Creaming, skimming and dumping: provider competition on the intensive and extensive margins

Randall P. Ellis *

Economics Department, Boston University, Boston, MA, USA

Received 30 September 1993; revised 27 February 1997; accepted 30 September 1997

Abstract

Reimbursement incentives influence both the intensity of services and who is treated when patients differ in severity of illness. The social optimum is compared to the private Cournot–Nash solution for three provider strategies: creaming—over-provision of services to low severity patients; skimming—under-provision of services to high severity patients; and dumping—the explicit avoidance of high severity patients. Cost-based reimbursement results in overprovision of services (creaming) to all types of patients. Prospectively paid providers cream low severity patients and skim high severity ones. If there is dumping of high severity patients, then there will also be skimming. © 1998 Elsevier Science B.V. All rights reserved.

Keywords: Creaming; Dumping; Skimming

1. Introduction

Given that payment incentives and competitive forces are increasingly being used to influence provider behavior, it is important to understand how payment

* Corresponding author. E-mail: ellisrp@bu.edu.

1 This is a substantially rewritten version of a paper entitled ‘Creaming, Skimming, and Dumping: Provider Competition for Patients’.

0167-6296/98/$19.00 © 1998 Elsevier Science B.V. All rights reserved.

PII S0167-6296(97)00042-8
incentives and competition among providers may influence patterns of treatment. This paper examines the implications of different payment incentives to hospitals that compete for patients who vary in their levels of severity of illness and hence need for health services. Attention is focused on variations in patient severity that are not fully reflected in the payment system, and hence more severely ill patients are assumed to be more costly to treat but receive more benefits from treatment than less severely ill patients. The model captures competition both on the intensity of services provided to patients of a given severity level (the intensive margin) and decisions made by providers about who is treated (the extensive margin).

As reflected in the title, attention is focused on three strategies that providers may adopt in response to reimbursement system incentives: ‘creaming’, the over-provision of services to low cost patients, ‘skimming’, the under-provision of services to high cost patients, and ‘dumping’, the explicit avoidance of high cost patients. Using the first-best and second-best social optimums as a benchmark, the intensity and extent of treatment are modeled for the case of two competing Cournot–Nash health care providers. Levels of treatment are shown to deviate from the social optimum under alternative ways of reimbursing providers: fully prospective payment, traditional ‘cost-based’ reimbursement, and ‘mixed payment systems’ that lie in between. The model is used to derive new insights into how the payment system affects the patterns of treatment when patient heterogeneity cannot be observed by the payer and hence payments cannot reflect patient severity.

Newhouse (1996) highlighted that competition between health care providers that are placed at risk for the full cost of treating patients will take the form of cost reducing/quality enhancing competition, or whether it will take other less desirable forms, such as competition to avoid unprofitable patients, or competition to attract patients for whom the benefits of treatment are less than the costs. This paper highlights that while prospective payment will reduce the quantity of services provided to a given severity level relative to cost-based reimbursement, the decrease is disproportionately on the most severely ill patients, who may be dumped or skimmed. Prospective payment exacerbates the incentive of cost-based reimbursement for providers to compete to attract low severity patients, for whom the benefits of treatment may not exceed the costs.

The model is explicitly designed to reflect competition between hospitals that are competing for patients under prospective payment, a payment system in which hospitals receive a lump sum payment for each patient that depends upon the patient’s diagnosis at time of discharge, but does not depend on the level of services provided to that patient. An example of this type of payments system in the United States is the Medicare program’s prospective payment system, which pays most hospitals on the basis of the patient’s Diagnosis Related Group (DRG)
at time of discharge. The model can also be seen to correspond to competition between Health Maintenance Organizations (HMOs) when they receive a fixed premium that does not vary according to the level of services provided, and to competition between nursing homes that are paid a fixed per diem payment, which does not depend upon the severity of illness of their patients. For clarity of presentation, I focus the discussion on the example of hospital prospective payment.

The current paper builds upon the model used in Ellis and McGuire (1986). In that earlier paper we model the behavior of a single provider whose utility depends upon a weighted sum of provider profits and patient benefits. We demonstrated using a normative model that if the first-best quantity of services can be achieved, then in general it will be through the use of what we called a ‘mixed system’ of reimbursement, which includes both a lump sum payment and a variable payment that is a proportion of cost. Ellis and McGuire (1990) extended Ellis and McGuire (1986) to show that even when patients and providers bargain over the level of services, the optimal payment system should minimize demand side cost sharing and in most cases should be neither fully prospective nor fully cost-based.

An important criticism of Ellis and McGuire (1986) is that it ignores competition. Ellis and McGuire model patients as strongly tied to their provider, so that providers need not worry about losing patients by providing too few services. This paper responds to this criticism, by modeling providers in a market setting in which competition to attract patients is a concern. This paper uses the same provider objective function as in Ellis and McGuire (1986), but adds patient heterogeneity in severity of illness, competition between providers, and provider dumping as a possible strategy.

The paper contributes to two strains of research on provider payment incentives. One strain focuses on how provider payment influences the intensity of health services, sometimes parameterizing intensity as functions of quality or provider effort. Pope (1989), Dranove and Satterthwaite (1992), Hodgkin and McGuire (1994), and Rogerson (1994) each examine the impact of provider payment using models in which providers choose an intensity measure, but do not choose which patients to treat. A second strain of the literature (Feldman and Dowd, 1982; Ellis and McGuire, 1987; Dranove, 1987; Dranove et al., 1992) treats the intensity of treatment as exogenous, and focuses attention on provider selection of who will be treated, i.e., the extent of treatment. This literature shares with the seminal work of Rothschild and Stiglitz (1976) the feature that in some, but not all equilibria, the market is segmented with some but not all consumers being insured or treated.

2 McClellan (1993) presents considerable evidence that the DRG system currently in place in the US is far from fully prospective, since its outlier adjustments and classification using procedures rather than diagnoses are retrospective, not prospectively. Within each DRG, the US system still creates incentives for providers to attract low cost patients and avoid high cost ones to the extent that they can.
In a useful review of provider payment and incentives, Dranove and White (1994) identify two articles which simultaneously consider both the intensity and extent of treatment. Allen and Gertler (1991) model the quality decisions of a single provider facing two types of patients, each of which have price and quality responsive demands. Ma (1994) models provider choice of effort when faced by a continuum of patient types that differ in costliness, and finds that a piecewise linear reimbursement system can achieve the first-best levels of effort even when dumping is a possible strategy. Neither Allen and Gertler nor Ma incorporate provider competition for patients in their models, and neither consider the assumption made here that providers may differentiate the intensity of services across patients of different severities.

The main contribution of the paper is to an understanding of how the payment system jointly influences the intensity of services and the extent of who is treated, in a context where differentiation between patients of different severities is possible. As the previous literature has established, cost-based reimbursement leads to excessive intensity of services to all types of patients. In this paper, cost based reimbursement not only results in too much intensity, but also in too many patients being treated, since cost-based reimbursement encourages providers to compete to attract patients even if the benefits of medical treatment do not exceed the costs. A further distortion is that providers who attach disutility to patient travel costs will tend to oversupply in relation to the second-best social optimum, which takes into account both treatment costs and travel costs.

Paying providers through fully prospective payment rather than using cost-based reimbursement reduces the intensity of services provided to all types of patients in relation to cost-based reimbursement, but leave other problems. Prospective payment exacerbates rather than improves the incentive for health plans to compete to attract profitable, low severity patients. In addition, prospective payments based on the average cost of diverse patients create incentives to undersupply care to high severity patients. If an overall profitability constraint is binding, then providers will dump the most severely ill. A new and interesting finding is that if there is dumping, then there will also be skimping. Finally, although the model is too general to be solved analytically for the optimal payment system, comparative statics demonstrate that a mixed payment system is superior to either fully prospective or fully cost-based reimbursement.

The remainder of this paper is organized as follows. The next section presents the formal model. Section 3 derives and discusses analytical results. Section 4 provides a brief conclusion.

2. Analytical model

The model is a three stage, complete information, noncooperative game in which two health care providers compete to attract patients. At the first stage the
payer chooses the provider reimbursement system. It is assumed that the payer cannot observe patient severity, or the cost of treating individual patients and hence only payments that are a function of total costs and number of patients are examined.

At the second stage two identical competing providers each announce a schedule of services for patients of each severity level. Each competing provider also announces a severity level above which that provider will not treat a patient. This ability of providers to differentiate across severity levels is a key assumption. This differentiation might take the form of hospitals competing by offering more nursing time, more diagnostic testing, or more ‘high tech services’ to the less severely ill in relation to their need. It might also take the form of hospitals not offering certain services that are valued by high severity patients. All of this diversity is collapsed into a single dimension of ‘providing more intensity of services’. Providers do not differentiate across patients according to their travel distance or tastes, but offer the identical treatment to all patients of a given severity.

At the third stage patients select a particular provider after observing each provider’s schedule of services and dumping threshold. Knowing her own severity level, each patient chooses one of the providers so as to maximize the benefits of treatment minus travel costs. Once the patient chooses a provider, the provider treats the patient according to the announced schedule of services.

2.1. Patients

Patients are assumed to be fully insured, so that all costs are paid for by the payer. Patients are assumed to be uniformly distributed over the unit square. The two dimensions of the square correspond to severity of illness, \( s \), and distance, measured in travel time, \( t \), along a line between the two competing providers. Providers are indexed by \( j = 1, 2 \). The level of services provided by provider \( j \) to a patient of type \( s \) is \( X_j(s) \). Providers are assumed not to discriminate across patients according to location, hence treatment offered is independent of patient location. Patient benefits of treatment are written as:

\[
B_j(s, X_j(s))
\]  

(1)

with \( B() \) strictly concave and \( B_x > 0 \) and \( B_s > 0 \)

Independent of patient severity, patients are uniformly distributed along a line of unit length connecting the two providers. Patients are indexed in this dimensions by \( t \), the distance (measured by travel time) from provider 1. Although formally the variable \( t \) will be considered distance, this variable might also be interpreted as patient disutility from seeking a certain style of health care, with the two providers differing along that dimension. Thus the model formally incorporates both vertical and horizontal differentiation between patients.
Patients take into account both the benefits from treatment and travel costs, and choose the provider who gives them the highest total utility. Travel costs per unit of travel time are assumed to be $1/\delta$, and hence patient benefit including travel costs are $B_0 - t/\delta$. A patient of type $s$ located at distance $t = N_i$ from provider 1 will be distance $1 - N_i$ from provider 2. The patient will be just indifferent between the two providers if:

$$B_1 - \frac{N_i}{\delta} = B_2 - \left(1 - N_i\right)/\delta$$

(2)

Eq. (2) does not imply that all patients will be treated by one of the two providers. Low severity consumers will only find it worthwhile seeking treatment if the total benefits from treatment are positive. A patient at location $N_i < 1/2$ will be indifferent between treatment and no treatment if:

$$B_1 - \frac{N_i}{\delta} = 0$$

(3)

Eqs. (2) and (3) define two different demand curves, which depend upon the level of patient severity. These functions are rearranged as Eqs. (4) and (5) below. For low severity patients, each provider can act as a monopolist in choosing the level of treatment, while for high severity patients the two providers interact and will need to act strategically.

**Monopoly:** $N_i = N_i(s, X_1(s)) = \delta B_1$

(4)

**Duopoly:** $N_i = N_i(s, X_1(s), X_2(s)) = \frac{1}{2} + \frac{\delta}{2}(B_1 - B_2)$

(5)

The parameter $\delta$ can be interpreted directly as a measure of the responsiveness of demand to the difference between total benefits offered by the two providers. This is convenient in that even with only a duopoly model, one can examine different degrees of competitiveness according to how sensitive demand is differences in patient benefits. The two limiting cases are $\delta = 0$, in which case demand is perfectly inelastic to patient benefits from treatment; and $\delta \to \infty$, in which case demand is perfectly elastic to treatment benefits. Note that the demand curve facing the monopolist is more responsive to benefits than that facing the duopolists. This has an important bearing on the results below.

### 2.2. Payment system

The payment system parameters are chosen by the payer. As in Ellis and McGuire (1986, 1990), I focus attention payment formulas that are linear functions of the per patient cost of treatment, where it is also assumed that there are no fixed costs or economies of scale or scope across different patient severity levels. Linear reimbursement systems, which include cost-based reimbursement, fully prospec-
tive payment, and the mixed payment system as special cases, are reasonable payment formulas when the payer cannot observe patient severity. Profit from a single patient receiving services $X_j(s)$ from provider $j = 1, 2$ is:

$$\pi_j = R + (r - 1)C(X_j(s))$$  \hspace{1cm} (6)$$

where $\pi_j$ is per patient profits, $R =$ lump sum reimbursement amount, $r =$ marginal reimbursement amount, $0 \leq r \leq 1$, $C(X_j(s)) =$ per patient cost to provider $j$ of level of health services at severity $s$.

The values $R = 0$, $r = 1$ correspond to cost-based reimbursement; $R > 0$, $r = 0$ correspond to a fully prospective system, and $R > 0$, $0 < r < 1$ represent a mixed payment system. $R$ and $r$ are the same for all types of patients.

2.3. Provider objectives

There are two competing providers, $j = 1, 2$. Any patient not treated by these two providers remains untreated, and has a utility normalized to zero. Providers (or their decision-making agents) are assumed to care about profits, $\pi_j$, and patient benefits from treatment, $B_j$ for each patient. Patient travel costs (which might be unobserved by providers) do not enter into the provider utility functions. For notational convenience, the arguments of the behavioral functions will generally be dropped. Following Ellis and McGuire (1986), but using the normalization of Glazer and McGuire (1994), provider $j$’s utility from a single patient of type $s$ is assumed to be:

$$u_j = v(s, X_j(s)) = \alpha B_j + (1 - \alpha)\pi_j$$  \hspace{1cm} (7)$$

where $\alpha$ and $(1 - \alpha)$ are the weights attached to patient benefits and profits, respectively. Provider 1 serves $N_1$ patients of type $s$, hence total utility to provider 1 from serving all patients of type $s$ is $v_1 N_1$, where $N_1$ is as defined in Eq. (4) or Eq. (5), depending on the level of $s$.

2.4. Provider dumping

In order to reduce losses on the most severely ill, one strategy is for providers to skimp on services. On the one hand, this saves money, and deters patients from seeking treatment. On the other hand, even with skimping a hospital may still lose money at the level of utility which maximizes the objective function (Eq. (7)) above.

A second strategy is for providers to avoid treating high cost cases altogether, which is defined here to be ‘dumping’. Although certain forms of dumping are illegal in the United States and many other countries, other forms are not. For example, if a hospital does not offer the more specialized forms of treatment for
high cost cases (e.g., intensive care units, cardiac catheterization units, and distinct psychiatric wards) then a hospital can expect that on average the patients that it attracts will be of low severity. Even in the US, it is acceptable for a provider to turn away a high cost patient on the grounds that necessary specialized facilities are not available.

Dumping is incorporated in the current model in the following manner. Each of the two providers is assumed to be able to take two types of actions for patients of each severity level. They can either treat patients of that severity, in which case the provider utility from that type of patient is as specified in Eq. (7), or they can dump the patients, in which case the provider receives zero utility from patients of that type.

There are many possible criteria that could be used to reflect the dumping decision. If hospitals were pure profit maximizers, then one would expect dumping to be driven solely by the profitability of a single patient. Empirical evidence from Frank and Lave (1989) and Dranove and White (1994) suggests that some hospitals are not pure profit maximizers, that considerable charity and uncompensated care is provided, and that hospitals receiving more donations provide more uncompensated care. Hodgkin and McGuire (1994) also present evidence of income effects: the impact of PPS appears not only to be due to marginal incentives, but also the levels seem to matter. It appears plausible that skimping and dumping decisions may be influenced by the overall profitability of the hospital, not simply the profitability of individual patients.

Reflecting this empirical evidence, the assumption made here is that dumping is motivated by overall hospital profitability, not individual patient profitability or the utility of providers. This assumption is especially plausible if one imagines that it is the hospital administration that decides on admission criteria (i.e., dumping), while doctors decide upon treatment intensity of specific patients. Provider 1 is assumed to dump patients of severity above $s_1^*$, where $s_1^*$ satisfies:

$$\int_{0}^{s_1^*} \pi_i(X_1(s)) N_i(s, X_1(s), X_2(s)) ds = \pi_i^{\min}$$

where $\pi_i^{\min}$ is the minimum profit that Provider 1 requires in order to stay in operation.

In addition to affecting the profitability constraint, the dumping threshold also affects total provider utility. Assuming that the provider gets no utility (or

---

3 A possible alternative assumption is that providers dump only patients for whom the utility of treating is below some threshold (the natural choice is zero). For many benefit functions $B(s, X_i)$ this implies that there will not be any dumping: there will exist some low level of treatment, $X_1$, above which provider utility of treatment is strictly positive. For example, if $B(s,0) = C(s,0) = 0$ and $B_i(s,0) > C_i(s,0)$ for all $s$, then dumping would never occur: providers would simply reduce the level of services until utility was strictly positive.
disutility) from dumped patients, the total utility to provider one of treating all patients of severity type 0 through severity $s^*$ can be written as:

$$V^1 \int_0^{s^*} = v(s,X_1(s)) N_1(s,X_1(s),X_2(s)) ds$$  \(9\)

Provider 1's objective is to choose $s_1^*$ and $X_1(s)$ so as to maximize (Eq. (9)) subject to (Eq. (8)). Provider 2 faces a symmetric problem.

In the discussion of results, I focus on the first order conditions, while assuming that the second order conditions are satisfied, and that well-behaved benefit and cost functions satisfying the usual properties assure that there is only one equilibrium value for the dumping threshold and a unique Cournot output for each severity type. Examination of the second order conditions (not presented here) suggests that as with other Cournot models, assuming that the product of the benefit function and the number of people treated (whether monopoly or duopoly) is strictly concave for all values of $s$, and that total costs of treating each type of patient are strictly convex are sufficient but not necessary conditions for a unique symmetric equilibrium of the constrained private optimum.

3. Results

3.1. First-best social optimum

Social welfare is taken here to be the sum, across severity levels and over all people receiving treatment, of all benefits, $B_0$, minus the treatment costs, $C(X(s))$, minus travel costs. The first-best social optimum can be achieved if the social planner is able to set not only the intensity of treatment for each severity type, but also the maximum distance for which each severity of patient is permitted to travel. Even though the first-best is not in general achievable using any payment system that ignores patient travel costs, it provides a useful benchmark for assessing the extent of over- or under-provision of services. The first-best social objective is assumed to be to choose a dumping threshold, $s^*$, schedule of treatment, $X_1(s)$, and number of people to be treated at each severity level, $N_1(s)$, for each of the two competing providers which maximizes social welfare. Assuming symmetry of the two providers, subscripts for the two individual providers are dropped. Under these assumptions, the first-best social objective function, $V^{FB}$, can be written as:

$$V^{FB} = \max_{s^*,X(s),N(s)} 2 \int_0^{s^*} \left[ \int_0^{N(s)} B(s,X(s)) - C(X(s)) - \frac{t}{\delta} \right] dt \right] ds$$  \(10\)

The solution to this problem is $s^* = 1$ (no dumping) and

$$\begin{cases}  
X^{FB}(s) = 0, N = 0 & \text{for } s \text{ such that } \delta B - \delta C < 0 \\
B_x - C_x = 0, N = \delta B - \delta C & \text{for } s \text{ such that } 0 < \delta B - \delta C \leq 1/2 \\
B_x - C_x = 0, N = 1/2 & \text{for } s \text{ such that } \delta B - \delta C > 1/2 
\end{cases}$$  \(11\)
For low severity patients, no treatment should be provided, regardless of the patient's location. For intermediate levels of severity, only patients for whom benefits exceed the costs should be treated; and for high levels of severity, all patients should be treated. All treated patients should receive treatment up until the point where marginal benefits, $B_x$, equal marginal costs, $C_x$.

3.2. Second-best social optimum

The first-best social optimum is generally not feasible to achieve with fully insured patients, since fully insured patients will be willing to travel for treatment as long total benefits of treatment are greater than travel costs, not only when net benefits exceed travel costs. It is interesting to therefore look at the second best, in which the number of patient seeking treatment is demand-determined rather than chosen by the social planner.

\[
V_{SB} = \max_{s^*, X(s)^0} \int_{0}^{s^*} 2 \left[ B(s, X(s)) - C(X(s)) - \frac{N(s, X(s))}{2\delta} \right] \times N(s, X(s)) ds
\]

Note that $N(s, X(s))$ is defined according to Eqs. (4) and (5). Integrals over distance, $t$, are already been taken and are reflected in $N(t)$. The constant of two outside of the brackets arises because there are two firms in the market, while the two in the denominator of $N$ is there because the average distance travelled is $N/2$, not $N$.

The solution to (Eq. (12)) is relatively straightforward. With regard to the choice of the optimal level of dumping, as long as benefits exceed costs, then no dumping of high cost patients is optimal, and hence $s^{SB} = 1$. The optimal intensity function $X_{1SB}(s)$ has three parts. Letting $B = B(s, X_{1SB}(s))$, $C = C(s, X_{1SB}(s))$, and letting subscripts denote partial derivatives, then $X_{1SB}(s)$ is implicitly defined by the following:

\[
\begin{align*}
X_{SB} &= 0 & \text{for } s \text{ such that } \delta B < \delta C \\
(B_x - C_x) - CB_x &= 0 & \text{for } s \text{ such that } \delta C \leq \delta B \leq 1/2 \\
B_x - C_x &= 0 & \text{for } s \text{ such that } \delta B > 1/2
\end{align*}
\]

The first part, with zero services provided to low severity patients, is appropriate when total benefits of treatment are less than total costs, even as $X \to 0$. The third part of the expression captures cases where severity is such that the patients located midway between the two providers obtain net benefits that are strictly

\footnote{One can find concave benefit functions and convex cost functions such that costs exceed benefits for high severity of illness patients, and hence social dumping is optimal, however such assumptions are uninteresting and will be ignored here.}
positive. In this case travel costs are irrelevant to the choice of services, and marginal benefits should be equated to marginal costs. The second term, which is more interesting, corresponds to cases where travel costs are such that the marginal patient is indifferent between seeking treatment or not. In this case, offering one more unit of health services not only increases treatment costs, but it also increases travel costs. Instead of equating marginal treatment benefits and marginal treatment costs, their difference should be equated to \( C B_s / B \), which is strictly positive. Given extensive insurance and the importance of travel and convenience costs for many health services, this simple point has received surprisingly little attention in the literature.

3.3. The Cournot–Nash solution

The solution of the two providers' optimization problem can be derived once an assumption is made about the strategies taken. I focus on symmetric equilibria in which both providers act as Cournot–Nash competitors. This implies that Provider 1 will choose a dumping threshold \( s_1^* \) and a schedule of services \( X_1(s) \) while assuming \( s_2^* \) and \( X_2(s) \) will remain unchanged. Provider 2 behaves symmetrically. The problem can be set up as a Lagrange multiplier problem.

\[
\begin{align*}
\text{Max} & \quad L_{s_1^*, X_1(s)} \\
& = \int_0^{s_1^*} \left[ \alpha B(s, X_1(s)) + (1 - \alpha) \pi(X_1(s)) \right] N_1(s, X_1(s), X_2(s)) ds \\
& + \lambda \left[ \int_0^{s_1^*} \pi(x_1(s)) N_1(s, X_1(s), X_2(s)) ds - \pi_{1}^{\text{Min}} \right]
\end{align*}
\]  

(14)

Note that the problem is set up so that the Lagrange multiplier will be positive. The Kuhn–Tucker conditions for a maximum of Eq. (14) (presented in Appendix A) are in the form of three pairs of inequalities. For each pair at least one weak inequality must hold with equality. The second row of equations characterizes an entire set of solutions for \( X_1(s) \), not just a single value.\(^5\) In the following discussion, I use abbreviated notation and assume symmetry between the two providers.

\(^5\) Since this solution concept may seem unusual to some readers, note that this is not an optimal control problem in which there is any dependence between neighboring values of \( s \). The problem could equally well have been set up as a discrete rather than a continuous number of patient severity types, and it would then be clear that we are maximizing the utility of each severity type, while imposing one constraint on profitability for the whole set. This alternative arrangement is used by Rogerson (1994), who develops a model of a large number of discrete patient types, without incorporating dumping or provider competition.
3.4. Provider choice of the dumping threshold

Assuming an interior solution for $s^*_t$, the first order condition for the dumping parameter, $s^*_t$, can be solved for the Lagrange multiplier, which can be interpreted as the marginal provider utility of spending one more dollar on treatment of the most severely ill patient:

$$\lambda = \frac{-v_1(s^*_t, X(s^*_t))}{\pi_t(s^*_t, X(s^*_t))} \geq 0$$  \hspace{1cm} (15)

Note that $v_1(0)$ will always be nonnegative at the maximum: if it were negative, then the provider’s objective function could be increased by dumping more patients. If $\lambda$ is to be positive, implying that the profitability constraint is binding, then the denominator must be negative.

Two interesting results follow from Eq. (15). The first result is that only unprofitable patients will be dumped. In order for $\lambda$ to be positive, a provider must be making a loss on the marginal, dumped patient. This is plausible: if the marginal patient is still profitable, then total profits could be increased, and the profitability constraint relaxed if additional patients are treated. The second result follows from the fact that only if $\alpha$ is positive can $v_1$ and $\pi_t$ be of opposite signs. The result is that if both providers are pure profit maximizers (i.e., $\alpha = 0$) then the Nash solution to Eq. (14) will not entail any dumping. Instead, pure profit maximizers will only skim. It is only when there is a divergence between the provider’s utility function and the administrator’s dumping criteria (Eqs. (8) and (9)) that both dumping and skimming can occur.

3.5. Provider choice of the $X(s)$ schedule

The inequality characterizing the optimal choice of the $X(s)$ schedule (Eq. (A2)) after imposing symmetry and using letter subscripts to denote partial derivatives, can be rewritten as:

$$\left[ \alpha B_x + (1 - \alpha + \lambda) \pi_x \right] N + \left[ \alpha B + (1 - \alpha + \lambda) \pi \right] N_x \leq 0$$  \hspace{1cm} (16)

Note that the Lagrange multiplier enters into the equation exactly analogously to $(1 - \alpha)$, the utility weight attached by providers to profits. This way of writing the first order condition highlights that when the profitability constraint is binding, it forces the providers to place greater weight on the profitability of individual patients relative to patient benefits when choosing $X(s)$. Hospitals that are dumping will be more profit oriented than hospitals that are not dumping.

The solution to Eq. (16) depends upon the demand function $N(0)$, which in turn depends upon the value of $s$. In general, three types of solutions to Eq. (16) will appear. For sufficiently low values of $s$, each provider can act as a monopolist, and choose a level of services without worrying about the competitor since travel costs prevent patients from switching. In this case the monopolist can choose $X(s)$
while assuming that the number of patients will be as specified in Eq. (4). For high
values of s, the two providers will compete as Cournot competitors, and choose
output X(s) consistent with Eq. (5). The third set of solutions are those in which
the provider retains some monopoly power, but is partially constrained by the
other provider. In these solutions N1 = N2 = 1/2, and the optimal strategy for the
providers is to offer a level of services such that patients located at t = 1/2 are
just indifferent between seeking care or not seeking care. This implies that output
should be chosen such that Eq. (4) is satisfied with N1 = 1/2. A specific example
is given below for prospective payment.

3.6. Cost-based reimbursement

Under cost-based reimbursement, provider profits are zero regardless of the
level of services provided, and hence providers can act as if they care about only
patient benefits. Unless π\text{min} < 0, there will be no dumping. They will therefore
provide services up to the point where the marginal benefit is zero, and ‘cream’ to
attract all types of patients. Even patients of the lowest severity will be offed
services, and only patients of low severity who are deterred by travel costs (or
tastes) will remain untreated. This type of solution is often argued to correspond to
health care provision in the United States.

3.7. Fully prospective payment

Now consider the provider’s optimal choice of X(s) when payment is fully
prospective. Assume that the profitability constraint is binding so that dumping
occurs. Recall that the monopoly solution, which occurs for low levels of s,
corresponds to the case where consumers are more responsive to the level of
services than the duopoly solution. Somewhat counterintuitively, for low levels of
severity, the monopoly solution will entail higher intensity of services than the
duopoly solutions. For high severity types, the opposite will be true: a monopolist
will provides fewer services than competing duopolists.

Under fully prospective payment, the three types of solutions mentioned above
can occur, and the choice will satisfy the following, which are derived from Eq.
(16) once the appropriate derivatives are substituted, symmetry is assumed, and
non-zero constant multipliers deleted.

\[
\begin{align*}
&\left[2\alpha + \frac{(1 - \alpha + \lambda) \delta(R - C(X(s)))}{\delta B}\right] \\
&B_x - (1 - \alpha + \lambda)C_x = 0 \text{ for } \delta B(s, X^{\text{MON}}(s)) < 1/2 \\
&\delta B = 1/2 \text{ for } \delta B(s, X^{\text{DUO}}(s)) \leq \delta B(s, X^{\text{MON}}(s)) \\
&\left[(1 + \delta B)\alpha + (1 - \alpha + \lambda) \delta(R - C(X))\right] \\
&B_x - (1 - \alpha + \lambda)C_x = 0 \text{ for } \delta B(s, X^{\text{DUO}}(s)) > 1/2
\end{align*}
\]
First compare the monopoly and duopoly solutions, as reflected in Eqs. (17) and (19). The two expressions differ in multipliers for their first terms \((2 > (1 + \delta B))\) and in the extra denominator of the second term \((\delta B < 1)\). Assuming profits of the lowest severity will be positive, then both terms will cause the monopolist to place more importance on marginal benefits, \(B_x\), relative to marginal costs, \(C_x\), and hence to provide more services than the duopolists.

Next consider the implications of Eq. (17) for the very lowest severity types of patients \((s = 0)\). For \(R > 0\), then profits on these patients will be positive. The benefits can be made arbitrarily small in the denominator term, which ensures that the marginal benefit term will have a large positive weight relative to marginal costs. Hence the two duopolists, acting as monopolists for this severity of patients, will wish to provide strictly positive levels of services to the lowest severity type patients. Prospective payment does not cause providers to totally neglect costs, as does cost-based reimbursement, however it still maintains the inefficiency of oversupplying to low severity types. Providers will unambiguously try to ‘cream’ by over-providing services to low severity patients.

Finally, consider the implications of Eq. (19), which corresponds to severity levels where patients located at the midpoint of the distance dimension are being treated, and the two providers are competing to attract these patients. Using the solution for \(\lambda\) as shown in Eq. (15), Eq. (19) can be simplified to the following:

\[
(\alpha) B_x - (1 - \alpha + \lambda) C_x = 0
\]  
(20)

Eq. (20) implies that the first-best social optimum, with \(B_x = C_x\) will only be chosen by the two competing providers if \(\alpha = 1 - \alpha + \lambda\). Although possible, the more likely scenario is for \(\alpha < 1 - \alpha + \lambda\). Two forces will tend to work in this direction. First, it is natural to assume that \(\alpha < (1 - \alpha)\), which is to say that providers attach more weight on profits than patient benefits. (Note that a pure profit maximizing provider will have \((\alpha = 0)\). Second, when the dumping constraint is binding, the positive value of \(\lambda\) will cause providers to place too much weight on costs.

If it is true that \(\alpha < (1 - \alpha - \lambda)\), then this implies that \(B_x > C_x\) at the Cournot solution for the highest severity patient that is treated. This implies that under relatively general conditions if dumping is occurring, then skimping of patients will also be occurring, i.e., providers will under-supply services relative to the social optimum. Even if providers use the correct social welfare weights of \(\alpha = 1 - \alpha = 1/2\), so that benefits and profits are weighted equally in the provider’s utility function, then if the profitability constraint is binding, providers will still skimp on services provided to high cost patients in relation to the social optimum.

3.8. Comparative statistics

The following comparative statistics analysis gives useful insights. Consider first the impact of relaxing the profitability constraint by giving a lump-sum
subsidy to each of the two providers. In the current model, this will tend to increase the services provided to all types of patients. The dumping threshold will increase, so that some patients previously dumped will now be treated. The subsidy will also lower the shadow price of treatment of the marginal dumped patient, \( \lambda \), since the profitability constraint will be less binding. Providers will now find it attractive to compete somewhat more intensely for all types of patients, thereby reducing skimping and increasing overprovision to those who are less severely ill. It is interesting to note that this income effect of the lump sum subsidy only results if providers care about patient benefits as well as profits. If providers are pure profit maximizers, then there is never any dumping (as shown above) and the subsidy will have no effect on the level of services provided.

Consider next the impact of the travel cost parameter, \( \delta \), which is the inverse of travel costs. This parameter is particularly interesting, since it can also be used as an index of the competitiveness between the two providers: High values of \( \delta \) mean that patients will switch providers more readily. I will consider the impact of decreasing travel costs, which is to increase \( \delta \). Recall that providers are assumed not to care directly about travel costs of their patients: they only care to the extent that travel costs influence the number of patients willing to travel for treatment.

Raising \( \delta \) will have no impact on treatment levels under a cost-based reimbursement, although it will increase the number of people seeking treatment for low severity types. Under prospective payment, raising \( \delta \) will have a more complex effect, as can be seen in Eqs. (15), (17)–(19). The first impact is that even with no changes in service offerings, more low severity patients will seek treatment. This will raise provider profits and relax the profitability constraint. Raising \( \delta \) will also lower the severity thresholds at which the providers switch from being monopolists to being partially-competing and then fully-competing Cournot competitors. Since monopolists provide more services to low severity patients than competing Cournot competitors, these changes tend to improve provider profits. The net effect of the changes is likely to be that dumping and skimping of high cost patients will decrease, more patients will seek treatment, and hence low severity overprovision will worsen. For certain intermediate severity patients the level of services may actually fall, where providers who had previously acted as monopolists switch over to providing few services once they are Cournot competitors.

Consider next the impact of increasing \( \alpha \), the weight attached by providers to benefits relative to profits. It can be seen that the direct effect of this change is to increase the weight attached to marginal benefits and lower the weight attached to marginal profits. Hence providers will tend to increase the intensity of service offered to all patients. This will worsen profitability, however, and will tend to require more dumping. Because increased dumping will tend to raise \( \lambda \), the net effect on services to patients of all types is ambiguous.

Finally, consider reimbursement system changes. Specifically, consider a change in the reimbursement system such that initially the system is fully prospective, and
afterwards a mixed payment system is implemented, i.e., one in which some reimbursement is prospective and some cost-based. Assume that there is initially some dumping, and that the payment system change is budget neutral, so that if the same level of treatments were provided to all patients, total profits would be the same. By construction, total profitability would be the same, so the profitability constraint (Eq. (8)) would initially still be exactly binding. The incremental profitability of the marginal dumped patient will have become less negative, however, raising the shadow price of dumping (see Eq. (15)). It will therefore be optimal to dump fewer patients. Hence one direct impact of the payment system change will be to raise the dumping threshold and treat more patients. The second impact of a change toward a mixed system will be to reduce the profit on low severity patients and raise it on high severity patients. This implies that fewer services will be offered to low severity patients and more offered to high severity patients. This last effect will be dampened by the increase in $\lambda$, necessitated in order to afford less dumping. In sum, from an initial system which is fully prospective, a movement toward a mixed system will tend to reduce dumping, reduce creaming, and probably reduce skimming.

An argument can also be made that starting out in a system of cost-based reimbursement, a balanced budget shift toward a mixed system of reimbursement will also be welfare improving. Given that there is no opportunity for providers to further increase the intensity of services, a small amount of cost sharing will reduce the oversupply on high cost patients, and cannot increase intensity for low severity patients. Since this can initially be done with no increase in dumping, then such a shift must be welfare improving.

Although these arguments are only intuitive and may not be fully general, it appears that a shift toward a mixed system from either a fully prospective payment system or a cost-based payment system will be welfare improving. It follows that if here is an optimal payment system, then it will be some type of a mixed system, not either of the two extremes considered here. This appears likely to be a very general result.

4. Discussion and empirical implications

The model used here is too general for deriving the optimal reimbursement system. However it is clear that given private behavior and the social objective function, neither the first-best nor the second-best social optima will in general be achievable. Given that a mixed payment system has only two parameters, ($R$ and $r$) it cannot hope to achieve efficiency in provision and dumping when there are a continuum of patient types and travel costs. This finding is in contrast with Ellis and McGuire (1986), who derive conditions under which a mixed system can achieve the first best.
The more general model used in this paper suggests however that mixed payment systems can still be used to achieve a ‘third best’ solution, even when there is patient diversity, market competition for patients, and patient dumping. Just as with demand side cost sharing only achieving a second best (Zeckhauser, 1970), supply side cost sharing is similarly limited. The ‘third best’ achieved by a (two parameter) mixed system will still be superior to the that achieved by the (one parameter) fully prospective system, and movements away from a fully prospective system appear to be welfare improving.

Is the finding that prospective payment may result in undesirable creaming, skimming, and dumping merely a theoretical possibility or is it one that has been found empirically to occur? Without attempting a complete review of the empirical literature, it is easy to point to real world examples of such behavior. With regard to hospital payment under Medicare’s prospective payment system, Newhouse and Byrne (1988) found that the observed reduction in length of stay at short term, acute care hospitals paid prospectively was more than offset by increases in average length of stay at long term and specialty hospitals that continued to be paid under ‘TEFRA’ or cost-based reimbursement, suggesting serious dumping. Newhouse (1989) presents evidence that Medicare’s prospective payment system increased the proportion of unprofitable patients were treated in public hospitals (dumping). There is growing evidence that Medicare patients are being discharged ‘quicker and sicker’ (skimming). Using Medicaid data, Frank and Lave (1989) present evidence that the distribution of average length of stay has changed in a manner consistent with creaming and skimming, with increasing days of care for less severe patients and decreasing days of care for more severe patients.

Acknowledgements

The author is grateful for comments from Joyce Huber, Albert Ma, Tom McGuire, Joseph Newhouse, Mark Pauly, Michael Riordan, Robert Rosenthal, and participants in the BU/Harvard/MIT Health Economics Seminar and the Northeast Regional Health Economics Conference.

Appendix A. Kuhn Tucker conditions for the solution to Eq. (14)

\[
\frac{\partial L}{\partial s_i^*} = \left[ \alpha B(s_i^*, X_i(s_i^*)) + (1 - \alpha) \pi(X_i(s_i^*)) \right] N_i(s_i^*, X_i(s_i^*)) ,
\]

\[
X_2(s_i^*) + \lambda \pi(X_i(s_i^*)) N_i(s_i^*, X_i(s_i^*), X_2(s_i^*)) \geq 0 \quad s_i^* \leq 1
\]

(A1)
\[
\frac{\partial L}{\partial X_i(s)} = \left[ \alpha \frac{\partial B_i}{\partial X_i} + (1 - \alpha) \frac{\partial \pi_i}{\partial X_i} \right] N_i(s, X_i(s), X_2(s)) + \left[ \alpha B_i(s, X_i(s)) \right] \\
+ (1 - \alpha) \pi_i(X_i(s)) \frac{\partial N_i}{\partial X_i} + \lambda \left[ \frac{\partial \pi_i}{\partial X_i} N_i(s, X_i(s), X_2(s)) \right] \\
+ \pi_i(X_i(s)) \frac{\partial N_i}{\partial X_i} \right] \leq 0, \ X(s) \geq 0 \text{ for } s \in [0,1] \tag{A2}
\]

\[
\frac{\partial L}{\partial \lambda} = \int_0^{s_1} \pi_i(X_i(s)) N_i(s, X_i(s), X_2(s)) ds - \pi_{\text{Min}} \geq 0, \ \lambda \geq 0 \tag{A3}
\]

References


