Physician Practice and MCO Negotiation:

The impact of time sensitive supply and demand

Daniel Ludwinski*

September 11, 2014

Abstract

Anecdotal evidence seems to indicate that modifying payer mix is a major mechanism that practices employ to maximize profits in the face of differing insurer reimbursements, limited capacity and stochastic domain. The practice/insurer interactions included in the current body of literature do not yet address this behavior. In this paper I develop a model that captures practice profit maximization under this regime and illustrate how it explains some recently validated empirical facts.

* Cornell University. Please send all questions, comments or corrections to DL679@Cornell.edu

1 Introduction

Many markets feature stochastic and time sensitive consumer demand along with supplier capacity that is static in the short term and non-storable. A prominent example includes the market for live performances such as concerts or sporting events. A common feature of these types of markets is the use of auctions as clearing mechanisms. The price is allowed to vary with contemporaneous demand. More popular events or times have higher equilibrium prices. Without price flexibility the result is typically excess demand (sell outs) or excess capacity (empty seats).

Though rarely applied to this context, the market for physician services also features stochastic and time sensitive consumer demand along with static, non-storable provider capacity. However, the market for physician services has both supply and demand side factors that do not allow a similar demand clearing mechanism. Prices for physician services are quite rigid. Medicare, the largest insurance provider in the United States, sets prices nationally. These prices are non-negotiable. Similarly, Medicaid prices are generally set by states and are also a take it or leave it proposition. Reimbursement rates between physicians and private insurers are set through a complex and opaque process of bilateral negotiations.

Furthermore, consumer demand for physician services is not particularly responsive to price. There are three main reasons that drive this. First, a substantial portion of the cost of care is covered by insurance. In fact, often patients are only required to pay a flat co-pay. This leads to a disconnect between the price charged and end price paid by the consumer. Second, the demand for physician services is generally fairly inelastic. For non-preventative care there are often no good substitutes available. Finally, even for patients that might be particularly cost sensitive, prices are often unknown and not easily discoverable. More complex mechanisms than price alone are needed to clear the market for physician services.

There is a body of literature on provider market power and MCO provider bargaining. However, this literature currently does not address the above mentioned features which are the mechanism through which a provider can leverage market power to receive higher prices. Empirical work has been done to examine how one payer's price impacts the bargained price for another payer, for example changes in Medicare's prices impacting private prices. But most theoretical models assume independent bargaining outcomes and price does not explicitly depend on the market structure of the MCOs.

The goal of this paper is to add to the existing literature by examining these features of the market for physician services – stochastic, time sensitive consumer demand and static, non-transferable supplier capacity in the face of rigid price structures and inelastic consumer demand – play out in terms of the bargaining relationship between multiple managed care organizations and physicians. This paper proceeds as follows:

First, I give more background and motivation to justify and support the development of my approach. I show where I am building on the relevant literature, and contrast my approach with was has been done previously. Second, I derive from profit maximizing behavior a function of the average excess capacity, given stochastic patient demand over a time period. I then use this to develop a model of the physician's decision to accept or reject a Managed Care Organization, given the expected price from that MCO. Third, I incorporate this model of physician behavior into a simple bargaining model and present some of the model's predictions. Finally, I examine several methods to empirically test the predictions of this model.

2 Background & Related Literature

An important assumption made in this paper is that in the short and medium term physician and practice supply is relatively fixed. For practices, the intuition is that the main production inputs of space, equipment, and support staff cannot be easily varied day to day or week to week. For individual physicians, the idea is that their services are labor intensive. Physician labor responds to a price increase with competing income and substitution effects. While this assumption can be relaxed, the main formation of the model assumes that the effects cancel out and there is no aggregate supply response to price.

This assumption is not contradicted by the current literature. In an important early work looking at physician behavior McGuire and Pauly (1991) provide a theoretical model to test whether physicians have a target income or seek to maximize profits. They found that the strength of physicians' income effect controls their behavior. More recently Kantarevic, Kralj and Weinkauf (2008) used reforms to the physician threshold system in Ontario, Canada to study this empirically. They find that, as expected, both the income effect and substitution effects are present with the expected signs. However, for different services, different effects dominate and there is no predominant aggregate supply effect.

There have been empirical studies (Gaynor and Gertler 1995, McGuire 2000) that find supply is impacted by supply side variables such as opportunity cost and incentives price. Ketcham, Nicholson, Unur and Casalino (2014) find that a one-percent increase in Medicare payments for a service results in a 0.15 percent increase in the supply of that service to Medicare patients. However, to my knowledge most of these are looking at the supply of specific services (by HCPCS code) and/or the supply to specific a specific payor (Medicare, Medicaid, private) and do not examine the aggregate supply of a physician's time.

The interplay between a practice and multiple payers, including Medicare, is an important mechanism in the model. A branch of the literature has sought to explain the response of private prices to changes in Medicare prices. Hospital administrators have advocated for "cost-shift theory", that is, lower prices from one insurer will need to be made up somewhere to meet cost, and will then be shifted to other insurers. In a 2011 review of the literature, Frakt finds some evidence that cost shifting may occur, however the effects

seem to be mild. In a more recent White (2013) finds the opposite effect – lower Medicare rates in hospitals resulted in lower private rates.

For physicians, Clemens and Gottleib (2013) found consistent positive effects on private payer rates from increases in Medicare payments. These effects are larger both when Medicare makes up a larger share of the market and also when insurers have more relative market power. Ketcham, Nicholson, Unur and Lawrence (2014) similarly finds a positive relationship.

There is large existing literature covering MCO bargaining with providers for inclusion in a network. Town and Vistnes (2001) and Capps, Dranove and Satterthwaite (2003) use a logit demand model to construct a patient's willingness-to-pay for inclusion of a provider based on observed provider and patient characteristics. These papers firmly established the WTP concept as a measure of market power as well as the connection between that measure, profits, and prices. While originally focused on hospitals, these models have recently been applied to physicians as well (Carlson et al 2013). These papers, however, only employ a simple, reduced-form bargaining model and as such cannot speak to the impact on prices stemming from different configurations of MCO market power.

More recent research has incorporated more sophisticated bargaining models. Ho and Lee (2013) studies the price impact of insurer consolidation, focused on two competing forces. Increased insurer competition lowers premiums. Lower premiums reduce the surplus available to split between hospital and insurers, resulting in reduced prices. However, increased insurer competition give hospitals more leverage to raise prices. They specify a general bargaining model in which price is determined by insurers' premiums and payments to other hospitals, and hospitals' costs and reimbursements from other payers. Lewis and Plum (2014) also develop a hospital, MCO bargaining model. Their innovation is to separately look at bargaining position (value of the hospital or network) and bargaining position (ability to obtain a higher share of the surplus).

I add to these bargaining models by making the value of a MCO to a provider more explicit. By introducing capacity constraints I am able to model two important provider-side considerations: the risk capacity will be unused, and the risk that a low paying patient will displace a higher paying patient. Neither of these two effects have been previously captured in the bargaining literature, which typically has featured fixed marginal costs.

3 Model of Practice MCO Negotiation

Below I develop a model of practice-MCO bargaining. I explicitly specify the benefit of contracting for both the MCO and the practice.

First, I specify how providers choose capacity given expectations about patient demand, the expected marginal cost and expected payment for patient (not conditioned patient type). Adjusting this capacity is costly and fixed in the short and medium term. This leads to an optimal average excess capacity, or the propensity that a given time slot is unfilled (λ_0).

Second, I specify the value to a practice of accepting patients of a particular type (taking prices as given). While this can be generalized to include any patient types that can be observable and discriminated, the focus here is on patients from different MCOs. Every MCO k has a price (p_k) and a propensity (λ_l) – which can be thought of as the probability that a patient of type k takes a given time slot, given the provider accepts patients from all MCOs.

This expected value of including plan type k depends on the prices of other accepted MCOs, their propensities, and propensity that a given time slot is unfilled (λ_0). This gives the value of the MCO to the provider.

The value of the provider to the MCO is captured by the change in willingness-to-pay for a patient to have the provider in the network – as developed by Capps et al (2003). For use in bargaining, this is converted from utils to dollars and standardized to WTP per time slot to be comparable to the value of the MCO to the provider.

The MCO and provider reach a deal to include the provider in the MCO network if there is a price between the lowest price the provider would accept, the expected value of a timeslot without the provider, and the highest price the MCO would pay, which is the willingness-to-pay. If they do reach an agreement they choose a price that splits the gains from inclusion by a constant fraction.

Unlike previous work, the explicit specification of the providers value function allows me to solve the system of equations and derive a formula for equilibrium prices that are determined simultaneously, depend on the both the provider and MCO competitive landscape.

A practice of a given size is opened with expectations about the expected value of a patient time slot, expected average demand, and the direct marginal cost of seeing a patient.

Capacity

In this section I derive the average excess capacity (λ_0) which comes about through three avenues: the profit maximizing behavior of the physician/practice, the uncertainty about how many patients will arrive in a given period of time, and the fixed cost associated with capacity.

First, let the physician have a belief about the expected value of a given unit of capacity (EV_c) which is expected net price (expected price minus expected variable cost). Second, denote capacity by S (size)

and let the cost for every unit of capacity be fixed at c_S. Finally, let the number of patients in a given time period be approximated by a Poisson distribution with mean and variance x.

The physician then chooses capacity S to maximize the following profit function:

$$\Pi = -c_S S + EV_p \sum_{i=0}^{S} i \frac{e^x x^i}{i!} + EV_p \left(\sum_{i=S+1}^{\infty} S \frac{e^x x^i}{i!} \right)$$

And the change in expected profit from an extra unit of capacity is:

$$\frac{\Delta \Pi}{\Delta C} = -c_S + EV_p \left[S \left(\frac{e^x x^S}{S!} - \frac{e^x x^{S+1}}{(S+1)!} \right) + \sum_{i=S+2}^{\infty} \frac{e^x x^i}{i!} \right]$$

Therefore, the rule to maximize profit is to add a unit of capacity if (subject to positive overall profit):

$$c_S < EV_p \left[S \left(\frac{e^x x^S}{S!} - \frac{e^x x^{S+1}}{(S+1)!} \right) + \sum_{i=S+2}^{\infty} \frac{e^x x^i}{i!} \right]$$

This provides the optimal level of capacity as a non-linear function of the unconditional average number of patients in a time period (x) and the ratio between the cost of an extra unit of capacity and the expected value of a patient and also allows the average excess capacity (λ_0) to be calculated.

$$S^* = S\left(x, \frac{c_s}{EV_p}\right)$$

$$\lambda_0 \left(x, \frac{c_s}{EV_p}\right) = \frac{1}{S^*} \sum_{i=0}^{S^*} (S^* - i) \frac{e^x x^i}{i!}$$

Optimal capacity is increasing in the unconditional mean (x) and decreasing in the cost/expected value ratio (holding constant EV higher cost of capacity will lead to less capacity). Lambda_0 is decreasing both in the unconditional mean and capacity, and increasing in the capacity cost to expected value ratio. To give an example, with a fixed cost to expected value ratio of 10% a provider facing a patient distribution with an unconditional mean of 20 will have a capacity of 30 and an average excess capacity of 33.4% percent, while a provider facing a patient distribution with an unconditional mean of 455 will have a capacity of 500 and an average excess capacity of only 9.4% percent.

An important note on the definition of a time period, as it relates to capacity: In this context, a time period should be thought of as a period in which once a patient realizes their health state, they can be flexible. The time period in question, therefore, will differ by type of service, and type of patient. The relevant time period for a cardiac intensive care unit may have a time period of 30 minutes, while the correct

time period for primary care office may be a week. Furthermore, there may be other objectives than profit maximization in play. Especially in critical care situations, with few good substitutes, the cost of excess demand may include severe negative health outcomes.

Provider's Selection of MCOs

This model's goal is to explain the choice of some practices to accept only certain types of insurance. The agent is the physician practice. An important way that physicians differ from hospitals is that physician offices have more severe capacity constraints. For new patients especially, the availability of a convenient time slot not too far in the future can be a major determinate of choosing a doctor.

Anecdotal evidence indicates that physicians take this into account when deciding whether to accept patients from a low paying insurer. For new managed care contracts, the Practice Management Resource Group encourages practices to evaluate "How the added patients will impact your payer-mix. Will these patients increase or decrease your expected collections? Will they displace higher paying patients?"[2] Similarly, a popular book "Mastering Patient Flow"[3] discourages closing practices fully to new patients due to the fact that it will decrease the practice's ability to alter the payer mix. The alternative suggested to alleviate capacity issues is to end participation with insurance companies that pay less.

In this model, the physician practice (indexed by j) faces K types of patients which it can either choose to accept or not accept – while this can be generalized to include any patient types that can be observable and discriminated, the focus for this exposition will be on patients from different MCOs.

Each slot is then filled with a patient of type k with a probability ($Prob_{k,j}$). Also, with positive probability, the time slot is not filled (denoted by $Prob_{0,j}$). I make the assumption that, conditional on acceptance, the practice cannot discriminate between patient types. So, the practice's problem is to choose which patient types to accept, conditional on the prices - which can generalized to mean expected payment minus any variable or administrative cost.

Note that price should be thought of not as the list of transacted price for that patient, but full net expected payment taking into considerations the cost of working with that type of patient or insurance company.

Alternatively, the MCO selection problem could be formulated in a format similar to the capacity problem above. That is, for every time period a number of patients from each accepted patient type demand the practices services. And each type is characterized by an independent Poisson distribution with varying

[2] http://www.medicalpmrg.com/payor-mix-analysis.html (last accessed April 17, 2014)
[3] Woodcock, Elizabeth W. *Mastering Patient Flow* (MGMA, 2009) 3rd edition

means. The appendix includes a sketch of this method, but I have made simplifying assumptions for tractability, wherein a slot only includes one patient and the type probabilities are independent across time.

It is relatively straight forward to add in the physician's choice of labor supply. While a simple formulation is included in the appendix, this addition adds further complications without impacting the insights. The utility maximizing problem including labor supply results in the same selection of patient types: the set that maximizes the expected value of a timeslot:

$$\max_{K_j} EV_{K_J} = \max_{K_j} \sum_{k \in K_j} Prob_{k,j} p_k$$

Probability of type k:

If $Prob_{k,j}$ is exogenous to the choice of K_j (no capacity constraints), then all plans will be included, as in this formation cost is ignored, but in reality the probabilities are not.

A patient type with a low expected value (a poor paying MCO) can take the capacity away from a patient type with a higher expected value (good paying MCO). Furthermore, if there were no chance that a slot was not filled (excess capacity) then there would be no reason to accept any plan except for the highest paying. The tradeoff then is balancing the probability that no one takes the slot, with the probability that a patient with a lower paying plan keeps a patient with a higher paying plan from coming.

This tradeoff can formalized by denoting the unconditional probability of patient type k (the probability if all types are included) by λ_k. Let λ_0 be the unconditional probability that there are no patients in that time period. Then for a set of plans K_j the probability of patient type k is:

$$Prob_{k,j} = \frac{\lambda_k}{\lambda_0 + \sum_{\kappa \in j} \lambda_\kappa}$$

Expected Value of a Time Slot

Therefore, the expected value of a time slot can be expressed as follows:

$$EV_{K_J} = \sum_{k \in K_j} \lambda_\kappa p_k \Big/ \left[\lambda_0 + \sum_{k \in K_j} \lambda_\kappa \right] \tag{1.0}$$

Maximizing this leads to the rule that patients of type δ should be included iff:

$$p_\delta > \left[\sum_{k \in K_j} \lambda_\kappa p_k \right] \Big/ \left[\lambda_0 + \sum_{k \in K_j} \lambda_\kappa \right] = EV_{K_J}$$

It is notable that the decision to include a particular type of patient does not depend on how many patients there are of that type (propensity). All that matters is the comparison between the expected value of the patient compared to the expected value of the set of currently accepted patients.

With the provider's problem now solved, we can examine some of the dynamics predicted by the set-up of the model.

Addition of a Provider δ (holding hours constant):

Using this formulation, the increase in the expected value of a time slot from provider i adding insurer, given other accepted insurers K and prices is then:

$$V_i(\delta|K_j,P) = \left[\lambda_\delta p_\delta + \sum_{k\in K_j}\lambda_\kappa p_k\right] / \left[\lambda_0 + \lambda_\delta + \sum_{k\in K_j}\lambda_\kappa\right] - \left[\sum_{k\in K_j}\lambda_\kappa p_k\right] / \left[\lambda_0 + \sum_{k\in K_j}\lambda_\kappa\right]$$

$$= \frac{\lambda_\delta}{\lambda_0 + \sum_{k\in K_j}\lambda_\kappa}\left(\lambda_0(p_\delta - 0) + \sum_{k\in K_j}(p_\delta - p_k)\lambda_\kappa\right) / \left(\lambda_0 + \lambda_\delta + \sum_{k\in K_j}\lambda_\kappa\right) \qquad (2.0)$$

The weighted price difference between δ and the existing prices, normalized to a time slot, and multiplied by the percent increase in lambda that δ brings.

MCO's Willingness-to-Pay for a Provider

In order to estimate the patients in an MCOs willingness-to-pay to have access to a particular provider I leverage the framework developed by Capps et al. (2003). A patient i has ex post (that is, after the revelation of a health diagnosis requiring treatment) expected utility for the services from provider j given by the following form:

$$U_{ij} = \alpha R_j + H_j'\Gamma X_i + \tau_1 T_{ij} + \tau_2 T_{ij} X_i + \tau_3 T_{ij} R_j - \gamma(X_i)P_j(Z_i) + \varepsilon_{ij}$$

$$= U\left(H_j, X_i, T_{ij}\right) - \gamma(X_i)P_j(Z_i) + \varepsilon_{ij}$$

Where H_j are the provider characteristics, X_i are the patient characteristic and T_ij is the geographical location of the patient in relation to the provider. If the error term is logit, and we assume there are no meaningful out of pocket cost differentials between providers, then a patients utility of having access to a network G of providers is:

$$V^{IU}\left(G, Y_i, Z_i, T_{ij}\right) = E\max_{g\in G}\left[U\left(H_g, Y_i, Z_i, T_{ij}\right) + \varepsilon_{ig}\right] = ln\left[\sum_{g\in G}\exp U\left(H_g, Y_i, Z_i, T_{ig}\right)\right]$$

And the additional utility derived from the inclusion of provider j is:

$$\Delta V_j^{IU}(G, Y_i, Z_i, \lambda_i) = ln\left(\frac{1}{1 - s_j(H_j, Y_i, Z_i, T_{ij})}\right)$$

This is the willingness to pay, in utils, for patient i to have provider j in network G. The willingness for the MCO to pay to have the provider to in the system is calculated by summing this additional utility over all of patients in the MCO. In order to be used for my purposes, and compared to price, this WTP is then normalized as WTP per visit, and converted to dollars.

It is important to note that even if we assume that patient preferences do not differ systematically across MCOs – that is preferences only differ through the observed characteristics included in the utility function – the willingness-to-pay measures for a given provider can be different. Two main things drive this difference - the MCOs network and the composition of patients.

Both ΔWTP and λ_0 (average excess capacity) reflect a provider's desirability, but it is important to recognize how they are different in this model. The important difference is that in this formation ΔWTP is normalized to a patient time slot, to correspond to price, and therefore does not depend on the size of the population. In contrasts λ_0 depends on the interplay between the number of patients, the number of other practices, and the size of the practice. If the number of patients increased (with no change in characteristics), ΔWTP normalized to a patient time slot would not change but λ_0 would decrease.

Provider-MCO Bargaining

Above, I explicitly specified the value of a contract between insurer j and provider j for both the provider and the MCO. A contract between MCO i and provider j will happen if the insurer's ΔWTP_{ij} is greater than the value of the provider to the insurer.

I now apply a simple bilateral bargaining framework, in which the parties choose a price that splits the bargaining surplus (normalized to a per time period amount) with constant parameter $\alpha \in (0,1)$. Recent work by Lewis and PFlum (2014) has suggested the bargaining power and the share of the surplus captured by the two parties may differ systematically across providers. I, however, restrict the bargaining parameter to be constant at least for a given provider. This leads to the following price equation:

$$p_{ij} = \alpha \mathrm{WTP_{ij}} + (1-\alpha) EV_{K_j/i}$$

$$= \alpha \Delta \mathrm{WTP_{ij}} + (1-\alpha)\left[\sum_{k \in K_j/i} \lambda_\kappa p_k\right] / \left[\lambda_0 + \sum_{k \in K_j/i} \lambda_\kappa\right] \qquad (3.0)$$

In contrast to previous bargaining models which assume independence, through $EV_{K_j/i}$ these prices are interdependent and determined simultaneously. This set up can be explicitly solved if the $\lambda' s$ are taken

as given – for each provider j we have I equations with I unknowns (where I is the total number of MCOs). The explicit solution for price follows for some configurations of insurers.

Monopolist

If insurer δ is a monopolist then $EV_{K_j/\delta}$ is 0, and the equilibrium price equation is:

$$p_\delta = \alpha \Delta \text{WTP}_{\delta j} + (1-\alpha) EV_{K_j/\delta} = \alpha \Delta \text{WTP}_{\delta \text{j}}$$

This is effectively the lowest price between insurer δ and provider j.

Two Private MCOs and Medicare

Consider the situation with two private insurers (indexed with 1 and 2), and Medicare (indexed by m). Assume that Medicare prices are exogenous. This leads to the following equilibrium prices:

$$p_{1j}^* = \alpha \left[1 - \alpha^2 \left(\frac{\lambda_1 \lambda_2}{(\Lambda - \lambda_2)(\Lambda - \lambda_1)}\right)\right]^{-1} \left[\Delta WTP_{1j} + \alpha \left(\frac{\lambda_2}{\lambda_0 + \lambda_2 + \lambda_m}\right) \Delta WTP_{2j} + \left(\frac{\lambda_m}{\lambda_0 + \lambda_2 + \lambda_m}\right)\left(1 + \alpha \frac{\lambda_2}{\lambda_0 + \lambda_1 + \lambda_m}\right) p_m \right]$$

This characterizes prices as a function of the provider competitive landscape, through the willingness to pay measure, and the insurer competitive landscape.

The coefficient on ΔWTP_1 is the amount of an increase in WTP that is reflected in the change in price.

Coefficient on p_m:

$$\left(\frac{\lambda_2}{\lambda_0 + \lambda_2 + \lambda_m}\right) \alpha \left(\frac{\lambda_m}{\lambda_0 + \lambda_1 + \lambda_m}\right) + \left(\frac{\lambda_m}{\lambda_0 + \lambda_2 + \lambda_m}\right) = \left(\frac{\lambda_m}{\lambda_0 + \lambda_2 + \lambda_m}\right)\left[1 + \alpha \left(\frac{\lambda_2}{\lambda_0 + \lambda_1 + \lambda_m}\right)\right]$$

Note, the case without Medicare is the same as if $\lambda_m = 0$.

4 Numerical Examples and Predictions

Below I flesh out the theoretical model by providing some examples as to how these unstudied dynamics play out. I gave examples of how the price between an insurer and a provider is impacted by relative size, and λ_0 (average excess capacity), willingness-to-pay and the willingness-to-pay of other MCOs. These examples provide a first test as to whether this model is valid.

Relationship between MCO Size and Price

Current hospital bargaining literature does not predict a consistent difference in prices arising solely from the size of the MCO in terms of number of patients (not the MCO network). In the above model, if two insurers have the same WTP then the ratio of prices is:

$$p_1^*/p_2^* = \left[1+\frac{1}{2}\left(\frac{\lambda_2}{1-\lambda_1}\right)\right] / \left[1+\frac{1}{2}\left(\frac{\lambda_1}{1-\lambda_2}\right)\right] = \frac{2\lambda_0\lambda_0 + 3\lambda_2\lambda_1 + \lambda_0(3\lambda_2+2\lambda_1)}{2\lambda_0\lambda_0 + 3\lambda_1\lambda_2 + \lambda_0(3\lambda_1+2\lambda_2)}$$

If $\lambda_1 > \lambda_2$ then the denominator is smaller, and insurers 1 pays less. The mechanism is that the expected value to the provider without insurer 1 is smaller than the expected value without insurer 2. To see how this plays out numerically here are few hypothetical examples for $\Delta WTP_1 = \Delta WTP_2 = 1$ and $\lambda_1 = 0.9$, $\lambda_2 = 0.1$:

Lambda_0	33.50%	9.40%
Price High	0.557	0.823
Price Low	0.680	0.713
Size mark up	22%	15%
Average MCO surplus	43.1%	18.8%

The markups are considerable, and depend on and λ_0.

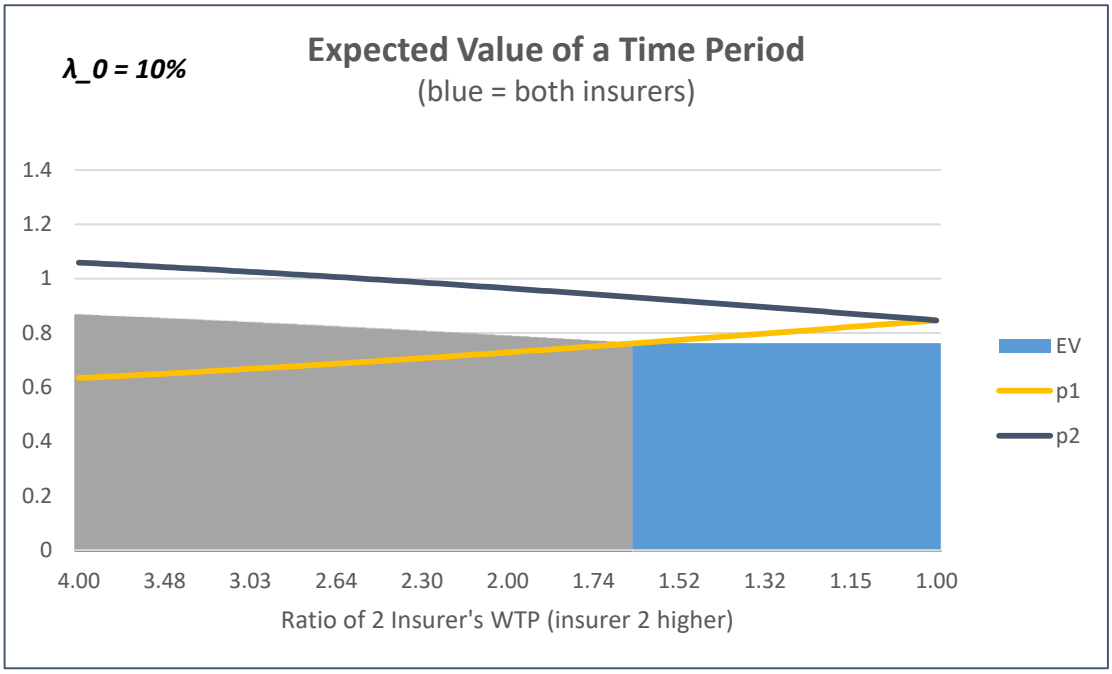
Demand Increase – Excess Capacity and Price

Here is an example of how an increase demand, through the Poisson parameter of average number of visits, works its way through the system resulting in higher prices – even while ignoring any impacts from the increase in willingness-to-pay. If average demand is currently 40 and the cost/EV ratio is 0.25 then optimal capacity=54 and lambda_0 = 18.9%. If there is a 10% increase in average demand (to x=44) and no corresponding change to capacity (in the medium term) then lambda_0 drops to 12.0%. Also, the average number of patients actually seen in a time period changes 37.5 to 38.9 (increases 4%).

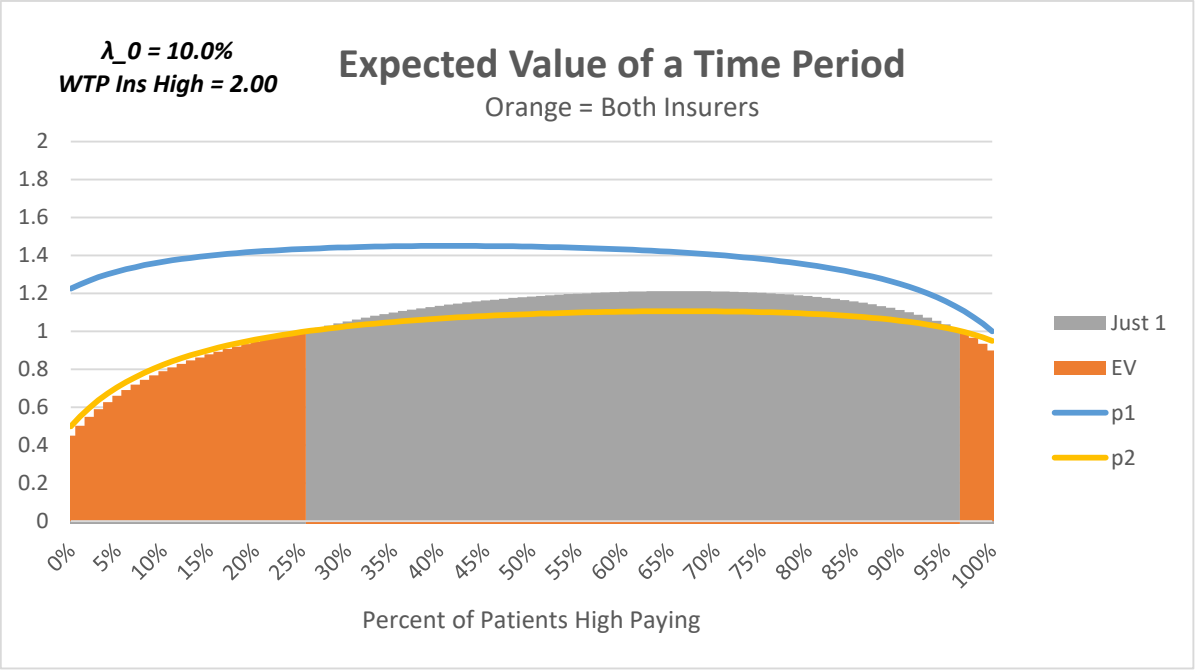
Using the model of prices with two insurers (no Medicare), setting WTP = 1 and $\lambda_1 = \lambda_2$ prices increase from 0.759 to 0.824 (8.5%). With a larger and smaller insurer ($\lambda_1 = 0.9$, $\lambda_2 = 0.1$) the price increase substantially for the larger insurer - from 0.612 to 0.667 (9.1%) for the large insurer, 6.3% for the small insurer (from 0.743 to 0.790), and the weighted average increased 8.7% (from 0.625 to 0.679).

According to my model a 10% increase in underlying demand, without any increase in WTP per visit, increases visits by 4%, prices by around 8.5% and profits by 12.8%.

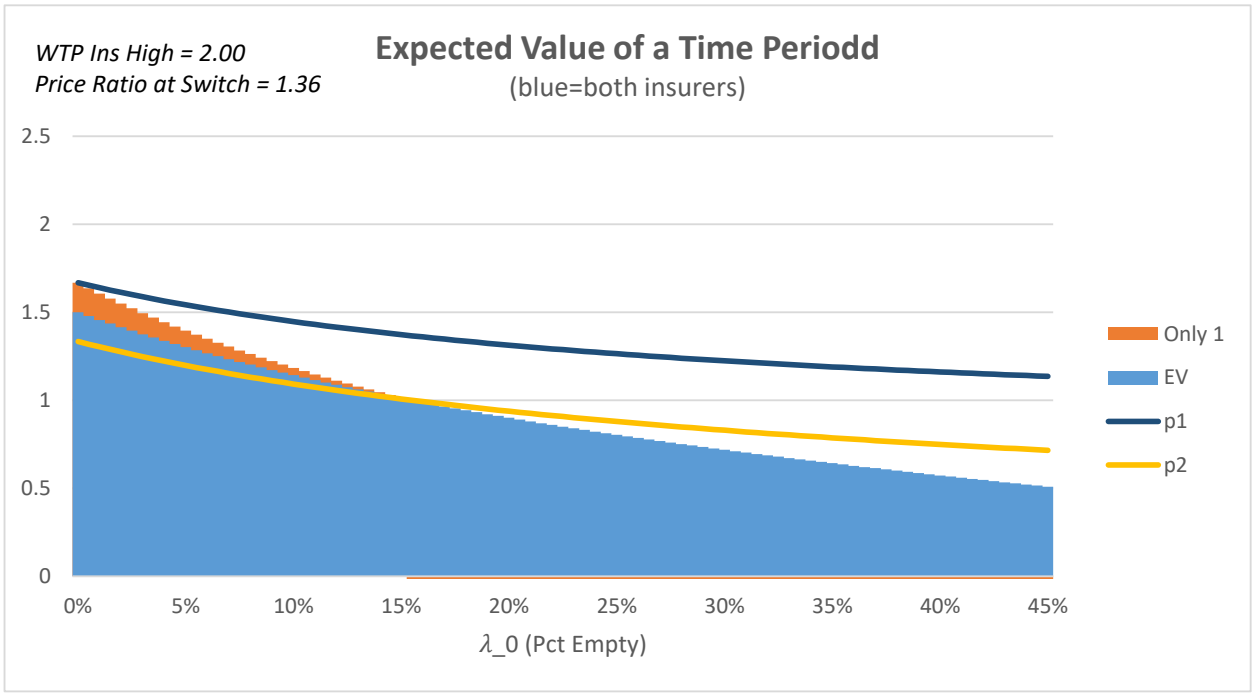
The following charts provide a visualization of the price dynamics with two insurers. I show the price for each MCO, the expected value of a time slot, along one of three dimension: the ratio of the insurers willingness-to-pay, the ratio of the size of the insurers, or and λ_0 (average excess capacity).



In the above graphic, the WTP ratios are varied in such a way to keep the average WTP constant at 1. The patient populations of both MCOs are held constant and equal. The expected value to the provider decreases as the WTP ratio head to one. The provider should accept patients from both MCOs unless the WTP ratio is higher than 1.62.



In figure two the two insurers have different WTP. The WTP for insurer that values the provider less is 1 and the WTP is 2 for the other insurer. λ_0 is held constant at 10%. The interesting thing here is that at the two extremes the provider should accept both patients, but in the middle the provider should only accept the higher paying patient types.



In figure three, both the size of the MCO patient population and the MCOs WTP are held constant. What varies is λ_0. For low value of λ_0 the provider should only accept patients from the high paying MCO.

5 Empirical Application

Above, I proposed a model that incorporates stochastic, time sensitive consumer demand and static, non-transferable supplier capacity into an MCO bargaining model. I believe these features are an important mechanism through which a provider can leverage market power to receive higher prices. There are a variety of ways that this can be empirically tested with data, the two main varieties use cross section vs panel data and employ different assumptions.

In the first stage I would calculate the willingness to pay measures for each MCO/provider pair. As noted above, WTP will vary by MCO due to variation in networks and patient composition. This willingness to pay metric will still be in utils, and as such cannot be directly used in the pricing equations above.

If I assume the preferences are similar across MCO patient populations, then the MCO lambdas can be assumed to be fixed across providers at an appropriate geographical level. The lambdas will then correspond to either the number of enrollees or number of relevant visits in that geographical area. Lambda 0 (the amount of extra capacity) will vary by provider.

Data on transacted prices at the MCO/provider level could be obtained through a state's all payer data set (such as New Hampshire or Massachusetts). With these assumptions and data I could run a non-linear model to estimate α, γ (conversion from utils to dollars) and each provider's λ_0. The accuracy of this model, and the reasonableness of the estimates would validate or reject the propositions proposed above.

6 Conclusion

In this paper I propose a model for capturing the fact that providers have a limited ability to service patients, and patient demand is time sensitive and variable. I show how these dynamics can lead to higher prices for MCOs with more patients – even when willingness-to-pay per visit is held constant. Furthermore, I give a mechanism by which changes in one MCOs willingness-to-pay result in an increase in the equilibrium price for all payers. These mechanisms have not previously been incorporated fully in to MCO provider bargaining frameworks and may be of particular importance for physician practices.

References

Brunt, Christopher S., and Gail A. Jensen. "Medicare payment generosity and access to care." *Journal of Regulatory Economics* 44, no. 2 (2013): 215-236.

Capps, Cory, David Dranove, and Mark Satterthwaite. "Competition and market power in option demand markets." *RAND Journal of Economics* (2003): 737-763.

Clemens, Jeffrey, and Joshua Gottlieb, 2013, "Bargaining in the Shadow of a Giant: Medicare's Influence on Private Payment Systems," *NBER working paper 19503*

Cutler, D. M., Ilckman, R. S., & Landrum, M. B. (2004). The Role of Information in Medical Markets: An Analysis of Publicly Reported Outcomes in Cardiac Surgery. The American Economic Review, 94(2), 342-346.

Dunn, Abe, and Adam Hale Shapiro. "Do Physicians Possess Market Power?."*Journal of Law and Economics* 57, no. 1 (2014): 159-193.

Frakt, Austin B. "How much do hospitals cost shift? A review of the evidence." Milbank Quarterly 89, no. 1 (2011): 90-130.

Gaynor, Martin, Kate Ho, and Robert Town. *The Industrial Organization of Health Care Markets.* No. w19800. National Bureau of Economic Research, 2014.

Ho, Katherine. "Insurer-Provider Networks in the Medical Care Market." *American Economic Review* 99, no. 1 (2009): 393-430.

Ho, Kate, and Robin S. Lee. "Insurer competition and negotiated hospital prices." Working Paper No. w19401. *National Bureau of Economic Research*, (2013).

Ketcham, Jonathan, Sean Nicholson, A. Sinan Unur, Lawrence Casalino. "Relative Prices, Payer Mix and Regional Variations in Medical Care". *Working Paper* (2014)

Lewis, Matthew, and Kevin Pflum. "Diagnosing hospital system bargaining power in managed care networks." Forthcoming *American Economic Journal: Economic Policy* (2014)

Li, Jinhu, Jeremiah Hurley, Philip DeCicca, and Gioia Buckley. "Physician Response To Pay-For-Performance: Evidence From A Natural Experiment." *Health Economics* (2013).

McGuire, Thomas G. "Physician Agency" Handbook of Health Economics, vol 1. Ed A. J. Culyer and J. P. Newhouse. Elsevier. 2000. pp461-536.

McGuire, Thomas G., and Mark V. Pauly. "Physician response to fee changes with multiple payers." *Journal of Health Economics* 10, no. 4 (1991): 385-410.

Woodcock, Elizabeth W. Mastering Patient Flow (MGMA, 2009) 3rd edition

Appendix 1: Alternative Provider Problem

A more complex set up of the provider's problem incorporates the size decision directly. In this formulation the number of patients of each type that demand the practices services is given by an independent Poisson distribution. Each patient type has a separate mean and separate expected value. The provider than chooses the size of the practice and the patients to accept to maximize the expected value. This expected value with two patient types is given by:

$$EV(S|P,K)\sum_{i=1}^{S}P_1(i)\left(\sum_{j=0}^{S-i}P_2(j)(i*p_1+j*p_2)\right)+\sum_{i=0}^{\infty}P_1(i)\sum_{j>S-i}^{\infty}P_2(j)*S*\frac{i*p_1+j*p_2}{i+j}$$

While this is likely a more accurate representation of the provider's problem, this formulation is not tractable. While the above presentation of the problem is a simplification presented above, it does capture the desired tradeoffs faced in provider's decision.

Appendix 2: Including Variable Costs

If variable cost are included then the expected value of a time slot is:

$$EV_{K_J}=\frac{\sum_{k\in K_j}\lambda_\kappa(p_k-c_j)}{\lambda_0+\sum_{k\in K_j}\lambda_\kappa}=\frac{\sum_{k\in K_j}\lambda_\kappa p_k}{\lambda_0+\sum_{k\in K_j}\lambda_\kappa}-c_j\frac{\sum_{k\in K_j}\lambda_\kappa}{\lambda_0+\sum_{k\in K_j}\lambda_\kappa}$$

And the change in expected value of time slot from including patients of type δ is:

$$\left(\frac{\lambda_\delta(p_k-c_j)+\sum_{k\in K_j}\lambda_\kappa(p_k-c_j)}{\lambda_\delta+\lambda_0+\sum_{k\in K_j}\lambda_\kappa}\right)-\left(\frac{\sum_{k\in K_j}\lambda_\kappa(p_k-c_j)}{\lambda_0+\sum_{k\in K_j}\lambda_\kappa}\right)$$

This is very similar to the above formation. The inclusion rule is very similar, it only now includes costs explicitly. This does change the amount of total surplus that the MCO and provider negotiate over, and therefore can change the predictions about the price. The provider's negotiation mechanism based on expected value without the MCO, however, remains mostly unchanged.

Appendix 3: Physician's Problem Including Time & Leisure

$$u_j(c,q)=\log\left(\sum_{k=1}^{K}p_kq_k\right)-\alpha_j\log\left(X-\sum_{k=1}^{K}q_k\right)$$

$$q=\sum_{k=1}^{K}q_k$$

$$\max_{K_j,q} E\left[u_j(q,K_j)\right] = \max_{K_j,q}\left(-\alpha_j \log(X-q) + \sum_{k\in K_j} q_k * p_k * Prob_k\right)$$

Note that price should be thought of not as the list of transacted price for that patient, but full net expected payment taking into considerations the cost of working with that type of patient or insurance company.

For simplicity let $q_k = 1, \forall k$

FOC q:

$$\frac{\partial EU}{\partial q} = \frac{\alpha_j}{X-q} + \sum_{k\in K_j} p_k * Prob_k = 0$$

$$q^* = X - \frac{\alpha_j}{\sum_{k\in K_j} p_k * Prob_k}$$

Then using this to calculate the expected utility of accepting the set of patients K_j:

$$E\left[u_j\left(q * |K_j\right)\right] = \alpha_j \log\left(X - (X - \frac{\alpha_j}{\sum_{k\in K_j} p_k * Prob_k})\right) + (X - \frac{\alpha_j}{\sum_{k\in K_j} p_k * Prob_k}) \sum_{k\in K_j} p_k * Prob_k$$

$$= \alpha_j \log\left(\frac{\alpha_j}{\sum_{k\in K_j} p_k * Prob_k}\right) + X \sum_{k\in K_j} p_k * Prob_k - \alpha_j$$

$$= X \sum_{k\in K_j} p_k * Prob_k - \alpha_j \left[1 - \log(\alpha_j) + \log\left(\sum_{k\in K_j} p_k * Prob_k\right)\right]$$

Appendix 4: Partial derivatives for Physician problem

For the following derivatives prices are held constant.

Hours Worked, wrt p_l:

$$q^* = X - \alpha_j \frac{\lambda_0 + \sum_{k\in K_j} \lambda_\kappa}{\sum_{k\in K_j} \lambda_\kappa p_k} = X - \alpha_j \frac{\lambda_0 + \sum_{k\in K_j} \lambda_\kappa}{\lambda_l p_l + \sum_{k\in K_{j/l}} \lambda_\kappa p_k}$$

$$\frac{\partial q^*}{\partial p_l} = -\alpha_j \frac{\lambda_0 + \sum_{k\in K_j} \lambda_\kappa}{\left(\sum_{k\in K_j} \lambda_\kappa p_k\right)^2} = -\frac{\alpha_j}{\sum_{k\in K_j} \lambda_\kappa p_k}\left(\frac{1}{EV(K_j)}\right) < 0$$

Note that an increase in λ_l can be interpreted as an increase in demand by patients of type l.

Hours Worked, wrt λ_l:

$$q^* = X - \alpha_j \frac{\lambda_0 + \sum_{k\in K_j} \lambda_\kappa}{\sum_{k\in K_j} \lambda_\kappa p_k} = X - \alpha_j \frac{\lambda_l}{\lambda_l p_l + \sum_{k\in K_{j/l}} \lambda_\kappa p_k} - \alpha_j \frac{\lambda_0 + \sum_{k\in K_{j/l}} \lambda_\kappa}{\lambda_l p_l + \sum_{k\in K_{j/l}} \lambda_\kappa p_k}$$

$$\frac{\partial q^*}{\partial \lambda_l} = -\alpha_j \left[\frac{\sum_{k\in K_{j/l}} \lambda_\kappa p_k}{\left(\sum_{k\in K_j} \lambda_\kappa p_k\right)^2} - \frac{p_l \left(\sum_{k\in K_{j/l}} \lambda_\kappa\right)}{\left(\sum_{k\in K_j} \lambda_\kappa p_k\right)^2} \right]$$

$$= -\alpha_j \left[\frac{\sum_{k\in K_j} \lambda_\kappa (p_k - p_l)}{\left(\sum_{k\in K_j} \lambda_\kappa p_k\right)^2} \right]$$

$$= -\alpha_j \left[\frac{\sum_{k\in K_j} \lambda_\kappa p_k}{\left(\sum_{k\in K_j} \lambda_\kappa p_k\right)^2} - \frac{\sum_{k\in K_j} \lambda_\kappa p_l}{\left(\sum_{k\in K_j} \lambda_\kappa p_k\right)^2} \right]$$

$$= -\frac{\alpha_j}{\sum_{k\in K_j} \lambda_\kappa p_k} \left[1 - p_l \frac{\sum_{k\in K_j} \lambda_\kappa}{\sum_{k\in K_j} \lambda_\kappa p_k} \right]$$

$$= -\frac{\alpha_j}{\sum_{k\in K_j} \lambda_\kappa p_k} \left[1 - p_l * 1/EV(K_j) \right]$$

$$= \frac{\alpha_j}{\sum_{k\in K_j} \lambda_\kappa p_k} \left[\frac{p_l}{EV_{K_j}} - 1 \right]$$

So, if p_l is higher than the expected value of the set then hours worked increases. Else, it decreases.

Since all included p_l's must be higher than the expected value (assuming ability to discriminate on types), then for all l ne 0, work will increase.

$$\frac{\partial q^*}{\partial \lambda_l} > 0$$

If we're talking about λ_0, than p is 0 so because a_j is greater than 0 the derivative is negative (less work). So in this simple model, doctors work more in respect to positive demand shocks, and less in response to negative demand shocks (as expected). Substitution effect dominates.

Patients Seen (wrt λ_l):

Patients seen =

$$q^*(1 - prob_0) = q^* \left(1 - \frac{\lambda_0}{\lambda_0 + \sum_{\kappa\in j} \lambda_\kappa} \right) = q^* \left(\frac{\sum_{\kappa\in j} \lambda_\kappa}{\lambda_0 + \sum_{\kappa\in j} \lambda_\kappa} \right)$$

q* rises (number of slots), and patients per slot (fill rate) rises drops as well, so patients seen rises.

Expected Value (wrt λ_l):

$$\frac{\partial EV_{K_J}}{\partial \lambda_l} = \frac{p_l\left[\lambda_0 + \sum_{k\in K_j/l}\lambda_\kappa\right]}{\left[\lambda_0 + \sum_{k\in K_j}\lambda_\kappa\right]^2} - \frac{\left[\lambda_0 + \sum_{k\in K_j/l}\lambda_\kappa p_k\right]}{\left[\lambda_0 + \sum_{k\in K_j}\lambda_\kappa\right]^2} = \frac{\lambda_0 p_l + \sum_{k\in K_j/l}\lambda_\kappa p_l - \lambda_0 - \sum_{k\in K_j/l}\lambda_\kappa p_k}{\left[\lambda_0 + \sum_{k\in K_j}\lambda_\kappa\right]^2}$$

$$= \frac{\sum_{k\in K_j}\lambda_\kappa\,(p_l - p_k)}{\left[\lambda_0 + \sum_{k\in K_j}\lambda_\kappa\right]^2} + \frac{\lambda_0(p_l - 1)}{\left[\lambda_0 + \sum_{k\in K_j}\lambda_\kappa\right]^2}$$

$$= \frac{1}{\sum_{k\in K_j}\lambda_\kappa}\left[\frac{\left(\lambda_0 + \sum_{k\in K_j}\lambda_\kappa\right)p_l}{\lambda_0 + \sum_{k\in K_j}\lambda_\kappa} - \frac{\left(\lambda_0 + \sum_{k\in K_j}\lambda_\kappa\right)p_k}{\lambda_0 + \sum_{k\in K_j}\lambda_\kappa}\right]$$

$$= \frac{1}{\sum_{k\in K_j}\lambda_\kappa}\left[p_l\frac{\lambda_0 + \sum_{k\in K_j}\lambda_\kappa}{\lambda_0 + \sum_{k\in K_j}\lambda_\kappa} - \frac{\sum_{k\in K_j}\lambda_\kappa\, p_k}{\lambda_0 + \sum_{k\in K_j}\lambda_\kappa}\right]$$

$$= \frac{p_l - EV_{K_J}}{\sum_{k\in K_j}\lambda_\kappa} > 0$$

Not surprisingly, an increase in demand increases the expected value of a time slot. The magnitude of the increase depends on the difference between the price of that type and the expected value.

Note: this does not take into account large changes in demand that potentially could impact which patient types are included. This happens if the increase pushes the expected value higher than the price for the lower patient types.

Expected value wrt p_l:

$$\frac{\partial EV_{K_J}}{\partial p_l} = \frac{\lambda_\kappa}{\lambda_0 + \sum_{k\in K_j}\lambda_\kappa} = Prob_{k,j} > 0$$

Note: this does not take into account large changes in price that potentially could impact which patient types are included. This happens if the increase pushes the expected value higher than the price for the lower patient types.