep the model tractable and are able to focus on the interactions between UR and physician compensation. At the end of our study, we consider how our results may change if there is imperfect UR agency.

Various models of physician behavior posit that physicians may be altruistic, striking a balance between personal financial goals and the health care benefits of treatment.⁹ We make the same assumption, allowing the relative weighting of financial goals and health care benefits to vary. We assume a very simple compensation structure in which the insurer chooses a single parameter whose level may cause the physician to tend to either overprescribe or underprescribe care. We compute the optimal value of the compensation parameter and the resulting optimal UR program.

We also examine UR in the event that the compensation parameter is not chosen optimally. This is important, because there is considerable evidence in the health economics literature that (a) compensation incentives vary dramatically by insurer; and (b) insurers often choose “extreme” compensation schemes such as pure fee-for-service or pure capitation even though economic models suggest that “in-between” schemes are optimal.¹⁰ This seeming departure from optimality may reflect transactions costs not captured in existing models. In any event, as insurers experiment with their compensation schemes, our model indicates how they should alter their UR programs.

The Model

A patient presents with a condition for which there is one potential course of treatment. This treatment is characterized by two variables: the patient's value of the treatment conditional upon success, v , and the probability of success, p . In practice, the physician and the UR agency may have some information about both v and p .¹¹ However, we capture the presence of private information in a simple way by assuming that only the physician observes information about v and only the UR agency observes information about p . Our main intuitions are robust to more general information assumptions.

The physician privately observes the patient's value of the treatment, v . This valuation is drawn from a density $f(v)$ on the interval $[0, \infty)$. This distribution has mean v_0 and cumulative distribution function $F(v)$. The patient obtains this value, v , if and only if the

⁹ This is fundamental to many models of provider decision making, such as Ellis and McGuire (1996), and Ellis (1998).

¹⁰ See, for example, Ma (1994) and McGuire, T. (2000).

¹¹ Indeed, it would be unusual if the physician did not have some opinion about v .

treatment succeeds. The UR agency privately observes the probability of success, p .¹² This probability is drawn from an independent distribution $g(p)$ on the interval $[0, 1]$ and has mean p_0 and cumulative distribution function $G(p)$. The expected value of treatment to the patient, above and beyond the value of the alternative (no treatment) is pv ; thus, both the physician and UR agency possess information that collectively determine the value of treatment.¹³ The social cost of treatment is denoted by $c \in (0, 1)$, so the overall social value of treatment is $pv - c$.

When making treatment recommendations, the physician maximizes the sum of health benefits and income. The expected health benefit of treatment is $E(pv)$. We model the physician's income as follows. The physician has some income that is independent of the treatment performed. We assume the physician is risk neutral, so that this component of total income drops out of all calculations. There is also a financial transfer between physician and insurer conditional on the treatment being performed. We denote the transfer m from physician to insurer, and normalize this so that $m \in (0, 1)$.¹⁴ This normalization simplifies notation yet captures the essential incentive issues we are addressing. Specifically, when m is near zero, the physician largely ignores medical costs. This corresponds to the incentives for overtreatment associated with fee-for-service medicine. When m is near 1, the physician puts considerable emphasis on costs. This corresponds to capitation and incentives to undertreat.

When we do comparative statics on this parameter, m , we will assume that the social cost of treatment, c , remains unchanged. This is formally justified when either (i) changes in m reflect changes in monetary transfers to the physician, or (ii) the physician's costs are not included in the social welfare function. The parameter m is assumed to be common knowledge.

The timing of the model is as follows. First, the UR agency chooses a threshold, \hat{p} . This threshold represents a commitment to accept the doctor's recommendation for treatment if and only if $p > \hat{p}$. Next, the doctor observes v and either makes a recommendation for treatment or declines to make a recommendation. Following Dranove (1988), we do not permit the physician to directly inform the UR agency of the exact value of v .¹⁵ If the physician recommends treatment then the case goes to the UR agency. The agency then observes the probability of success, p , and approves the doctor's request if and only if p is above their threshold, \hat{p} . Thus, the patient receives treatment if and only if the physician requests it and the UR agency approves. This model may be solved by backwards reasoning.

If the UR agency is a Stackelberg leader, then given the UR agency's commitment to \hat{p} and his own observation of the value of treatment, v , the doctor will recommend treatment if and only if $\int_{\hat{p}}^1 (pv - m)g(p)dp \geq 0$. In other words, he recommends treatment when the expected value of treatment exceeds his cost. This condition implicitly defines a threshold for

¹² Alternatively, one can think of p as an independent measure of the value of treatment, where the actual value equals pv .

¹³ The multiplicative structure lends tractability to the model. The main intuitions are not dependent on this structure, however.

¹⁴ The normalized range of m between $[0, 1]$ is arbitrary and does not drive any results. One may think of m as including a fixed payment for the opportunity cost of the physician's time. If so, this cost washes out of the analysis, with the slight change that the optimal m must include the physician's opportunity cost.

¹⁵ We could extend the model to situations in which physicians can report v within a range. As long as the decision to recommend treatment is itself informative, our main results should remain intact.

the physician which is a function of the UR agency's threshold, \hat{p} , and the physician's cost of treatment, m .

$$\hat{v}(\hat{p}, m) = \frac{m [1 - G(\hat{p})]}{\int_{\hat{p}}^1 pg(p)dp}. \quad (1)$$

The first lemma states that when the physician faces “capitation-like” incentives (m is higher) his standards for treatment are higher and he only takes the sickest patients. This is unsurprising and confirms our interpretation of m as an indicator of payment incentives. The lemma also states that if the UR agency commits to higher standards (\hat{p} higher) then the doctor becomes more liberal, accepting more cases. When the UR agency has higher standards, the doctor is willing to recommend treatment for marginal patients figuring that they will be screened out a later stage if the probability of success is low.

Lemma 1: $\partial \hat{v}(\hat{p}, m) / \partial m > 0$ and $\partial \hat{v}(\hat{p}, m) / \partial \hat{p} < 0$.

Proof: All proofs appear in the Appendix.

Social Welfare Analysis

Given the physician's strategy from equation (1), social welfare may be written:

$$W[\hat{p}, \hat{v}(\hat{p}, m)] = \int_{\hat{p}}^1 \int_{\hat{v}(\hat{p}, m)}^{\infty} (pv - c) f(v) g(p) dv dp. \quad (2)$$

This section explores the properties of this social welfare function and compares welfare under UR to welfare in its absence.

No Utilization Review

Consider first a world without UR. This is equivalent to a system with $\hat{p} = 0$ -- all recommendations are approved. The physician will recommend treatment whenever the expected benefit of treatment, p_0v , exceeds his expected cost, m . This gives a threshold where the physician recommends treatment when:

$$v \geq \hat{v}(0, m) = m / p_0. \quad (3)$$

The social welfare under this system is given by

$$W^{MD}(m) = W[0, \hat{v}(0, m)], \quad (4)$$

where social welfare is defined in (2) above. The next lemma states that social welfare is highest when $m = c$ and declines monotonically as m deviates in either direction from c . Thus, physicians make socially optimal treatment decisions (conditional on $\hat{p} = 0$) whenever they exactly weigh the social cost of treatment against the benefits. Put another way, when m increases, the physician's cutoff, $\hat{v}(0, m)$, increases. This will increase social welfare if the cutoff was initially too low, and will reduce social welfare if the cutoff was initially too high.

Lemma 2: Social welfare in the absence of UR, $W^{MD}(m)$, is increasing in m when the doctor is more liberal than the social planner, $c < m$, is decreasing in m when the doctor is more conservative than the social planner, $c > m$, and is maximized when $m = c$.

Utilization Review

Introducing UR has a profound effect on social welfare. Consider the change in welfare from a small increase in \hat{p} :

$$\frac{dW[\hat{p}, \hat{v}(\hat{p}, m)]}{d\hat{p}} = \left[\frac{\partial W(\hat{p}, \hat{v}(\hat{p}, m))}{\partial \hat{p}} + \frac{\partial W(\hat{p}, \hat{v}(\hat{p}, m))}{\partial \hat{v}} \frac{\partial \hat{v}(\hat{p}, m)}{\partial \hat{p}} \right]. \quad (5)$$

When \hat{p} increases there are two effects. The first term shows the *direct effect*, holding the physician's threshold \hat{v} fixed: raising \hat{p} weeds out cases whose probability of success is marginal. The second term shows the indirect effect: raising \hat{p} makes the physician more liberal in his recommendations, $\partial \hat{v}(\hat{p}, m) / \partial \hat{p} < 0$, by Lemma 1. The next lemma evaluates the direct and indirect effects of UR, starting from the point where $\hat{p} = 0$ (no UR).

Lemma 3: When $\hat{p} = 0$, the change in social welfare from a small increase in \hat{p} is:

$$\frac{dW[0, \hat{v}(0, m)]}{d\hat{p}} = c[1 - F(m/p_0)]g(0) + (m - c)(m/p_0)f(m/p_0)g(0). \quad (6)$$

The *direct effect* is positive and the *indirect effect* is positive if and only if $m > c$.

Starting at $\hat{p} = 0$, the *direct effect* of introducing UR is unambiguously positive because UR screens out the cases that have a zero probability of success. Since treatment has social cost c , it makes sense to deny treatment to these patients. The *indirect effect*, however, may be either positive or negative. When $m > c$ the physician was under-prescribing care to begin with so social welfare increases when the physician recommends more cases. When $m < c$, however, the physician was over-prescribing care, so social welfare falls when the physician recommends more cases. These several and potentially offsetting effects will continue to be present as \hat{p} increases. However, signing the effects is potentially ambiguous because the sign also depends on the densities f and g .

The UR agency chooses its acceptance threshold, $\hat{p} = p^*(m)$, to maximize social welfare taking into account both the direct and indirect effects.

$$p^*(m) = \arg \max_{\hat{p}} W[\hat{p}, \hat{v}(\hat{p}, m)]. \quad (7)$$

Social welfare under UR may be written:

$$W^{UR}(m) = W[p^*(m), \hat{v}(p^*(m))]. \quad (8)$$

Lemma 4: Social welfare under UR, $W^{UR}(m)$ is increasing in m when the doctor is more liberal than the social planner, $c < m$, and is decreasing in m when the doctor is more conservative than the social planner, $c > m$, and is maximized when $m = c$.

This lemma implies that if the UR agency is maximizing social welfare, it is still socially desirable for physicians to exactly weigh the social cost of treatment against the benefits. Moreover, if physicians do not weigh social costs appropriately, the UR agency can not adjust its approval threshold to fully offset the social loss. An optimal precommitment by the UR agency does not fully compensate for the physician's bad incentives.

It is interesting to note that although this proposition was proven under the assumption that the UR agency precommitted to the socially desirable threshold, $p^*(m)$, the result would also be true in if the threshold was a fixed number, not depending upon m at all. The proofs would be virtually identical to those above, with one difference: we would not have to appeal to the envelope theorem to argue that we need not consider the indirect effect of a change in m on $p^*(m)$.

Social Welfare Comparison

Although social welfare declines under both regimes as the physician's incentives deviate from the social optimum, the rate at which welfare declines differs in the two regimes: in general, *welfare falls off faster when there is no utilization review*. This implies that UR has *greater relative value* when the doctor's preferences deviate from those of the social planner. That is, UR is a more valuable instrument when the physician has incentives to greatly over-prescribe or under-prescribe care. The next Proposition identifies sufficient conditions on $f(v)$ so that the relative value of UR increases as the physician's incentives deviate further from the social optimum.

Proposition 1: If $v^2 f(v)$ is increasing in v , then $W^{UR}(m) - W^{MD}(m)$ is strictly decreasing in m when $m < c$ and increasing in m when $m > c$ and is minimized at the point where the physician's incentives are aligned with society's, $m = c$.¹⁶

This result tells us that the nominal value of UR increases as physician incentives grow more distorted. An implication is that UR is more desirable in conjunction with either strict forms of capitation (where m is much larger than c) or fee-for-service (where m is much smaller than c), especially if physicians heavily emphasize these financial incentives relative to the social welfare. UR provides relatively less value when the doctor's incentives are naturally aligned with those of society. Perhaps MCOs have abandoned UR for the simple reason that their compensation systems have more closely aligned physician and societal interests.

The finding that UR has greater value when physicians have strong fee-for-service incentives is probably unsurprising. But the finding that UR also has greater value when physicians are strongly capitated may be unexpected. Although UR may seem to be a substitute for capitation – another way to reduce utilization – it enables enable the capitated physician to achieve the same level of utilization on a more appropriate patient mix. This enhances social welfare.

It is important to note that in this simple model $W^{UR}(m) \geq W^{MD}(m)$ for all m -- UR always (weakly) dominates having no UR. This is true for the simple reason that the UR agency's cutoff, $p^*(m)$, was chosen to maximize social welfare. Since $\hat{p} = 0$ was in the UR agency's choice set to begin with, revealed preference tells us that UR must provide a level of welfare that is at least as high. If UR had social costs associated with it, however, then UR will be preferred if and only if the physician's incentives are sufficiently distorted.

¹⁶ This assumption that $v^2 f(v)$ is increasing in v implies that the density does not decline “too rapidly”. It holds for many densities, such as the uniform density and most of the support of the normal density, and most of the support of the exponential. It is more than sufficient for this proposition. The density matters because any change in incentives causes physicians to change their threshold. The number of patients affected depends, of course, on the density at that threshold.

Proposition 2: Suppose that there is a fixed cost, $K > 0$, of establishing a UR system and that K is sufficiently large. Then there exist parameters \underline{m} and \bar{m} where $0 \leq \underline{m} < c < \bar{m}$ for which UR is the superior system when $m < \underline{m}$ or $m > \bar{m}$, and is inferior when $m \in [\underline{m}, \bar{m}]$.

Although we do not model it formally, it is straightforward to assess how our model changes if there is a marginal cost to the UR agency of performing UR. First, this reduces the value of the UR regime relative to the no-UR regime. Second, this gives the UR agency a further incentive to discourage physicians from recommending borderline treatments. Thus, we would expect the UR threshold to decrease. By committing to approve more cases, doctors recommend fewer treatments.

UR may also impose marginal costs on the physician. In addition to reducing the social value of UR, this makes the physician less inclined to recommend treatment, in a manner analogous to an increase in m . As shown below, the effect of increasing m on the optimal UR threshold is ambiguous. Imposing marginal costs on the physician adds a layer of complexity to the analysis of the optimal threshold, as this cost enters into the social welfare calculation, and is beyond the scope of the present analysis.

The Uniform Distribution

The results of the previous subsections can be shown graphically. Suppose that $f(v)$ and $g(s)$ are both uniformly distributed on the interval $[0,1]$. Given a threshold \hat{p} , the social welfare from UR may be written

$$\int_{\hat{p}}^1 \int_{\hat{v}(\hat{p},m)}^1 (pv - c) dv dp - K,$$

where $\hat{v}(\hat{p},m) = 2m/(1+\hat{p})$ from equation (1). Figure 1 plots social welfare as a function of m and \hat{p} for the case where $c = .25$ and $K = 0$.

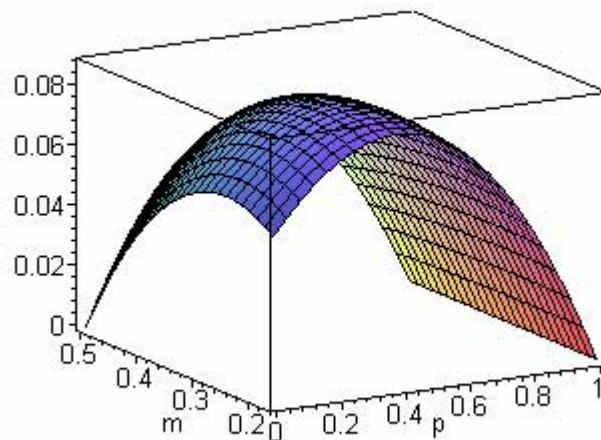


Figure 1:
Social Welfare Under
Utilization Review
 $c = .25$

Social welfare is nicely behaved and strictly concave in both m and \hat{p} . For any threshold, \hat{p} , social welfare is maximized when the doctor's incentives are aligned with those of society as a whole, $m = c = .25$. The figure also depicts the optimal threshold for the UR agency, $p^*(m)$. Properties of this function will be discussed in greater detail in the next section.

In the absence of UR, $\hat{p} = 0$ and social welfare is a function of m only: $W^{MD}(m) = \int_{\hat{v}(0,m)}^1 (pv - c)dv$. This function is plotted along with the more general social welfare function in Figure 2 below for the case where $c = .25$ and $K = .04$. (\hat{p} has been restricted to the $[.3, .6]$ range to simplify the visuals, although the results of course do not depend upon this.) The steeper of the two surfaces represents the case of no UR, while the flatter one represents a UR agency. Figure 2 illustrates the main result from the last section that UR is more valuable when the physician's incentives deviate from those of the social planner.

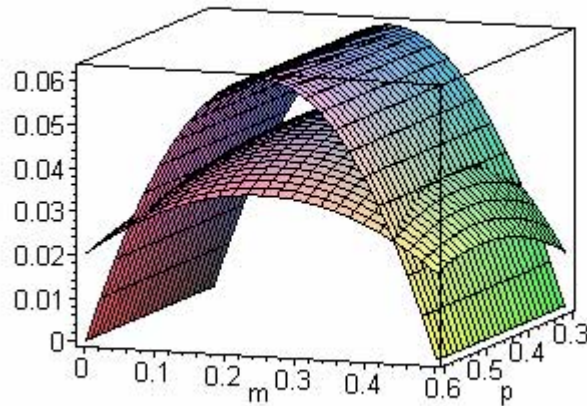


Figure 2:
Social Welfare Comparison
 $c = .25, K = .04$

Further Results and Extensions

The UR Agency's Optimal Threshold

We have identified the general set of circumstances under which UR will be desirable. In this section, we examine how the UR agency's strategy depends upon the physician's incentives. Specifically, we determine how the cutoff, $p^*(m)$, changes when m changes. Although one might intuitively expect that the cutoff will decrease as m decreases, (i.e., the UR agency will grow more conservative when financial incentives encourage physicians to be more liberal), this does not always turn out to be the case.

To determine how the UR agency responds to changes in physician incentives, we refer back to the implicit definition of p^* :

$$\frac{dW[p^*, \hat{v}(p^*, m)]}{dp} = 0.$$

Totally differentiating this expression with respect to p and m :

$$\frac{d^2W[p^*, \hat{v}(p^*, m)]}{dp^2} dp + \left[\frac{\partial}{\partial \hat{v}} \left(\frac{dW[p^*, \hat{v}(p^*, m)]}{dp} \right) \right] \left(\frac{\partial \hat{v}(p^*, m)}{\partial m} \right) dm$$

We may assume that the second (total) derivative of the social welfare function with respect to p is negative¹⁷, so the first term is negative. As for the second term, we have already established that $\partial \hat{v}(\hat{p}, m) / \partial m > 0$. The remaining term of indeterminate sign is

$$\frac{\partial}{\partial \hat{v}} \left(\frac{dW[p^*, \hat{v}(p^*, m)]}{dp} \right) = \frac{\partial}{\partial \hat{v}} \left(\frac{\partial W[p^*, \hat{v}(p^*, m)]}{\partial p} + \frac{\partial W[p^*, \hat{v}(p^*, m)]}{\partial \hat{v}} \frac{\partial \hat{v}(p^*, m)}{\partial p} \right)$$

The first term can be written:

$$\frac{\partial^2 W[p^*, \hat{v}(p^*, m)]}{\partial v \partial p} = [\hat{v}(p^*, m) p^* - c] f(v^*) g(p^*).$$

It can be shown that this is negative for m close to c . As for the second term, $\partial \hat{v}(\hat{p}, m) / \partial \hat{p} < 0$ from an earlier lemma, and $\partial^2 W(\bullet) / \partial \hat{v}^2$ should be negative under reasonable assumptions. So we have a negative first term and a positive second term. The overall sign is ambiguous.

The uniform distribution allows us to more precisely explore how the UR agency's optimal threshold varies with m , the physician's incentive parameter. Differentiating $W(\hat{p}, \hat{v}(\hat{p}, m))$ with respect to \hat{p} and rearranging terms gives us a first-order condition which defines p^* as an implicit function of m and c :

$$(1 + p^*)^2 \left(c - \frac{p^*}{2} \right) + 2m^2 - 4cm = 0.$$

Totally differentiating with respect to p^* and m we have

$$\frac{dp^*(m)}{dm} = \frac{8(m-c)}{(1+p^*)(1+3p^*-4c)}.$$

So long as c is not too large ($c < 1/4$ is a sufficient but not necessary condition) we have $dp^*(m)/dm > 0$ if and only if $m > c$. In other words, the utilization review agency becomes tougher when the physician deviates from the socially optimal preferences, regardless of the direction of this deviation. Figure 3 below shows the UR agency's optimal threshold $p^*(m)$ when $c = .25$.

One expects the UR agency to adopt a tough approval criterion when the physician overprescribes care due to fee-for-service compensation ($m \ll .25$), and this is confirmed in the figure. But the UR agency is also tough when the physician tends to underprescribe care due to capitation ($m \gg .25$). To understand why, recall that the physician becomes more liberal when the UR agency grows more conservative. By toughening its approval criterion, the UR agency encourages the otherwise conservative capitated physician to recommend more treatments. The result is that more patients receive high value treatments.

This finding points to a potential problem with empirical research about the impact of capitation on physician decision making. These studies, which generally compare levels of utilization among physicians receiving different forms of compensation, tend to ignore the restrictiveness of any associated UR. But our analysis shows that utilization depends on both capitation and UR, and that the restrictiveness of the latter may be correlated with that of the

¹⁷ We have shown that social welfare is increasing when $m > c$ and decreasing when $m < c$. So long as the function is continuously differentiable it will be concave when $m = c$.

former. If so, then researchers who fail to control for UR can easily misstate the effects of capitation.

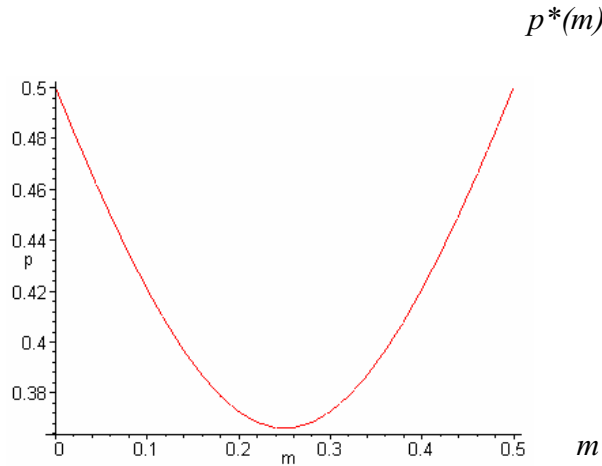


Figure 3:
The Utilization Review
Agency's Optimal Threshold
 $c = .25$

What if the Utilization Review Agency Cannot Precommit?

The result that the UR agency's threshold is increasing in the distance between m and c depends critically upon the ability of the UR agency to precommit to its cutoff $p^*(m)$. Suppose that the UR agency enjoys no such first mover advantage. Instead, it seeks to maximize social welfare *ex post* given its beliefs about the physician's decision rule. We can formally represent this new timing as a game where the UR agency and the physician choose their thresholds simultaneously and non-cooperatively. The thresholds, \hat{p} and \hat{v} , would be the solutions to the system of two simultaneous equations:

$$\hat{v}(\hat{p}, m) = \frac{m [1 - G(\hat{p})]}{\int_{\hat{p}}^1 pg(p)dp} \text{ and } \hat{p}(\hat{v}, c) = \frac{c [1 - F(\hat{v})]}{\int_{\hat{v}}^{\infty} vf(v)dv}.$$

The first of these equations is simply equation (1) from before, and the second is the analogous equation for the UR agency. One can easily establish analogous comparative statics for $\hat{p}(\hat{v}, c)$:

Lemma 5: $\partial \hat{p}(\hat{v}, c) / \partial c > 0$ and $\partial \hat{p}(\hat{v}, c) / \partial \hat{v} < 0$.

When the social cost of treatment increases, the UR agency chooses a more conservative threshold and denies treatment to more patients. If the UR agency believes that the physician's threshold, \hat{v} , is higher, they will be more liberal and approve more cases.

When f and g are uniformly distributed on $[0,1]$, the Nash equilibrium is given by the solution to the system of two simultaneous equations: $\hat{v}(\hat{p}, m) = 2m / (1 + \hat{p})$ and $\hat{p}(\hat{v}, c) = 2c / (1 + \hat{v})$. These two downward sloping reaction curves are shown in Figure 4 for the case when $c = .25$ and $m = .4$.

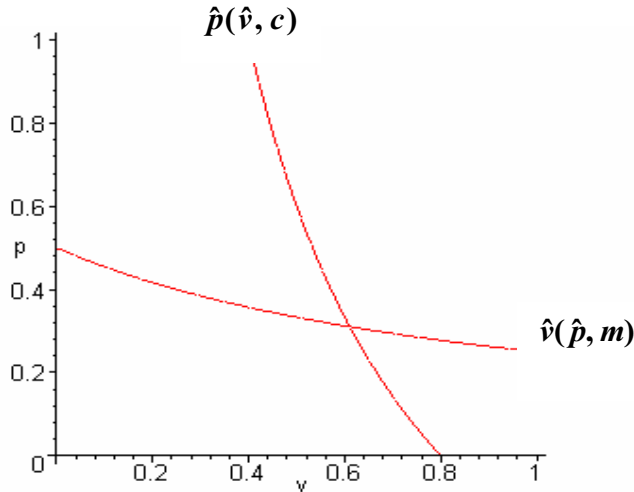


Figure 4:
Equilibrium of the
Simultaneous Move Game
 $c = .25, m = .4$

When m increase, the $\hat{v}(\hat{p}, m)$ locus shifts to the right and the equilibrium UR agency threshold, \tilde{p} , falls. When m is higher the physician becomes more conservative in their recommendations, and the review agency responds by becoming more liberal and approving of more cases. In other words, the UR agency threshold that emerges from this system of equations, $\tilde{p}(m)$, is a decreasing function of m . Figure 5 compares the threshold from this simultaneous move game to the precommitment game considered in the last subsection.

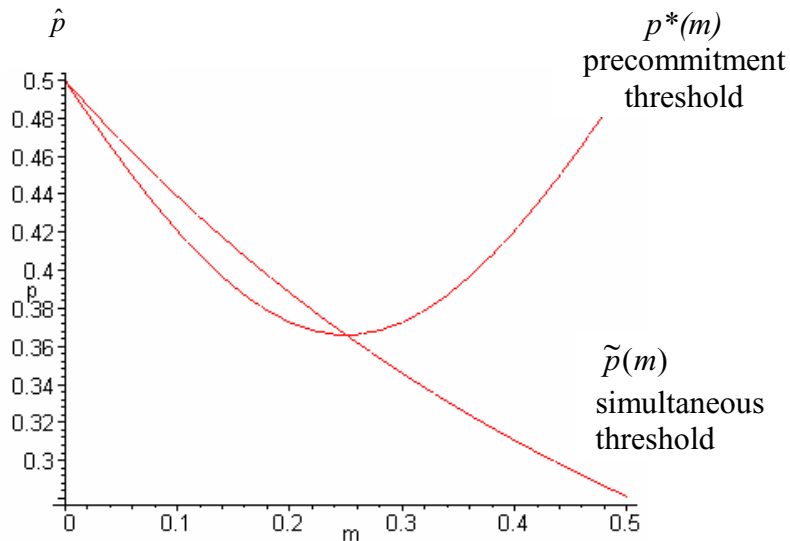


Figure 5:
Comparing UR Agency
Thresholds
 $c = .25$

Note that if physician incentives tend to lead to underprescription of treatment ($m > c$), then the UR agency's threshold in the precommitment game is bigger than it is in the simultaneous game. When it can precommit, the UR agency will appear *ex post* to have adopted an excessively tough approval threshold – it rejects some treatments that have a positive net expected surplus. The UR agency precommits to this seemingly overaggressive stance because of the positive indirect effect of encouraging capitated physicians to be more liberal.

Although the thresholds under the simultaneous and sequential timings are very different, the basic result concerning the desirability of UR is not. Figure 6 compares the social welfare functions under the assumptions the $c = .25$ and $K = .04$. Both social welfare functions are maximized when $m = c$, but UR gives rise to a flatter function. As before, UR has greater relative value the more distorted are the physician's incentives.

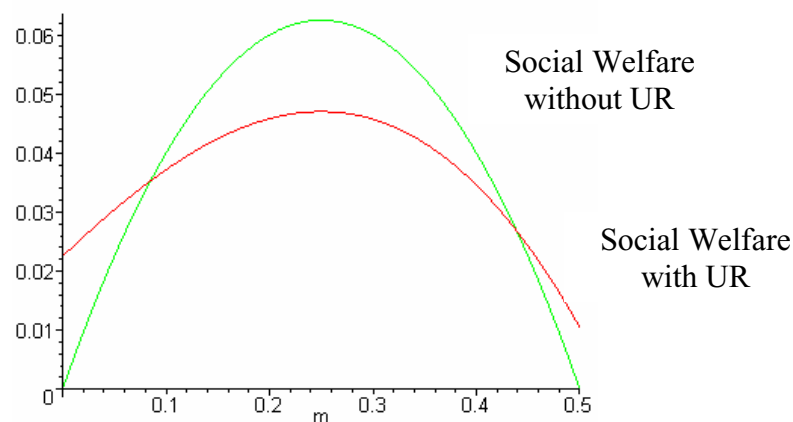


Figure 6:
Social Welfare Comparison
 $c = .25, K = .04$

What if the Utilization Review Agency Doesn't Maximize Social Welfare?

It is a legitimate question how our results would change if the UR agency did not represent the views of the social planner more generally. Perhaps they, like the physician, place different weights on the costs and benefits of care. Their incentives may make them too tough, thereby committing the agency to a threshold above $p^*(m)$. Alternatively, their incentives may be too soft, committing to a threshold below $p^*(m)$. This would not change the nature of our results, however. Going back to Figure 2, we see that UR review yields higher welfare in the extremes for a broad range of thresholds, not just at the optimal threshold, $p^*(m)$. Therefore UR is desirable when the physician's incentives are distorted *even when the utilization review agency's preferences are not aligned with those of the social planner*.

Conclusions

Managed care organizations use a variety of tools, including UR and provider financial incentives, to enhance the efficiency of health care delivery. Our study finds that the value of UR depends critically on the chosen financial incentives. Moreover, the relationship between the two is somewhat surprising. The value of UR increases as financial incentives cause

physicians to depart from the social optimum in either direction -- towards providing too little care or too much care. As MCOs increasingly use capitation to put providers at financial risk, they may find it desirable to carefully adjust the intensity of UR. If MCOs can get provider incentives "right", they may even find it optimal to abandon UR. Our findings also suggest that there is a U-shaped relationship between the restrictiveness of UR and physician decisions. As physician compensation moves from strong fee-for-service incentives ($m \ll c$) to strong capitation ($m \gg c$), the UR agency first decreases and then increases its restrictiveness.

By focusing on how UR agencies should respond to changes in physician payment rules, we have made several strong assumptions and have ignored many other important issues in health care markets. First, we do not attempt to solve for the optimal physician compensation rule. Formally, the physician's preferences over the medical benefits of treatment were captured by an *exogenous* parameter. Researchers such as Ma (1994) show that neither fee-for-service nor capitation is generally optimal, and that superior compensation mechanisms can be rather complex to design and implement. We expect that a more general model could endogenize these incentives, but at the cost of tractability. We believe that MCOs routinely struggle to find the appropriate financial incentives and that UR will continue to persist amidst a range of compensation systems, and therefore the results obtained here are very important in reality.

Second, we assumed that physicians could not assess the probability of treatment success and that UR agencies could not assess the value of successful treatment. This assumption is not intended as a literal description of the medical decision making process. However, it does capture the essential feature of UR that our model addresses, namely, the private information possessed by the physician and UR agency and the resulting conflict between the two. As we show, such conflict may arise even when UR and physician incentives are aligned, but will worsen as the gap in incentives grows.

Appendix

Proof of Lemma 1: Differentiating $\hat{v}(\hat{p}, m)$ in equation (1) with respect to \hat{p} and m :

$$\frac{\partial \hat{v}(\hat{p}, m)}{\partial m} = \frac{[1 - G(\hat{p})]}{\int_{\hat{p}}^1 pg(p)dp} > 0, \text{ and}$$

$$\frac{\partial \hat{v}(\hat{p}, m)}{\partial \hat{p}} = -mg(\hat{p}) \frac{\int_{\hat{p}}^1 (p - \hat{p})g(p)dp}{\left(\int_{\hat{p}}^1 pg(p)dp\right)^2} < 0. \quad (\text{A1})$$

Q.E.D.

Proof of Lemma 2: From the definition of $W(\hat{p}, \hat{v}(\hat{p}, m))$ in (2) and p_0 ,

$$W^{MD}(m) = \int_0^1 \int_{\hat{v}(0, m)}^1 (pv - c)f(v)g(p)dvdp = \int_{\hat{v}(0, m)}^1 (p_0v - c)f(v)dv.$$

Differentiating this expression with respect to m and rearranging terms gives:

$$\frac{dW^{MD}(m)}{dm} = -\frac{\partial \hat{v}(0, m)}{\partial m} [p_0 \hat{v}(0, m) - c] f(\hat{v}(0, m)).$$

Using the fact that $\hat{v}(0, m) = m / p_0$ from (3) above

$$\frac{dW^{MD}(m)}{dm} = \frac{1}{p_0} f[\hat{v}(0, m)](c - m), \quad (\text{A2})$$

and the statement of the lemma follows.

Q.E.D.

Proof of Lemma 3: First consider the direct effect. Taking the partial derivative of welfare with respect to \hat{p} in (2),

$$\partial W(\hat{p}, \hat{v}(\hat{p}, m)) / \partial \hat{p} = -g(\hat{p}) \int_{\hat{v}(\hat{p}, m)}^{\infty} (\hat{p}v - c)f(v)dv.$$

Using equation (3) above, $\hat{v}(0, m) = m / p_0$, so evaluating this direct effect at $\hat{p} = 0$ gives us the positive direct effect:

$$\partial W(0, \hat{v}(0, m)) / \partial \hat{p} = c[1 - F(m / p_0)]g(0) \geq 0. \quad (\text{A3})$$

This is positive because UR weeds out very low probability cases. The indirect effect has two parts. First,

$$\partial W(\hat{p}, \hat{v}(\hat{p}, m)) / \partial \hat{v} = -f(\hat{v}(\hat{p}, m)) \int_{\hat{p}}^1 (p\hat{v}(\hat{p}, m) - c)g(p)dp.$$

Evaluating this expression at $\hat{p} = 0$ and using the fact again that $\hat{v}(0, m) = m / p_0$ gives us

$$\partial W(0, \hat{v}(0, m)) / \partial \hat{v} = -f(m / p_0)(m - c). \quad (\text{A4})$$

The second part of the indirect effect is $\partial \hat{v}(\hat{p}, m) / \partial \hat{p}$ which is given in equation (A1) above.

Evaluating this expression at $\hat{p} = 0$,

$$\partial \hat{v}(0, m) / \partial \hat{p} = -g(0)(m / p_0). \quad (\text{A5})$$

Taken together, the indirect effect depends on whether $m > c$. Substituting (A3), (A4) and (A5) into (5) gives expression (6) in the lemma.

Q.E.D.

Proof of Lemma 4: Since $p^*(m)$ was chosen to maximize social welfare, the envelope theorem tells us that we need only consider the direct effect of a change in m :

$$\frac{dW^{UR}(m)}{dm} = \frac{\partial W[p^*(m), \hat{v}(p^*(m), m)]}{\partial \hat{v}} \frac{\partial \hat{v}(p^*(m), m)}{\partial m},$$

which is equal to

$$\frac{dW^{UR}(m)}{dm} = - \left(\int_{p^*(m)}^1 [p\hat{v}(p^*(m), m) - c] f[\hat{v}(p^*, m)] g(p) dp \right) \left(\frac{1 - G(p^*(m))}{\int_{p^*(m)}^1 pg(p) dp} \right).$$

Using the definition of $\hat{v}(p^*(m), m)$ from (1) and rearranging terms gives us:

$$\frac{dW^{UR}(m)}{dm} = \frac{\int_{p^*(m)}^1 pg(p) dp}{m^2} [\hat{v}(p^*(m), m)]^2 f[\hat{v}(p^*(m), m)] [c - m]. \quad (\text{A6})$$

The lemma follows from this expression.

Q.E.D.

Proof of Proposition 1: Rewriting (A2) above using the fact that of $\hat{v}(0, m) = m/p_0$ gives

$$\frac{dW^{MD}(m)}{dm} = \frac{p_0}{m^2} [\hat{v}(0, m)]^2 f[\hat{v}(0, m)] (c - m).$$

Combining this with expression (A6) gives:

$$\frac{\partial [W^{UR}(m) - W^{MD}(m)]}{\partial m} = \frac{(c - m)}{m^2} \left\{ \left(\int_{p^*(m)}^1 pg(p) dp \right) v_{UR}^2 f(v_{UR}) - p_0 v_{MD}^2 f(v_{MD}) \right\},$$

where $v_{UR} = \hat{v}(p^*(m), m)$ is shorthand for the doctor's threshold under UR, and $v_{MD} = \hat{v}(0, m)$ is shorthand for the doctor's threshold with no UR. We will now argue that the term in brackets is negative. Recall from our earlier lemma that $\partial \hat{v}(\hat{p}, m) / \partial \hat{p} < 0$, and so $\hat{v}(p^*(m), m) \leq \hat{v}(0, m)$ and we have $v_{UR} \leq v_{MD}$. Now we will make use of the assumption that $v^2 f(v)$ is increasing in v to get

$$v_{UR}^2 f(v_{UR}) - v_{MD}^2 f(v_{MD}) \leq 0.$$

Since $\int_{p^*(m)}^1 pg(p) dp \leq \int_0^1 pg(p) dp = p_0$, we have

$$\left(\int_{p^*(m)}^1 pg(p) dp \right) v_{UR}^2 f(v_{UR}) - p_0 v_{MD}^2 f(v_{MD}) \leq 0$$

and we are done.

Q.E.D.

References

- Arrow, K. (1963) "Uncertainty and the Welfare Economics of Medical Care" *American Economic Review* 53:941-973.
- Crawford, V. and J. Sobel (1982), "Strategic Information Transmission," *Econometrica* 50(6): 1431-1451.
- Dranove, D. (1993) "The Five W's of Utilization Review" in Helms, R. ed. *American Health Policy* Washington: AEI Press.
- _____ (1988) "Demand Inducement and the Physician-Patient Relationship" *Economic Inquiry* 26:281-298.
- Ellis, R. (1998) "Creaming, Skimping, and Dumping: Provider Competition on the Intensive and Extensive Margins" *Journal of Health Economics* 17(5): 537-555.
- Ellis, R. and T. McGuire (1996) "Hospital Responses to Prospective Payment: Moral Hazard, Selection, and Practice Style Effects: *Journal of Health Economics* 15(3): 257-278.
- Glied, S. (2000) "Managed Care" in Culyer, A. and J. Newhouse, eds. *Handbook of Health Economics* Amsterdam: North-Holland.
- Ma, A. (1994) "Health Care Payment Systems: Cost and Quality Incentives" *Journal of Economics and Management Strategy* 3(1):93-112.
- McGuire, T. (2000) "Physician Agency" in Culyer, A. and J. Newhouse, eds., *Handbook of Health Economics* Amsterdam: North-Holland
- Pauly, M., 2001, "Making Sense of a Complex System: Empirical Studies of Employment-based Health Insurance" *International Journal of Health Care Finance and Economics* 1(3/4): 333-339
- Pauly, M. and S. Ramsey (1999) "Would You Like Suspenders to Go with that Belt? An Analysis of Optimal Combinations of Cost Sharing and Managed Care" *Journal of Health Economics* 18(4): 443-458.
- Rosenberg, S. et al. (1995) "Effect of Utilization Review in a Fee-for-service Health Insurance Plan" *New England Journal of Medicine* 333(20): 1326-1330
- Wagner, T. and L. Wagner (1999) "Who Gets Second Opinions?" *Health Affairs* 18(3):137-145
- Wickizer, T., (1992) "The Effects of Utilization Review on Hospital use and Expenditures: A Covariate Analysis" *Health Services Research* 27(1): 103-119
- _____ and D. Lessler (1998) "Effects of Utilization Management on Patterns of Hospital Care Among Privately Insured Adult Patients" *Medical Care* 36(11): 1545-1554
- Wolinsky, A. (2002), "Eliciting Information from Multiple Experts," *Games and Economic Behavior* 41(1): 141-160.
- Wolff, N. and M. Schlesinger 1998 "Risk, Motives, and Styles of Utilization Review: A Cross-condition Comparison" *Social Science and Medicine* 47(7): 911-26.