WORKING PAPERS IN ECONOMICS

No. 10/03

KURT R. BREKKE, ROBERT NUSCHELER
AND ODD RUNE STRAUME

GATEKEEPING IN HEALTH CARE

Department of Economics

UNIVERSITY OF BERGEN

Gatekeeping in Health Care*

Kurt R. Brekke†, Robert Nuscheler‡, Odd Rune Straume§
July 4, 2003

Abstract

We study the competitive effects of restricting direct access to secondary care by gatekeeping, focusing on the informational role of gatekeeping general practitioners (GPs). We consider a secondary care market with two hospitals choosing the quality and specialisation of their care. GPs perfectly observe the diagnosis of a patient and the exact characteristics of the secondary care market. Patients are either informed or uninformed when accessing the hospital market. We consider two distinct cases: first, we let the fraction of informed patients be exogenous, implying that the regulator can only influence patients' decision of consulting a GP by making this compulsory ('direct gatekeeping'). Second, we endogenise this fraction by assuming GP consultation to be costly for the patient. Then the regulator can influence the GP attendance rate through the regulated price ('indirect gatekeeping'). A main finding of the paper is that strict gatekeeping may not be socially desirable, even if it is costless.

Keywords: Gatekeeping; Imperfect information; Quality competition; Product differentiation; Price regulation.

JEL classification: D82; I11; I18; L13

*We thank Pierre-Yves Geoffard, Kai A. Konrad, Daniel Krähmer, Kjell Erik Lommerud, the participants of the microeconomic colloquium at the Free University of Berlin, and the participants of the 4th European Workshop in Health Economics, Oslo 2003, for helpful comments. The usual caveat applies.

†University of Bergen, Department of Economics, Programme for Health Economics in Bergen (HEB). E-mail: kurt.brekke@econ.uib.no

‡Corresponding author. Wissenschaftszentrum Berlin für Sozialforschung (WZB), Reichpietschufer 50, D-10785 Berlin, Germany. Phone: ++49 30 25491-430, Fax: ++49 30 25491-400, E-mail: robert@wz-berlin.de

§Institute for Research in Economics and Business Administration (SNF) and Department of Economics, University of Bergen. E-mail: odd.straume@econ.uib.no

1 Introduction

The UK and the Scandinavian countries are examples of countries where general practitioners (GPs) have a gatekeeping role in the health care system. Patients do not have direct access to secondary care. They need a referral from their (primary care) GP to get access to a hospital or a specialist.[1] Restricting direct access to secondary care by giving GPs a gatekeeper role is currently on the political agenda in Germany, while in Sweden there has been some debate about whether patients should be able to approach a specialist or a hospital directly.[2] The current paper contributes to the discussion on gatekeeping by analysing the competition effects that arise when GPs are equipped with a gatekeeping role.

In general, there are two main arguments for introducing gatekeeping in health care markets (see Scott, 2000). Firstly, it is usually claimed that gatekeepers contribute to cost control by reducing 'unnecessary' interventions.[3] Second, it is argued that secondary care is used more efficiently since 'GPs usually have better information than patients about the quality of care available from secondary care providers' (Scott, 2000, p. 1177). In the present paper we focus on the second argument, highlighting the fact that making this information available to patients changes the nature of competition between secondary care providers, which in turn affects the social desirability of gatekeeping.

As pointed out in a seminal paper by Arrow (1963), uncertainty and asymmetric information make health care markets different from other markets. Uncertainty generates demand for health insurance, implying that non-price strategies are important in attracting patients as the consumption of medical care is paid for by a third party. Asymmetric information is present in the sense that consumers (patients) are typically less informed about their health conditions, and thus the appropriate treatment, than the providers of medical care. In this paper we stress both these features of health care markets.

Building on the familiar model of Hotelling (1929), we consider a secondary care market with two providers (hospitals). In order to attract patients, and thus obtain third party payments, the hospitals have

[1] In the US, several Health Maintainance Organizations also practice gatekeeping. Recently, however, some HMOs have relaxed the restrictions on access to specialists (see, e.g., Ferris et al., 2001).

[2] The local authority in Stockholm has recently allowed patients to have direct access to hospital care, while in the rest of Sweden patients still need a referral before receiving secondary care.

[3] Although this is a common argument for restricting access to secondary care, the empirical evidence that gatekeeping actually contributes to lower health care expenditures seems to be scarce (see, e.g., Barros, 1998).

two strategic variables at their disposal - location and quality of care. We refer to location as the specialisation or service mix at a hospital, though it may also be interpreted in geographical terms. Thus, hospitals engage in non-price competition in terms of both horizontal and vertical differentiation of services.

The major aim of the paper is to highlight the informational role of gatekeepers in such secondary care markets. Without gatekeepers, we assume that at least some of the patients are uninformed about both their own specific diagonsis and the exact characteristics of the secondary care market. Thus, with direct access to secondary care, patients' choices may be subject to substantial errors. First, a patient may end up in a poor match, i.e., he may choose the hospital that is less able to cure his disease. Second, he may decide to go to the specialist who provides the lower quality of care. By introducing GP gatekeeping we assume that all relevant information is transmitted to the patients, thereby enabling them to make informed choices.[4] Thus, GPs observe the actual disease of a patient with certainty and know which specialist is more able to cure a particular disease. Additionally, we assume that GPs obtain perfect quality signals. Both features are in line with the above mentioned second argument for introducing gatekeeping.

We analyse the informational role of gatekeepers by applying two different variants of the basic model. In the first part of the paper we consider an exogenously given number (fraction) of patients that are *a priori* fully informed about their disease and the most appropriate treatment for this condition. One possible interpretation is to think of these patients as the chronically ill who have obtained all relevant information through repeated consumption. Introducing GP gatekeeping is then simply equivalent to making the uninformed fraction of patients fully informed. Since gatekeeping can only be regulated directly, we refer to this variant as 'direct gatekeeping'. Although we consider gatekeeping to be costless, we find that introducing strict gatekeeping, i.e., making it compulsory to get a referral to secondary care, is not necessarily socially desirable. The reason is that more informed patients lead to more intense quality competition between hospitals, amplifying the hospitals' incentives to differentiate their services. Consequently, gatekeeping may induce too much quality and differentiation from the viewpoint of social welfare.[5] However, we show that, under second-best price regulation,

[4] We abstract from agency problems by assuming that the GPs truthfully convey their information to the patients.

[5] This result is related to Dranove et al. (2003), who empirically analyse whether public disclosure of patient health outcomes at the level of the individual physician or hospital ('report cards') is beneficial to patients and social welfare. They find that

gatekeeping is always socially beneficial. Thus, gatekeeping should be accompanied by proper price regulation.

One might ask, however, if an uninformed patient would not *voluntarily* consult a GP before seeking secondary care. We argue that consulting a GP may in itself involve costs for the patients, such as out-of-pocket payments, travelling and/or time costs. A patient would then have to compare the benefits of consulting a GP - in terms of reduced risk of choosing the less suitable hospital - against such costs. This situation is analysed in the second part of the paper, where we let the fraction of informed patients be endogenously determined by this trade-off. A crucial feature of this variant of the model is that GP consultation can be indirectly influenced through price regulation. Consequently, we will refer to this mechanism as 'indirect gatekeeping'. The endogeneity of the consultation decision alters hospitals' incentives. Although differentiation relaxes quality competition, it also increases the fraction of informed patients, since the (expected) benefits of consulting a GP now are higher. Thus, the incentives for differentiation of services are weakened. Moreover, there is now a real cost of introducing a strict gatekeeping regime, which is reflected in the individual costs of GP consultation. In this case, individual consultation incentives coincide with the social incentives, implying that there is no need to regulate gatekeeping directly. However, we show that second-best price regulation implies a *de facto* strict gatekeeping regime - in which every patient finds it beneficial to consult a GP before accessing the secondary care market - if quality costs or GP consulting costs are sufficiently low, or if mismatch costs are sufficiently high.

The paper relates to both the general literature on spatial competition and the literature on (imperfect) competition in health care markets. The interaction between quality and location choices has been investigated by Economides (1989) under price competition and Brekke et al. (2002) under price regulation. The present paper contributes to this literature by introducing imperfect information into the framework. As previously mentioned, we find that the hospitals' incentives to differentiate services are significantly altered by the presence of uninformed consumers. In particular, we find that uninformed consumers tend to soften the incentives for horizontal differentiation. In this respect our findings are in the spirit of Bester (1998), who shows that quality competition may induce minimum differentiation - i.e., agglomeration at the market centre - when consumers are uncertain about product quality

report cards led to both selection behaviour by providers and improved matching of patients with hospitals. However, on net this led to higher levels of resource use and to worse health outcomes (for sicker patients).

and use observed prices to ascertain the quality of goods.

In a related paper, Calem and Rizzo (1995) analyse hospitals' choices of quality and speciality mix (location) under exogenous prices. An incentive for closer locations is introduced by assuming that the hospitals cover a fraction of their patients' mismatch costs. Besides this particular assumption, their paper differs from ours in two important ways. Firstly, they are not concerned with imperfect information and the issue of gatekeeping and how this affects the nature of competition in the market for secondary care. Second, they do not consider the implications of optimal price regulation on the hospitals' incentives with respect to quality and location choice.[6]

The paper also relates to the more general literature on transparency in imperfectly competitive markets.[7] Increased transparency on the consumer side of the market typically leads to intensified price competition and thus to a more socially desirable market outcome. Our paper contributes to this literature by analysing the effects of improved transparency in markets that are characterised by *non-price competition*. In this case, more intense competition between firms does not necessarily improve social welfare. Improved market transparency consequently has ambiguous welfare effects.[8]

The remainder of the paper is organised as follows. The basic framework is presented in Section 2. In Section 3, we consider the case of 'direct gatekeeping', where the fraction of informed patients is exogenously given. In Section 4, we analyse the case of 'indirect gatekeeping', where the fraction of informed patients are endogenously determined by individual GP consultation decisions. In both sections we derive the quality and specialisation equilibria and analyse the social desirability of gatekeeping. We also discuss the issue of optimal price regulation. Section 5 provides concluding remarks.

[6] Two other related papers applied to the primary care market are Gravelle (1999) and Nuscheler (2003). Both papers address the issue of competition between physicians by investigating the interaction between quality and location choices when prices are regulated. Building on the seminal contribution of Salop (1979), they apply a circular model with attention directed towards entry of physicians into the market, so the focus of these papers is clearly quite different from ours.

[7] See, e.g., Varian (1980), Burdett and Judd (1983), Schultz (2002, 2003), Lommerud and Sørgard (2003).

[8] Another related paper in this strand of the literature is Baye and Morgan (2001), who analyse the competition effects of information gatekeepers on the Internet, where such gatekeepers create a market for price information by charging fees to firms that advertise prices and to consumers who access the list of advertised prices.

2 The model

There is a continuum of patients with mass 1 distributed uniformly along the Hotelling line $S = [0, 1]$. The location of a patient is denoted $z \in S$ and is associated with the disease he suffers from. A disease z can be seen as a realisation of a random variable Z which is uniformly distributed on S. All patients need one medical treatment to be cured. There are two health care providers - henceforth called hospitals - both able to cure all diseases. However, they are differentiated with respect to the disease they are best able to cure. Specialisation of a hospital is denoted $x_i \in \mathbf{R}$, $i = 1, 2$. Without loss of generality, we will assume throughout the paper that $x_1 \leq x_2$. Note that the degree of specialisation is not restricted to the disease space S. Thus hospitals may locate outside S.[9]

In addition to specialisation, there is a second strategic variable used by the hospitals to attract patients, namely the quality of care $q_i \geq 0$, $i = 1, 2$. Quality costs are assumed to be symmetric and quadratic, kq_i^2, where $k > 0$. These costs are considered to be fixed, i.e., they are independent of how many patients are actually treated. This implies that quality has the characteristics of a public good at each hospital. Examples of such quality investments are the cost of searching for and hiring more qualified medical staff, additional training of existing medical staff, and investments in improved hospital facilities, which can be related to both medical machinery and non-medical facilities such as room standard or catering quality. Without loss of generality, other fixed costs are set to zero. Marginal production costs are assumed to be constant and equal to zero. This cost structure stresses the importance of fixed costs which seems reasonable for the hospital market.[10] The price for one treatment is denoted $p \geq 0$ and is set by some regulatory authority.[11] As the price is independent of which hospital is actually attended it may alternatively be interpreted as a premium for a health insurance with full coverage. The profit function of hospital i is given by

$$\Pi_i = pD_i - kq_i^2, \tag{1}$$

where D_i is the demand for hospital i treatment.

A patient's (ex-post) utility when going to hospital i is given by

$$u_i(z; p) := u(q_i, x_i, z; p) = v + q_i - p - t(z - x_i)^2. \tag{2}$$

[9] This assumption is made for convenience, but does not qualitatively affect any of the results in the paper.

[10] The assumption of production-independent quality costs is widely used in the literature on quality competition in health care markets (see, e.g., Calem and Rizzo, 1995; Lyon, 1999; Gravelle and Masiero, 2000; Barros and Martinez-Giralt, 2002).

[11] All results we derive also hold for constant marginal costs $MC > 0$. Let $\widetilde{p}$ denote the mill price, then the mark-up is given by $p = \widetilde{p} - MC$.

The maximum gross willingness to pay for hospital treatment, v, is assumed to be sufficiently large for the entire market to be covered. Thereby, we preclude monopoly and kink equilibria and concentrate on competitive ones.[12] Notice that this assumption essentially means that all patients have access to hospital or specialist care, which seems reasonable, at least for developed countries (without waiting lists). The last term measures the mismatch costs incurred when treated by hospital $i = 1, 2$. The parameter $t > 0$ determines the importance of mismatch costs relative to the quality of care. Of course, mismatch costs would be zero if the patient suffers exactly from the disease for which the hospital he goes to is specialised. Mismatch costs are assumed to be quadratically increasing in distance.

To evaluate the effects of gatekeeping we consider two different patient types. The fraction $\lambda \in [0, 1]$ of the population is fully informed when accessing the hospital market, i.e., these patients know their own location (diagnosis) and the specialisation and quality provision of each hospital. In the first part of the paper (direct gatekeeping, Section 3) we will assume that the number of fully informed patients is exogenously given. In this case we can think of these patients as the chronically ill, who know exactly what disease they are suffering from and have obtained sufficient information about the hospital market through repeated consumption. In the second part of the paper (indirect gatekeeping, Section 4) we will endogenise λ by explicitly modelling patients' decision about consulting a GP before accessing the hospital market. We will assume that the GP has a gatekeeper role in the system and that he or she obtains all significant information. This information is then truthfully conveyed to those patients consulting the GP, making them fully informed about all relevant variables.

To simplify the analysis we assume that the fully informed patients are uniformly distributed on S. Members of the remaining part of the population, $1 - \lambda$, only know v, the distribution of Z, and that medical treatment is required. They cannot observe x_i, q_i, and z. For these patients secondary care is an experience good. Their ex-ante utility of attending hospital i is given by

$$Eu_i(Z; p) := Eu(q_i^e, x_i^e, Z; p) = v + q_i^e - p - tE(Z - x_i^e)^2, \qquad (3)$$

where the superscript e denotes the expected value of the respective variable. Patients learn their ex-post utility given by (2) only through actual consumption.

[12] In a circular model, Economides (1993) and Nuscheler (2003) make similar assumptions, whereas Salop (1979) and Gravelle (1999) study monopoly and kink equilibria in detail.

For the direct gatekeeping scenario we will consider that the regulator can only influence λ through direct regulation. In theory, it is possible to imagine that the regulator can influence the amount of information available to patients in the market through several different means. We will, however, focus on what is probably the most realistic regulatory instrument, namely introducing a strict gatekeeping regime, where all patients are required to consult a GP before accessing the hospital market. Thus, the scope for regulating λ is restricted to setting $\lambda = 1$. In the indirect gatekeeping scenario, the regulator can indirectly influence the endogenously determined value of λ through price regulation.

The impact of introducing gatekeeping to the market for hospital or secondary care is analysed in a 5-stage game:

- Stage 1: the regulator sets her available regulatory variables. These are one or both of p and λ. Regulation on the latter variable is restricted to setting $\lambda = 1$.
- Stage 2: the hospitals simultaneously decide on their specialisations, x_1 and x_2, where $x_1, x_2 \in \mathbf{R}$, and $x_1 \leq x_2$.
- Stage 3: the hospitals simultaneously set their quality levels $q_1 \geq 0$ and $q_2 \geq 0$.
- Stage 4: patient information about x_i, q_i, and z can be obtained by consulting a gatekeeping general practitioner who truthfully conveys information about the relevant variables. The choice of consulting a GP is reserved for the second version of the model (Section 4). In the first version (Section 3) the share of fully informed patients, λ, is exogenously given.
- Stage 5: the patients demand secondary care treatment.

The sequential structure of the game is argued by the different degree of irreversibility of strategic decisions. Clearly, the decision of whether to consult a gatekeeping GP and/or which hospital to go to is the most flexible decision to be taken in the entire game. Changing quality or specialisation requires more effort and investment. In both cases it may be necessary to replace some medical machinery and/or have the current staff undergo significant training, or even hire new staff. Although it may sometimes be hard to distinguish between quality investments and a change of specialisation, it seems logically consistent to assert that hospitals first decide what to produce (their service or speciality mix), and then determine the quality of services.[13] This sequential structure

[13] Calem and Rizzo (1995) discuss this in some more detail.

is common in models that combine horizontal and vertical differentiation (see, e.g., Economides, 1989; Calem and Rizzo, 1995; Bester, 1998; Gravelle, 1999).

That the regulator can determine λ and p at the beginning of the game essentially means that we consider commitment power on her side. This assumption is, of course, crucial as in most sequential games. With respect to λ, this may be justified since introducing a strict gatekeeping system ($\lambda = 1$) must be regarded as a major reform of the health care system. This may be less clear with the price. As in Brekke et al. (2002) and Nuscheler (2003) there will be an incentive to reoptimise after specialisations have been chosen. Nevertheless, since commitment is valuable for the regulator, one could argue that she should be able to obtain such commitment power, either through reputation or by creating institutional mechanisms that makes it costly, or otherwise difficult, to change the regulated price. In any case, as price regulation is not the major focus of the present paper we will concentrate on the commitment case.

3 Direct gatekeeping

In this section we will consider that $\lambda \in [0, 1]$ is exogenous and can only be regulated directly by setting $\lambda = 1$. Hence, the regulator determines whether or not to introduce a strict gatekeeping system, and thereby make all patients fully informed. The game is solved by backward induction.

3.1 The specialisation-quality game

3.1.1 The demand for secondary care

The share $1-\lambda$ of the population is uninformed about the actual quality levels and about specialisations. Moreover, these people do not know the disease they suffer from. To make a decision about which hospital to consult, patients have to evaluate their expected utility, given by equation (3), for both hospitals. Imposing symmetry, these patients are indifferent between hospitals in expected terms. Both hospitals receive one half of these patients, $(1 - \lambda)/2$.

This assumption is not necessary, but it eases the presentation of the main ideas dramatically. Actual demand depends on the patients' beliefs which influence expected utilities. Since these beliefs do not change the optimisation problem of the hospitals (see below), they can be neglected at the earlier stages of the game. Nevertheless, beliefs have to be confirmed in equilibrium and, as we concentrate on symmetric equilibria, beliefs will also be symmetric. Our assumption that each hospital

gets half of the uninformed patients is thus the outcome of a more general treatment.[14] Given the symmetry assumption, the decision about which hospital to attend reduces to flipping a fair coin - which seems not unrealistic.

In contrast, the informed fraction of the population, λ, is responsive to quality investments and specialisation decisions as both strategic variables and the own disease are observable. The informed patient who is indifferent between hospital 1 and hospital 2 suffers from disease $\overline{z}$, which is obtained by solving $u_1(z;p) = u_2(z;p)$ for z, where u_i is given by equation (2), yielding

$$\overline{z} = \frac{q_1 - q_2}{2t(x_2 - x_1)} + \frac{x_1 + x_2}{2}. \tag{4}$$

The demand for hospital 1 is thus $D_1 = \lambda\overline{z} + (1-\lambda)/2$. Hospital 2 receives the residual demand $D_2 = 1 - D_1 = \lambda(1-\overline{z}) + (1-\lambda)/2$.

3.1.2 Quality competition

We look for an equilibrium in pure strategies in the quality subgame.[15] If a pure strategy equilibrium exists, it is found by inserting demand from equation (4) into the profit function (1) and optimising with respect to q_1, which yields the optimal quality provision for both hospitals for given specialisations:

$$q^*(\Delta;\lambda,p) = \frac{p\lambda}{4tk\Delta}, \tag{5}$$

where $\Delta := x_2 - x_1$. The equilibrium quality levels depend only on the distance between hospitals' locations and not on their actual locations. An immediate implication is that optimal specialisations will be characterised by some certain distance and not by absolute locations.

From (5) we see that $\lim_{\Delta\to 0} q^*(\Delta;\lambda,p) \to \infty$. Since quality investments are costly, this means that (5) yields negative profits if Δ is sufficiently small. In other words, there exists a (small) range of $\Delta \in \left[0, \overline{\Delta}\right]$ where investment incentives are so strong that hospitals are led into 'ruinous competition'. Thus, in order to secure positive profits - and thus pure strategy equilibrium existence - we have to impose a restriction that the hospitals are not located too closely. Let Q be the set of all

[14] Although there is a game of incomplete information (the fraction $1-\lambda$ of patients do not know their disease type) and imperfect information (the fraction $1-\lambda$ of patients cannot observe qualities and specialisations), beliefs are irrelevant for the outcome. Subgame perfection is thus sufficient to obtain a unique symmetric (perfect Bayesian) equilibrium. This changes when λ is endogenised.

[15] The concept of mixed strategies does not seem to make much sense in the context of hospital quality investments, so we disregard this possibility by assumption.

location pairs (x_1, x_2) such that a Nash equilibrium exists in the quality subgame. Using (5), it is easily shown that $(x_1, x_2) \in Q$ if

$$k > \frac{p\lambda^2}{8t^2\Delta^2\left(1-\lambda+\lambda\left(x_1+x_2\right)\right)}. \tag{6}$$

Assuming that (6) is satisfied, the comparative static results are straightforward: the smaller product differentiation, i.e., the smaller Δ, the more intense is quality competition. Patients are more responsive to quality improvements when mismatch costs are small. Thus t is a measure of competition intensity. Not very surprisingly, an increase in the quality cost parameter k has an adverse effect on quality provision. The better medical treatments are paid, the higher the benefits of capturing additional market shares from the competitor. At this stage of the game the only means of competition is the quality of care and thus hospitals will improve their quality. The comparative statics with respect to λ are the same as with respect to p: more informed patients lead to higher quality provision.

Because of its exogeneity, the fraction $1-\lambda$ of the population cannot have any effect on competition, thus, λ can be interpreted as the density of patients that are distributed along the Hotelling line. When defining $\widehat{p} := \lambda p$ we obtain the same results as Brekke et al. (2002).

3.1.3 Specialisation

At this stage of the game hospitals decide on their specialisation, taking the effects on quality competition and demand into account. In order to obtain a perfect pure strategy equilibrium of the specialisation-quality game, we follow the approach taken in similar location models[16] and restrict the strategy space of the specialisation game to the set Q, for which a pure strategy equilibrium of the quality game obtains. Intuitively, it seems highly plausible to assume that the hospitals will not consider locations which trigger incentives for 'ruinous competition'. Following Economides (1986), we define the direction in which $\partial\Pi_i/\partial x_i$ is positive as the 'relocation tendency' of firm i. An equilibrium of the specialisation game must then be at the zero relocation locus, $\partial\Pi_1/\partial x_1 = \partial\Pi_2/\partial x_2 = 0$, and a perfect equilibrium of the specialisation-quality game is defined as the intersection between the zero relocation locus and the existence set Q. Formally, a specialisation equilibrium (x_1^*, x_2^*) exists if

$$\frac{\partial\Pi_i\left(x_1^*, x_2^*\right)}{\partial x_i} = 0; \quad \frac{\partial^2\Pi_i\left(x_1^*, x_2^*\right)}{\partial x_i^2} < 0; \quad \left(x_1^*, x_2^*\right) \in Q; \quad i = 1, 2.$$

[16] See, e.g., Economides (1984, 1986, 1989), Hinloopen and Marrewijk (1999), Lambertini (2001).

Inserting the optimal quality levels into hospital 1's profit function, we obtain the following partial derivative with respect to x_1:

$$\frac{\partial \Pi_1}{\partial x_1} = \frac{\lambda p}{2} - \frac{p^2 \lambda^2}{8t^2 k \Delta^3}. \tag{7}$$

As already mentioned, setting $\partial \Pi_1 / \partial x_1 = 0$ only yields Δ^*. There exists a continuum of locations fulfilling $x_2 - x_1 = \Delta^*$. Imposing symmetry leads to a unique equilibrium of the game, provided that $(x_1^*, x_2^*) \in Q$, where

$$x_1^*(\lambda, p) = \frac{1}{2}(1 - \Delta^*) \text{ and } x_2^*(\lambda, p) = \frac{1}{2}(1 + \Delta^*), \tag{8}$$

and

$$\Delta^*(\lambda, p) = \left(\frac{p\lambda}{4t^2 k}\right)^{\frac{1}{3}}. \tag{9}$$

It is easily shown that the second-order conditions are met. However, it remains to identify the exact condition for equilibrium existence. According to the specification of the game, two requirements must be met. First, we need to have that $(x_1^*, x_2^*) \in Q$. Second, it must not be a profitable strategy for either firm to deviate in the quality subgame by offering zero quality and only serve the uninformed consumers arriving in equilibrium. Using (9) and (5), and imposing symmetry in the profit function, it is straightforward to show that both requirements are met, thus guaranteeing the existence of a unique symmetric equilibrium, if

$$k > \frac{p\lambda^4}{32t^2}. \tag{10}$$

For the remainder of the analysis, we will assume that this condition is met.

The hospitals' location incentives are governed by two opposing forces. *Ceteris paribus*, each hospital can obtain a larger share of the market by moving closer to its rival. On the other hand, closer locations imply that quality competition is intensified, as can be seen from equation (5).

Consider an increase in the price p. This will strengthen the market share effect, since hospitals now receive a higher mark-up on each treatment. However, a price increase also means that quality competition is amplified. From (9) we see that the latter effect always dominates: a higher price implies that hospitals aim at dampening the resulting increase in quality competition by locating further apart. Indeed, as long as the fee for secondary care treatments exceeds marginal costs (and λ is strictly positive), quality competition among providers induces product differentiation.

An identical mechanism determines the relationship between patient information and locations. More informed patients will result in stronger quality competition and hospitals will respond by differentiating more.[17] A social planner thus faces a trade-off when setting the price or taking measures to improve information in the market. The improved quality has to be weighed against the change in aggregate mismatch costs.

We have already identified the mismatch cost parameter t as a measure of competition intensity. A low t boosts quality provision and - to dampen this effect - hospitals locate further apart. Finally, an increase in the quality cost parameter k reduces quality competition, resulting in less product differentiation. When inserting (9) into (5) we obtain the equilibrium quality levels of the game:

$$q^*(\lambda, p) = \left(\frac{p^2\lambda^2}{16tk^2}\right)^{\frac{1}{3}}. \tag{11}$$

3.2 Social Welfare

Consider a social planner who aims at maximising social welfare. Assuming symmetry in qualities and locations the social welfare function is given by

$$W = v + q - 2kq^2 - \frac{t}{12} + \frac{t}{4}\Delta(\lambda - \Delta). \tag{12}$$

Note that we consider that acquiring information about the market is costless, i.e., gatekeeping involves no costs.[18] We will relax this assumption when endogenising λ in Section 4.

3.2.1 The second-best optimum

Let us first consider the case where λ cannot be regulated by the social planner at all. In this sense the solution derived here may be called a 'constrained first-best', or simply the second-best. Quality provision is second-best efficient when

$$q^{sb} = \frac{1}{4k}. \tag{13}$$

Maximising the last term of equation (12) yields $\Delta^{sb} = \lambda/2$, which determines the second-best efficient specialisations

$$x_1^{sb} = \frac{1}{2} - \frac{\lambda}{4} \text{ and } x_2^{sb} = \frac{1}{2} + \frac{\lambda}{4}. \tag{14}$$

[17] This result is clearly dependent on the mode of competition. If we allow the firms (hospitals) to compete on prices, and not qualities, the opposite result would apply (cf. Schultz, 2002).

[18] If we interpret p as a per patient (or per treatment) reimbursement from a government agency, this particular specification of social welfare also relies on the assumption that the third party (i.e., the regulator) is able to raise the necessary funds in a non-distortionary manner.

The regulator faces the following fundamental trade-off: on the one hand, the mismatch costs incurred by the informed patients are minimised when hospitals locate at $\frac{1}{4}$ and $\frac{3}{4}$, respectively. These locations would obtain when the entire population is informed, $\lambda = 1$. On the other hand, as the uninformed patients choose a provider randomly, their mismatch costs are at a minimum when hospitals do not specialise and agglomerate at the market centre, i.e., at $\frac{1}{2}$. Minimum differentiation would obtain for $\lambda = 0$. Balancing these opposing effects leads to locations $x_1^{sb} \in \left[\frac{1}{4}, \frac{1}{2}\right]$ and $x_2^{sb} \in \left[\frac{1}{2}, \frac{3}{4}\right]$.

3.2.2 The first-best optimum

Now consider that the regulator has the available option of introducing a strict gatekeeping regime, which amounts to setting $\lambda = 1$. From equation (12) it is easily seen that $\partial W/\partial\lambda \geq 0$ for all feasible values. The social planner would thus implement a strict gatekeeping regime whenever $\Delta > 0$. The first-best solution is consequently given by[19]

$$\lambda^{fb} = 1,\ q^{fb} = \frac{1}{4k},\ x_1^{fb} = \frac{1}{4} \text{ and } x_2^{fb} = \frac{3}{4}. \tag{15}$$

3.3 Gatekeeping

The aim of this subsection is to show that introducing strict gatekeeping, i.e., setting $\lambda = 1$, is not necessarily socially beneficial when the price is exogenously given. This may be surprising at first sight since strict gatekeeping implies that additional information is acquired. Taking the competitive effects into account it may turn out that - although gatekeeping is costless - strict gatekeeping is harmful from a social welfare perspective. The relationship between social welfare and the share of informed patients is given by the following proposition:

Proposition 1 *For an exogenously given price, social welfare is maximised at*

(i) $\lambda = 1$ *if mismatch costs are sufficiently high,*

(ii) $\lambda \in (0, 1)$ *if mismatch costs are sufficiently low and quality costs are sufficiently high.*

Proof. Inserting (9) and (11) into (12) yields a welfare function $W(p, \lambda)$. We can easily calculate

$$\frac{\partial W}{\partial \lambda} = (2p)^{\frac{1}{3}} \left(\frac{1}{8} \left(\frac{2p}{\lambda t k^2} \right)^{\frac{1}{3}} - \frac{1}{6} \left(\frac{\lambda}{t^2 k} \right)^{\frac{1}{3}} (2p - t) \right) \tag{16}$$

[19] Notice that the solution $\Delta = 0$ and $\lambda = 0$ is always dominated by $\Delta = 1/2$ and $\lambda = 1$.

and

$$\frac{\partial^2 W}{\partial \lambda^2} = -(2p)^{\frac{1}{3}} \left(\frac{1}{24} \left(\frac{2p}{\lambda^4 t k^2} \right)^{\frac{1}{3}} + \frac{1}{18} \left(\frac{1}{t^2 \lambda^2 k} \right)^{\frac{1}{3}} (2p - t) \right) \qquad (17)$$

(i) We have that $\frac{\partial W}{\partial \lambda} > 0$ for all permissible values of λ if $t > 2p$. (ii) Assume that $t < 2p$. In this case we have that $\frac{\partial W}{\partial \lambda} > (<) 0$ if $k < (>) \overline{k} := \frac{27pt}{32\lambda^2(2p-t)^3}$. Since $\lim_{\lambda \to 0} \overline{k} \to \infty$ and $\frac{\partial^2 W}{\partial \lambda^2} < 0$ it follows that social welfare is maximised for a unique value of λ that lies strictly between 0 and 1 if $k > \frac{27pt}{32(2p-t)^3}$. ■

Ceteris paribus, more informed patients lead to more intense competition between the hospitals, which implies a higher provision of quality and more differentiation. If mismatch costs are high, the degree of competition between hospitals is low, which further implies that the incentives for horizontal differentiation are also low. In this case from a welfare point of view there is underprovision of quality and an insufficient degree of differentiation. A larger share of informed patients would thus increase efficiency with respect to both quality provision and horizontal differentiation.

However, more informed patients could lead to *excessive competition* if mismatch costs are sufficiently low. If, in addition, quality costs are sufficiently high, so that first-best quality provision is relatively low, a fully informed market would lead to both excessive differentiation *and* overprovision of quality. This could be sufficient to outweigh the benefits of increased patient information on aggregate mismatch costs, implying that social welfare is maximised in a situation where not all patients are fully informed.

The welfare implications of introducing a strict gatekeeping regime follows immediately:

Corollary 1 *For an exogenously given price, introducing a strict gatekeeping regime is detrimental to social welfare if (i) mismatch costs are sufficiently low, (ii) quality investments are sufficiently costly, or (iii) the fraction of a priori informed patients is sufficiently high.*

In other words, costless gatekeeping can reduce social welfare due to excessive competition between health care providers. In order better to illustrate the main mechanisms behind this result we provide a numerical example.

3.3.1 A numerical example

Let $p = 1$ and $k = 1$. Then the remaining parameters of the model are t and λ.[20] We illustrate the case of fairly intense competition with $t = 1/2$ (Case 1) and moderate competition with $t = 3/2$ (Case 2). In Table 1 we present the outcome of the location-quality game, with the associated level of social welfare, for different values of λ.

Table 1: Equilibrium outcomes for $p = 1$, $k = 1$

	Case 1, $t = 1/2$			Case 2, $t = 3/2$		
λ	q^*	Δ^*	$W^* - v$	q^*	Δ^*	$W^* - v$
0.1	0.11	0.46	0.022	0.07	0.22	-0.072
0.2	0.17	0.58	0.043	0.12	0.28	-0.043
0.3	0.22	0.67	0.051	0.16	0.32	-0.021
0.4	0.27	0.74	0.051	0.19	0.35	-0.002
0.5	0.31	0.79	0.046	0.22	0.38	0.015
0.6	0.36	0.84	0.035	0.25	0.41	0.030
0.7	0.39	0.89	0.021	0.27	0.43	0.043
0.8	0.43	0.93	0.003	0.30	0.45	0.054
0.9	0.47	0.97	-0.018	0.32	0.46	0.065
1	0.5	1	-0.042	0.35	0.48	0.075

As can be seen from equation (5), quality competition is intense for low values of the mismatch cost parameter t. Thus, hospitals provide higher quality in Case 1 than in Case 2. To mitigate costly quality competition, hospitals aim at making their products less substitutable. This incentive is clearly higher in Case 1, partially offsetting the competition effect. In Case 1, increasing the share of informed patients is beneficial for low values of λ. Besides the net benefits derived from higher quality provision, patients may also gain from reduced mismatch costs. As λ increases, though, the centrifugal force drives hospitals further away from the market centre, combined with an increase in quality provision. At $\lambda = 0.4$ we see that there are both overprovision of quality *and* too much differentiation, compared with the first-best solution, implying that a further increase in λ unambiguously reduces welfare.[21] In fact, since $\lim_{\lambda \to 0} (W^* - v) = 0$ we see that implementing a strict gatekeeping system would be socially detrimental even if there are no informed patients to begin with. This changes when Case 2 is considered, where

[20] Of course v is another not yet specified parameter. The actual size is irrelevant for the model (as long as v is sufficiently large), so we will keep this general.

[21] As hospitals still specialise within the disease space, our example shows that Proposition 1 does not rely on the assumption that hospitals are allowed to locate outside the disease space.

moderate specialisation incentives are at work. In this case it pays to generally demand a GP referral.

3.4 Price regulation

The results of the previous section hinge on the assumption that the price is exogenous. We will now consider the case where the regulator is able also to use the price as a regulatory instrument. Assuming second-best price regulation, the following result obtains:

Proposition 2 *With second-best price regulation, introducing a strict gatekeeping system always improves social welfare.*

Proof. Again, inserting (9) and (11) into (12) yields the welfare function $W(p,\lambda)$. By defining $\widehat{p} := p\lambda$ we can define a new welfare function $\widehat{W}(\widehat{p},\lambda)$. Maximising $W(p,\lambda)$ with respect to p and λ is then equivalent to maximising $\widehat{W}(\widehat{p},\lambda)$ with respect to $\widehat{p}$ and λ. Taking the partial derivative with respect to λ yields $\frac{\partial \widehat{W}(\widehat{p},\lambda)}{\partial \lambda} = \frac{t}{4}\left(\frac{\widehat{p}}{4t^2k}\right)^{\frac{1}{3}} > 0$. Thus, social welfare is maximised by setting $\lambda = 1$. ■

From (9) and (11) we know that p and λ have identical effects on equilibrium differentiation and quality provision. Thus, by using the price instrument properly, the regulator can induce exactly the same location-quality outcome for any given value of λ. Consider an increase in the share of informed patients in the market. The resulting effects - stronger quality competition and larger differentiation - can be exactly offset by reducing the price accordingly. This would, however, have an unambiguously positive effect on social welfare - even though differentiation and quality provision remain unchanged - since expected aggregate mismatch costs are reduced when fewer patients run the risk of attending the 'wrong' hospital. Thus, the regulator can maximise social welfare by introducing a strict gatekeeping system in order to make all patients fully informed, and then use price regulation to correct for the potential negative effects of increased information.[22]

4 Indirect gatekeeping

In this section we endogenise the share of informed patients, λ. We assume that patients have the choice of consulting a gatekeeping GP, thereby obtaining all relevant information, before accessing the hospital market. To obtain an interior solution for λ we consider cost heterogeneity with respect to GP consultation. Let $y \in [0,1]$ denote the cost

[22] A more detailed discussion of the optimal second-best price in the case of a strict gatekeeping system is presented in Brekke et al. (2002).

type of a patient. The associated costs are assumed to be ay, where $a > 0$. This heterogeneity can simply be justified by an opportunity cost argument, e.g., by varying time costs due to different wage earning abilities. To simplify the analysis we assume that patient types are uniformly distributed on the disease space S.

The GP consultation decision is based on the expected benefits of gatekeeping relative to a patient's cost type. Benefits are in expected terms as prior to consultation none of the patients can observe specialisation, quality and disease. So, in a game-theoretic sense, the consultation decision is simultaneous to specialisation and quality decisions. Since hospitals cannot observe patients' consultation decisions, and since they do not know a patient's cost type, they have to form beliefs about the actual consultation rate. We are solving for the perfect Bayesian equilibrium where expectations will be confirmed. Additionally, we require that beliefs have to be consistent out of equilibrium. This restriction is discussed in some detail below.

4.1 The specialisation-quality game

4.1.1 The demand for secondary care and GP consultation

The demand for hospital 1 is exactly the same as in the previous section for a given share λ of informed patients. So for this subsection it remains to determine the consultation decision.

When deciding whether to approach a (randomly chosen) hospital directly or to consult a gatekeeping GP first, a patient has to weigh the costs of going to a GP against the benefits. Imposing the same symmetry assumption as previously, the quality of hospital care is unimportant for this problem. The quality received is independent of whether a GP was consulted or not. Determining the benefits of gatekeeping simply requires ascertaining the expected reduction in mismatch costs. This requires to forming expectations Δ^e about the degree of product differentiation Δ in the market. We will assume that these expectations are symmetric, which seems plausible as patients are (except for consultation costs) ex-ante identical. As before we will assume that patients know that the equilibrium will be symmetric, i.e., that hospitals locate equidistantly from the market centre, but on opposite sides.

The expected mismatch costs when directly approaching a hospital are

$$MMC^e_{DA} = \frac{t}{2}\int_0^1 \left(z - \frac{1}{2}\left(1-\Delta^e\right)\right)^2 dz + \frac{t}{2}\int_0^1 \left(z - \frac{1}{2}\left(1+\Delta^e\right)\right)^2 dz. \tag{18}$$

When consulting a GP first, expected mismatch costs are reduced to

$$MMC^e_{GP} = t\int_0^{\frac{1}{2}} \left(z - \frac{1}{2}\left(1-\Delta^e\right)\right)^2 dz + t\int_{\frac{1}{2}}^1 \left(z - \frac{1}{2}\left(1+\Delta^e\right)\right)^2 dz. \tag{19}$$

The expected benefit of gatekeeping, $B^e := MMC^e_{DA} - MMC^e_{GP}$, is thus

$$B^e = \frac{t\Delta^e}{4}. \tag{20}$$

The equilibrium value of λ is obtained by equating the expected benefits of gatekeeping to its actual costs, $t\Delta^e/4 = a\lambda$, and solving for λ, yielding

$$\lambda = \frac{t\Delta^e}{4a}. \tag{21}$$

The comparative static results are intuitive: the higher the costs of consulting a GP, a, the lower the share of patients actually going to a GP. The benefits of gatekeeping are increasing in mismatch costs, since some costs may be avoided by obtaining information. Aggregate expected mismatch costs are determined by two different factors. For any given positive distance between the hospitals, these costs are obviously increasing in the mismatch cost parameter t. In addition, expected aggregate mismatch costs are increasing in the degree of horizontal differentiation. The further apart the hospitals are located, the more costly, in terms of mismatch costs, to attend the 'wrong' hospital.

4.1.2 Quality competition and specialisation

We now assume that for *any* patient consultation strategy hospitals' expectations about the fraction of informed consumers is equal to the actual fraction induced by that strategy. In other words, we require beliefs to be consistent out of equilibrium. With this a hospital's best response against patients strategies can be written in terms of the fraction of informed patients. Notice that the restriction on strategies that the lowest cost types demand GP referral, as used in equation (21), is irrelevant from a hospital perspective. Only the expected share of informed patients matters and not their actual composition.

The equilibrium of the quality subgame is thus simply obtained by substituting λ by λ^e in equation (5), yielding

$$q^{**}\left(\Delta, \lambda^e; p\right) = \frac{p\lambda^e}{4tk\Delta}. \tag{22}$$

The comparative static properties are comparable to those with direct gatekeeping. The same applies to product differentiation, which is obtained in a similar fashion:

$$\Delta^{**}\left(\lambda^e; p\right) = \left(\frac{p\lambda^e}{4t^2k}\right)^{\frac{1}{3}}. \tag{23}$$

4.1.3 The solution of the game

Imposing rational expectations, we obtain the solution of the game. One requirement is that hospitals' expectations are confirmed, $\lambda^e = \lambda$, and the other that patients' expectations are confirmed, $\Delta^e = \Delta$. By inserting equation (21) into equation (23), and solving for Δ, we obtain equilibrium product differentiation

$$\Delta_u^{**} = 0 \text{ and } \Delta_s^{**} = \frac{1}{4}\left(\frac{p}{tka}\right)^{\frac{1}{2}}, \tag{24}$$

where the subscript 'u' indicates that the equilibrium is unstable, and 's' that it is stable. This will be shown in the proposition below.

The corresponding quality levels are obtained by substituting equation (23) into (22), taking the relationship in equation (21) into account, yielding

$$q_u^{**} = 0 \text{ and } q_s^{**} = \frac{p}{16ak}. \tag{25}$$

Using the equilibrium product differentiation displayed in equation (24), we can solve for the share of GP patients by substituting the respective values into equation (21):

$$\lambda_u^{**} = 0 \text{ and } \lambda_s^{**} = \frac{1}{16}\left(\frac{pt}{ka^3}\right)^{\frac{1}{2}}. \tag{26}$$

Proposition 3 *With endogenous GP consultation there are two symmetric equilibria of the specialisation-quality-consultation game. (i)* $S_u^{**} = (\Delta_u^{**}, q_u^{**}, \lambda_u^{**})$, *and (ii)* $S_s^{**} = (\Delta_s^{**}, q_s^{**}, \lambda_s^{**})$, *where* $\Delta_i^{**}, q_i^{**}, \lambda_i^{**}$, $i = u, s$, *are given by equations (24), (25), and (26). The equilibrium given in (i) is unstable and the equilibrium shown in (ii) is stable.*

Proof. (i) That S_u^{**} is an equilibrium of the game is straightforward. Consider that all patients expect that the hospitals will not differentiate, $\Delta^e = 0$. Given these expectations, the benefits of gatekeeping are zero, $B^e = 0$. Since there are positive costs of consulting a GP, nobody will actually go to a gatekeeper, i.e. $\lambda = 0$. Hospitals correctly anticipate these expectations: they know that patients are completely uninformed and thus not responsive to quality investments. Consequently, hospitals set $q_u^{**} = 0$. Hospitals do not incur any costs in this scenario so they cannot do better than confirm expectations on specialisation and agglomerate at the same point. However, this equilibrium is unstable when expectations have to be consistent out of equilibrium. Then, wrong expectations on λ on the hospitals' side directly translate into wrong expectations on Δ when taking equation (21) into account. Specialisation is then

$$\Delta^{**} = \left(\frac{p\Delta^e}{16atk}\right)^{\frac{1}{3}}. \tag{27}$$

Now consider that $\Delta^e \in (0, \Delta_s^{**})$. Since $\Delta^{**}(\Delta_s^{**}) = \Delta_s^{**}$ and Δ^{**} is increasing and concave in Δ^e, this implies $\Delta^{**}(\Delta^e) > \Delta^e$. As expectations are not confirmed, Δ^e cannot be an equilibrium of the game. Moreover, since actual differentiation exceeds expectations, there is no force driving expectations back to zero, proving instability.

(ii) Here it remains to show that the equilibrium is stable. First, assume that $\Delta^e \in (0, \Delta_s^{**})$, then - as above - we have that $\Delta^{**}(\Delta^e) > \Delta^e$, driving expectations back to Δ_s^{**}. Second, consider that $\Delta^e > \Delta_s^{**}$; then we have $\Delta^{**}(\Delta^e) < \Delta^e$. This not only proves that $\Delta^e > \Delta_s^{**}$ can never be an equilibrium, but also that expectations will be driven back towards Δ_s^{**}. ■

Requiring consistent beliefs out of equilibrium enables us to prove instability of S_u^{**} and stability of S_s^{**}. Although it seems plausible to concentrate on S_s^{**} note that S_u^{**} can easily be supported as a stable equilibrium when consistency is not imposed. Consider that hospitals expect that nobody will become informed, $\lambda^e = 0$, independent of what patients' strategies would actually suggest. Then hospitals do not differentiate and there are zero benefits of gatekeeping. As patients correctly anticipate missing specialisation incentives, $\lambda^e = 0$ will always be confirmed.

As we concentrate on the stable equilibrium, we ease notation by suppressing the index 's' in the remainder of the paper. In order to analyse the comparative static properties of the equilibrium it is convenient to neglect the restriction $\lambda^{**} \in [0, 1]$. We will be more precise about that later.

The share of the population going to a gatekeeping GP increases in the mismatch cost parameter, t, as this drives up the benefits of gatekeeping. It also increases in price. This is an indirect effect stemming from specialisation. Price increases boost quality competition and, to dampen this effect, hospitals aim at reducing the substitutability of their products, increasing the benefits of gatekeeping. Obviously, λ^{**} is a decreasing function of a. The higher the disutility incurred by consulting a GP the lower the share of patients who actually consult one. Finally, an increase in the quality cost parameter, k, reduces quality competition and thereby differentiation incentives. This, in turn, reduces the benefits of gatekeeping.

Compared to the previous section, there are two major differences in the comparative static properties of quality. Firstly, the mismatch costs parameter has no effect. With direct gatekeeping, patients were more responsive to quality investments at lower values of t, amplifying quality competition. As can be seen from equation (22), this is also true here, but this effect is opposed by the consultation effect. A lower t reduces

the benefits of gatekeeping, resulting in a less competitive market. With linear costs of GP consulting these two effects exactly offset. Secondly, an increase in consulting costs, a, lowers λ^{**} and thereby softens quality competition. The latter effect is also present with respect to the specialisation decisions: high consulting costs weaken the competition effect, making product differentiation less desirable.

4.2 Social welfare

Subtracting the consulting cost term $a\int_0^{\lambda} xdx$, the social welfare function can be rewritten from equation (12),

$$W = v + q - 2kq^2 - \frac{t}{12} + \frac{t}{4}\Delta\left(\lambda - \Delta\right) - \frac{1}{2}a\lambda^2. \tag{28}$$

The first-order conditions for the first-best solution are obtained by differentiation, yielding

$$\Delta^{fb} = \frac{\lambda}{2},\ q^{fb} = \frac{1}{4k},\ \text{and}\ \lambda^{fb} = \frac{t\Delta}{4a}. \tag{29}$$

First-best quality is identical to the previous analysis, and depends only on the costs of quality investments. Specialisation is the second-best from Section 3.2.1, and is thus conditional on the share of GP patients. Most importantly, although not surprising, the first-best λ coincides with individual decisions, so there is no need to regulate gatekeeping directly.

Proposition 4 *The first-best efficient solution of the game with endogenous gatekeeping has quality* $q^{fb} = \frac{1}{4k}$ *and*
(i) $\Delta^{fb} = 0$ *and* $\lambda^{fb} = 0$ *for* $t < 8a$,
(ii) $\lambda^{fb} \in [0,1]$ *and* $\Delta^{fb} = \lambda^{fb}/2$ *for* $t = 8a$, *and*
(iii) $\Delta^{fb} = \frac{1}{2}$ *and* $\lambda^{fb} = 1$ *for* $t > 8a$.

Proof. The first-best solution in (i) is an interior solution where both first-order conditions, $\Delta^{fb} = \frac{\lambda}{2}$ and $\lambda^{fb} = \frac{t\Delta}{4a}$, are satisfied. As λ is restricted to the unit interval there are situations where $\lambda^{fb} = \frac{t\Delta}{4a}$ does not hold, i.e. when parameters are such that λ exceeds one. Considering the unrestricted Hotelling model the first order condition for Δ can always be met. Inserting $\Delta = \frac{\lambda}{2}$ into (28) and differentiating yields $\frac{\partial W}{\partial \lambda} = \lambda\left(\frac{t}{8} - a\right)$. For $t > 8a$ the regulator sets λ to its maximum, $\lambda^{fb} = 1$, and $\Delta^{fb} = 1/2$. Obviously, the regulator is indifferent between all feasible values of λ when $t = 8a$. ■

The mismatch cost parameter has two opposing effects on the benefits of gatekeeping. On the one hand, a higher value of t increases expected aggregate mismatch costs for every given pair of hospital locations. On the other hand, an increase in t reduces quality competition

and leads to less differentiation, which reduces the benefits of gatekeeping. Using (20) and (24) it is straightforward to show that the former (direct) effect always dominates, i.e., $\partial B^e/\partial t > 0$, implying that the benefits of gatekeeping are increasing in the mismatch cost parameter t. Consequently, when t is sufficiently low relative to GP consulting costs, $t < 8a$, the benefits of increased information, in terms of reduced expected mismatch costs, are outweighed by the costs of going to a GP.

4.3 Price regulation

When setting the optimal price, the regulator trades off inefficiencies along three different dimensions: quality provision, horizontal differentiation and GP consultation. In general, this will not result in a strict gatekeeping regime. The objective function is obtained by inserting the equilibrium values of the specialisation-quality-consultation game into the welfare function. The candidate second-best price is then found to be

$$p^{sb} = \frac{t}{8} + 3a. \tag{30}$$

The following equilibrium obtains:

$$\Delta^{sb} = \frac{\left[2tka\left(t+24a\right)\right]^{\frac{1}{2}}}{16tka},\ q^{sb} = \frac{24a+t}{128ak},\ \text{and}\ \lambda^{sb} = \frac{\left[\frac{2t(24a+t)}{ka^3}\right]^{\frac{1}{2}}}{64}. \tag{31}$$

The price equilibrium given in (30) exists if $\left(x_1^{**}\left(p^{sb}\right), x_2^{**}\left(p^{sb}\right)\right) \in Q$. Using (31), the existence condition (provided that the solution is interior) is given by

$$k > \frac{24a+t}{1024a^2}. \tag{32}$$

It is straightforward to show that the second-best price given by (30) is an interior solution for a subset of the parameter values, defined by

$$k > \overline{\overline{k}} := \frac{t\left(24a+t\right)}{2048a^3}.$$

Thus, if $k < \overline{\overline{k}}$ we have a corner solution with $\lambda^{sb} = 1$.[23] Given that $\partial\overline{\overline{k}}/\partial t > 0$ and $\partial\overline{\overline{k}}/\partial a < 0$, the implications for indirect gatekeeping follow immediately:

Proposition 5 *Second-best price regulation implies a de facto strict gatekeeping regime if quality costs or GP consulting costs are sufficiently small, or if mismatch costs are sufficiently high.*

[23] From (32) it follows that an interior solution always meets the existence condition if $a < t/2$.

The intuition is basically the same as before: gatekeeping is a way of reducing expected aggregate mismatch costs, and the social benefits of gatekeeping are consequently linked to these costs that are increasing in t and decreasing in k. Of course, the role of GP consulting costs is straightforward: the smaller a, the larger share of the population consults a GP to obtain information.

In the following we will only consider interior solutions.[24] Thus, GP consultation will generally be inefficient. The efficiency properties of the second-best interior solution are summarised as follows:

Proposition 6 *The second-best (interior) solution of the specialisation-quality-consultation game has the following properties:*

(i) for $t < 8a$ there is too much differentiation given λ^{sb} and inefficiently low quality provision,

(ii) for $t = 8a$ the first-best optimum is implemented, $\Delta^{sb} = \frac{1}{8}\left(\frac{2}{ak}\right)^{\frac{1}{2}}$, $q^{sb} = \frac{1}{4k}$ and $\lambda^{sb} = \frac{1}{4}\left(\frac{2}{ak}\right)^{\frac{1}{2}}$,

(iii) for $t > 8a$ there is insufficient differentiation given λ^{sb} and inefficiently high quality provision.

Proof. First-best specialisation, conditional on the share of GP-patients, requires $\Delta^{sb} = \lambda^{sb}/2$. From (31) we find that $\Delta^{sb} - \lambda^{sb}/2 = \frac{8a-t}{128}\left[\frac{2(24a+t)}{tka^3}\right] > (<)\,0$ if $t < (>)\,8a$. First-best quality is given by $q^{fb} = \frac{1}{4k}$. From (31) we have that $q^{sb} - q^{fb} = -\frac{(8a-t)}{128ka} < (>)\,0$ if $t < (>)\,8a$. ■

We have an interior solution if GP consultation costs are sufficiently high. From (25) we know that a high value of a implies that quality provision will be relatively low in equilibrium. We also know that a price above marginal costs is in any case a necessary condition to prevent under-provision of quality. If, in addition, mismatch costs are relatively low, the value of obtaining information will be limited and, consequently, GP consulting will be low in equilibrium. Since the first-best efficient level of quality provision is independent of mismatch costs, this implies that social welfare is maximised at a low degree of differentiation. In this case, $t < 8a$, the price that yields first-best differentiation is not high enough to generate efficient quality provision. Thus, higher quality can only be obtained at the expense of excessive differentiation, and these considerations are optimally traded off at a price which yields under-provision of quality and too much differentiation.

On the other hand, if mismatch costs are high, the first-best level of differentiation will be higher - closer to $\frac{1}{2}$ - due to higher GP consultation.

[24] For a discussion of optimal price regulation in the corner solution (strict gatekeeping), see Brekke et al. (2002).

In this case, $t > 8a$, the optimal degree of differentiation is obtained at a price that yields over-provision of quality. Consequently, optimal regulation implies accepting a less than optimal degree of differentiation in order to avoid too much over-investment in quality.

5 Concluding remarks

Equipping GPs with a gatekeeper role in the health care system is a major issue in the debate on health care reforms. Among politicians, the conventional wisdom is that gatekeeping contributes to cost control. This is somewhat surprising since evidence is lacking, as was demonstrated in an empirical study by Barros (1998). Economists are more concerned about efficiency arguments rather than fiscal ones. As GPs are usually better informed than patients about the characteristics of the secondary health care market, e.g. about quality and specialisation of hospitals, matching of patients to hospitals may indeed be improved by gatekeeping. However, this argument neglects the potential competitive effects in the hospital market. We presented a model that analyses the competitive effects of gatekeeping in the presence of non-price competition.

While prices were regulated, we allowed for competition in specialisation and quality. We considered two versions of the basic model, one in which the share of ex ante informed patients is exogenously given (direct gatekeeping), and another where the share of informed patients is endogenously determined (indirect gatekeeping).

In the direct gatekeeping scenario we assumed gatekeeping to be costless. We found that when the price is exogenously given strict gatekeeping does not necessarily improve social welfare. This is the case when the additional information acquired by GPs boosts competition to such an extend that excessive specialisation of hospitals occurs. In these cases, due to the endogeneity of specialisations, mismatch costs are higher with gatekeeping than without. This raises doubts about whether gatekeeping improves efficiency. Things change dramatically when allowing for second-best price regulation. We demonstrated that strict gatekeeping, i.e. GP consultation is compulsory before accessing the secondary care market, is always socially desirable. Thus, direct gatekeeping should always be accompanied by proper price regulation.

Gatekeeping was endogenised by introducing cost heterogeneity with respect to GP consultation. Since consultation decisions of patients are the same to what a social planner would implement, there is no need for direct regulation of gatekeeping. GP consultation can be indirectly influenced by price regulation. With second-best price regulation, a strict gatekeeping regime obtains if the benefits of gatekeeping are sufficiently

high (improved matching outweighs the potentially negative competitive effects) compared to its costs. When the share of GP patients is below one, the second-best outcome will, in general, be inefficient. Depending on the parameters, there may be too much differentiation and too low quality or vice versa. Direct implementation of a strict gatekeeping regime may again reduce social welfare.

The analysis demonstrates that efficiency gains that are usually attributed to GP gatekeeping cannot be taken for granted when non-price competition is incorporated into the analysis. In the short run, efficiency gains may indeed be obtained by better matches. However, quality provision may still be inefficient. In the long run, hospitals will adjust their specialisation so that differentiation increases and this counteracts the positive short run effect.

References

[1] Arrow, K., 1963. Uncertainty and the welfare economics of medical care. American Economic Review 53, 941-973.

[2] Barros, P.P., 1998. The black box of health care expenditure growth determinants. Health Economics 7, 533-544.

[3] Barros, P.P and X. Martinez-Giralt, 2002. Public and private provision of health care. Journal of Economics & Management Strategy 11, 109-133.

[4] Baye, M.R. and J. Morgan, 2001. Information gatekeepers on the internet and the competitiveness of homogeneous product markets. American Economic Review 91, 454-474.

[5] Bester, H., 1998. Quality uncertainty mitigates product differentiation. RAND Journal of Economics 29, 828-844.

[6] Brekke, K., R. Nuscheler and O.R. Straume, 2002. Quality and location choices under price regulation. Working Papers in Economics 24/2002, University of Bergen.

[7] Burdett, K. and K.L. Judd, 1983. Equilibrium price dispersion. Econometrica 51, 955-969.

[8] Calem, P.S. and J.A. Rizzo, 1995. Competition and specialization in the hospital industry: An application of Hotelling's location model. Southern Economic Journal 61, 1182-1198.

[9] Dranove, D., D. Kessler, M. McClelland and M. Satterthwaite, 2003. Is More Information Better? The Effects of "Report Cards" on Health Care Providers. Journal of Political Economy 111, 555-588.

[10] Economides, N., 1984. The principle of minimum differentiation revisited. European Economic Review 24, 345-368.

[11] Economides, N., 1986. Minimal and maximal product differentiation in Hotelling's duopoly. Economics Letters 21, 67-71.

[12] Economides, N., 1989. Quality variations and maximal variety differentiation. Regional Science and Urban Economics 19, 21-29.
[13] Economides, N., 1993. Quality variations in the circular model of variety-differentiated products. Regional Science and Urban Economics 23, 235-257.
[14] Ferris, T.G, Y. Chang, D. Blumenthal and S.D. Pearson, 2001. Leaving gatekeeping behind - effects of opening access to specialists for adults in a Health Maintenance Organization. The New England Journal of Medicine 345, 1312-1317.
[15] Gravelle, H., 1999. Capitation contracts: Access and quality. Journal of Health Economics 18, 315-340.
[16] Gravelle, H. and G. Masiero, 2000. Quality incentives in a regulated market with imperfect information and switching costs: Capitation in general practice. Journal of Health Economics 19, 1067-1088.
[17] Hinloopen, J. and C. van Marrewijk, 1999. On the limits and possibilities of the principle of minimum differentiation. International Journal of Industrial Organization 17, 735-750.
[18] Hotelling, H., 1929. Stability in competition. Economic Journal 39, 41-57.
[19] Lambertini, L., 2001. Vertical differentiation in a generalized model of spatial competition. The Annals of Regional Science 35, 227-238.
[20] Lommerud, K.E. and L. Sørgard, 2003. Entry in telecommunication: customer loyalty, price sensitivity and access prices. Information Economics and Policy 15, 55-72.
[21] Lyon, T.P., 1999. Quality competition, insurance, and consumer choice in health care markets. Journal of Economics & Management Strategy 8, 545-580.
[22] Nuscheler, R., 2003. Physician reimbursement, time-consistency and the quality of care. Journal of Institutional and Theoretical Economics 159, 302-322.
[23] Salop, S.C., 1979. Monopolistic competition with outside goods. Bell Journal of Economics 10, 141-156.
[24] Schultz, C., 2002. Market transparency and product differentiation. CIE Discussion Paper 2002-02. Centre for Industrial Economics, University of Copenhagen.
[25] Schultz, C., 2003. Transparency on the consumer side and tacit collusion. Mimeo, European Economic Review, forthcoming.
[26] Scott, A., 2000. Economics of general practice, in: Culyer, A.J. and J.P. Newhouse (Eds.), Handbook of health economics, Vol. 1B. North-Holland, Amsterdam, pp. 1175-1200.
[27] Varian, H., 1980. A model of sales. American Economic Review 70, 651-659.

Symbols in order of appearance in the text

$S = [0,1]$	Disease set
$z \in S$	Disease
$Z \sim U(S)$	A disease is random, uniform distribution on S
$x_i \in \mathbf{R}$, $i = 1,2$	Specialisation of secondary care provider i
$q_i \geq 0$, $i = 1,2$	Quality of secondary care of provider i
$k > 0$	Quality cost parameter
$p \geq 0$	Price for secondary care treatment
D_i, $i = 1,2$	Demand for treatment at provider i
Π_i, $i = 1,2$	Profit of provider i
u_i, $i = 1,2$	(ex post) utility when consulting provider i
v	Quality independent willingness to pay for secondary care treatment
$t > 0$	Mismatch cost parameter
$\lambda \in [0,1]$	Fraction of informed patients in the population
Eu_i	Expected utility when consulting provider i
$\overline{z}$	Disease of the (informed) patient who is indifferent between providers
$\Delta := x_2 - x_1$	Degree of product differentiation
Q	For $x_1, x_2 \in Q$ an equilibrium of the quality subgame exists
index "sb"	stands for "second best optimum"
index "fb"	stands for "first best optimum"
W	Social welfare function
$y \in [0,1]$	Consultation cost type of a patient
$a > 0$	Consultation cost parameter
MMC	Expected mismatch costs with direct access or gatekeeping
B	Benefits of consulting a gatekeeping general practitioner
S_i^*, $i = u, s$	Unstable (u) and stable (s) equilibrium of the indirect gatekeeping game