Abstract

This paper analyses whether GPs in a capitation system have incentives to provide quality even though health is a credence good. A model is developed where the quality of the service varies due to inherent differences between the GPs and rational patients make choices based on the outcome of treatment.

We find that it is difficult to provide appropriate incentives since the search activity of patients offsets direct effects of a change in reimbursement. Variation in the inherent ability of the GPs is good since it increases the search activity of the patients and the optimal reimbursement scheme is inversely proportional to the dispersion in types. Finally, we find that offering a menu of contracts can potentially increase social welfare above the level of a simple capitation regime, but it tends to lead to a higher variation in quality levels.

© 2006 Elsevier B.V. All rights reserved.

JEL classification: I11; I18; D43

Keywords: General practice; Quality; Asymmetric information

1. Introduction

This paper analyses the effects of competition between general practitioners (GPs) within a public health care system. In the 1990s, a common reform in countries with national health service (NHS) systems, was to provide GPs with incentives to compete for patients by introducing a capitation payment in the reimbursement system, and to encourage patients to shop around among the GPs. This was an element of the extensive changes in the NHS of the United Kingdom that were implemented in 1991 (Koen, 2000). Also in Sweden, where the health care system has many
similarities with the NHS, a so-called family doctor reform was introduced by the centre-right government in 1994. The reform changed the environment in which GPs operate from a relatively rigid system of salaried physicians to whom a certain number of patients are assigned (based on vicinity) into a system based on capitation (adjusted for age and gender). The reform was later partially reversed, but the most important features of the reform have been kept more or less constant in many regions (Anell, 1996).

So why should GPs be allowed and encouraged to compete for patients? The desirability of competition between GPs can be analysed in various dimensions. For instance, attaching a high value to consumer sovereignty per se speaks in favour of such a system. If, on the other hand, equality is a main concern of the policy maker, then introducing competition might be connected with problems, since it is difficult to design a reimbursement formula that completely compensates for heterogeneity among the patients. Hence, competition may very well give incentives to cream skimming and skimping which are not present in a salary-based system. One main rationale of the reforms, however, was the presumption that they would increase quality in health care. This effect can come about in at least two ways. One possibility is that quality is regarded as a fixed attribute of the provider, in which case competition may drive less productive GPs out of business (cf. McGuire, 1983). Alternatively, quality is endogenous and the result of investment on the part of the GPs. Then the effect of competition could be stronger incentives to invest in quality.

This paper focuses on the latter effect. We analyse how a policy-maker can use competition among GPs to provide them with proper incentives to deliver high quality services. Doing this means that the regulator uses the demand of a GP as the main source of information to evaluate their performance. In particular, we seek to answer three questions. The first question is how the reimbursement system should be designed to achieve optimal levels of quality. Secondly, we want to analyse the welfare losses of a capitation system compared to a first best setting where the regulator observes the quality level of different providers directly and pays them accordingly. This question is of great importance since it tells us whether it is worthwhile for the regulator to consider alternative regulatory tools. Thirdly, we compare the effects of a linear capitation rate with more complex reimbursement rules to find out how ‘costly’ the use of a linear capitation rate is. As a main alternative, we consider the situation where the regulator can offer the GPs a menu of linear contracts. Such a reimbursement scheme is typically more efficient but it is also connected with a higher implicit cost since it requires communication between the regulator and all GPs. Hence, our analysis aims at giving the policy maker some guidance in the decision of how far to go in terms of contractual complexity.

Our analysis suggests that GP competition might be a problematic instrument for assuring quality in primary care. We find that changes in consumer behaviour tend to offset the incentives provided by a capitation regime, and that this indirect effect sometimes dominates the direct effect completely. Furthermore, we find that equilibrium quality levels depend on the dispersion in ability of GPs, even if all players involved are risk neutral. Consequently, the optimal capitation rate has to take the diversity of providers into account, and should in fact be inversely proportional to the dispersion in ability. Hence, variation in quality, which is normally perceived as a negative thing, will indeed have beneficial effects on competition. Finally, we find that welfare can be improved by offering a menu of contracts. Again, the welfare gains depend crucially on the distribution of ability. These gains must of course be weighted against the potential administrative costs connected with running a more complex reimbursement scheme.

The rest of the paper is structured as follows. In the next section, we position our paper in the health economics and in the more general IO literature in an attempt to justify and explain the most important modelling choices. Thereafter, we present and formalise the fundamentals of the
model. In the following section, we analyse equilibria in the subgame between the GPs and their patients, and propose intuitions behind their properties. The fifth section turns to the normative issue of finding appropriate reimbursement schemes for the providers. The sixth and last section concludes.

2. Modelling considerations

Health care is different from other goods and services. As pointed out already by Arrow (1963), the market for health care services is a market where virtually none of the assumptions of the fundamental welfare theorems hold. One main complication is the informational asymmetry between the provider and the patient. First of all health care is a service which means that, as soon as there are quality differences, it is an experience good (Nelson, 1970). But, more importantly, there is an informational asymmetry as to whether the treatment provided was actually needed and accurate for the actual condition of the patient, and this asymmetry tends to persist even after the service has been consumed. This means that we are somewhere in the continuum between experience goods and credence goods.\(^1\)

The information asymmetry concerning the appropriate quantity of care has triggered an abundant literature on supplier induced demand (Evans, 1974; Wilensky and Rossiter, 1983; Tussing, 1983; Reinhardt, 1985; Rice and Labelle, 1989; Grytten et al., 1995). The main idea in this literature is that the physician will induce or reduce demand for his services in the pursuit of a certain target income. The theory has been tested many times with weak if any support of the predictions. A more serious criticism, however, is that it is difficult to distinguish the model from a situation where providers have market power as they are observationally equivalent (McGuire, 2000). A contribution by De Jaegher and Jegers (2001) address this problem by modelling demand inducement as a cheap talk game.

In this paper, we assume that demand inducement is not a problem, or that it has been taken care of by compensating the GP for the marginal cost of a visit. Instead, we model the problem as one of moral hazard, where patients have imperfect information concerning the endogenously chosen quality levels. This problem has previously been analysed by Gravelle and Masiero (2000), who model a situation where patients observe the quality of the service only after consuming it, whereas the prior perception of quality is a noisy observation. The authors conclude that the informational asymmetry does not necessarily prevent the regulator from achieving a first best quality level. Just as in Gravelle and Masiero (2000), we assume the quality levels to be fixed throughout the game. Hence, our model applies to a setting where success of treatment depends on fixed attributes (such as the GPs knowledge) rather than decisions which are taken at each visit to the surgery (such as the number of tests to take).

The Gravelle and Masiero model shows that multiplicative noise in patients’ perceptions potentially impedes the possibility to arrive at a first best solution, whereas switching costs and additive noise does not. In this paper, we study the same problem with a slightly different approach. Firstly, we assume realistically that the information problem is not completely resolved after the patient

\(^1\) There is a growing literature on markets for credence goods (cf. (Dulleck and Kerschbamer, 2006) for an overview). However, this literature is different from the approach taken here since prices are assumed to be an instrument of the providers and not of the regulator. It should be noted that the information asymmetry is not the only reason why competition might not achieve a higher level of social welfare than monopoly. Other sources of inefficiency are e.g. that providers may have intrinsic motivations that are hampered by monetary incentives, or competition may give rise to selection effects that reduce social welfare.
has visited the GP. Furthermore, the Gravelle and Masiero model requires irrational patients: once the patients realise that their beliefs are marred with systematic errors, they should rationally trust their accurate knowledge of the GP’s technology and not their inaccurate perceptions of quality. This paper, in contrast, treats the service as a credence good, where the patient is rational but uncertain of the quality even after having consumed it.

We develop the model as a game with three types of players: GPs, patients and the regulator. There are two GPs in a Hotelling style model that compete for patients. It is common in health economics to attribute objectives to health care providers which do not conform with the neoclassical paradigm (the already mentioned target income approach is one example, outright altruism is another one). In this paper, we stick to the mainstream assumption of profit maximisation, but claim instead, as argued above, that the main difference between health care and other services is in the information structure.

We assume that each GP makes a hidden investment decision at the beginning of the game and that this decision materialises as the subsequent probability of providing successful treatment. The same probability applies to all patients, but the individual draws are independent. GPs differ, however, in their ability to make this investment. A topical example of this is when new technical equipment is introduced in primary care. To some GPs it may be easy to acquire the skills necessary to operate the equipment, whereas others find it more difficult. This difference in ability then naturally translates into a difference in the success of treatment. Other examples would be things such as “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” (Gravelle and Masiero, 2000).

Our model does not allow for any variable to be used by GP’s as a signal of their quality level. The most natural candidate would be the price, which is, however, set by the regulator. Furthermore, it is a well established result in industrial organisation that there are no signalling equilibria when the level of quality is endogenous (Tirole, 1988). Instead, contractual performance has to be induced by means of reputation (Klein and Leffler, 1981; Mailath and Samuelson, 2001).

Concerning patients, we assume that they are a priori identical, with the exception of their location (representing physical distance or taste, for example). Patients are also assumed to be rational utility maximisers who use their limited information on the quality levels of the GPs to find their preferred provider. Hence, the main driving force behind patient movements is their expectations and their learning. We assume that the game goes on for two time periods and that patients visit the GP once in each time period. Since the average number of visits to GPs vary between 3 and 7 per year, depending on age group (Bajekal, 1998), this assumption might seem a bit conservative. There is no need, however, to assume that one period in the game corresponds to a calendar year; it might even be the case that the whole game takes place within one calendar year. The restriction to two periods only is mainly done for the sake of simplicity, but since we assume that the quality level remains constant for the duration of the game, it is reasonable to keep a relatively short time horizon.

Contrary to Gravelle and Masiero (2000), we do not allow for switching costs. The literature on switching costs (von Weizsäcker, 1984; Klemperer, 1987; Nilssen, 1992) has shown that they can have important implications for the market equilibrium and consumer welfare, and Gravelle and Masiero show that switching costs do reduce welfare on average (but not in each individual case). Still, there are good reasons not to analyse switching costs in the context of GPs competing for patients. Firstly, the most important effects observed by Klemperer (1987), Nilssen (1992) are due to firms changing their pricing strategies in order to exploit consumer loyalty. Secondly, for the context of health care there are reasons to believe that
learning costs (i.e., costs related to acquiring information about the provider) are more important than transaction costs (i.e., costs incurred at every switch), and the former type of costs have been shown by Nilssen (1992) to have smaller effects on provider strategies and welfare. Thirdly, and most importantly, it should be remembered that the Hotelling model, where all consumers incur transportation costs, could be interpreted as a model of switching cost.

The regulator, finally, has the objective of maximizing the value of the service less the transfers to the GPs. This transfer can be based on the number of patients attracted to a certain surgery only. The regulator’s problem is one of hidden action and hidden information: they observe neither the investment decision of the GP, nor their innate ability. The distribution of abilities is known, however. A further complication is that the regulator must take the responses of the patients into account when designing the reimbursement system.

Having motivated our modelling approach, we now turn to a formal presentation of the model.

3. The model

The game we consider consists of three types of players: the regulator, two GPs and a continuum of patients with mass normalised to one. The timing of the game is as follows. At the outset, the regulator announces a reimbursement scheme consisting of a capitation rate \( r \) and a base salary (or practice allowance) \( b \). Then the two GPs observe the reimbursement scheme and choose a level of investment in quality \( p \). After that, the patients decide to join the list of one of the two providers. They visit the GP once and then update their beliefs concerning their quality levels and then finally decide whether to stay with the current GP or switch to the other one.

We will now go through the three players of the model and outline their objectives, their strategy spaces and their knowledge.

3.1. The general practitioners

We assume that there are two general practitioners (GPs), situated at the opposite extremes of a line of unit length. Upon observing the reimbursement scheme announced by the regulator, GP \( i \) (where \( i = 0 (1) \) denotes the GP to the left (right)) decides on a quality level, \( p_i \in [0, 1] \) which is the probability of successful treatment that applies for the entire duration of the game and equally to all patients (but the actual outcome is an independent draw for each patient). The GP is a profit maximiser and the profit function is

\[
\Pi_i = D_i(p_i, p_j)r - c_i(p_i; \theta) + b
\]

where \( D_i(p_i, p_j) \) is total demand (hence adding up demand from both periods) when the GPs choose strategies (\( p_i, p_j \)), \( r \) is the capitation rate decided by the regulator and \( b \) the base salary, and \( c_i(p_i; \theta) \) is the cost of quality (assumed to be convex and continuously differentiable with respect to \( p_i \)). The parameter \( \theta \) is the type of the GP, which we refer to as ability henceforth. Only the individual GP knows his or her ability with certainty. To all the other players, including the competitor, it is a random variable with density function \( f(\theta) \) and support \( \Theta = [\theta, \bar{\theta}] \).

The following definition is useful for our analysis.

**Definition 1.** We refer to the variance-to-mean ratio of a variable \( \theta \), \( \Psi(\theta) = \text{Var}[\theta]/\text{E}[\theta] \) as the dispersion of that variable.
3.2. The patients

In each round of the game, patients visit at most one of the GPs for treatment. Utility from both periods carry equal weight. In each period \( j \) they choose an action \( a_j \in \{0, 1\} \) derive utility from treatment according to the function:

\[
u_j(d, a_j) = vI_j - t(d + a_j - 2a_jd)
\]

where \( j \in \{1, 2\} \) is the time index, \( d \) is the distance of the patient from the GP to the left, \( a_j \) is the GP chosen by the patient in that round (GP 0 is the GP to the left and GP 1 is the GP to the right), \( t \) is a travelling cost parameter and \( I_j \in \{0, 1\} \) indicates whether the treatment of GP \( a \) was successful or not (an event that occurs with probability \( p_a \)). Patients are furthermore characterised by their beliefs concerning the quality levels offered by the two providers. Patients know the distribution of ability \( f(\theta) \), but not the actual quality level of the two physicians at their disposal.

3.3. The regulator

We assume for simplicity that the regulator seeks to maximise the expected value of the service less the transfer to the providers (which is augmented by a loading factor representing distortionary taxation). Hence, the social welfare function is

\[
W = v \int_{\theta \in \Theta} f(\theta) p(\theta) d\theta - (1 + \lambda)(r + b)
\]

where \( \lambda \) is the loading factor. Furthermore, we assume that the regulator has to respect a participation constraint of the providers in the sense that the expected profit of each type must be at least zero.

4. The market equilibrium

In this section, we analyse the comparative statics of the equilibrium of the subgame between the GPs and the patients. Our equilibrium concept is perfect Bayesian equilibrium, according to the following definition.

**Definition 2.** A perfect Bayesian equilibrium is a set of provider actions \( p^* = (p_0^*(\theta), p_1^*(\theta)) \), consumer beliefs \( \hat{p} = (\hat{p}_0(\theta), \hat{p}_1(\theta)) \) and consumer strategies \( a_\tau(\mu_\tau(p_1), \mu_\tau(p_2), d) \) satisfying the following conditions:

\[
p_i^*(\theta) \in \arg \max \Pi_i(p_i, p_j, \hat{p}, a_\tau(\mu_\tau(p_1), \mu_\tau(p_2), d); \theta) \quad \forall \theta \in \Theta, \ i = 0, 1, \tau = 1, 2, j \neq i
\]

\[
\hat{p} = p^*
\]

\[
a_\tau(\mu_\tau(p_1), \mu_\tau(p_2), d) \in \arg \max E[u_\tau|\mu_\tau(p_1), \mu_\tau(p_2), d], \quad \tau = 1, 2
\]

\[
\mu_i(p_i) = \int_{\theta \in \Theta} f(\theta) \hat{p}_i(\theta) d\theta, \quad i = 0, 1
\]
The equilibrium conditions spelled out in Definition 2 state that providers maximise profits with respect to their quality levels $p_i$ (4), that patients hold rational beliefs $\hat{p}$ concerning the quality levels of the various types (5), and that they choose a provider $a_t \in \{0, 1\}$ in order to maximise their expected utility from treatment, given their beliefs and their location $d$ (6). The conditions (7) and (8) assure that patients hold rational expectations and that they update their beliefs according to Bayes’ formula. Since the patients have no further information concerning the quality of the GP not visited in the first round, the belief concerning that GP cannot change.

Remark 3. As long as $d \hat{p}_i(\theta)/d\theta \neq 0$, we will have $\mu_2(p_i|a_1 = i, 1) > \mu_1(p_i) - \mu_2(p_i|a_1 = i, 0)$, that is, patients who experience a good (bad) outcome increase (decrease) their expectations.

4.1. Patient strategies

We restrict attention to cases where all patients visit a GP so that the market is entirely covered. In the first round, this boils down to comparing prior beliefs concerning the quality levels of the two providers. Denoting by $d^*$ the patient who is indifferent between the two providers in the first round, we have:

$$d^* = \frac{v(\mu_1(p_0) - \mu_1(p_1)) + t}{2t} = \frac{v\int_{\theta \in \Theta} f(\theta)(\hat{p}_0(\theta) - \hat{p}_1(\theta)) d\theta + t}{2t}$$

which is very similar to the condition in the original Hotelling model. In the second round, patients update their beliefs. It is obvious by Remark 3 that patients who experienced a positive outcome get their belief reinforced and thus will not change their decision. Hence, the potential searchers are among those who experienced a failure. This gives us two different ‘indifferent consumers’. The left one is defined by

$$d^*_0 = \frac{v(\mu_2(p_0|a_1 = 0, I_1 = 0) - \mu_1(p_1))}{t} = \frac{1}{2} + \frac{v\left(E_0[(1 - \hat{p}_0)\hat{p}_0] - E_0[\hat{p}_1]E[1 - \hat{p}_0]\right)}{2t}$$

where $E_0$ refers to the prior expectation, i.e. $E_0[p] = \int f(\theta)p(\theta) d\theta$. Hence, the point of indifference between the two providers now depends on how the updated belief compares with the prior belief. The second equality is obtained by inserting Eqs. (7) and (8) from the definition of equilibrium. Similarly, the indifferent consumer to the right is now:

$$d^*_1 = \frac{v(\mu_2(p_1|a_1 = 1, I_1 = 0) - \mu_1(p_0))}{t} = \frac{1}{2} + \frac{v(E_0[\hat{p}_0]E[1 - \hat{p}_1] - E_0[(1 - \hat{p}_1)\hat{p}_1])}{2t}$$

A pictorial representation of the expected utility as a function of location is given in Fig. 1. The curve labelled $E[u_{11}(d, 0)]$ shows how the utility from visiting the GP to the left in the first round depends on location, and $E[u_{11}(d, 1)]$ shows the corresponding utility from visiting the GP to the right. The curves $E[u_{21}(d, 0)|1]$ and $E[u_{21}(d, 0)|0]$ show the corresponding utility derived from
visiting the GP to the left after having had a positive or negative experience in the first round. Accordingly, the locations $d^*$ and $d_0^*$ are points of indifference.

Hence, we can derive a demand function for the GP to the left in the case where the market is fully covered:

$$D_0(p_0, p_1) = d^* + p_0 d^* + (1 - p_0) d_0^* + (1 - p_1) (d_1^* - d^*)$$

$$= 1 + p_0 (d^* - d_0^*) + p_1 (d^* - d_1^*)$$

(12)

Since we assume the market being fully covered, the GP to the right get the remaining consumers: $D_1(p_0, p_1) = 2 - D_0(p_0, p_1)$.2

4.2. Provider strategies

Inserting the demand function (12) into the profit function gives us the maximisation problem of the GP to the left3:

$$\max_{p_0} \Pi_0(p_0, p_1, \theta) = (1 + p_0 (d^* - d_0^*) + p_1 (d^* - d_1^*)) r - c_0(p_0; \theta) + b,$$

s.t.

$$p_0 \in [0, 1]$$

(P)

Hence, the Lagrangian function of this problem reads:

$$\mathcal{L} = (1 + p_0 (d^* - d_0^*) + p_1 (d^* - d_1^*)) r - c_0(p_0; \theta) + b - \delta_1 p_0 + \delta_2 (p_0 - 1)$$

(12)

where $\delta_1$ and $\delta_2$ are the Lagrange multipliers associated with the constraints $p_0 \in [0, 1]$. In the sequel, we make use of the shorthand variables $\Delta_0 = (d^* - d_0^*)$ and $\Delta_1 = (d^* - d_1^*)$ for patients switching away from the left and the right GP, respectively. Notice that in a symmetric equilibrium, $\Delta_0 = \Delta_1$ so that where no confusion arises we will use the general notation $\Delta$ in such a case and

---

2 Notice that each patients visit a GP twice so the total demand in both periods equals 2.

3 Again, this maximisation problem is just a mirror image of the maximisation problem of the GP to the right.


707

refer to these patients as ‘searchers’. Furthermore,

\[ \Delta = v \int f(\theta) p(\theta) (p(\theta) - \int f(\theta) p(\theta) d\theta) d\theta \] 
\[ = \frac{v \text{Var}[p(\theta)]}{2t \int f(\theta)(1 - p(\theta)) d\theta} \]

(13)

We now separately derive equilibrium quality levels for two different specifications of the cost function: quadratic and hyperbolic costs.

4.2.1. Profit maximisation with quadratic costs

In this part, we assume that the cost function takes the shape:

\[ c_i(p_t; \theta) = \frac{\theta(p_t)^2}{2} \]

(14)

Then, the first order conditions of the maximisation problem are

\[ \frac{\partial L}{\partial p_0} = r\Delta_0 - \theta p_0 + \delta_1 - \delta_2 \leq 0 \] 
\[ \delta_1, \delta_2 \in \mathbb{R}_+ \]
\[ \delta_1 p_0 = 0 \]
\[ \delta_2(p_0 - 1) = 0 \]

(15) (16) (17) (18)

The following Lemma rules out a certain type of equilibria.

**Lemma 4.** There is no equilibrium where only a subset of the type support \( \Theta \) has positive quality levels.

**Proof.** Suppose that some types invest and some do not. Then \( r\Delta_0 > 0 \) since the updated belief (8) will be lower than the prior (7). But this implies that \( \partial L/\partial p_0|_{p_0=0} > 0 \) which means that \( p_0 = 0 \) cannot be an optimum for any \( \theta \).

Hence, an equilibrium will either be degenerate in the sense that no investment is undertaken, or it will have all GPs choosing positive levels of quality investments. Next, we characterise the unique equilibrium which is an interior solution to (P).

**Proposition 5.** For reimbursement rates:

\[ r \geq \frac{2t(1 - v\theta E(1/\theta))}{v\theta \text{Var}(1/\theta)} \]

(19)

there exists an equilibrium where the equilibrium quality level \( p^*(\theta) \) is interior. It is decreasing in the reimbursement rate \( r \), increasing in the travelling cost \( t \), decreasing in the variance of the type distribution and in the expected value of the type parameter \( \theta \).

**Proof.** In the interior solution to (13) the first order condition for an optimum (15) reduces to \( p^*(\theta) = r\Delta_0 \theta^{-1} \). Furthermore, we have by (13) that \( r\Delta_0 = rv(\text{Var}[\tilde{p}^*])(2t(1 - E[\tilde{p}^*]))^{-1} \). Replacing in (15) and solving for \( p^*(\theta) \) yields:

\[ p^*(\theta) = 2t \left( v\theta \left( 2tE \left[ \frac{1}{\theta} \right] + r\text{Var} \left[ \frac{1}{\theta} \right] \right) \right)^{-1} \]

(20)
The second derivative of the GP’s objective function (P) is
\[
\frac{\partial^2 L}{\partial p_0^2} = -\theta
\] (21)
so this is indeed a maximum. Finally, requiring \( p^*(\theta) \leq 1 \forall \theta \) gives condition (19). □

4.2.2. Profit maximisation with hyperbolic costs

Now we instead consider the alternative assumption that the cost function in (P) takes the shape:
\[
c_i(p_i; \theta) = \frac{\theta p_i}{1 - p_i}
\] (22)
In this case we obviously need not worry about the possibility that some types choose \( p_i = 1 \) and hence the Lagrangian reads:
\[
\frac{\partial L}{\partial p_0} = r\Delta_0 - \frac{\theta}{(1 - p_0)^2} + \delta_1 \leq 0
\] (23)
\[
\delta_1 \in \mathbb{R}_+
\] (24)
\[
\delta_1 p_0 = 0
\] (25)
It can be shown that there are potentially multiple equilibria in this case. Only one of them, however, has real-valued and positive quality levels. That equilibrium is characterised below.

**Proposition 6.** For reimbursement rates:
\[
r > \frac{2t(\bar{\theta})^3}{v\Psi(\sqrt{\theta})}
\] (26)
there is an equilibrium where the quality \( p^*(\theta) \) level is interior and increasing in the reimbursement rate \( r \), decreasing in the travelling cost \( t \) and increasing in the dispersion of abilities.

**Proof.** In the interior case the first order condition (23) reduces to \( p^*(\theta) = 1 - \sqrt{\theta/r\Delta} \). Furthermore, we have from (13) that
\[
\Delta = \frac{v(\text{Var}[\hat{p}^*])}{(2t(1 - E[\hat{p}^*]))} = \left( \frac{v\Psi(\sqrt{\theta})}{2t^{1/2}} \right)^{2/3}
\]
Replacing and solving for \( p^*(\theta) \) yields:
\[
p^*(\theta) = 1 - \sqrt{\theta \left( \frac{2t}{rv\Psi(\sqrt{\theta})} \right)^{1/3}}
\] (27)
Finally, requiring \( p^*(\theta) > 0 \forall \theta \) gives condition (26). □

4.2.3. Finding an intuition

Thus, we have found that a seemingly unimportant alteration of the cost function can reverse the comparative statics of the model completely. Besides, for a quadratic cost specification we get the counter-intuitive result that equilibrium quality is decreasing in the reimbursement rate \( r \). What is the intuition behind this result? It transpires that the counterintuitive equilibrium is the
result of a rather complex interaction between the actions of the providers and the beliefs of the patients. This effect will be explored in more detail in this section.

Clearly, the sensitivity of the result depends on the credence good character of the service. Due to their ignorance of actual quality levels, consumers will base their decisions on, amongst other things, the variance of quality within the population of providers. The shape of the cost function determines how much a change in the reimbursement rate translates into a change in the variance, and this in turn has implications for the consumers’ response in their search activity. Formally, notice that the unique interior optima of the GP’s problems are (in the case of quadratic and hyperbolic costs, respectively):

\[ p^{q*}(\theta) = \frac{r\Delta}{\theta} \]  

(28)

\[ p^{h*}(\theta) = 1 - \sqrt{\frac{\theta}{r\Delta}} \]  

(29)

Since the number of searchers (\( \Delta \)) also is a function of \( r \), the total differential is of (28) and (29) is

\[ \frac{dp^*(\theta)}{dr} = \frac{\partial p^*(\theta)}{\partial r} \frac{d\Delta}{dr} + \frac{\partial p^*(\theta)}{\partial \Delta} \frac{d\Delta}{dr} = (1 + \eta_{\Delta r}) \frac{\partial p^*(\theta)}{\partial r} \]  

(30)

where \( \eta_{\Delta r} \) is the elasticity of the number of searchers with respect to the reimbursement rate. Eq. (30) holds for both specifications. Furthermore, the number of searchers equals:

\[ \Delta = \frac{v \text{Var}[p(\theta)]}{2t \text{E}[1 - p(\theta)]} \]  

(31)

In equilibrium with a quadratic cost function, this amount is equal to

\[ \Delta^q = \frac{2t}{r(v \text{Var}[1/\theta] + 2t \text{E}[1/\theta])} \]  

(32)

and the corresponding elasticity is

\[ \eta^q_{\Delta r} = -1 - \frac{rv \text{Var}[1/\theta]}{rv \text{Var}[1/\theta] + 2t \text{E}[1/\theta]} < -1 \]  

(33)

hence, with a quadratic cost function, the response to a change in the reimbursement rate is very strong and completely offsets the direct effect. For the hyperbolic function, the equilibrium number of searchers is

\[ \Delta^h = \left( \frac{v \Psi(\sqrt{\theta})}{2t} \right)^{2/3} r^{-1/3} \]  

(34)

Hence, the corresponding elasticity \( \eta^h_{\Delta r} \) is constant and equal to \(-1/3\).

We have found that the elasticity of consumer switching with respect to the reimbursement rate is negative: the reactions of the consumers to a change tends to offset the direct effect on the marginal revenue of the providers. In both specifications, an increase in the reimbursement rate leads to a reduction in the variance of provider quality, which in turn discourages consumer search. Only with a quadratic cost function, however, is this effect strong enough to offset the direct effect.

We have here encountered a phenomenon that is known in the literature as a Bertrand supertrap; a case where the strategic effect (the reaction of consumers) not only counteracts but also dominates
the direct effect (Cabral and Villas-Boas, 2005). However, since the hyperbolic function seems to be more reasonable for the problem under consideration (perfection should be infinitively costly), and since the regulatory problem is trivial in the case where quality decreases in the reimbursement rate, we restrict our attention to the hyperbolic function in what follows.

Now we have characterised equilibria in the subgame between GPs and patients. In the next section, we turn to the normative issue of optimal reimbursement.

5. Normative analysis

In this section, we analyse the normative issues involved. We do this in two different ways. Firstly, we look at the comparative statics of the subgame between GPs and patients, and compare with the benchmark of perfect information. In this part, it is also useful to refer to the results of Gravelle and Masiero (2000) for comparison, since the implications are quite different. Secondly, we analyse the regulatory problem by deriving an optimal capitation rate and contrasting it with the alternative of offering a menu of contracts.

5.1. Perfect versus imperfect information

First, we contrast the equilibrium quality levels with and without perfect information, and then we contrast with the findings of Gravelle and Masiero (2000).

Consider the equilibrium quality level in the subgame between GPs and patients, with imperfect information:

\[ p^* (\theta) = 1 - \sqrt{\theta} \left( \frac{2t}{rv\Psi(\sqrt{\theta})} \right)^{1/3}. \]

(35)

Solving for the equivalent perfect information game is trivial. The corresponding quality level is

\[ p^{PI} (\theta) = 1 - \sqrt{\theta} \left( \frac{t}{rv} \right)^{1/3}. \]

(36)

This way, we arrive at the following proposition:

**Proposition 7.** The quality loss from imperfect information is increasing in GP ability, but decreasing in the dispersion of ability (\( \sqrt{\theta} \)).

**Proof.** Taking the difference between imperfect information quality (35) and full information quality (36) and rearranging yields:

\[ p^{PI} (\theta) - p^* (\theta) = \sqrt{\theta} \left( \frac{t}{rv} \right)^{1/3} \left( \left( \frac{2}{\Psi(\sqrt{\theta})} \right)^{1/3} - \left( \frac{t}{rv} \right)^{1/6} \right). \]

(37)

which clearly increases in \( \sqrt{\theta} \) and decreases in \( \Psi(\sqrt{\theta}) \). □

We now contrast these findings with the results of Gravelle and Masiero (2000). There are three main differences between their model and ours: firstly, in their model, the range of the quality variable is assumed to be unbounded and with equal realisations for every patient. Patients have imperfect information, however, since they observe the actual quality level only with a noise. Gravelle and Masiero find that additive noise does not change the equilibrium quality levels, whereas multiplicative noise does. Hence, we focus on the multiplicative case below. A third
difference is that there is no heterogeneity amongst providers in their model so the parameter \( \theta \) is the same for all providers.

First, the equilibrium quality level with imperfect information is

\[
q^* = \frac{r(1 - m/4)}{2\theta} 
\]

where \( m \) is the coefficient of the multiplicative noise. The corresponding quality level with perfect information is

\[
q_{PI} = \frac{r}{2\theta} 
\]

So we can quantify the difference as

\[
q_{PI} - q^* = \frac{rm}{8\theta}. 
\]

Accordingly, the quality difference in this case is *increasing* in the dispersion of the perceived quality levels; and *decreasing* in the ability of the GP.

The result of Gravelle and Masiero requires some qualification, however, since it relies on the assumption that patients do not realise that their quality perceptions are biased. In our model, on the contrary, patients are fully rational and make efficient use of all the information available when they make their decision. Our finding that the dispersion of ability, \( \Psi(\sqrt{\theta}) \), is such an important parameter, comes from the fact that consumer search is triggered by variance in quality levels. It indicates that the regulator is facing an interesting dilemma when designing the reimbursement system, since variance in quality levels is generally seen as something bad in the health care sector.

Now we turn to a detailed analysis of the regulator’s problem.

5.2. Benchmark: the first best

Henceforth, we define the first best as a situation where the regulator can observe the individual GP abilities \( \theta \). In the first best, the optimisation problem is

\[
\max_{p(\theta)} \int f(\theta)(vp(\theta) - (1 + \lambda)b(\theta))d\theta, \quad \text{subject to } E[\Pi(p^*(\theta), \theta)] \geq 0 \ \forall \theta \in \Theta \tag{R1}
\]

where the expectation in \( E[\Pi(p^*(\theta), \theta)] \) is taken with respect to the distribution of the competitor’s type. The solution to this problem is one where the regulator gives the GPs zero expected profits and equates the marginal benefits of investment with marginal costs.

**Proposition 8.** The first best level of quality is

\[
p^{FB}(\theta) = 1 - \sqrt{\theta(1 + \lambda)v^{-1}} 
\]

which gives a social welfare of

\[
W^{FB} = v - 2E[\sqrt{\theta}]\sqrt{v(1 + \lambda) + (1 + \lambda)E[\theta]}.
\]

Furthermore, we define the second best as a scenario where \( \theta \) is not observed but the regulator can offer a menu of contracts \((r(\theta), b(\theta))\). The third best, finally, is when the regulator cannot observe \( \theta \) and is furthermore constrained to offer a single capitation contract \((r, b)\) only.
The quality level can be induced by a contract offering the GPs a transfer:

\[ b(\theta) = \sqrt{\theta} (\sqrt{v(1 + \lambda)} - \sqrt{\theta}). \]  

(43)

**Proof.** This is straightforward from the first order conditions applying to (R1). □

5.3. The optimal capitation formula

Now we introduce asymmetric information (i.e. the regulator cannot observe \( \theta \)) and assume that the instruments at the regulator’s disposal are a uniform capitation rate \( r \) and a uniform base salary \( b \) only. In this case, the regulator’s maximisation problem is

\[
\max_{b,r} v \int f(\theta) p^*(\theta) \, d\theta - (1 + \lambda)(b + r), \quad \text{subject to } E[\Pi(p^*(\theta), \theta)] \geq 0 \quad \forall \theta \in \Theta (R3)
\]

The following Lemma is a standard result from Principal-Agent models.

**Lemma 9.** Provider profits are decreasing in \( \theta \).

**Proof.** Differentiating profits with respect to the type parameter yields:

\[ \frac{d\Pi(p^*(\theta), \theta)}{d\theta} = \frac{\partial \Pi(p^*(\theta), \theta)}{\partial \theta} + \frac{\partial \Pi(p^*(\theta), \theta)}{\partial p^*(\theta)} \frac{dp^*(\theta)}{d\theta} \]  

(44)

and by the envelope theorem the second term is equal to zero. The first term, \( \partial \Pi(p^*(\theta), \theta)/\partial \theta = -(p^*(\theta)/1 - p^*(\theta)) \) is indeed negative for all \( p^*(\theta) \in (0, 1) \). □

It follows from Lemma 9 the participation constraint is binding for the worst type only, and, since profits are costly to the regulator, the profits of this type will be kept at zero. Furthermore, since the total payments to the two GP’s always equals \( 2(b + r) \) we have:

\[ r + b = \int_{\Theta} f(\theta)(\Pi(p^*(\theta), \theta) + c(p^*(\theta); \theta)) \, d\theta \]  

(45)

Replacing in the regulator’s problem (R3) yields:

\[
\max_r v \int_{\Theta} f(\theta) p(\theta) \, d\theta - (1 + \lambda) \left( \int_{\Theta} f(\theta)(\Pi(p(\theta), \theta) + c^*(p(\theta); \theta)) \, d\theta \right) \] 

subject to \( \Pi(p^*(\bar{\theta}), \bar{\theta}) = 0 \) (R3’)

The solution to the regulator’s problem (R3’) is given in the following proposition.

**Proposition 10.** When the regulator does not observe the cost parameter of the providers, the optimal capitation formula is

\[ r^* = \left( \frac{v}{1 + \lambda} \right)^{3/2} \left( \frac{2t}{v \Psi(\sqrt{\theta})} \right) E(\sqrt{\theta}) \left( \int_{\Theta} \frac{(F(\theta) + \theta f(\theta))}{\sqrt{\theta}} \, d\theta \right)^{-1} \]  

(45)

which induces a quality level of

\[ p^C(\theta) = 1 - \sqrt{\theta} \left( \int_{\Theta} \frac{(F(\theta) + \theta f(\theta))}{\sqrt{\theta}} \, d\theta \right) \left( \frac{1 + \lambda}{v E(\sqrt{\theta})} \right) \]  

(46)
With this reimbursement scheme, the regulator can achieve social welfare equal to

\[ W^C = v - 2\sqrt{(1 + \lambda)v} \sqrt{E(\sqrt{\theta})} \int_\Theta \frac{(F(\theta) + \theta f(\theta))}{\sqrt{\theta}} \, d\theta \]

\[ + (1 + \lambda) \int_\Theta (F(\theta) + \theta f(\theta)) \, d\theta. \]  

(47)

**Proof.** First notice that by setting the profit of the worst type equal to 0 and having profit following the law of motion (Lemma 9):

\[ \dot{\Pi} = -p^*(\theta) \left( \frac{1}{1 - p^*(\theta)} \right) \]  

(48)

expected profits equal:

\[ \Pi(\theta) = \int_{\theta}^{\hat{\theta}} \frac{p^*(\theta)}{1 - p^*(\theta)} \, ds \]  

(49)

Integrating this expression by parts, we get expected profits equal to

\[ \int_\Theta f(\theta)\Pi(\theta) \, d\theta = \int_\Theta F(\theta) \frac{p^*(\theta)}{1 - p^*(\theta)} \, d\theta \]  

(50)

Combining with expected costs, the total transfer to GPs is (according to (45)):

\[ r + b = \int_\Theta (F(\theta) + \theta f(\theta)) \left( \frac{p^*(\theta)}{1 - p^*(\theta)} \right) \, d\theta \]  

(51)

Inserting in the objective function (R3′) we get the maximand:

\[ W^C = v \int_\Theta f(\theta)p^*(\theta, r) \, d\theta - (1 + \lambda) \int_\Theta (F(\theta) + \theta f(\theta)) \frac{p^*(\theta, r)}{1 - p^*(\theta, r)} \, d\theta \]  

(52)

Recalling from Section 4.2.2 that

\[ p^*(\theta) = 1 - \sqrt{\theta} \left( \frac{2t}{rv\Psi(\sqrt{\theta})} \right)^{1/3} \]  

(53)

it is easily confirmed that \( p^*(\theta) \) is concave in \( r \) and that the objective function is concave in \( p^*(\theta) \). Hence, the first order condition is necessary and sufficient for a maximum. Taking the FOC and applying the chain rule yields:

\[ \frac{\partial W^C}{\partial r} = v \int_\Theta f(\theta) \frac{\partial p^*(\theta, r)}{\partial r} \, d\theta - (1 + \lambda) \int_\Theta \frac{(F(\theta) + \theta f(\theta))(\partial p^*(\theta, r)/\partial r)}{(1 - p^*(\theta, r))^2} \, d\theta = 0 \]  

(54)

Inserting the derivative of the quality level with respect to the capitation rate:

\[ \frac{\partial p^*(\theta, r)}{\partial r} = \frac{\sqrt{\theta}}{3} \left( \frac{2t}{r^4v\Psi(\sqrt{\theta})} \right)^{1/3} \]  

(55)

the solutions stated above are achieved. □

Hence, we have found that the optimal capitation rate is increasing in travelling cost \( t \), in the value of the service, \( v \), and decreasing in the shadow cost of public funds, \( \lambda \) and the dispersion of
ability, $\Psi(\sqrt{\theta})$. The optimal quality level, however, is independent of the travelling cost, and so is the welfare level attained in this equilibrium.

5.4. The second best: introducing communication

Suppose instead that the GPs are asked to state their type parameter before the game begins and that the regulator prescribes a certain quality level $p(\hat{\theta})$ where $\hat{\theta}$ is the stated type. Since the demand of a certain GP is uniquely determined by their own and their competitor’s investment level, the regulator can observe from the realised demand whether one of the GPs (or both) have deviated from the prescribed quality level. Hence the contract could stipulate that a GP who reports type $\hat{\theta}$ should make quality investment $p(\hat{\theta})$ and attract a number of patients $D(\hat{\theta}_i, \hat{\theta}_j)$. We also make the assumption that any deviation from this realisation of demand can be punished by an arbitrarily large punishment.

Under these assumptions, the problem is a standard hidden information problem, and we need to consider solutions that satisfy incentive compatibility as well as participation constraints. The incentive compatibility constraint is

$$\theta \in \arg \max \frac{1}{\hat{\theta}} \Pi(\theta, p(\hat{\theta})) = \arg \max \frac{1}{\hat{\theta}} T(\hat{\theta}) - c(\theta, p(\hat{\theta})) \quad (56)$$

where $T(\hat{\theta})$ is the transfer granted to a GP who states type $\hat{\theta}$. These IC constraints require the first order condition:

$$\dot{T}(\theta) - c_2(\theta, p(\theta))p(\theta) = 0 \quad (57)$$

(where $c_2$ denotes differentiation with respect to the second argument of the cost function) and hence:

$$\dot{\Pi}(\theta) = -c_1(\theta, p(\theta)) \quad (58)$$

Thus, profits are again decreasing in $\theta$ which means that the participation constraint of the worst type is binding. The Hamiltonian of this problem reads:

$$H = f(\theta)(vp(\theta) - (1 + \lambda)(\Pi(\theta) + c(\theta))) - \delta(\theta)c_1(\theta, p(\theta)) \quad (59)$$

where $\delta(\theta)$ is the Pontryagin multiplier for the law of motion of the state variable. Solving this dynamic optimisation problem leads us to the next proposition:

**Proposition 11.** When communication between the GPs and the regulator is allowed, the optimal contract entails a quality level:

$$p_{II}(\theta) = 1 - \frac{1 + \lambda}{v} \frac{f(\theta)\theta + F(\theta)}{f(\theta)} \quad (60)$$

With this contract, a welfare level of

$$W_{II} = v - 2\sqrt{v(1 + \lambda)} \int_{\theta} \sqrt{f(\theta)(f(\theta)\theta + F(\theta))}d\theta + (1 + \lambda) \int_{\theta} (F(\theta) + \theta f(\theta)) d\theta \quad (61)$$

is attained.
Proof. The optimality conditions are
\[
\frac{\partial H}{\partial p(\theta)} = f(\theta) \left( v - (1 + \lambda) \frac{\theta}{(1 - p(\theta))^2} \right) - \delta(\theta) \frac{1}{(1 - p(\theta))^2} = 0
\] (62)
\[
\dot{\Pi}(\theta) = -c_1(\theta, p(\theta))
\] (63)
\[
\frac{\partial H}{\partial \Pi(\theta)} = -f(\theta)(1 + \lambda) = -\dot{\delta}(\theta)
\] (64)

Hence, the Pontryagin multiplier is
\[
\delta(\theta) = \int_{\theta}^{\tilde{\theta}} f(\theta)(1 + \lambda) = (1 + \lambda)F(\theta)
\] (65)

Inserting (65) in (62) and solving for \( p(\theta) \) gives (60). Welfare is given by
\[
W^{\Pi} = v \int_\theta^{\tilde{\theta}} f(\theta)p^{\Pi}(\theta) d\theta - (1 + \lambda) \int_\theta^{\tilde{\theta}} f(\theta)(\Pi(\theta) + c(\theta)) d\theta.
\] (66)

Inserting:
\[
\Pi(\theta) = \int_{s=\theta}^{\tilde{\theta}} \frac{p(\theta)}{1 - p(\theta)} ds
\] (67)

from Eq. (58) above, and using integration by parts, we get:
\[
W^{\Pi} = v \int_\theta^{\tilde{\theta}} f(\theta)p^{\Pi}(\theta) d\theta - (1 + \lambda) \int_\theta^{\tilde{\theta}} (F(\theta) + \theta f(\theta)) \frac{p^{\Pi}(\theta)}{1 - p^{\Pi}(\theta)} d\theta
\] (68)

Finally, inserting the value \( p^{\Pi}(\tilde{\theta}) \) gives the welfare level in (61). □

It seems unlikely, however, that a contract of the kind just described would be offered in reality. Firstly, there is probably uncertainty concerning the relationship between the quality performance and demand for the services of the providers. Furthermore, there are probably random elements in demand that are not attributable only to the decisions made by the GPs. Hence, it is not realistic to assume that the regulator can ‘impose’ a certain level of demand for a given combination of types. Our next proposition, however, assures that this needs not be problem. If we define \( G(\theta) \) as the inverted hazard ratio, i.e. \( G(\theta) = F(\theta)/f(\theta) \) we can state the following result.

**Proposition 12.** The second best allocation can be achieved if the regulator offers a menu of contracts, with a capitation rate:
\[
r(\theta) = \frac{v}{(1 + \lambda)\Delta} \frac{\theta}{G(\theta) + \theta}
\] (69)
and a base salary equal to
\[
b^* = \sqrt{\frac{v}{(1 + \lambda)}} (2\sqrt{\theta} - \int f(\theta)(\sqrt{G(\theta) + \theta}) d\theta) - r(\tilde{\theta}) - \tilde{\theta}
\] (70)

Proof. According to Proposition 11, second best quality level is
\[
p^{\Pi}(\theta) = 1 - \sqrt{\frac{1 + \lambda}{v} (\theta + G(\theta))}
\] (71)
Setting this quality level equal to
\[ p^*(\theta) = 1 - \sqrt{\frac{\theta}{r(\theta)\Delta}} \] (72)
gives the capitation rate \( r(\theta) \) in (69). Quality levels and rents in (59) only depend on the capitation rate \( r(\theta) \) and not on the base salary; hence, the base salary is flat and set in order to assure that the participation constraint of the worst type (\( \bar{\theta} \)) is satisfied (profits are decreasing in \( \theta \) as established in (58)). Inserting the equilibrium values for \( r^*(\theta), p^*(\theta) \) and \( \Delta \) into the profit function (1) of the worst type and setting it equal to zero yields the result in (70).

Thus, we have seen that the second best can be achieved with a relatively simple reimbursement scheme. Of course, the results in Proposition 12 require that the regulator wishes to keep even the least efficient GP type active in the second best.

5.5. Welfare comparisons

Next, we analyse how welfare depends on the shape of the reimbursement scheme. Obviously, the simple linear capitation rate is simpler and hence cheaper to operate than a menu of contracts, which require communication between regulator and individual GPs. This loss has to be compared with the welfare gain from a more efficient scheme. If we compare the welfare levels given for the capitation scheme (47) with the welfare under the second best (61) we find that the difference equals:
\[ 2\sqrt{v(1+\lambda)} \left( \sqrt{\int_{\Theta} f(\theta)\sqrt{\theta} d\theta} \int_{\Theta} \frac{(F(\theta)+\theta f(\theta))}{\sqrt{\theta}} d\theta \right) - \left( \int_{\Theta} \sqrt{f(\theta)(f(\theta)\theta + F(\theta))} d\theta \right) \] (73)

Hence, the welfare gain is proportional to the value of the service, \( v \), the shadow cost of public funds, \( \lambda \), and apart from that depends only on the distribution of the types, where a higher variance in the distribution implies greater gains from introducing communication.

5.6. Summary of results

In this section, we have derived optimal results for three different types of information structures. Now we briefly summarise the findings.

5.6.1. Reimbursement rates

First, notice that the optimal capitation rate derived above can be simplified as
\[ r^* = \left( \frac{v}{1+\lambda} \right)^{3/2} \frac{1}{\Delta} \left( \frac{E\sqrt{\theta}}{2\sqrt{\bar{\theta}} - E\sqrt{\theta}} \right)^{3/2} \] (74)
Hence, the capitation formula consists of three terms; the first one reflects the balance between the value of the service and the marginal cost of public funds. The second term is the inverse of the number of patients that search. The third term reflects the spread in the cost of the types between the worst type and the average provider. Hence, the formula has a straightforward interpretation; the first term reflects the (first best) ratio between marginal benefit and marginal cost. The second term compensates for the number of searchers. An increase in the number of searchers makes \( r \)
more effective in providing incentives, and hence the reimbursement rate can go down when the number searchers increases. The third term represents the shadow cost of keeping the worst type at zero profit. An increase in \( r \) relaxes this constraint and hence the loading factor is less than one. It is useful to compare this formula with the (type-dependent) formula of the second best:

\[
\theta^*(\theta) = \frac{\theta}{(1 + \lambda)\Delta (G(\theta) + \theta)}
\] (75)

This formula has a similar structure, but here there is a different loading factor, which represents the shadow cost of the incentive compatibility constraints.

### 5.6.2. Quality levels

Now compare the quality levels in the three different scenarios:

\[
p^{\text{FB}}(\theta) = 1 - \sqrt{\frac{\theta(1 + \lambda)}{v}}
\] (76)

\[
p^{\text{C}}(\theta) = 1 - \sqrt{\frac{(1 + \lambda)}{v} \left( \theta + \frac{2\theta(\sqrt{\theta} - E\sqrt{\theta})}{E\sqrt{\theta}} \right)}
\] (77)

\[
p^{\text{II}}(\theta) = 1 - \sqrt{\frac{1 + \lambda}{v} (\theta + G(\theta))}
\] (78)

The first quality level, \( p^{\text{FB}}(\theta) \), is simply one that equals marginal value with marginal social cost. Quite reasonably, it increases in the marginal value of the service, \( v \), and decreases in marginal cost and the shadow cost of public funds. The third one, \( p^{\text{II}}(\theta) \), is an adaptation of a well known result from incentive regulation (Laffont and Tirole, 1993), that the deviation from the first best level of quality is increasing in the marginal cost. Hence, the best type will deliver first best quality, whereas all others will have a quality level that balances the trade-off between incentives and rent extraction, which is represented by the hazard ratio \( G(\theta) \). The third best quality, \( p^{\text{C}}(\theta) \), deviates from both. It does not, however, entail a gradual deterioration in quality as in the first best case. It is rather a shift of the entire function, which means that relatively good types will deliver lower quality than in the second best, whereas relatively bad types might well deliver a higher level of quality than in the second best.\(^5\) This reflects that the capitation scenario has no incentive compatibility problem and hence the quality levels of bad types will not have to be kept down for this reason. It is rather the participation constraint of the worst type that calls for a general reduction in quality. Hence, the deviation from first best quality is larger the larger the difference between the worst type and the average.

This difference between the capitation scenario and the second best does have important implications, since the regulator might also have preferences over the variation in quality. The results above suggest that the variance in quality will be lower in a capitation system, and hence this system might be preferred by an inequality-averse regulator, despite having a lower average quality level.

\(^5\) Comparing the functions (77) and (78) for the extreme points \( \theta \) and \( \tilde{\theta} \), we find that the best type (\( \tilde{\theta} \)) always provides higher quality in the second best compared with the capitation scheme. The worst type (\( \theta \)) will provide lower quality under capitation whenever (a) their frequency in the population, \( f(\tilde{\theta}) \) is relatively low and (b) their cost \( \tilde{\theta} \) is relatively high compared to the average.
6. Conclusion

The question of how to provide incentives to achieve high quality in primary care remains an open issue. The vast diversity in the design of reimbursement systems in different countries suggests that the answer is far from straightforward. In this paper, we have left the traditional issue of supplier induced demand, which certainly is of relatively more importance in a system where providers are allowed to set prices of their services. Instead, we have focused on whether patient search can be used as a regulatory tool despite the information asymmetries that are typical for health care. In particular, we have analysed equilibrium outcomes where profit maximising providers compete for patients who are rational but base their decision on a noisy observation of the actual quality. The model has delivered a series of policy relevant results.

Firstly, this paper makes clear that the outcome in a regulated market for an experience good depends crucially on the expectations of the consumers and not only on the actions taken by the providers. We have found that even if consumers are completely rational and make efficient use of the information they have at hand, the equilibrium outcomes can behave in a way that seems a priori completely counterintuitive. One result of the paper is that the comparative statics of the model depend crucially on a seemingly unimportant assumption on the shape of the cost function of the providers. Thus, if the marginal cost of quality is not rising steeply enough, equilibrium quality will be decreasing in the capitation rate. This result obtains because the strategic interaction between provider strategies and consumer beliefs turns in a negative direction: patients reduce their search effort to an extent that offsets the incentives initially generated by the change in capitation rates. In plain words, such an equilibrium offers too strong incentives to increase quality of care, and consumers have to adjust their beliefs downwards to compensate. This is a so-called Bertrand supertrap, where the strategic effect actually dominates the direct effect.

For the case where the model behaves in an intuitive way, i.e. quality increases in the reimbursement rate, we have found that equilibrium quality increases in the dispersion of ability, even if patients are risk neutral. This is due to the fact that it is the variance in quality which motivates search. Hence, in contrast with Gravelle and Masiero (2000), a higher dispersion in the perceived quality levels is actually beneficial to patients. This poses an interesting dilemma to the policy maker, since variance in quality is typically seen as a bad thing.

Some of the model’s assumptions are restrictive, and it is useful to consider how results would change if they were relaxed. Amongst the critical assumptions are, firstly, that the game is confined to two periods; that there is no congestion, i.e., the marginal cost of a visit is constant irrespective of the number of patients, and, finally, that there is no communication between patients. Most probably, relaxing these assumptions would make a “normal” equilibrium (i.e. quality increasing in the reimbursement rate) more likely. If the game goes on for more than two periods, patients will get an ever more accurate estimate of the actual quality level; congestion would indirectly increase the marginal cost of increasing the quality (as the payment per patient remains constant), and communication between patients would also tend to improve their information. Hence, it seems that relaxing our assumptions would make the intuitive result (quality increasing in the reimbursement rate) more likely to occur in practice. Nevertheless, this paper points at a potential problem in public sector voucher systems where the provider’s incentives depend on the consumer beliefs and consumer beliefs depend on the incentives of providers.

Concerning the normative issue as to how the reimbursement scheme should be designed, we have found that the capitation rate should be increasing in the value of the service and decreasing
in the marginal cost of public funds. More importantly, however, the capitation rate should be inversely proportional to the number of patients who search. The plain reason is that a high search activity augments the incentives for GPs to invest in quality. We have found that the number of searchers is increasing in the variance of quality and decreasing in the expected quality level in the market. Thus, the reimbursement system should take local conditions, such as the density of GPs, into account since such conditions affect the search activity.

If communication between the regulator and the GPs is allowed, a higher level of welfare can potentially be attained by letting GPs self-select in a menu of contracts. These contracts have a shape that is similar to the universal capitation rate, but vary according to the distribution of abilities in the GP population. This way, the regulator can extract more rents than with a uniform capitation rate. It is difficult, however, to make outright welfare comparisons between these two scenarios, since the welfare properties depend on the distribution of abilities in the population. What we do find is that the second best (i.e. where communication is allowed) entails an increasing deviation from the first best levels of quality as the marginal cost of the GP increases, whereas the capitation scenario entails a general shift downwards in the quality levels compared to the first best. The reason is that in the second best, the regulator wants to increase the variance in the quality of the service more than what is implied by a uniform capitation rate, in order to increase the patients’ incentives to search. Hence, the variation in quality levels will typically be higher in the second best. This finding highlights a serious trade-off that a regulator who is concerned about equality faces: variation in quality levels is considered a bad thing and at the same time keeps the market working. Hence, a concern for equality speaks in favour of a capitation system or even going for alternative regulatory mechanisms.

A further aspect that should be noted is that we have constrained how the regulator uses verifiable information. In principle, since the identity of patients is known, payments could actually be made dependent on the entire switching history of a patient. Such a payment system would probably be asymmetric in the sense that outward switches (where the decision depends on experienced quality) are punished more severely than inward switches are rewarded (since that decision depends on expected quality only). Hence future research could analyse the welfare properties of such a reimbursement system. Further alterations to the reimbursement system which might be considered are time-dependent payments (which would not qualitatively change results) or a more direct form of performance-based pay, possibly by means of surveys amongst patients. This latter form of reimbursement might become more useful as information processing becomes less expensive, but still there is the question whether the information acquired is reliable. In summary, the simplifying assumptions we have made concerning the reimbursement rule seem not to be very restrictive in practice.

There are several questions open for future research. One such is the already mentioned equity dimension, which is important in most health care systems, but has been disregarded here. One simple way to incorporate this in the model would be to let the regulator have disutility from variation in the quality levels. Such an assumption could change some of the results derived here, since variation in the quality levels is what keeps the system going. Furthermore, a richer model would not abstract from the issue how single visits at the GP surgeries are reimbursed. This would also open for the possibility of different quality dimensions, one to which the GP commits and one that is decided on at each meeting with a patient. Another topic for future research, finally, would be to allow for switching costs. Such a feature might reduce the competitiveness of the market and hence patient welfare, but if, as in Gravelle and Masiero (2000) switching costs are endogenous, the effects on the market equilibrium could be important.
References


