Quality incentives in a regulated market with imperfect information and switching costs: capitation in general practice

Hugh Gravelle * Giuliano Masiero †

Published in *Journal of Health Economics*, 19 (2000)
Accepted: 1 June 2000

Abstract

We model a system akin to the British National Health Service in which general practictioners (GPs) are paid from general taxation. GPs are horizontally and vertically differentiated and compete via their imperfect observed quality. We focus on the way in which patient uncertainty and switching costs interact and the implications for GP's choice of quality. We show that for any given capitation fee quality is lower and the incentive effects of the fee on quality are smaller. There are diminishing welfare gains from improving consumers information but increasing welfare gains from reducing switching costs. GPs do not act efficiently to improve consumer information via advertising or to reduce the costs of switching.

Keywords: switching costs. Imperfect information. Quality. Product differentiation. Capitation. General practice.
JEL Nos.: I1, L13

*National Primary Care Research and Development Centre, Centre for Health Economics, University of York, Heslington, YO10 5DD; email: hg8@york.ac.uk; fax: 01904 433664; phone 01904 433663. Support from the Department of Health to the NPCRDC is acknowledged. The views expressed are those of the authors and not necessarily those of the Department of Health.

†Doctoral student, Departement of Economics and Related Studies, University of York; email: gm109@york.ac.uk

1 Introduction

In the British National Health Service (NHS) patients join the list of a general practitioner (GP) who is paid a tax financed capitation fee for each registered patient. Care is provided free of charge to the patient. In addition to providing primary care GPs are the gatekeepers for hospital care. As part of the reforms of the NHS in 1990 capitation fees were increased and it was made easier for patients to switch from one GP to another (previously a patient's current GP had to consent formally to the transfer).

The intention of the reforms was to provide greater incentives for GPs to improve the service offered to patients. GPs can vary the quality of their service, for example increasing their surgery opening hours, employing more practice nurses to provide additional services, being more willing to make home visits, or keeping their medical knowledge up to date. Higher capitation fees make it more profitable to attract additional patients by raising the quality of the service provided.

There are three reasons why the reforms might not have the desired effect of improving quality. First, it has been suggested that a patient's choice amongst practice is determined mainly by their distance from the patient's home and that differences in quality will have a very minor impact. This objection does not appear to be valid at either theoretically or empirically. Simple product differentiation models show that, even when all patients choose the nearest practice, practices compete via quality for the marginal patients at the boundary between practices (Gravelle, 1999). Further, most patients do not in fact choose the practice nearest their home (Dixon et al, 1997): their choice of practice appears to be influenced by other practice characteristics including the number of clinics and opening hours which are attributes of practice quality.

Second, patients are unlikely to be very good judges of quality. The extensive literature on doctor-patient agency problems attests to the prevelence of the belief that patients are imperfectly informed about the quality of their doctors. However, it can be argued that many aspects of quality in primary care which may not be obvious when choosing a practice can be judged by patients once they have experienced them. Examples range from the interpersonal aspects of consultations to the ease of getting appointments or out of hours visits. Thus at least some aspects of the practice are experience goods. Since on average patients consult their GPs around six times a year (General Household Survey, 1998) patients may learn about such aspects over time.

But, third, even if patients become better informed about their practice they face costs in switching to another GP. Their new GP will be initially

less well informed about them than their current doctor. Medical records are an imperfect substitute for personal contact and are transferred with a significant delay. Thus in addition to the time and trouble involved in changing registrations, switching to another GP imposes costs in the form of a lower initial level of care *ceteris paribus*.[1]

We investigate these arguments with a simple model which enables us to consider the extent to which switching costs and imperfect patient information about quality interact to blunt incentives for quality. We examine the implications for the level of quality at a given level of the capitation fee, for the incentive to improve quality when capitation is increased and for the welfare maximising quality.

Patient information and switching costs are to some extent endogenous in that GPs can advertise their services and can reduce switching costs by greater effort in acquiring information about new patients when they register. We are also interested in whether competition between GPs leads to appropriate levels of information and switching costs or whether additional regulation is required.

The implications of switching costs and imperfect information for incentives for quality in a regulated market have not to our knowledge been analysed, though there are related papers in the industrial and health economics literature. In a variation of the Salop (1979) circular horizontal differentiation model Economides (1993) shows when there are fixed costs of quality the market equilibrium has inefficient quality. In Gravelle (1999) there are no fixed quality costs but it is shown that quality is efficient only if consumers preferences are weakly separable in distance costs, consumers have zero income elasticity of demand for quality, and firms costs are linear in quantity.

The implications of imperfect consumer information about product quality under horizontal differentiation are considered in Wolinksy (1984, 1986) but consumers have imperfect information about the horizontal characteristics of firms, rather than their vertical quality. Riorden (1986) has variable and imperfectly perceived quality but does not consider switching costs. He shows that when prices act as signals the equilibrium quality tends to the full information solution as the number of firms increases. Bester (1998) uses the Hotelling model, keeping the number of firms fixed, as in the current paper, and examines the implications of imperfectly observed quality for the location rather than their number.

We apply the standard Hotelling horizontal product differentiation model

[1] Using the classification in Nilssen (1992) we are concerned with switching costs that are "transaction costs" since they are incurred on every switch.

to the market for primary care by incorporating switching costs and imperfect information about practice quality. Of the many switching cost models (Klemperer, 1995) ours is perhaps closest to the two period model of Klemperer (1987). We introduce additional features (endogenous product quality, experience goods). Since we are examining a regulated market in which prices (capitation fees received by GPs) are fixed by a regulator and quality is an investment good we can avoid the analytical complications which arise when producers can vary their price from period to period to exploit locked in consumers. We also do not need to consider the implications of prices signalling quality since prices are regulated.

In Section 2 we introduce a two-period model and derive the demand faced by GPs when patients may make errors when initially choosing a GP when young but learn by experience and consider when old whether it is worth incurring switching costs and changing GPs. In Section 3 we examine the equilibrium of a tax financed capitation system with regulated capitation and consider how quality is affected by patients' errors and switching costs. We then consider the implications for the incentive effects of capitation on quality. Section 4 discusses the welfare properties of the regulated market. In section 5 the implications of GPs being able to change patient information and switching costs are examined. Section 6 concludes.

2 The model

GPs receive a capitation payment for each patient who joins their list. The fee is financed from taxation rather than paid directly by patients, and so patients care only about the quality of the practice they join and about its location. Quality is endogenous and determined by an initial investment at the start of the first period and is constant over the two periods.

At the beginning of the first period n patients are located uniformly along a street of unit length. At the end of the period $\gamma^o n$ ($\gamma^o \in [0,1]$) old patients leave the market and a new generation of young patients $\gamma^y n$ ($\gamma^y \in [0,1]$) enters. At the end of the second period all patients leave and none enter.

We assume that preferences, costs and technology are time invariant and such that the market is covered. Full coverage is ensured by assuming that there is a utility r from joining either practice which is independent of the quality of the practice chosen and the patient's location. We assume that r is sufficiently large so that all patients prefer to join some practice rather than none. The *full coverage* assumption is standard in the switching cost literature. We adopt it here to make our results comparable and also because we are interested in the efficiency of GPs' choice of quality. When the market

is not fully covered reductions in quality could drive some patients from the market rather than to the other GP. It is well known (Spence 1975, Economides 1993, Gravelle 1999) that quality and price are inefficient in such circumstances. We wish to separate out the effects of switching costs on the quality of the experience good from other sources of inefficiency.

A practice is located at each end of the street and a patient's location between them determines his preference for the service characteristics. Patient distance $d \in [0,1]$ can be interpreted as geographical distance or as the difference between the level of some horizontally differentiated service characteristic of the practice and the level which would maximize the utility of that particular patient. td is the patient's disutility of being located at a distance d from GP A if he joins that practice. $t(1-d)$ is the distance cost if he joins GP B. Patients are *ex ante* identical except for location and age.

2.1 Patient information

GP i provides a service of quality q_i All patients would agree, if correctly informed, that the GP was providing a more valuable service if q_i increased. Before joining a list a young patient has imperfect information about the quality of both practices and observes the quality provided by GP i with an error $\tilde{e}_i$:

$$\tilde{q}_i = q_i + \tilde{e}_i. \tag{1}$$

An old patient has a perfect knowledge of the quality provided by the list he decided to join in the first period. He does not acquire any information about the quality of the other practice: he makes the same error about it as he did when young.

The errors which young patients make in observing practice quality are identically and independently symmetrically distributed and have zero mean. To keep the analysis tractable we adopt a simple error structure

$$\tilde{e}_i \in \{e_i, -e_i\}, \quad \Pr[\tilde{e}_i = e_i] = \Pr[\tilde{e}_i = -e_i] = 1/2,$$

$$e_i = a_i + m_i q_i, \quad a_i \geq 0, \quad m_i \in [0,1), \quad i = A, B. \tag{2}$$

The formulation allows for both purely additive ($a_i > 0, m_i = 0$) and multiplicative ($a_i = 0, m_i > 0$) errors as well as mixed types. Purely additive errors may be somewhat implausible since they imply that the range of perceived qualities does not vary with actual quality. The assumption is sometimes useful for generating unambiguous results.

For the moment we assume that the parameters in the error distributions are exogenous and the same for both GPs ($a_i = a$, $m_i = m$). We relax the assumption in Section 5.2 where GPs may advertise to inform patients.

Young patients do not realise that their initial observation are subject to error and never expect to revise their beliefs about quality. This assumption simplifies the derivation of the results and is perhaps not unrealistic in the context of health care.

2.2 Demand from young patients

In the first period when all patients are young, a patient is located a distance d from GP A and perceives benefits $r + \tilde{q}_A - td$ and $r + \tilde{q}_B - t(1-d)$ from joining the list of GPs A and B, compared with joining no list. Care is financed from taxation and practices do not charge prices to their patients.

Since young patients do not realise that their perceptions of quality may be in error, a patient will choose GP A rather than B if and only if $\tilde{q}_A - td \geq \tilde{q}_B - t(1-d)$. Hence, of the patients whose realised errors are $(\tilde{e}_A, \tilde{e}_B)$, GP A will get those whose distance from her is no more than

$$\delta(q_A, q_B, \tilde{e}_A, \tilde{e}_B) = \frac{\tilde{q}_A - \tilde{q}_B + t}{2t} = \frac{w + \tilde{e}_A - \tilde{e}_B}{2t}, \tag{3}$$

where $w = q_A - q_B + t$. GP B gets the remainder with $d \in (\delta(q_A, q_B, \tilde{e}_A, \tilde{e}_B), 1]$.

From the distribution assumptions on patients' errors, the realised values of the errors defines four groups of young patients each of size $n/4$ (See Table 3.) For example, patients in group 1 are those who overestimate the quality of both GPs. Demand from patients in the first period in group 1 for GP A is $n\delta_1/4$ where

$$\delta(q_A, q_B, e_A, e_B) = \frac{w + e_A - e_B}{2t} \equiv \delta_1.$$

By contrast the errors made by patients in group 2, who overestimate the quality of GP A and underestimate the quality of GP B, are not offsetting and GP A gets young patients in group 2 whose distance is less than

$$\delta(q_A, q_B, e_A, -e_B) = \frac{w + e_A + e_B}{2t} \equiv \delta_2.$$

GP A gets a larger proportion of group 2 ($\delta_1 > \delta_2$) than GP B because patients overestimate her quality and underestimate the quality of GP B.

Using (3) and making the appropriate substitutions for the error terms for all the patients' groups, the first period demand for GP A is

$$D_1^A = \sum_{j=1}^{4} \frac{n}{4}\delta_j = n\frac{w}{2t}. \tag{4}$$

GP B gets the remainder of the patients: $D_1^B = n - D_1^A$.

Provided that the error distribution is symmetrical and unbiased for each GP the errors made by young patients will be offsetting in total. The size of the error parameters has no effect on the total demand for either GP. But notice that some of the patients in group 2 and 3 whose errors are not offsetting make the wrong choice of GP even when observations of both GPs quality are subject to the same error distribution.

2.3 Second period demand

In the second period a proportion (γ^o) of the first period cohort leaves the market and $\gamma^y n$ new patients enter. Old patients have experienced the service actually provided by the practice they joined in the first period. They now evaluate the quality of that GP correctly. They do not acquire any further information about the quality of the GP they did not choose.[2] Old patients who decide to change to GP i incur a switching cost of s_i. Initially we assume that switching costs are the same for both practices and relax this assumption in Section 5.3.

In period 2 old patients must decide whether to switch GPs. If an old patient originally underestimated the quality of his current GP he will not switch when old. He has now revised the estimate of the quality of the practice chosen upward and has not changed his estimate of the quality of the other GP. Only those patients who overestimated the quality of their GP when young will consider switching to another GP. Patients in groups 2 and 4 who chose GP B when young never switch to GP A when old because they revise their estimate of the quality of GP B upward. Similarly, patients in group 3 and 4 who chose GP A when young never switch to GP B.

For a patient who overestimated the quality of his chosen GP i the perceived gain from practice i compared with practice j falls by $e_i - s_j$ He has revised his estimate of the quality of practice i downward by e_i but if he moves to j he incurs a switching cost s_j. A patient who was just indifferent between the two practices will switch provided that $e_i - s_j > 0$. We assume that switching costs are less than the error parameter so that some old patients who overestimated the quality of their chosen GP will switch in the second period. Other patients who overestimated the quality of their GP may switch but will have had a positive preference for the GP chosen because they are closer to the practice. They will switch if they are not too far from the other GP.

[2] The assumption is a simple tractable case of the more general and plausible assumption that patients learn more about the quality of their current GP than the other GP.

Consider, for example, old group 1 patients of GP A. They now know that they will have a utility from GP A of $r+q_A-td$ and perceive the utility from GP B, net of the cost of switching, as $r+q_B+e_B-s_B-t(1-d)$. Only those group 1 patients of GP A whose distance is greater than

$$\delta_1^{AB}=\frac{w-e_B+s_B}{2t}<\delta_1 \tag{5}$$

will switch to GP B. Since GP A had all of group 1 whose distance was no more than δ_1 she loses $\frac{n}{4}(1-\gamma^o)(\delta_1-\delta_1^{AB})$ of her group 1 old patients to GP B.

GP B will also lose some of her old group 1 patients who overestimated her quality and believe that the increase in quality, net of any change in distance cost, from switching to GP A outweighs the cost of switching: $\tilde{q}_A-td-[q_B-t(1-d)]>s_A$. GP B will loose those group 1 patients for whom

$$d<\delta_1^{BA}=\frac{w+e_A-s_A}{2t}>\delta_1 \tag{6}$$

so that they are close enough to GP A to make a switch worth while. Since GP B had all group 1 patients for whom $d>\delta_1$ the number of old group 1 patients who switch to GP A is $\frac{n}{4}(1-\gamma^o)(\delta_1^{BA}-\delta_1)$.

Proceeding similarly for the patients in group 2 who chose GP A and the patients in group 3 who chose GP B we get Table 3 which shows the number of old patients who switch in and out in each group. Of the group 2 patients only those patients who chose GP A consider switching when old. Those with $d\in(\delta_2^{AB},\delta_2)$ will switch, where $\delta_2^{AB}=(w+e_B+s_B)/2t$. Of the group 3 patients only those who chose GP B consider switching and those who do switch have $d\in(\delta_3,\delta_3^{BA})$ where $\delta_3^{BA}=(w-e_A-s_A)/2t$.

Adding the proportion of old patients who switch in (S^{BA}) and deducting the proportion of those who switch out (S^{AB}) to the installed base ($(1-\gamma^o)D_1^A$) gives the demand for GP A from old patients as

$$\begin{aligned}D_2^{Ao}&=(1-\gamma^o)\left[D_1^A+S^{BA}-S^{AB}\right]\\&=(1-\gamma^o)\left[D_1^A+\frac{n\left(e_B-s_A\right)}{4t}-\frac{n\left(e_A-s_B\right)}{4t}\right]\end{aligned} \tag{7}$$

Note the effects of errors and switching costs on the demand for GP A from old patients:

(a) increases in the error parameter e_A increase the number switching into the list of GP B

(b) increases in the cost of switching to GP A reduce demand from old patients.

Thus, as we will see in section 5, each GP has an incentive to reduce the costs of patients switching in and to provide information to reduce patient errors about her practice.

The new young patients in period 2 behave in the same way as the young patients in period 1. Since quality is the same in the two periods demand from young patients for GP A in period 2 is given by $\gamma^y D_1^A$. Adding (7) gives the total second period demand for GP A

$$D_2^A = D_2^{Ay} + D_2^{Ao} = (1 + \gamma^y - \gamma^o)D_1^A + (1 - \gamma^o)\left(S^{BA} - S^{AB}\right) \tag{8}$$

The demand for GP B is $D_2^B = n(1 + \gamma^y - \gamma^o) - D_2^A$.

3 Regulated market

3.1 Equilibrium quality

The NHS is a regulated market where the tax financed capitation fee (p) per patient on the GP's list is set by the government and patients face a zero price for joining a practice list. Quality is the only way in which GPs can compete for patients. The regulator cannot control quality directly and we are interested in the extent to which he can influence it indirectly via the regulated capitation fee.

GPs have identical cost functions and incur a constant unit cost per patient in each period of c. Practices make an investment in quality at a cost of βq^2 before the young patients in period 1 decide which practice to join. Practice quality is constant over the two periods and is an excludable public good in that its cost is independent of the number of practice patients. Examples are investment by the GP in a computer system for patient records or good practice facilities or in undergoing training (for example in minor surgery).

The discounted expected profit of GP i is

$$V^i = (p - c)(D_1^i + kD_2^i) - \beta q_i^2 + f \tag{9}$$

where $k \in (0, 1]$ is the discount factor on future earnings and f is remuneration which does not vary with the number of patients.[3] We assume that f

[3] In the NHS these include payments related to the age of the GP and the training status of the practice.

is always large enough to ensure non-negative V^i so that the GPs are always willing to participate.

Doctors take their competitor's choices as given and non-cooperatively maximize expected discounted profit by their investment in quality at the beginning of period 1. We consider only pure strategy Nash equilibria and, since the GPs have identical preferences, cost and demand functions, look for a symmetric solution. The obvious way to proceed is set the partial derivative of $V^i(q_i, q_j)$ with respect to q_i equal to zero, impose $q_A = q_B$ and solve for the equilibrium quality $\hat{q}$. Provided that $V^A(\hat{q}, \hat{q}) > V^A(q_A, \hat{q})$ for all $q_A \neq \hat{q}$, and analogously for GP B, the procedure would yield the unique symmetric Nash equilibrium $\hat{q}$. Unfortunately establishing that $V^A(\hat{q}, \hat{q}) > V^A(q_A, \hat{q})$ for all $q_A \neq \hat{q}$, and analogously for GP B, is not straightforward despite the apparent simplicity of the demand functions and the convexity of cost in quality. The demand functions are piecewise linear in quality so that the marginal revenue from quality is a step function. Worse, marginal revenue could be stepwise increasing and then stepwise decreasing, so that the objective function is not concave in quality.

Suppose that when $q_B = \hat{q}$ and GP A sets $q_A = 0$, she gets no young patients, even from group 2 patients who overestimate her quality and underestimate the quality of GP B. Hence $a < \hat{q} - \hat{e} - t$ where $\hat{e} = a + m\hat{q}$, so that even group 2 patients for whom $d = 0$ prefer GP B. As GP A raises her quality above zero with $q_B = \hat{q}$ she at first has no patients but further increase in q_A enable her to capture some of the young group 2 patients who overestimate her quality and underestimate the quality of GP B. Her marginal revenue increases discontinuously at this point. Increases in quality enable her start capturing young patients from groups 1 and 4 who make offsetting errors. Her marginal revenue steps up again. As quality increases further she is able to serve some of the group 3 patients who underestimate her quality and overestimate q_B. Marginal revenue is now at its maximum since she gains patients in all four groups as q_A increases. Eventually she will have all the patients in group 2 and marginal revenue will drop discontinuously. As her quality increases further she will gain all the consumers in groups 1 and 4, leading to a further drop in marginal revenue. Finally she captures all the group 3 and her marginal revenue drops to zero. Thus the marginal revenue from young consumers is stepwise increasing and then decreasing and their demand function has a piecewise linear "S" shape. Allowing for the effect of quality on the numbers of old patients complicates the story but yields the same conclusion.

To avoid the complications resulting from the non-concavity of the GP objective functions we place restrictions on the parameters of the model. These ensure that the marginal revenue function has only downward steps

so that $V^i(q_i, \hat{q})$ is concave in $\hat{q}$ and the symmetric solution is the only Nash equilibrium.

Proposition 1 *There is a unique Nash equilibrium in qualities*

$$\begin{aligned} \hat{q}(p, \cdot) &= \frac{n(p-c)\omega}{4\beta t}, \qquad p \geq c \\ &= 0, \qquad\qquad\quad p < c \end{aligned} \tag{10}$$

where $\omega = 1 + k[(1-\gamma^o)\left(1 - \frac{m}{2}\right) + \gamma^y)]$ *provided that the* uniqueness condition $t > \hat{q}(1+m) + 2a$ *is satisfied.*

Proof 1 *(Sketch) The quality* $\hat{q}$ *defined by* $V^i_{q_i}(\hat{q}, \hat{q}) \leq 0$, $q_i \geq 0, V^i_{q_i}(\hat{q}, \hat{q})q_i = 0$, $i = A, B$ *is the only Nash equilibrium if* V^i *is strictly concave in* q_i*. The uniquenss condition* $t > \hat{q}(1+m) + 2a = \hat{q} + \hat{e} + a$ *means that GP A has young patients in group 3 who underestimate her quality and overestimate the quality of GP B even when* $q_A = 0$*. Hence, she has young patients in all groups at* $q_A = 0$*. The old group of patients who value her services least and are therefore the first to be monopolised by the GP B are those in group 1 who chose her practice when young. At* $q_A = 0$ *GP A will retain some of this group, who now correctly perceive her quality but still overestimate* q_B*, provided that* $r > \hat{q} + \hat{e} - t - s_B$ *which is implied by the uniquenss condition. Hence if the uniqueness condition is satistified GP A will have patients from all groups, young and old, and her marginal revenue from quality increases is constant until she starts to monopolise groups, at which points marginal revenue is discontinuous downward. The same argument applies to GP B. Thus both GPs have identical objective functions which are strictly concave in their own quality. The uniqueness condition can be written as the requirement that a quadratic function* $H(t)$ *is positive. It is easy to show that* $H(t)$ *is convex and achieves a minimum at* $t = a$ *with* $H(a) < 0$*. Hence there exists a* $t_1 > a$ *such that* $H > 0$ *for* $t > t_1$ *and the uniqueness condition is satisfied.*

3.2 Comparative statics

The comparative static properties of the regulated equilibrium are straightforwardly derived from (10)

Proposition 2 *The regulated equilibrium quality is increasing in the proportion of patients who enter the market in the second period* (γ^y)*, the discount factor* (k)*, and decreasing in the proportion of patients who leave the market after the first period* (γ^o)*, the cost of quality* (β)*, distance costs* (t)*, and the multiplicative error* (m)*. Additive errors* (a) *and switching costs* (s) *have no effect on quality.*

Switching costs (s) and additive errors (a) have no effect on equilibrium quality. They enter additively into GPs' demand functions and are equal for the two GPs and therefore offsetting. They do not interact with quality and do not affect the marginal revenue from quality changes and so have no effect on the profit maximizing quality.

If the errors which patients make in judging quality are multiplicative the equilibrium is affected by patient misperceptions: multiplicative errors reduce the equilibrium quality for any given capitation fee. Quality is lower even though, on average, patients estimate quality correctly before they have joined a practice and can observe quality perfectly after experiencing it. The greater the error the more likely are old patients to switch to the other GP (see (7)). Since higher quality leads to greater errors the gain to a GP from increasing quality is reduced.

Quality increases with the size of the total population of consumers, parameterised by n. The cost of quality is independent of the number of patients but marginal revenue from quality increases with n. When there are multiplicative errors equilibrium quality also depends on the mix of young and old patients since the demand from old patients depends on the error and the error increases with quality. Increasing the proportion of old patients $(1-\gamma^o)$ reduces quality because demand from young patients is independent of quality.

Quality is also lower the more consumers care about their distance from the practice ie the larger is their distance cost parameter t. A higher t means that patients place more weight on location relative to quality when comparing practices, thereby reducing practices' incentives to compete via quality.

3.3 Incentive effects of capitation

We can also use (10) to investigate the arguments about the implications of imperfect information and switching costs on the ability of a regulator to influence quality by raising the capitation fee.

Proposition 3 *Increases in the capitation fee increase quality. The marginal effect of the capitation fee on equilibrium quality decreases with multiplicative error (m), distance cost (t), the marginal cost of quality (β), and the proportion of patients leaving the market after the first period (γ^o), increases with the size of the population (n), the provider discount factor (k) and the proportion of new patients entering the market (γ^y), and is unaffected by additive errors and switching costs.*

Increasing capitation fees to make patients more valuable for practices does indeed lead to higher quality even when patients also care about distance, are imperfect judges of quality and face costs in switching when they become better informed. However, the positive impact of the fee on quality is affected by distance costs and multiplicative errors so that a higher capitation fee is necessary for any required level of quality.

One might expect that increases in the capitation fee would increase practice profit and thus ease the participation constraint $\hat{V}^i \geq 0$. In fact substituting the equilibrium quality $\hat{q}(p,\cdot)$ into V^i shows that discounted practice profit is quadratic in p and is decreasing for sufficiently high capitation fee. The apparent exception to the envelope theorem arises because although the capitation fee increases V^i for given q_i it also induces the other practice to increase its quality and increases in q_j reduce V^i.

4 Welfare

The welfare function is

$$W = n(q-c)[1+k(1-\gamma^o+\gamma^y)] - nT - 2\beta q^2 \\ - \lambda\left\{2f + n\left[1+k(1-\gamma^o+\gamma^y)\right]p\right\}$$

where T is the average patient distance and switching cost (to be derived shortly) and λ is the marginal deadweight loss from the taxation required to finance payments to GPs. Welfare is the sum of patients' surpluses and GPs' profits less the cost of taxes levied to finance payments to GPs. Equivalently, since any payments to doctors are exactly offset by payments by taxpayers, welfare is the sum of patients' willingness to pay for the quality of service received less their distance and switching costs, the costs of providing the service and of tax financing the payments to providers.

The welfare function is paternalistic in that welfare is assumed to depend on actual realised patients benefits, not their perceived benefits. It also implies that individuals are not considered the best judges of their own welfare because of their mistaken beliefs.

Since the equilibrium is symmetric with GPs taking the same decisions and we are interested in regulation of those decisions we evaluate the welfare function at $q_A = q_B = q$. Since the market is always covered so that every patient joins a list the total gross benefit to young and old generations is $n[1+k(1-\gamma^o+\gamma^y)]q$,[4] where k is the social discount factor.

[4]Patients get utility of $r+q$ from their practice (gross of distance costs) but since r is

4.1 Distance, error and switching costs

Patients incur distance costs and some of them also incur switching costs. These costs differ with the errors made by patients and with their generation. Patients in group 1 overestimate the quality of both GPs. There are two subgroups defined by the GP chosen. When they are old and have acquired better information about the quality of their current GP some of them switch to the other GP in period 2. There are four subgroups of old group 1 patients defined by the GP chosen when young and whether the patient stays or switches to the other GP.

Patients in group 2 overestimate the quality of GP A and underestimate the quality of GP B when they are young. Their choice of practice defines two young subgroups. When they are old and have acquired information about the quality of the practice chosen those who chose GP A and overestimated her quality may decide to switch to GP B. Those who chose GP B never switch because they revise upward their beliefs about her quality. There are three old subgroups: those who stay with practice A, those who move to practice B and those who stay with practice 2. Similarly there are five subsets of young and old group 3 patients.

Group 4 patients underestimate the qualities of both GPs when young. Since they revise their beliefs about the quality of the chosen practice upward and do not change their beliefs about the quality of the other practice none of them switch when old. There are two subgroups of young and old patients defined by their choice of practice.

The costs incurred by patients who are young in the first and second period are nT^y and $n\gamma^y T^y$ respectively, where T^y the average cost per young patient is[5]

$$T^y(e;t) = \left[\frac{t}{4} + \frac{t}{2}\left(\frac{e}{t}\right)^2\right]. \tag{11}$$

The total costs for old patients are $n(1-\gamma^o)T^o$, where

$$T^o(e,s;t) = \left[\frac{t}{4} + \frac{1}{4t}(e^2+s^2) + \frac{1}{2t}(e-s)s\right]. \tag{12}$$

The first term inside the square brackets in each equation is the distance cost which would be incurred if there was perfect information: patients of each

constant and patients always join some list, we ignore it in the welfare analysis of quality levels.

[5] The details of the derivation of the distance and switching costs for the various groups of young and old patients are in the appendix

generation would choose correctly and would on average be located 1/4 units of distance away from their chosen practice and incur average distance costs of $t/4$.

The second terms are the welfare losses arising from poor information. Some patients choose the wrong GP and incur too great a distance cost. These mismatch costs increase with the errors made by patients. They also increase with the switching costs of old patients which prevent some old patients switching to a GP with smaller distance costs. The third term in T^o is the cost of switching: the proportion of old patients switching multiplied by the cost per switch.

Adding up the costs of young and old patients gives the total per capita discounted distance, error and switching costs as

$$T = T^y + k[\gamma^y T^y + (1 - \gamma^o)T^o] \tag{13}$$

4.2 Costly switching and welfare

The ability of patients to switch GPs has different effects on the welfare of different patient groups. Group 1 patients who overestimate the quality of both GPs are made worse off by the ability to switch. Since their initial errors were offsetting they chose the correct practice when young. Acquiring perfect information about their chosen practice leads them to revise their quality estimate downward and some of them incur switching costs to switch to the other practice. Given that they have overestimated the quality of the GP they switch to, they choose the wrong GP and incur unnecessary distance and switching costs.

Group 2 and 3 patients have reinforcing rather than offsetting errors: they overestimate one GP's quality and underestimate the other GP's. Those who chose the practice whose quality they overestimated consider switching when old and better informed. Those who do switch will in fact be better off since the actual quality at the new doctor is better than their expectations. Indeed, not enough of them switch because of their underestimate of the quality of the alternative practice. Thus group 2 and 3 patients are better off if switching is feasible and not too costly.

Group 4 patients make offsetting underestimates and make the correct choice of GP. Since their experience reinforces their choice they never switch and hence do not gain from the ability to switch.

If patients were not allowed to switch old patients would not incur switching costs but would have mismatch costs equal to those for young patients: $T^o(e, \infty; t) = T^y(e; t)$. Comparison of (11) and (12) shows that $T^y(e; t) > T^o(e, s; t)$ provided that $e > s$. Hence we have

Proposition 4 *Switching is welfare increasing even though imperfect patient information leads some patients to switch a practice which is worse for them.*

4.3 Optimal quality

The regulator cannot observe quality directly but knows the equilibrium quality function $\hat{q}(p,\cdot)$ and chooses the capitation fee p and the lump sum payment f to maximise W subject to the GP participation constraint $\hat{V}^i = V^i(\hat{q},\hat{q}) \geq 0, i = A, B$.

Setting up the Lagrangean $W + \phi\hat{V}^i$ and solving the first order conditions gives the optimal p^* and f^*. From $\hat{q}(p^*,\cdot)$ we have optimal quality

$$\hat{q}^* = \frac{2tn[1 + k(1 - \gamma^o + \gamma^y)] - nm[a\zeta + sk(1 - \gamma^o)]}{(1 + \lambda)8\beta t + nm^2\zeta}. \tag{14}$$

where $\zeta = 2 + k(1 - \gamma^o + 2\gamma^y)$.

Optimal quality is smaller the more costly it is to produce (the greater is β) and the greater the marginal deadweight loss from the taxation required to finance its production. If errors are purely additive ($m = 0$) optimal quality is unaffected by the patient errors, distance costs or switching costs. However, in the more plausible case in which errors vary with the level of quality, the optimal quality is affected by imperfect information and switching costs.

Proposition 5 *Optimal quality is reduced by multiplicative error. If there is multiplicative error quality is smaller the higher is switching cost and the smaller is distance cost*

The marginal social benefit from an increase in quality depends on the gain to patients from increased quality and the marginal cost of producing extra quality. If patients' errors vary with quality, a third factor must be taken into account: increases in quality leads to larger errors and hence larger error costs. Hence, the greater the multiplicative error parameter m, the lower is socially optimal quality. Switching costs also reduce optimal quality when errors are multiplicative because the greater the error the more patients who will make costly switches. The positive effect of distance cost on optimal quality arises because quality becomes relatively less important to consumers in their choice of practice when distance cost is greater. Practices' incentives to increase quality are reduced and so is the amount of error made by patients.

5 Endogenous error and switching costs

5.1 Welfare, error and switching costs

There are welfare gains to reducing patient errors and switching costs. Practices have an incentive to reduce the errors made by patients and the costs they incur in switching to them. This raises the question of whether practices will choose the socially optimal level of a costly activity which reduces errors or switching costs when the capitation fee is set at the level which induces optimal quality.

To simplify the analysis assume that errors are additive ($m = 0$) and that there is no marginal deadweight loss from taxations ($\lambda = 0$). Since $\hat{V}^i(\hat{q}, \hat{q})$ is not affected by the additive error a or switching costs, the participation constraint is neither relaxed or tightened by changes in a or s and their marginal social value arises solely from their effects on the per capita costs of error and switching incurred by patients T, as given in (13). We have

$$T_a = \frac{a\zeta + sk(1-\gamma^o)}{2t} > 0, \tag{15}$$

$$T_s = \frac{k(1-\gamma^o)(a-s)}{2t} > 0, \tag{16}$$

where $\zeta = 2 + k(1 - \gamma^o + 2\gamma^y) > 0$, and $T_{ss} < 0, T_{sa} > 0, T_{aa} > 0$. Not only are the marginal values of reductions in error and switching costs positive but the marginal value of reductions in errors is increasing with the level of switching costs and vice versa. Welfare is concave in the error parameter: the marginal value of error reductions declines with the error. However W is convex in switching costs: the marginal value of a reduction in the cost of a switch is greater the smaller is s. The explanation is that the amount of switching is larger when s is smaller so that total switching costs fall more with a given reduction in the cost per switch.

5.2 Information provision

We first consider the incentives of GPs to improve patient information by advertising (say by providing practice leaflets). The marginal value to a GP of a reduction in error parameter a_i is

$$-(p-c)\left(\frac{\partial D_1^i}{\partial a_i} + k\frac{\partial D_2^i}{\partial a_i}\right) = \frac{nk}{4t}(1-\gamma^o)(p-c). \tag{17}$$

Errors by young patients are symmetric: they have no effect on demand in the first period when all patients are young. However, reducing the error

made by the old patients when young, makes them less likely to switch out when old. Thus the gain from providing better information depends on the number of old patients $n(1-\gamma^o)$.

Suppose that the error parameter for GP i is $a_i = a_0 - \eta h_i$ where h_i is effort in advertising chosen by GPs at the beginning of the first period at a cost of h_i^2. We assume that advertising can be regarded as a public good: the costs of informing potential patients does not vary with the number informed.

Because the expected profit function is separable in advertising and quality, the first order conditions on quality and advertising can be solved independently of each other and the equilibrium level of advertising is [6]

$$\hat{h} = \frac{n\eta k}{8t}(p-c)(1-\gamma^o) \tag{18}$$

The welfare function is, remembering that we assume $\lambda = 0$,

$$W = n(q-c)[1+k(1-\gamma^o+\gamma^y)] - nT - 2\beta q^2 - 2h^2. \tag{19}$$

Using (15) the socially optimal level of advertising is[7]

$$h^* = \frac{n\eta\{a_0\zeta + (1-\gamma^o)sk\}}{8t + n\eta^2\zeta}. \tag{20}$$

The optimal level of advertising is increasing in the switching cost since higher switching costs mean that errors are more costly.

With purely additive error and no deadweight taxation costs, the regulator could achieve a first best level of quality by setting $p^* = c + t$ but substitution in (18) and comparison with (20) shows that

Proposition 6 *It is impossible to achieve optimal information provision and quality solely by regulating the fee.*

The regulator can raise both quality and the amount of information provided by practices by increasing the capitation fee but she will require additional instruments to achieve optimal information and quality. At the fee inducing the socially optimal quality practices have inappropriate incentives for information provision. Comparison of (18) and (20) shows for example

[6] $V^i(q_i, h_i, \hat{q}, \hat{h})$ is strictly concave in q and h and $V^i_{qh} = 0$.

[7] $W(q,h)$ is concave in q and h: $W_{qq} = -4\beta < 0$, $W_{hh} = -\left(4 + \frac{n\eta^2}{2t}\zeta\right) < 0$ and $W_{qq}W_{hh} - W_{qh}W_{hq} = 4\beta\left(4 + \frac{n\eta^2}{2t}\zeta\right) > 0$. Thus, the first order conditions are sufficient to define a maximum.

that GPs will take no account of switching costs or of the error costs of young patients in choosing the amount of effort to put into informing patients.

A first best can only be achieved if the regulator has an additional instrument to influence the level of information provision. For example, as in the NHS GPs are required to provide information about their practices to prospective patients and Health Authorities also provide information directly to patients about the characteristics of practices in their areas.

5.3 Endogenous switching costs

Suppose now that GPs can reduce the costs of patients who switch to them by additional effort. From the demand functions for young (4) and old patients (7), the marginal benefit of a reduction in the cost of switching to GP i is

$$-(p-c)\left(\frac{\partial D_1^i}{\partial s_i}+k\frac{\partial D_2^i}{\partial s_i}\right)=\frac{nk}{4t}(1-\gamma^o)(p-c)>0 \tag{21}$$

which is increasing in the capitation fee, so that the regulator can control s_i via the capitation fee.

When the planner controls both GPs' switching costs s the social value of a reduction in switching costs is

$$-W_s=\frac{nk(1-\gamma^o)(a-s)}{2t}>0 \tag{22}$$

At the capitation fee $p^*=c+t$ which induces socially optimal quality it is apparent that

Proposition 7 *It is impossible to achieve socially optimal quality and effort to reduce switching costs solely by regulating the fee.*

One possible means to reduce patient switching cost is to reimburse some of costs of those who switch. Let σ be the subsidy paid to patients who switch in either direction (as we will see it does not matter whether the subsidy is a reimbursement by the GP or by the regulator). The total distance costs of young patients are unaffected by the level of switching costs. The average social cost of an old patients is

$$T^o=\frac{t}{4}+\frac{1}{4t}\left[e^2+(s-\sigma)^2\right]+\frac{1}{2t}(e-s+\sigma)s \tag{23}$$

instead of (12). σ reduces the distance costs of the additional old patients who are induced to switch (the second term). There is no change in the social

cost per switch s and the increase in the number of switchers therefore raises social cost of switching (the last term). On balance the marginal increase in the social cost of switching outweighs the reduction in distance costs:

$$T^o_\sigma = \left[-\frac{1}{2t}(s-\sigma) + \frac{s}{2t}\right] = \frac{\sigma}{2t} > 0.$$

Since σ is a pure transfer payment its only welfare effect is T^o_σ and so

Proposition 8 *It is always socially sub-optimal to reimburse patients' switching costs.*

There are no externalities in a patient's decision to switch GPs since the patient bears the switching cost and receive the perceived benefit from switching. On average their perceptions of the benefit are correct, though some over-estimate and some underestimate it. Unless a regulator can improve the information available to patients or reduce the social cost per switch the privately optimal switching decisions are also socially optimal.

6 Conclusions

In the market for primary care patients improve their knowledge about the characteristic of the practice they join after experiencing its services. Patients make initial errors in judging quality and switching costs lock some of the mistaken patients into the wrong GP.

It has been suggested in the health services research literature (Salisbury 1989; Thomas, Nicoll and Coleman 1995) that the fact that not many people (around 1.5% per annum) change their GP without a change of address means that GPs do not need to compete for patients. We have shown that patient errors and switching costs do not eliminate the incentives for GPs to increase quality to compete for patients when the regulated fee increases. Moreover this conclusion does not depend on the the proportion of the population switching which in our model could be arbitrarily small.[8]

Regulation of the capitation fee received per patient can yield a welfare maximising level of quality, despite the fact that patients are imperfect judges of quality and incur switching costs. In a system like the NHS where the price of care received by GPs is paid from taxation, errors and switching costs do however have a real effect because they increase the cost to taxpayers of

[8]The proportion of the second period population who switch is $(1-\gamma^o)(e-s)/2t(1-\gamma^o+\gamma^y)$ and our results merely require that $e > s$.

inducing a required level of quality and hence, if there is a deadweight loss from taxation the socially optimal regulated quality is reduced.

Errors and switching costs also have direct welfare consequences. Errors lead some patients to choose the wrong practice. A reduction in the dispersion of the error distribution will make some patients better off and none worse off. Switching costs reduce the number of patients who switch when they revise their estimate of the quality of the GP they have chosen downward. Some of these patients would be better off as result of switching because the other GP really does have higher quality. But some of them will be worse off because they overestimate the quality of the other GP. Thus, although on average patients gain from a reduction in switching costs, some of them are made worse off.

The regulated market in which the regulator's sole instrument is the capitation fee will not lead GPs to choose simultaneously the socially optimal levels of effort to reduce patient errors and quality or switching costs and quality. If the regulated fee is utilized to achieve the socially optimal level of quality, GPs have incentives to reduce patient errors about their own service and to reduce the costs of patients switching to them but these do not reflect the marginal welfare effects. GPs may choose to advertise too much or too little for optimality. Similarly GPs have incentives to reduce the costs of consumers switching to them but these do not reflect the marginal welfare effects. When information and switching costs are endogenous the optimal capitation fee will be a second best compromise balancing its effects on quality, information and effort to reduce switching costs. There will also be scope for additional policy instruments to influence quality, information provision and switching costs.

References

Bester H. "Quality uncertainty mitigates product differentiation", *RAND Journal of Economics*, 1998, 29, 828-844.

Dixon, P. Gravelle, H., Carr-Hill, R. and Posnett, J. *Patient Movements and Patient Choice*, Report for National Health Service Executive, York Health Economics Consortium, December 1997.

Economides N. "Quality variations in the circular model of variety-differentiated products", *Regional Science and Urban Economics*, 1993, 23, 235-257.

General Household Survey, Living in Britain 1995. 1998, HMSO

Gravelle H. "Capitation contracts: access and quality", *Journal of Health Economics*, 18, 1999, 315-340.

Klemperer P. "The competitiveness of markets with switching costs", *The RAND Journal of Economics*, 1987, 18, 138-150.

Klemperer P. "Competition when consumers have switching costs: an overview with applications to industrial organization, macroeconomics and international trade", *Review of Economic Studies*, 1995, 62, 515-539.

Nilssen T. "Two kinds of consumer switching costs", *The RAND Journal of Economics*, 1992, 23, 579-589.

Riorden MH. "Monopolistic competion with experience goods", *Quarterly Journal of Economics*, 1986, 101, 265-279.

Salisbury C. "How do people choose their doctor?", *British Medical Journal*, 1989; 299: 608-610.

Salop SC. "Monopolistic competition with outside goods", *Bell Journal of Economics*, 1979, 10, 141-156.

Schmalensee R. "Product differentiation advantages of pioneering brands", *American Economic Review*, 1982, 72, 349-365.

Thomas K, Nicholl J and Coleman P. "Assessing the outcome of making it easier for patients to change general practictioner: practice characteristics asssociated with patient movements", *British Journal of General Practice*, 1995, 45, 581-586.

Wolinsky A. "Product differentiation with imperfect information", *Review of Economic Studies*, 1984, 53-61.

Wolinksky, A. "True monopolistic competition as a result of imperfect information", *Quarterly Journal of Economics*, 101, 1986, 493-511.

Appendix

Patients' distance and switching costs. The average distance and switching costs of young and old patients in group 1 are

$$\begin{aligned} T_1^y &= \frac{1}{4}\left[\int_0^{\delta_1} t\delta d\delta + \int_{\delta_1}^1 (1-\delta)td\delta\right] \\ T_1^o &= \frac{1}{4}\left\{\int_0^{\delta_1^{AB}} t\delta d\delta + \int_{\delta_1^{AB}}^{\delta_1} \left[(1-\delta)t+s\right] d\delta \right. \\ &\quad \left. + \int_{\delta_1}^{\delta_1^{BA}} (t\delta+s)d\delta + \int_{\delta_1^{BA}}^1 (1-\delta)td\delta\right\}. \end{aligned}$$

T_1^y is the distance costs of the two young subgroups of patients. T_1^o defines the costs of the four old subgroups: those who stay with GP A, those who switch to GP B from GP A, those who switch to GP A from GP B, and those who stay with GP B.

The average costs of group 2 patients in each period are

$$\begin{aligned} T_2^y &= \frac{1}{4}\left[\int_0^{\delta_2} t\delta d\delta + \int_{\delta_2}^1 (1-\delta)td\delta\right] \\ T_2^o &= \frac{1}{4}\left\{\int_0^{\delta_2^{AB}} t\delta d\delta + \int_{\delta_2^{AB}}^{\delta_2} \left[(1-\delta)t+s\right] d\delta + \int_{\delta_2}^1 (1-\delta)td\delta\right\}. \end{aligned}$$

Old patients either stay with the GP chosen when old (the first and third terms in T_2^o) or switch from GP A to GP B (the second term in T_2^o). Given the symmetry assumptions, the costs for group 3 patients who underestimate the quality of GP A and overestimate the quality of GP B is equal to the costs of group 2.

Total distance costs of young and old group 4 patients in each period who remain with the GP chosen when young are

$$T_4^y = T_4^o = \frac{1}{4}\left[\int_0^{\delta_4} t\delta d\delta + \int_{\delta_4}^1 (1-\delta)td\delta\right].$$

Using table 2 we evaluate the integrals to give the expressions for T^y and T^o given in the text.

p	capitation fee
d	patient's distance to GP A
t	cost per unit of distance
q_i	quality of GP i
$\tilde{q}_i = q_i + \tilde{e}_i$	perceived quality of GP i
$\tilde{e}_i \in \{e_i, -e_i\}$	equiprobable errors of uninformed patients
$e_i = a_i + m_i q_i$	error parameter
s_i	cost to patient of switching to GP i
δ_j	market segment of GP A amongst young patients in group $j = 1, ..., 4$
δ_j^{AB}	location of group j patient who is indifferent between switching from GP A to GP B
D_1^i	first period demand for GP A
$D_2^i = D_2^{iy} + D_2^{io}$	second period demand for GP A
V^i	expected discounted profit of GP A
h_i	advertising effort of GP i
$e_i = a_0 - \eta h_i$	effect of advertising on patient errors
λ	marginal deadweight loss from taxation
βq_i^2	cost of quality

Table 1: Notation

Groups	Errors	Critical distance for young patients	Critical distance for old patients who switch to GB A	switch to GP B
1	(e_A, e_B)	$\delta_1 = \frac{w+e_A-e_B}{2t}$	$\delta_1^{BA} = \frac{w+e_A-s_A}{2t}$	$\delta_1^{AB} = \frac{w-e_B+s_B}{2t}$
2	$(e_A, -e_B)$	$\delta_2 = \frac{w+e_A+e_B}{2t}$	$\delta_2^{BA} = \frac{w+e_A-s_A}{2t}$	$\delta_2^{AB} = \frac{w+e_B+s_B}{2t}$
3	$(-e_A, e_B)$	$\delta_3 = \frac{w-e_A-e_B}{2t}$	$\delta_3^{BA} = \frac{w-e_A-s_A}{2t}$	$\delta_3^{AB} = \frac{w-e_B+s_B}{2t}$
4	$(-e_A, -e_B)$	$\delta_4 = \frac{w-e_A+e_B}{2t}$	$\delta_4^{BA} = \frac{w-e_A-s_A}{2t}$	$\delta_4^{AB} = \frac{w+e_B+s_B}{2t}$

Table 2: Distances

Groups	Errors	Demand from young patients	Old patients who switch in	switch out
1	(e_A, e_B)	$\frac{n}{4}\delta_1 = \frac{n}{4}\frac{w+e_A-e_B}{2t}$	$\frac{n(1-\gamma^o)}{4}\frac{(e_B-s_A)}{2t}$	$\frac{n(1-\gamma^o)}{4}\frac{(e_A-s_B)}{2t}$
2	$(e_A, -e_B)$	$\frac{n}{4}\delta_2 = \frac{n}{4}\frac{w+e_A+e_B}{2t}$	0	$\frac{n(1-\gamma^o)}{4}\frac{(e_A-s_B)}{2t}$
3	$(-e_A, e_B)$	$\frac{n}{4}\delta_3 = \frac{n}{4}\frac{w-e_A-e_B}{2t}$	$\frac{n(1-\gamma^o)}{4}\frac{(e_B-s_A)}{2t}$	0
4	$(-e_A, -e_B)$	$\frac{n}{4}\delta_4 = \frac{n}{4}\frac{w-e_A+e_B}{2t}$	0	0
		$D_1^A = n\frac{w}{2t}$	$n(1-\gamma^o)\frac{(e_B-s_A)}{4t}$	$n(1-\gamma^o)\frac{(e_A-s_B)}{4t}$

Table 3: Demand for GP A