Intermediary Bargaining for Price Insensitive Consumers

Guy Arie∗ Shiran Rachmilevitch†‡
April 8, 2015

Abstract

An essential feature of the bargaining problem between insurers and hospitals is that once the hospital network and insurance-terms are set, the insured ignore the cost of medical service when they decide which hospital to go to. We show that under such price-insensitivity, prices derived from standard bargaining models are significantly inflated. Under assumptions that are common in current structural empirical research, surplus-maximizing insurers end up paying more for every patient-service than the value of the service to the patient. We propose an alternative model and the corresponding equations for estimation, in the equilibrium of which prices are guaranteed to be lower than the value of the service. In addition, we show that a commonly-assumed price-monotonicity property may be violated in a variety of standard models and propose a version of the property that is satisfied in our model.

Keywords: Collective bargaining; Health economics; Insensitivity to prices.

1 Introduction

Intermediaries often bargain collectively with various upstream providers over prices, letting the final customers make their purchasing choice based (at least in part) on other qualities of the product. A clear example of this phenomenon—and the one we will focus on in the sequel—is hospital-insurer bargaining. In such a market, patients choose hospitals within their insurer network, but otherwise generally ignore the price that the insurer will pay.

∗University of Rochester, Simon Business School: guy.arie@simon.rochester.edu
†University of Haifa: mailto:shiranrach@con.haifa.ac.il
‡We thank participants in UECE Lisbon Meetings, Simon School junior lunch and Paul Ellickson, Jeanine Miklos-Thal, Greg Shaffer and Gerry Wedig for helpful comments. All errors are our own.
A key desired property of price bargaining models is what we will call *positive-surplus prices* (PSP)—the price paid per service is lower than the average value of the service to the patient. We show that under common assumptions, the “de-facto standard” method used in the literature—an application of the asymmetric simultaneous Nash bargaining solution in Kalai (1977)—necessarily violates this property whenever hospitals have sufficiently high bargaining power.

Four important features of the market work together to cause this result. First, the hospital-insurer bargaining is mostly based on fee-per-service. More complicated contracts could, in theory, solve some problems, but give rise to other issues. Second, consumers pay only a small fraction (often zero) of the cost of specific services they use, and instead pay a “subscription” fee (insurance premium). Third, the consumer’s choice of insurers cannot guarantee that the consumer will switch insurers to have ex-post access to their most favored hospital, in case the bargaining ends with some hospitals not in the insurer’s network. Fourth, insurers often consider the surplus they generate rather than simply their current net revenue.

These features are common to the current empirical assumptions for the US healthcare market. The recent handbook chapter by Gaynor and Town (2011) (henceforth GT) treats the first three assumptions as standard for the related empirical work. The literature is split with regards to the fourth assumption. Most recently, GT assume insurers maximize current net revenue while Lewis and Pfum (2015) assume surplus maximizing insurers. Gowrisankaran et al. (2015) allow for insurers to consider either net revenue or surplus and find that both explain the data equally well.

The standard approach assumes that prices are set as if each hospital bargains with the insurer holding the contracts with all other hospitals fixed. As in any bargaining setting, price is determined by comparing the network’s surplus with and without the hospital. The key insight of our analysis is that when a certain hospital, say hospital $A$, leaves an insurer’s network, some patients who would eventually prefer $A$ remain insured in the network. These patients will now see an inferior (for them) hospital, generating lower utility when treated. However, when bargaining with $A$, the insurer’s price for the remaining hospitals implicitly assumes these patients see their preferred hospital, $A$.

To see the problem, consider some other hospital, $B \neq A$. If, for example, $B$’s price is sufficiently close to its surplus per patient when $A$ in the network, then it may well be higher...
than the value of $B$'s service to $A$'s patients. When $A$ leaves, $A$'s "would-be" patients that remain with the insurer and use $B$ will therefore generate negative surplus (net of prices). The net value to a surplus-maximizing insurer by adding $A$ to the network is therefore the sum of the value of $A$'s service and the negative surplus that is avoided. With a sufficiently large bargaining power, $A$'s price can surpass the value of $A$'s service. Equilibrium forces exacerbate the effect and result in prices that are higher than value of service for each hospital.

Therefore, if the hospitals’ bargaining power is sufficiently high the only solution to the multi-hospital bargaining game with the aforementioned assumptions violates the insurer's "individual rationality" as a collective bargainer. The insured public will be strictly better off using spot market prices and limiting the insurer to provide insurance without price negotiation.

This dynamic and its implications are avoided if the model assumes that the insurer bargains sequentially with the hospitals. Under a fixed bargaining sequence, the first hospital (more generally, a hospital which is positioned early in the sequence) receives a low price, and the insurer's surplus is positive. For such a course of action to be feasible, the order of negotiations has to be non-negotiable. In particular, it is important that the insurer be able to commit to a negotiations order. Whether the insurer can commit to not reopen failed negotiations with a hospital is therefore an empirical question of significant importance. Our impression is that insurers have sufficient mechanisms to make such commitment. For example, a negotiating unit that is limited by capacity to only negotiate with a small fraction of the industry's providers at a time would accomplish this.

As the order of negotiations matters for sequential bargaining, an important issue that comes about is how to determine this order. We show that this question is cleanly resolved in a repeated game version of our model, in which our sequential bargaining story takes place in every stage game. We construct the equilibrium of the repeated sequential setting and show that if the hospitals are sufficiently patient then a hospital's price converges to its price as the first-in-the-sequence hospital. This, in turn, depends on its value as a standalone provider to the network.

By replacing the simultaneous bargaining assumption with sequential bargaining, our model provides alternative (with respect to GT) equations for estimation that guarantee the per-service prices paid by the insurer are are lower than the value of the service to the insured. All other assumptions currently used in the literature may be maintained common assumptions.
The aforementioned two-hospital illustration of the *bargaining interdependency* that we focus on can therefore be understood as closely related to, and as a by-product of, the insurer’s lack of commitment ability. The analysis then implies that market inefficiencies should be expected and that prices may be higher than the value of the service. This may provide a micro-economic foundation to the view that health-care services in the US are overpriced, especially by large providers (e.g., the reportedly high cost of inpatient care in the US compared to the rest of the developed world). To the extent that the insured do cover these costs through premiums, this may provide an explanation for the relatively low insurance rate in the US by consumers who are not subsidized, either by their workplace or a government program.

Our model suggests that merger can have significant implications on efficiency. In the aforementioned two-hospital illustration (and in the more general and more formal setting to be described shortly) inefficiency in the absence of commitment is driven by the possibility that the insurer is “stuck” with directing patients to hospitals that provide them with too low a value. In the extreme case that all hospitals are merged, this possibility never arises and the outcome, consequently, is efficient. If insurers cannot commit and hospitals’ bargaining power is sufficiently high, two hospitals in a market as above necessarily generate less welfare than one.

Insurers may mitigate the inefficiency by committing to have a limited network. The resulting networks will be smaller than the surplus-maximizing one, but may nevertheless generate relatively high surplus. However, holding any network agreements fixed, the insurer increases the surplus it generates by adding hospitals to its network. Moreover, [Wedig (2013)](#) finds that consumers insurance choice is consistent with significant welfare costs for having access to a smaller hospital network. Hence commitment to a smaller network is an issue. We view commitment to an order of negotiations as weaker than an a-priori commitment to a “legitimate candidates pool”. This weaker form of commitment is enough for delivering our good news: sequential negotiations results in positive surplus and lowers average prices.

We also consider another property which is central to price bargaining models. GT (p. 530, emphasis added) observe that: “…a hospital’s price will be increasing in its costs, bargaining ability, the prices of other competing hospitals, and, importantly, the net value that the hospital brings to the insurer’s network.” We show that the emphasized part of the citation

---

1In the extreme case that the insurer bargains with a single hospital, the aforementioned bargaining interdependency problem obviously cannot arise.
is violated under plausible assumptions. The reason is that the (asymmetric Nash) bargaining solution implies only that a hospital’s overall profit increases in its net value. However, since increases in value may imply an increase in the number of patients, per-patient price need not increase in hospital net value. The consumers’ insensitivity to prices implies that even a small difference in net value between hospitals may generate significant differences in ex-post market shares. A hospital that provides higher net value will typically enjoy a high overall profit, but serve a large fraction of the market. Spreading the overall profit over the larger market share may result in a lower price per patient.

To illustrate, consider again the case of two hospitals, $A$ and $B$, which are identical in terms of their costs and bargaining power, but differentiated in terms of their value. The market that these hospitals serve looks as follows. There is a unit mass of identical (price insensitive) patients. The utility for a patient from being treated at hospital $j \in \{A,B\}$ is $v_j - 1_{\{\text{hospital } j \text{ is overcrowded}\}} \times d$, where a hospital becomes overcrowded if (and only if) the mass of patients that visit this hospital exceeds $\alpha$, where $\alpha \in (\frac{1}{2}, 1)$ is a capacity constraint parameter. The parameter $d > 0$ stands for the disutility associated with being treated in a crowded hospital.

Suppose that $v_B = V$ and $v_A = kV$, for some $k > 1$. If $k$ is not too far from one, and if $d$ is sufficiently large, then when both hospitals are in the network the equilibrium of this simple model is such that a mass $\alpha$ of the population visits hospital $A$, the rest go to $B$, and there is no congestion. The total value that the hospitals provide in this full-network scenario is therefore $\alpha kV + (1 - \alpha)V$. Hospital $j$’s value (its marginal contribution) is the difference between the aforementioned number and the value that would have been obtained by the system when $j$ is out of the network. It is obvious that $A$’s value exceeds $B$’s. However, since total compensation is proportional to the value (under the Nash bargaining solution), $B$ will fetch higher per-patient price if $\alpha$ is sufficiently large; namely, if, with the full network in place, most patients go to $A$.²

Our repeated game framework provides an alternative form of price monotonicity. The insurer “threatens” each hospital that it would be in an unfavorable position in the sequence-of-negotiations, starting from tomorrow onwards, if it deviates today. With sufficiently patient hospitals, this threat bases each hospital’s price on its stand-alone contribution. Thus, hospital prices are monotonic in the net value of the hospital as an isolated network.

²The far-reaching implications of price insensitivity are sharply seen in this example when $k = 1 + \epsilon$, for a small $\epsilon > 0$. In this case, there is almost no difference in net values (the hospitals are almost symmetric) but since $A$ serves more patients its per-patient price will clearly be smaller than $B$’s.
While our analysis focuses on hospital-insurer bargaining, it is also relevant to other settings. Large organizations (firms, hospitals, military, etc.) often have dedicated "purchasing departments" that negotiate with multiple "approved vendors", while the actual product choice is done by professionals (e.g., doctors, engineers, military commanders) based on the products' features on a case-by-case basis. Another application is negotiation between content distributors (e.g., Cable, Netflix, Amazon) and content providers (e.g., CBS, Disney, HBO), whenever the payment is based on ex-post viewership. In all of these settings the eventual consumer is either unable to change an intermediary when a specific supplier quits the network (e.g., the "approved vendors" case) or is unlikely to (as in the cable network case). Thus, our analysis explains why such purchasing departments or content distributors may lead to higher prices in the long run.

The next Section reviews the related bargaining literature in general, and in the specific health economics aspect; Section 3 provides the main result in a simple model that is generalized and analyzed in Section 4; Section 5 considers the repeated game version of the model; Section 6 considers price-monotonicity; Section 7 concludes.

2 Literature Review

In terms of the theoretical framework, our paper belongs to a strand of literature that concerns bilateral bargaining in vertically-structured markets. In this regard, the main reference to our work is that of Horn and Wolinsky (1988), who apply the asymmetric Nash bargaining solution of Kalai (1977) to determine input prices in a duopoly model in which two firms acquire inputs from a common upstream supplier. More recently, Iozzi and Valletti (2014) considered a similar setting, in which each of $N$ identical downstream firms buys its input from a single supplier at a first stage, which is followed by competition (among the firms who made a non-zero input purchase) at a second stage. As in Horn and Wolinsky (1988), Iozzi and Valletti (2014) assume that each of the $N$ input prices is determined through bilateral Nash bargaining. Iozzi and Valletti (2014) consider the case where the second-stage competition is either a Bertrand competition or Cournot competition.

The central difference between our work and the aforementioned papers is in the bargain-
ing stage\footnote{Iozzi and Valletti (2014) also distinguish between (1) the case where negotiation breakdown between the upstream monopolist and a particular firm $i$ is observable by $i$'s rivals and (2) the case where it is not observable. Our analysis does not rely on downstream (insurer) competition. However, if such competition exists, we follow the literature and assumes that the negotiation outcome (i.e. insurer network structure) is common knowledge.}. Both in \cite{HornWolinsky1988} and \cite{IozziValletti2014}, the upstream-downstream negotiations are modeled as a collection of (simultaneous) independent bargaining problems. In our model, by contrast, there are significant interdependencies between the problems. Indeed, the main contribution of our analysis is the realization that the interdependencies are so strong that even the most basic intuitive properties of a bargaining solution may be violated. Crucially, the objective of the upstream entity of our model—the insurer, or, to use a more general term, the collective bargainer—is not a linear revenue-function; instead, the collective bargainer's payoff (the expected welfare of the mass population it represents) is sensitive to the composition of the downstream market.

From the theoretical bargaining literature, the central work about interdependencies among bargaining problems is by Bennett (1997). She considers the case where $n \geq 3$ players form (possibly several) coalitions, bargaining takes place within each coalition, and the interdependencies among the different bargaining problems are reflected in the endogenous determination of the outside options: the disagreement payoff of player $i$ in the bargaining problem in which he participates is calculated on the basis of what would have happened had that player belonged to another coalition (i.e., what payoff he would have obtained then). There are two important differences between Bennett's work and ours. First, in her work each player can participate in at most one coalition. In our work, by contrast, the collective bargainers bargains with each hospital separately; that is, he simultaneously belongs to several coalitions. Secondly, at the conceptual level, one of the main issues in Bennett (1997) is that of coalition formation. In our work, by contrast, this is not an issue, due to exogenously given vertical structure: it is a priori clear what are the possible coalitions to look at, and, moreover, it is clear what is the coalition-configuration to focus on—the one corresponding to the full hospital-network.

Our application of the Nash bargaining solution in the context of health economics closely follows the handbook chapter by Gaynor and Town (2011), which builds on numerous works from the health economics literature. Regarding the Nash bargaining that takes place between an insurance provider and a hospital, Gaynor and Town write (p. 530) "if the net surplus from the hospital-insurer match is not greater than zero, then bargaining does not take place and the
hospital is not in the insurer's network." One informal way to interpret our negative-surplus result (Theorem 1 below) is that the aforementioned circumstances are actually plausible in many settings.

Our analysis highlights the importance of another question in the health economics literature—do insurers maximize surplus (or a fraction of it) or their own profit? We show that the implications for expected prices and surplus are drastic and unexpected (profit maximizing insurers may provide more surplus). The model in Gaynor and Town (2011) assumes that insurers maximize their own profits which are calculated as the premium revenue less treatment costs. However, the literature is inconclusive on which assumption (surplus or profit maximizing) best models insurers. Town and Vistnes (2001), Capps et al. (2003) and Lewis and Pfum (2015) assume surplus maximizing insurers, in part arguing that this is a better proxy for the insurer's long run profits. In contrast, Gal-Or (1997) and Ho (2009) assume short run profit maximization. Most recently, Gowrisankaran et al (2015) estimate their model for both surplus and profit maximizing insurers. We consider the implications for profit maximizing insurers in section 4.1.

3 The basic idea: two symmetric hospitals

A symmetric setting provides a simple illustration of the underlying dynamic resulting in violation of the positive-surplus prices (PSP) property. The next section generalizes the result.

An insurer negotiates with two hospitals, \( j \in \{A, B\} \), serving a specific market. Wlog, normalize the cost of serving a patient to zero. The market has two types of patients, denoted \( \theta \in \{a, b\} \). Assume for now that patients of type \( a \) (resp. \( b \)) have a value of 10 from being served by hospital \( A \) (resp. \( B \)) and 5 from being served by the other hospital. With both hospitals in the network, the insurer expects a unit mass of patients for each hospital. However, if hospital \( A \) leaves the network, only a fraction \( \alpha \in (0, 1) \) of the \( a \) patients and \( \gamma \in (0, 1) \) of the \( b \) patients remain in the network. Similarly, if \( B \) leaves, \( \alpha \) of the \( b \) patients remain and \( \gamma \) of the \( a \) patients. That is, the hospitals are exactly symmetric.

The insurer maximizes surplus to the patients (or a fraction of it). Let \( V \) denote the patient surplus of the full network and \( V^{-j} \) the same without hospital \( j \). Negotiations are over prices \( p^j \) and set each hospital's bargaining power at \( \beta \in (0, 1) \). Thus: \( V = 20 - p^A - p^B \) and \( V^{-A} = \alpha \cdot (5 - p^B) + \gamma \cdot (10 - p^B) \).
Under Nash bargaining, if we hold $p^B$ fixed then $p^A$ solves

$$
\max_p (V - V^{-A})^{(1-\beta)} \cdot p^B
$$

(3.1)

The price response is given by:

$$
p^A = \beta \cdot [10 (1 - \alpha) + \alpha (5 + p^B) + (1 - \gamma) (10 - p^B)].
$$

(3.2)

Equation (3.2) shows that $A$ obtains a fraction $\beta$ of the surplus it generates. The $(1 - \alpha)$ new patients each account for 10 utils. The $\alpha$ patients that instead would have went to $B$ only gain 5 directly from going to their preferred hospital, but also save the payment of $p^B$. Finally, $A$ also enjoys the additional $(1 - \gamma)$ patients that would leave the insurer without $A$ but end up using $B$.

Of these three different consumer segments $(1 - \alpha, \alpha, \text{and } 1 - \gamma)$, prices may surpass surplus only because of the second (\alpha) group. In particular, whenever $p^B > 5$, the insurer actually generates negative surplus (net of prices) serving these patients without $A$ in the network. The surplus that $A$ generates to these patients is then higher than its ex-post per-patient value of 10. If the hospitals' bargaining power is sufficiently high so that hospitals obtain most of the surplus they generate, price will then be higher than the ex-post value.

Formally, prices are obtained by solving (3.2) and the symmetric equation for $p^B$. This obtains:

$$
p^j = 5\beta \frac{4 - \alpha - 2\gamma}{1 + \beta(1 - \alpha - \gamma)},
$$

(3.3)

for both $j \in \{A, B\}$. Recalling that with both hospitals in the network the value of each service is 10, the following is easy to verify:

$$
p^j \leq 10 \iff \beta \leq \frac{2}{2 + \alpha}
$$

(3.4)

That is, as long as some patients stay with the insurer despite ex-post preferring a hospital that opts out of the network (i.e. $\alpha > 0$), there is some lower bound on the bargaining power parameter, $\bar{\beta} < 1$, such that for any $\beta > \bar{\beta}$, the surplus generated by the insurer is negative – each service’s price is higher than its value.
Figure 3.1: Symmetric example for $\gamma = 1$ and $\alpha = 0.25, 0.5$ or $0.75$

Figure 3.1 illustrates the resulting price per patient when $\gamma = 1$. Even if just a quarter of the patients stay with the insurer even if their favored hospital leaves the network, prices exceed value when $\beta \geq 0.88$. As the share of patients that would not adjust their insurer choice increases, equilibrium prices significantly increase and may well be more than double the value per patient.

The example is easily generalized. Let the 2-hospital symmetric model be defined as follows: $v$ is the value of a hospital to the patients that would go to it if both hospitals were in the network (i.e., 10 in the example) and $\lambda \cdot v$ be the value per patient of going to their second best hospital (i.e., $\lambda = 0.5$ in the example). Then again derive $V^i, V^{-i}$ and the bargaining solution for each hospital as in (3.2) to obtain:

Proposition 1. In the 2-hospital symmetric model, the insurer generates negative surplus per patient (i.e., $p^i > v$) iff

$$\beta \geq \frac{1}{1 + \alpha(1 - \lambda)}.$$  \hspace{1cm} (3.5)

In particular, for any valid parametrization $(\alpha, \lambda, \gamma \in (0,1))$, there is some $\bar{\beta} < 1$ such that for any $\beta > \bar{\beta}$ the price of each service is higher than its value.

4 A general model

This section uses a variation on the handbook model of Gaynor and Town (2011) (GT) to generalize the violation of the PSP property. The next section then provides the alternative,
repeated-sequential model that recovers PSP. To simplify the mapping between our result and the GT model, we fully adopt their notation, but slightly simplify the model. GT's model allows for multiple insurers and hospitals, all potentially differentiated in costs and value of service. The main qualitative difference in our model relative to GT is that we assume insurance maximize their insured surplus rather than profit. As we reviewed in Section 2, the literature seems split on which is correct. However, an important implication of the current analysis is that this distinction has greater implications than perhaps was previously realized (see e.g., footnote 22 in GT).

Our model takes on the cooperative approach to bargaining. Namely, we do not specify a concrete extensive form, but only describe the economy in terms of its payoff-relevant information. We assume (as in the example from the previous Section) that prices are set as to maximize the relevant Nash products. In terms of the terminology and phrasing that we use in the text, we allow ourselves a little degree of informality, which manifests itself in two ways: (1) when we write "in equilibrium..." we do not refer to an equilibrium of any specific game, but simply to the situation in which all the prices in our model are set at their Nash-product-maximizing values; (2) when we write descriptive statements such as "when hospital j leaves the network," or "the insurer proposes a price to the hospital," this is only meant for the sake of illustration—it is not a description of a move in some extensive form.

The model is as follows. There is a measure of \( \frac{1}{\rho} \) patients, of which a fraction of \( \rho \) will need hospital service. There is a fixed (finite) set \( J \) of ("candidate") hospitals. A network is a non-empty subset of \( J \). A generic network will be written as \( J_h \) (following GT, the index \( h \) stands for the insurer, and \( J_h \) is the insurer's network). Should patient \( i \) need hospital service, she obtains utility \( u_{ij} = f(x_j, z_i, d_{ij}) + e_{ij} \) from using hospital \( j \), where \( z_i \) and \( x_j \) are, respectively, patient and hospital specific parameters, \( d_{ij} \) captures joint parameters (e.g., distance from patient to hospital) and \( e_{ij} \) is an unobserved logit error term (i.e., type-1 EV) that is only revealed to the patient when requiring hospital service. The patient can also choose an out-of-network hospital, which provides zero utility by assumption.\(^5\) Compared to GT, the simplification here is that we bundle all possible diagnosis to one (instead of \( M \) possible diagnosis in GT). This has no qualitative implications.

\(^5\)Both the logit assumption and the out-of-network assumption are made in GT and used here to simplify the exposition. Our model is agnostic about these. The formal assumptions required for the result are explicitly identified below.
The logit assumption implies that the ex-ante value of a network $J_h$ to a patient is:

$$W_i^{J_h} = \ln \left( 1 + \sum_{j \in J_h} \exp \left( f(x_j, z_i, d_{ij}) \right) \right)$$

In the GT framework, each insurer $h$ agrees with each hospital in its network on a fee per patient service $p_{jh}$ and then sets premiums knowing all networks and assuming the standard logit model for patient insurance choice with $W_i$ as patient utility and some premium sensitivity parameter $\alpha$. This results in insurer gross revenue $F_h(J_h)$ and hospital ex-post quantity demanded from the insurer $q_{jh}$. The same can be calculated for the network without any single hospital $j$, which will be denoted as $J_{h-j}$. Let $c_j$ denote the hospital’s cost per patient-service, $r_j$ the hospital’s outside option and $cm_{jh}$ some hospital-insurer specific fixed costs. Then the Nash bargaining solution provides equation (9.7) in GT as the equilibrium price-response for hospital $j$ from insurer $h$ (the equivalent of equation 3.2). We provide it here verbatim, multiplying both sides by $q_{jh}$ for clarity.

$$p_{jh}q_{jh} = (1 - \beta) (q_{jh}c_j - r_j) + \beta \left[ F_h(J_h) - F_h(J_{h-j}) - cm_{jh} + \sum_{l \neq j} p_{lh} \cdot (q_{lh}^{J_{h-j}} - q_{lh}) \right] \tag{4.1}$$

The hospital’s total revenue is a bargaining-power weighted average of its costs and the net benefit to the insurer, holding all other prices fixed. The difference $F_h(J_h) - F_h(J_{h-j})$ is the increase in revenue for the insurer. The sum at the end is, as in our simpler model from the previous Section, the change in payments to other hospitals $l$ due to patients that go to these hospitals without $j$ in the network but otherwise would not. Both $r_j$ and $cm_{jh}$ are exogenous variables that do not contribute to the theoretical analysis. We therefore simply set these to zero.

In addition to allowing for multiple hospitals and insurers, the GT model assumes insurers maximize revenue both in their network choice and their pricing choice. Whether insurers maximize profits or surplus is an open empirical question. We assume for the main analysis that insurers are surplus-maximizing. Therefore, we replace the $F_h$ interpretation in 4.1 with the expected value for the patients. This allows us to abstract away from the patient’s decision to pay the insurer premium. We discuss profit-maximizing insurers separately below. To evaluate surplus, we let $w_{h,j}$ denote hospital $j$’s average value to $h$’s patients that choose $j$ in equilibrium
As in our simple 2-hospital benchmark, the key assumption in the analysis is that some patients remain with an insurer even if their ex-post preferred hospital leaves the network. These patients will now use a different hospital, implying substitution between hospitals. Denote by \( w^1_{h,j} \) the average value from \( j \) for \( h \)'s patients that choose \( l \) in the equilibrium network but would choose \( j \) if \( l \) is not in the network. To capture the notion of "second best," we assume that patients generally value their second-best hospital less than the patients that freely choose it. That is:

**Assumption 4.1.** For any hospital network, for all insurers:

1. Hospitals are substitutes: \( q^l_{l,h} \geq q^l_{h} \) with the inequality strict at least for one \( l \) for every insurer-hospital pair \( h, j \).
2. Second best values are lower: \( \forall j, l: \ w_{h,j} \geq w^1_{h,j} \).

We make an additional "independence of irrelevant alternatives" (henceforth, IIA) assumption that greatly simplifies the exposition. While stronger than required, the assumption is natural and would otherwise be replaced by cumbersome restrictions on patient preferences:

**Assumption 4.2.** IIA: If hospital \( j \) leaves a network, the hospital choice of patients that would have went to hospital \( l \neq j \) with \( j \) in the network does not change. Call the above model the general model. The main result for this Section is this:

**Theorem 1.** In the general model, there is a \( \bar{\beta} < 1 \) such that for any \( \beta > \bar{\beta} \) the surplus generated by each insurer is negative.

**Proof.** For simplicity, we omit the \( h \) subscript. Let \( x^l_i = q_i^j - q_i \) denote the quantity of \( j \)'s patients that move to \( l \) if \( j \) leaves the network.

Using (4.1):

\[
p_j q_j = (1 - \beta) q_j c_j + \beta \left( F - \sum_{l \neq j} p_l \cdot x^l_i \right) \tag{4.2}
\]

By construction, \( F = \sum_l q_l \cdot w_l \) and \( F_{-j} = \sum_{l \neq j} (q_l \cdot w_l + x^l_i \cdot w^j_l) \). This simplifies (4.2) to

\[
p_j q_j = (1 - \beta) q_j c_j + \beta \left( q_j w_j + \sum_{l \neq j} x^l_i \cdot (p_l - w^j_l) \right) \tag{4.3}
\]
What we have here is \( |J| \) equations in \( |J| \) unknowns, \( \{p_j\}_{j \in J} \). If all \( x^j_l = 0 \), then as \( \beta \to 1 \) we have that \( p_l \) is arbitrarily close to \( w_l > w^1_l \). Now increase all of \( x^j_l \) to their true value. As \( p_l > w^1_l \), this strictly increases \( p_l \) so that \( p_1 > w_1 > w^1_1 \) without affecting any other prices. Repeating the procedure for every \( j \) completes the proof.

\[ \square \]

4.1 Discussion

Theorem 1 applies to insurers whose charter is to maximize surplus (such as Medicare). As presented above, whether most insurers maximize surplus or consider only revenues and costs is an open empirical question. In recent work, Gowrisankaran et al. (2015) perform their analysis in parallel for both cases, reaching qualitatively similar results. Town and Vistnes (2001) suggest that even profit maximizing insurers attempt to maximize surplus because, in the long term, patient choice reacts to surplus and as a result, an insurer’s long term profit is ultimately best approximated as a fraction of the surplus it generates.

Short-term-profit maximizing insurers are immune from the negotiation tactic we described above. The insurer does not “care” if its patients will use a hospital that provides less value than reflected in the premium, as long as the patients pay the premium. In addition, if we assume that patients only choose an insurer if the premium is lower than the expected value of service, positive expected surplus for those choosing insurance is guaranteed by assumption. The main implication of our analysis in this Section is therefore that whether insurers maximize surplus or short-term profits is of significant importance in predicting prices and prescribing market policies.

Another important consideration is the ex-post inefficiency associated with partial insurer networks. For patients to avoid getting “stuck” with a second-best hospital, they must switch to the insurer network that provides them the most value conditional on their predictions and premiums. Handel (2013) found that inertia (or switching costs) plays a significant role in consumer insurance choice. This implies also that each hospital contributes very little to the insurance revenue and prices. As a result, hospital margins and prices will be low.\(^6\)

\(^{6}\)The surplus implications of such high switching costs can be significant, both in terms of high insurance premiums and sub-optimal insurance choice by consumers.
switch out of smaller networks. [Wedig (2013)] found that such concerns made smaller networks unprofitable for insurers. At the extreme, smaller networks are never chosen by consumers and the net value of each hospital to the insurer in the simultaneous bargaining approach (as in GT) is the entire revenue per patient, leading to a market breakdown.

This highlights the important of both empirical and theoretical research regarding the behavior of intermediaries and consumers. Our analysis shows that if consumer insurer choice may be sub-optimal and hospitals have sufficient bargaining power, profit maximizing insurers may well generate more surplus than surplus maximizing ones. In particular, surplus maximizing insurers will generate negative surplus, while profit-maximizing insurers will generate positive surplus.

This insight is also important for other settings to which our theory applies. To be specific, consider a purchasing department that contracts with various competing providers over inputs to be used by several downstream "user" departments (e.g., a university's purchasing department negotiating with several laptop makers). If the department is evaluated as a profit center based on some transfer price (or willingness to pay) from each of the downstream departments and the downstream users must make their purchases through the upstream purchasing department, then the organization will be subject to the same problem as the surplus maximizing insurer and may well generate negative surplus. We expand on this in follow-up work.

A second implication is that if insurers are surplus maximizing, it may be more efficient to let hospitals merge. If all hospitals in a market merged, the merged hospital will provide each patient the best option. The insurer is never threatened with negative patient surplus and as a result, the merged hospital can never obtain more than the entire ex-post surplus it generates.

A direct implication of Theorem 1 is that for sufficiently high $\beta$, this is the only market structure that protects a surplus-maximizing insurer from generating negative surplus.

Corollary 1. In the general model, there is a $\bar{\beta} < 1$ such that for any $\beta > \bar{\beta}$, surplus from a surplus-maximizing insurer is maximized if and only if all hospitals merge.

5 Recovering positive-surplus prices with sequential bargaining

The root of the overpricing which gives rise to the negative surplus from Theorem 1 lies in the hypothetical "disagreement event" associated with each negotiation. For example, in the
two-hospital case, when $A$ is out of the network (i.e., there is disagreement with $A$) its entry-contribution exceeds its true value because it saves some too-high payments that would otherwise be directed to $B$. Pushing $A$'s price down, to a level that would result in a positive surplus for the insurer, crucially depends on turning down the influence of this hypothetical disagreement scenario.

One way to achieve lower prices is to make it impossible for $A$ to “go back” into the network if and when it is out of it. This can be achieved by sequential bargaining: the hospitals are ordered in a (commonly known) sequence, and once there is negotiation breakdown with some hospital early in that sequence this hospital goes “out of the game” for the duration of the negotiation phase, and all subsequent negotiations effectively ignore the dropped out hospital.

It is intuitive that the order of negotiations affects hospital payoffs. For example, consider the two-hospital case where $A$ is the first in the sequence. Then, disagreement with $A$ automatically makes $B$ a monopolist, whereas disagreement with $B$ cannot have such a favorable effect on $A$'s bargaining position, since disagreement with $B$ can only happen after the interaction with $A$ has concluded; specifically, either (i) a deal with $A$ has already been signed and so $A$'s price is fixed, or (ii) $A$ has dropped out, and is no longer in the network.

It is also rather easy to see that the insurer's surplus is positive under sequential negotiations. Consider again the two-hospital case where $A$ going first and $B$ going second. Given any outcome of negotiations with $A$, the bargaining problem with $B$ must result in a non-negative addition to the insurer's overall surplus (or else the insurer will not sign a deal with $B$). Also, one can map any possible outcome in the negotiations with $A$, say $o$, to the subsequent bargaining problem that will be played with $B$, say $P(o)$. Since, as we just observed, the insurer's surplus in $P(o)$ is non-negative given any possible $o$, the bargaining problem with $A$ boils down to a standard Nash bargaining problem in which both parties make positive profits. This idea generalizes to any length of hospital-sequence.

In terms of formalism, the economic environment that we consider is the same as in the previous Section, with the only difference being that the Nash products now reflect the order of negotiations. Also, we add a technical assumption about off-equilibrium beliefs. We assume that the insurer and hospitals assume that if two hospitals leave the network then the patients that had those two hospitals as their top two choices, also leave the network. This allows us
to ignore "high-order" bargaining effects\footnote{That is, the hospitals and insurers assume that if \( A \) leaves the network, each patient that had \( A \) as its preferred hospital and remains in the insurer's network would leave the network if their second option hospital also leaves the network.} that would be negligible when \( \beta \) is sufficiently large in any case. Technically, this keeps the steps of the proof from growing linearly in the number of hospitals. We note that the assumption is vacuous in a two-hospital setting. Call this the \textit{sequential model}.

The positive surplus result (for the insurer) and the favorable status of late bargaining positions (for the hospitals) are summarized in the following results (which are proved in the appendix).

\begin{proposition}
Assume the sequential model. For any \( \beta \in (0, 1) \), the surplus generated by each insurer is positive. Moreover, the insurer's surplus is independent of the order of negotiations with the hospitals.
\end{proposition}

Appendix \appendixa.3 provides the closed form solution for the sequential model \textendash; prices per hospital in the insurer's expected value on which the propositions are based.

That the insurer is indifferent regarding the order may result in lower prices for better hospitals. We resolve this issue by moving from a single sequential interaction to an infinite horizon setting, in which the aforementioned interaction\textendash;namely, the sequential model\textendash;takes place in each and every period. Call this the \textit{repeated sequential model}.

We will assume throughout that the punishment is implemented by placing the hospital first in the order of negotiation. This is consistent with standard economic intuitions as described above. For a general model, proposition \ref{prop:repeated-sequential} provides a formal result for a two-hospital setting or for the case that all hospitals are symmetric in value per patient and expected number of patients (as in section \sectiona).\footnote{That is, patients differ in their preference over hospitals, but at the aggregate all hospitals are similar in terms of demand and value per visiting patient.}

\begin{proposition}
Assume the sequential model with two hospitals or with \( J \) symmetric hospitals, if \( \beta \to 1 \), negotiating later in the order increases a hospital's price and profits.
\end{proposition}

As before, we specify only payoff-relevant data, not an extensive form game. Within each given period, payoffs are as we described above. The only difference is that now each agent (insurer, hospitals) seeks to maximize the discounted sum of its periodic payoffs from each time period onwards. We make use of the following notation.

\footnotesize
\begin{itemize}
  \item That is, the hospitals and insurers assume that if \( A \) leaves the network, each patient that had \( A \) as its preferred hospital and remains in the insurer's network would leave the network if their second option hospital also leaves the network.
  \item That is, patients differ in their preference over hospitals, but at the aggregate all hospitals are similar in terms of demand and value per visiting patient.
\end{itemize}
1. \( \delta \in (0,1) \) denotes the hospitals’ (common) discount factor.

2. \( F_{jh} \) is the patient surplus for insurer \( h \) (or expected insurer revenue) with only hospital \( j \) in the network.

3. \( F_h \) is, as before, the value of the insurer’s objective (either patient surplus or profit).

4. \( W_j \) is the increase in the insurer’s objective from adding hospital \( j \) to the network, given all other hospitals’ prices (i.e. the bracketed term in any of 4.1 4.2 or 4.3).

5. \( p_j \) is hospital \( j \)'s price.

6. \( \hat{p}_j \) is the price the hospital would obtain if the insurer ordered the negotiation in the worst possible way for that hospital.

7. \( q_j \) is hospital \( j \)'s expected number of patients.

8. \( q_{jh}^j \) is hospital \( j \)'s expected number of patients if it is the only hospital in \( h \)'s network.

Though we do not study any particular extensive form, we are guided by game-theoretic reasoning, which we now describe for illustrative purposes. Suppose that in each period, when the insurer bargains with a specific hospital, the interaction between them assumes the following form: the insurer makes a price offer to the hospital, which the hospital either accepts or rejects. If a certain hospital, say \( j \), rejects an offer instead of accepting it, the insurer punishes that “deviating” hospital by implementing, in each period from tomorrow onwards, the negotiations-order which is worst for hospital \( j \). In every period in this perpetual punishment phase, \( j \) receives the price \( \hat{p}_j \). We show in the appendix how to derive \( \hat{p}_j \). While tedious, this is a purely technical exercise.

The important qualitative result is this:

Proposition 4. Assume the repeated sequential model. If hospitals have sufficiently large bargaining power, the punishment price converges to the hospital’s per patient value if it was the only hospital in the network: For every \( \varepsilon > 0 \) there is a \( \bar{\beta} < 1 \) such that for all \( \beta > \bar{\beta} \), \( |\hat{p}_j - \frac{F_{jh}}{q_{jh}^j}| < \varepsilon \)

Underlying proposition 4 is the realization that, if \( j \) is the first in the negotiation order, with sufficiently large hospital bargaining power, any value above \( F_{jh} \) that hospital \( j \) creates will be captured by some hospital that follows it in the negotiation order.
By assumption 4.1, the per-patient value a hospital provides is lowest when it is the only hospital in the network. This is because removing a rival hospital \(j'\) from a network only adds to \(j\) some patients that would prefer \(j'\). Then \(j\)'s per-patient value for these patients is lower than for patients that choose \(j\) regardless of \(j'\). Thus, starting from the full network and iteratively removing hospitals until only \(j\) remains in the network decreases \(j\)'s per-patient value at each iteration. Moreover, the “punished” hospital will obtain even less than its standalone value \(F_{jh}\). This is because \(j\)'s equilibrium quantity \(q_j\) is smaller than its standalone quantity \(q_{jh}^{jh}\).

**Corollary 2.** In the repeated sequential model, the profit for a “punished” hospital is strictly lower than the average per-patient value that the hospital provides.

Now suppose that the insurer and the hospitals play some stationary equilibrium of the repeated game. Suppose, moreover, that this equilibrium is a “Nash reversion” equilibrium, which means that once a hospital deviates, an absorbing punishment phase ensues from the next period onwards. Intuitively, the insurer will agree to a triggered-by-deviation-profit for the hospital exactly up to \(W_j\). This gives rise to the following IC constraint for hospital \(j\):

\[
W_j \cdot (1 - \delta) + \delta \hat{p}_j q_j \leq p_j q_j. \tag{5.1}
\]

The insurer’s problem is to choose the lowest possible prices \(p_j\) subject to these \(J\) constraints. We refer to the solution of this problem as the equilibrium of the (repeated sequential) model. As \(W_j\) and the insurer’s objective are linear in prices, all IC must bind, yielding:

**Proposition 5.** In the equilibrium of the repeated sequential model, hospital prices are given by

\[
p_j = \delta \hat{p}_j + (1 - \delta) \frac{W_j}{q_j}. \tag{5.2}
\]

Equation (5.2) therefore provides the alternative for the estimating equation (9.7) in GT. Note that, as in GT, the term \(W_j\) is endogenous as it includes the equilibrium prices for all hospitals except \(j\). Solving simultaneously the \(|J|\) equations for the \(|J|\) hospitals provides a closed form solution.

If hospitals are myopic, then the equilibrium of the repeated sequential model is the same as the one of GT: each hospital takes all other hospitals prices as fixed and acts “as if” it is last. However, if hospitals are very patient, prices converge to the “punishment” levels. Since
punishment levels provide each hospital with less profit than the value it generates (lemma 2), the insurer is guaranteed a positive surplus (or profit).

Proposition 6. In the repeated sequential model, if hospitals are sufficiently patient, the insurer’s per-period surplus (or profit) is strictly positive. If hospitals have sufficiently large bargaining power, the insurer’s profit converges to $F_h - \sum_j \left( F_{jh} \cdot \frac{a_j}{q_j} \right)$.

Figure 5.1 compares the prices for the symmetric setting used as the motivating example in section 3 under three bargaining models: the standard model, the one-shot sequential bargaining, and the repeated sequential bargaining for patient hospitals. Moving from the standard to the sequential model restores positive surplus. However, as $\beta \to 1$, the hospitals are able to capture all the surplus. In particular, the second hospital in the negotiation obtains prices that are higher than its value, while the first hospital obtains prices that are lower. Moving to the repeated sequential bargaining model shifts the prices for both hospitals to the “first hospital” price, leaving the insurer a strictly positive margin even when the hospitals have all the formal bargaining power ($\beta \to 1$).

6 Price Monotonicity

We now turn to consider price monotonicity. GT (p. 530, emphasis added) observe that: “...a hospital’s price will be increasing in its costs, bargaining ability, the prices of other competing hospitals, and, importantly, the net value that the hospital brings to the insurer’s network.” We show that the emphasized part of the citation is violated under plausible assumptions if prices are determined by standard (simultaneous) Nash bargaining, but is recovered...
in the repeated sequential model.

The violation occurs whether the insurer maximizes surplus or revenue. The reason for this is simply that while the higher-net-value hospital provides more value overall, it may see many more patients. It is quite easy to construct examples in which, by virtue of treating more patients, the higher value hospital obtains lower prices. We construct such an example in the Subsection below; then, in the following Subsection, we suggest a solution to the non-monotonicity problem.

6.1 Violating price monotonicity – a simple example

A small modification of the example in Section 3 is sufficient. As there, consider two differentiated hospitals $A$ and $B$. The population mass is divided into three equally sized types, denoted $a, b, c$; there is a unit mass of each type. The $a$ and $b$ types have the same preferences as before:

$v^A_a = v^B_b = 10$ and $v^B_a = v^A_b = 5$. The $c$ type, however, has a weaker preference among the hospitals: $v^A_c = 10, v^B_c = 8$. As before, assume that a fraction $\alpha \in [0, 1]$ of either type $a$ or $b$ stay with the insurer if their favored hospital is not in the network. To keep the example simple, assume that patients do not leave if their preferred hospital remains in the network and that the $c$ type patients remain in the network as long as it has either hospital, but always go to $A$ (their preferred hospital) if it is in the network.

It is clear that hospital $A$ provides more net value to the network – this is because of the symmetry between $a$ and $b$, and the fact that the $c$ types strictly obtain more value from $A$. Thus, the price monotonicity condition requires that the price paid to $A$ is at least as high than the price paid to $B$.

To see why price-monotonicity is violated in the GT framework (i.e., under standard Nash bargaining), assume first that $\alpha = 0$. That is, types $a$ and $b$ would leave the insurer if their hospital leaves. In this case, adding $B$ to the network simply adds all of the $b$ type patients (who obtain a value of 10 each) and has no effect on the other patients. Thus, $B$'s effect on surplus does not depend on $A$'s price and $B$'s price is simply $p_B = \beta \cdot 10$.

Now consider adding $A$ to the network that only has $B$. This creates a value of 10 for each $a$ type patient, but only 2 to each $c$ type patients. On average, added value per treated patient is

\[9\text{For a revenue maximizing insurer, one needs to adjust instead the expected premium, but this is straightforward.}\]

\[10^9p_B \text{ is the maximizer of } p^\beta \cdot (10 - p)^{1-\beta}.\]
just 6, while B added an average value of 10 per patient it treats. If prices reflected only value, this is sufficient to violate price monotonicity: A's price would be lower than B's for any $\beta$.

However, prices reflect surplus per patient, which for A also accounts for the fact that the insurer no longer pays B for the c type patients. Therefore, A's per-patient surplus increase is $6 + \frac{p_B}{2}$ (recall that only half of A's patients would see B in the network without A while the other half simply leave the network). If $p_B < 8$ (which happens if $\beta < 0.8$), the surplus per patient created by A is smaller than the surplus per patient created by B (which is always 10) and so A's price will be lower.

For a general $\alpha$ in the example above, it is easy to verify that the GT equilibrium prices obtain price monotonicity if and only if $\beta > \frac{\alpha-5\alpha}{10-\alpha}$. Therefore, if $\beta < 0.8$ or if $\alpha \to 1$, the hospital with the lower net value obtains a higher price per patient.

### 6.2 Repeated sequential negotiations and price monotonicity

The price monotonicity problem may be resolved in the repeated sequential model in some special cases. Proposition 6 obtains that for sufficiently high bargaining power and discount values the price of each hospital converges to the value per patient the hospital generates as the sole hospital in the network, $\frac{F_{jh}}{q_{jh}}$. Thus, price monotonicity in the repeated sequential bargaining is with respect to this ratio.

If each hospital expects the same quantity of patients as the sole hospital in the network (i.e. $q_{jh}^h = \bar{q}$), then the repeated model obtains price monotonicity with respect to the total value of the hospital as a single network $F_{jh}$. One likely setting for this is therefore in markets with a single dominant insurer. In this case we would expect the effect of a hospital leaving the network to be on consumer premiums rather than on the number of consumers that choose the insurer (See also Gowrisankaran et al. (2015)). Another likely setting is when hospitals are symmetric along the dimension that determines demand (say location), but not along the dimension that determines value (say quality).

Finally, if there are only two hospitals for such a network, the hospital with the larger net

---

11. See appendix for details.
12. In the repeated sequential model, it is easy to show that if the insurer negotiates first with the hospital that provides the least net value to the network, and continues in increasing order, the resulting equilibrium prices will satisfy price monotonicity. However, this is an arbitrary solution to the problem.
13. The same applies for the purchasing department example as the total quantities demanded are fixed regardless of the number of providers.
value $V_{jh}$ is also the hospital with the larger standalone value $F_{jh}$ by construction. This is because the value of a network with both is fixed and so

$$F_{jh} + V'_{jh} = F'_{jh} + V_{jh}.$$ 

Therefore, if $q_{jh} \approx q'_{jh}$ the punishment and equilibrium prices are both monotonic in $F_{jh}$ and $V_{jh}$.

Proposition 7. In the repeated sequential model, if (1) the insurer is negotiating with only two hospitals, (2) hospitals are sufficiently patient and have sufficiently large bargaining power, and (3) both hospitals expect the same number of patients from a single hospital network, then equilibrium price for a hospital is increasing in the hospital's net value $V_{jh}$

Proof. See above.

7 Conclusion

This paper has focused on the analysis of multilateral bargaining settings in which the total surplus from a suppliers' network is non-linear in the (equilibrium) quantity sold by each supplier. This is particularly common when an intermediary (insurer or purchasing department) bargains with suppliers (hospitals or input producers) on behalf of downstream users.

We showed that if the intermediary cannot commit to a specific network, suppliers can charge unit prices that surpass the unit value because the intermediary must also consider the potential negative surplus from directing its users to their second-best suppliers.

This dynamic has significant normative implications. First and foremost, “forcing” the intermediary to commit not to reopen a failed negotiation guarantees that prices are below the users' values. If this is not enforceable, encouraging profit (rather than surplus) maximization and reducing frictions in the downstream markets become very important.

Our proposed repeated sequential model describes the outcome that would result under a moderate (and reasonable) commitment power of the insurer—power which is sufficiently great to enforce sequential negotiations (and Nash reversion); the model can be used for estimation for both surplus and profit maximizing insurers.
Finally, we considered price monotonicity. Our analysis shows that, as stated, this is a harder property to obtain than perhaps previously assumed, but that under some specific conditions it is guaranteed in our repeated sequential model.

Our analysis has several implications for empirical work. First, it identifies the important implications of the implicit “no-commitment” assumption common to the existing models. To the extent that this assumption is maintained, other assumptions – positive surplus, monotonic prices – and expectations – inefficient mergers increase prices – should be doubted. Second, it provides an alternative estimation approach based on the commitment assumption.

A  Proofs

Note: proofs for Theorem 1 and Proposition 7 are in the text.

A.1 Preliminary Lemma

We first prove the following lemma. Suppose that the insurer is bargaining with hospital $j$ and that:

1. The insurer’s profit without $j$ in the network is $V_0$.
2. Adding $j$ to the network at unit price $p$ increases the insurer’s profit by $K - p \cdot y$.
3. The hospital’s unit cost is $c$.

Lemma 1. If there is a price $p$ that solves the bargaining game defined above, then it is:

\[ p = \beta \frac{K}{y} + (1 - \beta) c. \]  

(A.1)

The insurer’s profit is $V_0 + (1 - \beta) (K - cy)$ and the hospital’s profit per unit is $\beta \cdot \frac{K - cy}{y}$.

Proof. Suppose the hospital and the insurer expect the hospital’s quantity to be $q$. Then the bargaining problem is

\[ \max_p [K - py]^{(1-\beta)} [x (p - c)]^\beta \]  

(A.2)

The interior solution must satisfy:

\[ (1 - \beta) y [K - py]^{-\beta} [x (p - c)]^\beta = \beta \cdot x \cdot [K - py]^{(1-\beta)} [x (p - c)]^{(\beta - 1)} \]  

(A.3)
Simplifying:

\[ p \cdot [(1 - \beta) y + \beta y] = \beta K + c \cdot (1 - \beta) y \]  

(A.4)

Simplifying again obtains the desired price. the insurer and hospital profits are straightforward.

\[ \square \]

A.2 Two Hospitals Sequential Solution

Consider the negotiation with the second hospital given the first hospital’s price \( p_1 \). Denote by \( q_{j,i} \) the number of patients that would go to hospital \( j \) if it was in the network but instead go to hospital \( i \neq j \) because \( j \) isn’t in the network.

Without hospital 2, \( V^1 = F_1 - (q_1 + q_{2,1})p_1 \). For any price \( p_2 \) the insurer’s value is \( V^{1,2} = F_{1,2} - q_1 p_1 - q_2 p_2 \). Applying Lemma 1 for the bargaining gain \( V^{1,2} - V^1 \) obtains:

\[
p_2 = \beta \frac{F_{1,2} - F_1 + p_1 q_{2,1}}{q_2} + (1 - \beta) c_2 \\
V^{1,2} = F_1 - (q_1 + q_{2,1})p_1 + (1 - \beta)(F_{1,2} - F_1 + p_1 q_{2,1} - q_2 c_2) \\
= (1 - \beta)(F_{1,2} - q_2 c_2) + \beta F_1 - p_1 (q_1 + \beta q_{2,1}).
\]

(A.5)

Observe that as \( q_{2,1} > 0 \), \( p_2 \) is increasing in \( p_1 \). Moreover, as by assumption it is surplus increasing to add the second hospital, \( F_{1,2} - F_1 > c_2 q_2 \) and so \( p_2 > c_2 \).

Without hospital 1, \( V^2 = (1 - \beta)(F_2 - (q_2 + q_{1,2})c_2) \). Therefore, the bargaining gain from 1 is

\[
V^{1,2} - V^2 = \beta (F_1 - p_1 q_{2,1}) + (1 - \beta)(F_{1,2} - F_2 + q_{1,2} c_2) - p_1 q_1.
\]

(A.6)

Again applying lemma 1

\[
p_1 = \beta \frac{\beta F_1 + (1 - \beta)(F_{1,2} - F_2 + q_{1,2} c_2)}{q_1 + \beta q_{2,1}} + (1 - \beta) c_1.
\]

(A.7)

To see that \( p_1 > c_1 \), set \( c_2 = 0 \) \( (p_1 \) increases in \( c_2 \). Again, as long as \( F_{1,2} - F_2 > q_1 c_1 \) (adding the hospital to the network is efficient) and \( F_1 > c_1 \cdot (q_1 + q_{2,1}) \) (the hospital is efficient in a network by itself) , we have that \( p_1 > c_1 \).
We can now place $p_1$ in $V^{1,2}$ from equation (A.5) to obtain the insurer’s surplus:

$$V^{1,2} = (1 - \beta)(F_{1,2} - c_2 q_2) + \beta F_1 - (1 - \beta) [c_1 (q_1 + \beta q_{2,1})]$$  
$$- \beta (1 - \beta)(F_{1,2} - F_2 + q_{1,2} c_2) - \beta^2 F_1$$  

(A.8)

Simplifying:

$$V^{1,2} = (1 - \beta)(F_{1,2} - c_2 q_2 - c_1 q_1) + \beta (1 - \beta) [F_1 - c_1 q_{2,1} + F_2 - c_2 q_{1,2} - F_{1,2}]$$

$$= (1 - \beta)^2 (F_{1,2} - c_2 q_2 - c_1 q_1) + \beta (1 - \beta) [F_1 - c_1 (q_1 + q_{2,1}) + F_2 - c_2 (q_2 + q_{1,2})]$$  

(A.9)

A.3 $J$ Hospitals Sequential Solution

Note that, with a slight abuse of notation, we use $J$ to refer to the number of hospitals, the set of hospitals and the last hospital in the negotiation. We use the following notation for this section:

1. $\{1, \ldots, J\}$ is the list of all relevant hospitals, in order of negotiation.

2. $V(\{p_1, \ldots, p_i\}; \{i + 1, \ldots, J\})$ is the expected value on equilibrium for the insurer after negotiating with the first $i$ hospitals and still having all the rest to negotiate with.

3. The value for a patient that would go to hospital $j$ if it was in the network but instead goes to hospital $i \neq j$ because $j$ isn’t in the network is $v_{j,i}$.

4. The number of patients that would go to hospital $j$ if it was in the network but instead go to hospital $i \neq j$ because $j$ isn’t in the network is $q_{j,i}$.

We make the following simplifying assumption: All consumers would leave the insurer if the two top hospitals in their list wouldn’t be in the network. That is, there is no “third choice” hospital. Formally, if only hospital $j$ leaves the network, the insurer loses $q_j - \sum_{k \neq j} q_{j,k}$ patients. If hospitals $j$ and $i$ leave the network, then the insurer loses $q_j - \sum_{k \neq j,i} q_{j,k} + q_i - \sum_{k \neq j,i} q_{i,k}$ patients.

Lemma 2. The expected value for the insurer after negotiating with $i$ of $J$ hospitals
and obtaining prices \(\{p_1, \ldots, p_i\}\) is given by:

\[
V(\{p_1, \ldots, p_i; \{i+1, \ldots, J\}\}) = \sum_{j=1}^{i} q_j(v_j - p_j) + \beta \sum_{k=i+1}^{J} q_{k,j}(v_{k,j} - p_j)
\]

\[+ (1 - \beta) \sum_{j=i+1}^{J} q_j(v_j - c_j) + \beta \sum_{k=i+1, k \neq j}^{J} q_{k,j}(v_{k,j} - c_j) \]

(A.10)

The price for insurer number \(i\) in the order of negotiations is given by:

\[
p_i = \beta q_i v_i + \beta \sum_{k=i+1}^{J} q_k i_k v_{k,i} - \sum_{j=1}^{i-1} q_i,j (v_{i,j} - p_j) - (1 - \beta) \sum_{j=i+1}^{J} q_i,k (v_{i,k} - c_k)
\]

\[q_i + \beta \sum_{k=i+1}^{J} q_i k \]

(A.11)

**Proof.** By backward induction. Suppose the hospitals are ordered 1..\(J\). The value after \(J\) negotiations is:

\[
V(\{p_1, \ldots, p_J; \emptyset\}) = \sum_{j=1}^{J} q_j(v_j - p_j).
\]

When negotiating with the last hospital, the insurer’s outside option is

\[
V(\{p_1, \ldots, p_{J-1}; \emptyset\}) = \sum_{j=1}^{J-1} [q_j(v_j - p_j) + q_{J,j} \cdot (v_{J,j} - p_j)]
\]

The gain from bargaining with \(J\) is therefore

\[
W_J(\{p_1, \ldots, p_{J-1}; \emptyset\}) = V(\{p_1, \ldots, p_{J}; \emptyset\}) - V(\{p_1, \ldots, p_{J-1}; \emptyset\}) = q_J \cdot (v_J - p_J) - \sum_{j=1}^{J-1} q_{J,j} \cdot (v_{J,j} - p_J)
\]

Applying Lemma 1

\[
p_J = \beta \left( v_J - \frac{\sum_{j=1}^{J-1} q_{J,j} (v_{J,j} - p_j)}{q_J} \right) + (1 - \beta)c_J
\]

Then we can place \(p_J\) to obtain the value after negotiating only with the \(J-1\) hospitals:

\[
V(\{p_1, \ldots, p_{J-1}; \{J\}\}) = \sum_{j=1}^{J-1} q_j(v_j - p_j) + q_J v_J - q_J p_J
\]

\[
= \sum_{j=1}^{J-1} [q_j(v_j - p_j) + \beta q_{J,j} (v_{J,j} - p_j)] + (1 - \beta)q_J(v_J - c_J)
\]
We now have a formulation for both $p_J$ and the insurer's value given the first $(J - 1)$ prices.

Moving to the second to last bargaining game, if the bargaining succeeds and price is $p_{J-1}$ the value to the insurer is given in $V\left(\{p_1, ..., p_{J-1}\};\{J\}\right)$ above. If the bargaining fails, we can then continue to the $J$ negotiation step with only the first $(J - 2)$ hospitals, accounting for the patients that would have went to hospital $(J - 1)$ but instead go to either an earlier hospital or to hospital $J$. The resulting value to the insurer is

$$V\left(\{p_1, ..., p_{J-2}\};\{J\}\right) = \sum_{j=1}^{J-2} \left[ q_j(v_j - p_j) + q_{J-1,j} \cdot (v_{J-1,j} - p_j) + \beta q_{J,j}(v_{J,j} - p_j) \right]
+ (1 - \beta) \left[ q_J(v_J - c_J) + q_{J-1,J} \cdot (v_{J-1,J} - c_J) \right].$$

The gain from bargaining with $J - 1$ is therefore:

$$W_{J-1}\left(\{p_1, ..., p_{J-1}\};\{J\}\right) = V\left(\{p_1, ..., p_{J-1}\};\{J\}\right) - V\left(\{p_1, ..., p_{J-2}\};\{J\}\right)
= \sum_{j=1}^{J-2} \left[ q_j(v_j - p_j - 1) + \beta q_{J,j-1}(v_{J,j-1} - p_{J-1}) - (1 - \beta) q_{J-1,j} \cdot (v_{J-1,j} - c_J) \right]
- \sum_{j=1}^{J-2} q_{J-1,j} \cdot (v_{J-1,j} - p_j).$$

Applying Lemma 1:

$$p_{J-1} = \frac{q_{J-1}v_{J-1} + \beta q_{J,j-1}v_{J,j-1} - (1 - \beta) q_{J-1,j}v_{J-1,j} - c_J - \sum_{j=1}^{J-2} q_{J-1,j}v_{J-1,j} - p_j}{q_{J-1} + \beta q_{J,j-1}} + (1 - \beta)c_{J-1}.$$

Placing $p_{J-1}$ in $V\left(\{p_1, ..., p_{J-1}\};\{J\}\right)$:

$$V\left(\{p_1, ..., p_{J-2}\};\{J - 1, J\}\right) = \sum_{j=1}^{J-2} \left[ q_j(v_j - p_j) + \beta \sum_{k=J-1}^{J} q_{k,j}(v_{k,j} - p_j) \right]
+ (1 - \beta) \sum_{j=J-1}^{J} \left[ q_j(v_j - c_j) + \beta \sum_{k=J-1,k \neq j}^{J} q_{k,j}(v_{k,j} - c_j) \right].$$

To generalize, suppose that for a “future hospitals” list of length $J - i$, the expected value
for the insurer given prices \( \{p_1, ..., p_i\} \) is given by:

\[
V(\{p_1, ..., p_i\}; \{i + 1, ..., J\}) = \sum_{j=1}^{i} \left[ q_j(v_j - p_j) + \beta \sum_{k=i+1}^{J} q_{k,j}(v_{k,j} - p_j) \right] \\
+ (1 - \beta) \sum_{j=i+1}^{J} \left[ q_j(v_j - c_j) + \beta \sum_{k=i+1, k \neq j}^{J} q_{k,j}(v_{k,j} - c_j) \right].
\]

Then:

\[
V(\{p_1, ..., p_{i-1}\}; \{i + 1, ..., J\}) = \sum_{j=1}^{i-1} \left[ q_j(v_j - p_j) + q_{i,j}(v_{i,j} - p_j) + \beta \sum_{k=i+1}^{J} q_{k,j}(v_{k,j} - p_j) \right] \\
+ (1 - \beta) \sum_{j=i+1}^{J} \left[ q_j(v_j - c_j) + q_{i,j}(v_{i,j} - c_j) + \beta \sum_{k=i+1, k \neq j}^{J} q_{k,j}(v_{k,j} - c_j) \right].
\]

So the bargaining with \( i \) is over:

\[
V(\{p_1, ..., p_{i-1}\}; \{i + 1, ..., J\}) - V(\{p_1, ..., p_{i-1}\}; \{i + 1, ..., J\}) = q_i(v_i - p_i) + \beta \sum_{k=i+1}^{J} q_{k,i}(v_{k,i} - p_i) - \sum_{j=1}^{i-1} q_{i,j}(v_{i,j} - p_j) - (1 - \beta) \sum_{j=i+1}^{J} q_{i,j}(v_{i,j} - c_j).
\]

Again applying Lemma 1 obtains equation A.11.

We can now derive \( V(\{p_1, ..., p_{i-1}\}; \{i, ..., J\}) \) to complete the recursion. This can be done in two ways (that of course provide the same answer as above): either place \( p_i \) in \( V(\{p_1, ..., p_i\}; \{i + 1, ..., J\}) \) or apply Lemma 1. Here are the steps using the first approach:

\[
V(\{p_1, ..., p_{i-1}\}; \{i, ..., J\}) = \sum_{j=1}^{i} \left[ q_j(v_j - p_j) + \beta \sum_{k=i+1}^{J} q_{k,j}(v_{k,j} - p_j) \right] \\
+ q_i v_i + \beta \sum_{k=i+1}^{J} q_{k,i} v_{k,i} - (q_i + \beta \sum_{i+1}^{J} q_{k,i}) p_i \\
+ (1 - \beta) \sum_{j=i+1}^{J} \left[ q_j(v_j - c_j) + \beta \sum_{k=i+1, k \neq j}^{J} q_{k,j}(v_{k,j} - c_j) \right].
\]
Placing $p_i$ the second line changes to:

$$(1 - \beta)q_i(v_i - c_i) + (1 - \beta)\beta \sum_{k=i+1}^{J} q_{k,i}(v_{k,i} - c_i) + \beta \sum_{j=1}^{i-1} q_{i,j}(v_{i,j} - p_j) + \beta(1 - \beta) \sum_{j=i+1}^{J} q_{i,j}(v_{i,j} - c_j)$$

Moving the $\beta$ term to the first line we get and the $(1 - \beta)$ term to the third line:

$$V(\{p_1, ..., p_{i-1}\}; \{i, ..., J\}) = \sum_{j=1}^{i-1} \left[ q_j(v_j - p_j) + \beta \sum_{k=i+1}^{J} q_{k,j}(v_{k,j} - p_j) \right] + \beta(1 - \beta) \sum_{k=i+1}^{J} q_{k,i}(v_{k,i} - c_i) + \sum_{j=i+1}^{J} q_{i,j}(v_{i,j} - c_j)$$

Collecting the $\beta(1 - \beta)$ terms in the second and third lines obtains the formulation in equation A.10.

For the two hospital case:

$$V(\emptyset, \{1, 2\}) = (1 - \beta) [q_1(v_1 - c_1) + q_2(v_2 - c_2) + \beta (q_{2,1}(v_{2,1} - c_1) + q_{1,2}(v_{1,2} - c_2))] .$$

Setting $F_{1,2} = q_1v_1 + q_2v_2$ and $F_j = q_jv_j + q_{i,j}v_{i,j}$ this exactly equals the result in A.9.

A.4 Proposition 2

Assume the sequential model. For any $\beta \in (0, 1)$, the surplus generated by each insurer is positive.

*Proof. We use the notation of section A.3. Observing equation A.10 the expected value for the insurer before any negotiations is:

$$V(\emptyset, \{1, ..., J\}) = (1 - \beta) \sum_{j=1}^{J} \left[ q_j(v_j - c_j) + \beta \sum_{k\neq j}^{k} q_{k,j}(v_{k,j} - c_j) \right] .$$

By assumption, $v_j > v_{k,j} > c_j$ for all $j$ and so the sum is positive.*
A.5 Proposition 3

Assume the sequential model with two hospitals or with $J$ symmetric hospitals, if $\beta \to 1$, negotiating later in the order increases a hospital’s price and profits.

**Proof.** We use the notation of section A.3. The order of negotiations only affects the price (and not quantity), and so it is necessary and sufficient to prove that the hospital’s price is higher if it is later in the negotiations. We consider only the limit as $\beta \to 1$, in the two hospital setting, $A$’s price if it is first

$$p_A^1 = \frac{q_A v_A + q_{B,A} v_{B,A}}{q_A + q_{B,A}}.$$  

If $A$ is second, it’s price is

$$p_A^2 = \frac{q_A v_A + q_{B,A} (p_B^1 - v_{B,A})}{q_A}.$$  

Then

$$p_A^2 - p_A^1 = \frac{q_A q_{B,A} (v_A - v_{B,A}) + q_{A,B} (q_A + q_{B,A})(p_B^1 - v_{A,B})}{q_A (q_A + q_{B,A})} \tag{A.12}$$

As $v_A > v_{B,A}$ by construction and $p_B^1 > v_{A,B}$ when $\beta \to 1$ (see $p_A^1$ above), the difference must be positive.

For $J$ hospitals, using equation [A.11]

$$\lim_{\beta \to 1} p_j = \frac{q_j v_j + \sum_{k=j+1}^{J} q_{k,j} v_{k,j}}{q_j + \sum_{k=j+1}^{J} q_{k,j}} + \frac{\sum_{k=1}^{j-1} q_{k,j} (p_k - v_{j,k})}{q_j + \sum_{k=j+1}^{J} q_{k,j}}.$$  

The first fraction increases if $j$ increases as $v_j > v_{k,j}$. For the second fraction to increase, a sufficient condition is that $\lim_{j \to 1} p_k \geq v_{j,k}$.

This is immediate by induction:

$$\lim_{\beta \to 1} p_1 = \frac{q_1 v_1 + \sum_{k=2}^{J} q_{k,1} v_{k,1}}{q_j + \sum_{k=2}^{J} q_{k,1}} \in (v_{k,1}, v_1).$$  

For any $j > 1$, if for all $k < j$ it holds that $p_k > v_{j,k}$, then $p_j - v_{j,k}$ is even larger than it is for $p_1$, which completes the proof.[14]

---

[14] In fact, for $J$ symmetric hospitals, suppose that $q_{i,j} = q_j \cdot \alpha$ then using software we find that for every $j > 3$:

$$p_{j+1} - p_j = \frac{(p_j - p_{j-1}) \cdot \frac{1 + \alpha (J - (j + 1))}{1 + \alpha (J - (j - 2))}}{1 + \alpha (J - (j - 2))} \tag{A.13}$$
A.6 Proposition 4

Assume the repeated sequential model. If hospitals have sufficiently large bargaining power, the punishment price converges to the hospital’s per patient value if it was the only hospital in the network: For every $\varepsilon > 0$ there is a $\beta < 1$ such that for all $\beta > \beta$, $|\hat{p}_j - \frac{F_{jh}}{q_j}| < \varepsilon$

Proof. We use the notation of section A.3 Using equation A.11

$$\lim_{\beta \to 1} p_1 = \frac{q_1 v_1 + \sum_{k=2}^{J} q_k v_k}{q_1 + \sum_{k=2}^{J} q_k} = \frac{F_{jh}}{q_j^n}$$

A.7 Proposition 5

In the equilibrium of the repeated sequential model, hospital prices are given by equation 5.2:

$$p_j = \delta \hat{p}_j + (1 - \delta) \frac{W_j}{q_j}.$$  

Proof. Given $J$ hospitals in the network, the insurer’s problem is:

$$\min_{p_1, \ldots, p_J} \sum_{j=1}^{J} q_j p_j$$

s.t. $\forall j: p_j q_j - W_j (1 - \delta) \geq \delta \hat{p}_j q_j$

As $W_j$ is a linear combination of all the prices, the problem is a standard constrained linear optimization problem with $J$ unknowns and $J$ constraints and thus all constraints bind.

A.8 Proposition 6

In the repeated sequential model, if hospitals are sufficiently patient, the insurer’s per-period surplus (or profit) is strictly positive. If hospitals have sufficiently large bargaining power, the insurer’s profit converges to $F_h - \sum_j \left( F_{jh} \cdot \frac{q_j}{q_j^n} \right)$.
Proof. By proposition 5, hospital prices $p_j$ are given in 5.2. If $\delta \to 1$ then this converges to $\hat{p}_j$. Then proposition 4 implies that $p_j < v_j$, proving the first sentence. The first claim follows directly from proposition 2. The second claim follows from proposition 4. \hfill \Box

A.9 Example for section 6

The values for the insurer from a network of either both hospitals ($\{A, B\}$) or just one is given by:

\begin{align*}
V^{A,B} &= \frac{30 - 2p_A - p_B}{3} \\
V^A &= \frac{20 + 5\alpha - p_A(2 + \alpha)}{3} \\
V^B &= \frac{18 + 5\alpha - p_B(2 + \alpha)}{3}
\end{align*}

The differences $V^{A,B} - V^A$ and $V_{A,B} - V^B$ determine price reactions:

\begin{align*}
p^A &= \beta \frac{12 - 5\alpha + (1 + \alpha)p_B}{2} \\
p^B &= \beta (10 + \alpha(p_A - 5))
\end{align*}

Monotonicity requires $p_A \geq p_B$ which, after algebraic simplifications is equivalent to

$$\beta \geq \frac{8 - 5\alpha}{10 - 7\alpha}$$

as stated in the text.

References


