Gatekeeping in health care

Kurt R. Brekke [a], Robert Nuscheler [b,*], Odd Rune Straume [c]

[a] *Department of Economics, Health Economics Bergen (HEB), University of Bergen, Herman Fossgate 6, N-5007 Bergen, Norway*
[b] *Wissenschaftszentrum Berlin für Sozialforschung (WZB), Reichpietschufer 50, D-10785 Berlin, Germany*
[c] *Department of Economics, School of Economics and Management, University of Minho, Braga, Portugal and Health Economics Bergen (HEB), Norway*

Received 1 May 2004; received in revised form 1 September 2005; accepted 1 April 2006
Available online 4 August 2006

Abstract

We study the competitive effects of restricting direct access to secondary care by gatekeeping, focusing on the informational role of general practitioners (GPs). In the secondary care market there are two hospitals choosing quality and specialization. Patients, who are ex ante uninformed, can consult a GP to receive an (imperfect) diagnosis and obtain information about the secondary care market. We show that hospital competition is amplified by higher GP attendance but dampened by improved diagnosing accuracy. Therefore, compulsory gatekeeping may result in excessive quality competition and too much specialization, unless the mismatch costs and the diagnosing accuracy are sufficiently high. Second-best price regulation makes direct regulation of GP consultation redundant, but will generally not implement first-best.

JEL classification: D82; I11; I18; L13

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

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

* Corresponding author. Tel.: +49 30 25491 408; fax: +49 30 25491 400.
E-mail addresses: kurt.brekke@nhh.no (K.R. Brekke), robert@wz-berlin.de (R. Nuscheler), o.r.straume@eeg.uminho.pt (O.R. Straume).

specialist.[1] In the US, several health maintenance organizations (HMOs) also practice gatekeeping. Recently, however, some HMOs have relaxed the restrictions on access to specialists (see, e.g., Ferris et al., 2001). In Germany, patients need a referral to get access to a hospital and it has been on the political agenda to also restrict direct access to specialist care by giving GPs a gatekeeper role. The international experience with gatekeeping thus appears to be mixed: while some countries relax gatekeeping regulations (e.g., the US), others seem to move towards stricter rules (e.g., Germany). The current paper contributes to the discussion on gatekeeping by analyzing 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). First, it is usually claimed that gatekeepers contribute to cost control by reducing 'unnecessary' interventions.[2] 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 various informational problems make health care markets distinctly different from most other markets. The present paper stresses the importance of non-price competition between health care providers, as well as the role of imperfect information in the relationship between patients and providers. Building on the familiar model of Hotelling (1929), we consider a secondary care market with two providers (hospitals). In order to attract patients (and obtain third party payments) the hospitals have two strategic variables at their disposal: location and quality of care. We refer to location as the specialization 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 GP gatekeepers in secondary care markets. We assume that patients are ex ante uniformed about their specific diagnosis and the exact characteristics of the hospitals. Thus, if they access secondary care providers directly, their 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 hospital that provides the lower quality of care. To reduce the risk of choosing the 'wrong' hospital, patients may therefore (at some costs) consult a GP first. The GPs are informed agents (middlemen) and convey accurate information about hospital characteristics, i.e., quality and specialization. They also give attending patients a noisy diagnosis. Thus, the GPs are imperfect agents in the sense that diagnosing accuracy is not perfect.[3] We abstract from any moral hazard problems that may originate in the agency relationships between players.[4] When deciding whether to consult a GP or to approach a hospital directly, patients simply weigh the consulting costs against the reduction in (expected) mismatch costs due to better information.

[1] In Sweden, though, individuals have direct access to hospital outpatient care, but still need a referral if hospitalization is required.

[2] 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).

[3] Diagnosing accuracy may be determined by several factors like a GP's skills, a GP's effort, a patient's disease type, etc.

[4] The physician agency literature analyzes in detail strategic reasons for GPs to make false reports or to exercise inappropriate levels of (diagnosis) effort (see McGuire, 2000, for an overview). Below we discuss the part of this literature which is relevant for gatekeeping.

The analysis is focused on two basic questions. (i) How does GP gatekeeping affect hospitals' incentives to specialize and to invest in quality? (ii) Is strict gatekeeping – i.e., no access to secondary care without a GP referral – socially desirable? The answers to these two questions are closely connected. Concerning the first question, we show that a higher GP attendance rate amplifies quality competition and induces the hospitals to specialize their services. The former is explained by the fact that informed patients are sensitive to quality differences, while uninformed patients are not. The latter is due the fact that hospitals can dampen quality competition by specializing their services.[5]

Interestingly, the other information variable – diagnosing accuracy – has the exact opposite effect. When diagnosing accuracy is low, patients attending a GP put a larger weight on quality differences than hospital specializations, since the probability of a wrong diagnosis is high. As a consequence, improved diagnosing accuracy tends to weaken quality competition and, in turn, the corresponding incentives for specialization. However, improved diagnosing accuracy also increases the benefit of consulting a GP, leading to higher GP attendance, which, in turn, increases hospital competition. Thus, when the patients' decision of whether or not to attend a GP is endogenized, the latter (indirect) effect of improved diagnosing accuracy on hospital competition tends to counteract the former (direct) effect.[6]

Regarding the second question, numerical simulations of our model suggest that strict gatekeeping is detrimental to welfare unless mismatch costs and diagnosing accuracy are sufficiently high. The reason is that both low mismatch costs and low diagnosing accuracy trigger hospital competition. Since higher GP attendance has the same directional effect on competition, as explained above, strict gatekeeping tightens hospital competition even further. As a consequence, hospitals engage in excessive competition, resulting in too high quality and too much specialization from a welfare perspective.[7]

The regulator (payer) determines the hospital reimbursement by setting a (prospective) price per treatment (or patient). We show that if second-best price regulation is available, then there is no scope for direct regulation of GP attendance. Thus, the treatment price is a sufficient instrument to induce second-best optimal quality and specialization of hospital care. Finally, we characterize the second-best equilibrium, showing that first-best is generally not achievable for the regulator.

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. (2006) under price regulation.[8] The present paper contributes to this literature by introducing

[5] A completely analogical feature is present in the location–price game by D'Aspremont et al. (1979), where firms differentiate (specialize) to soften price competition. Like in their paper, the dampening-of-competition effect dominates the countervailing market-expanding effect of locating closer to your rival. For a more detailed discussion, see Brekke et al. (2006).

[6] In our specific model, with linear GP consultation costs, these two effects exactly offset, so that equilibrium hospital specialization and quality provision are unaffected by the degree of diagnosing accuracy. However, under (enforced or de facto) strict gatekeeping, where every patient attends a GP before receiving secondary care, the indirect effect is eliminated and improved diagnosing accuracy will dampen hospital competition.

[7] This result is related to Dranove et al. (2003), who empirically analyze 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 report cards led to both selection behavior 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).

[8] 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

imperfect information into the framework. As previously mentioned, we find that the hospitals' incentives to differentiate services crucially depend on the degree of information in the market. In particular, we find that the presence of uninformed consumers tends 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 center – when consumers are uncertain about product quality and use observed prices to ascertain the quality of goods.

The paper also relates to the more general literature on transparency in imperfectly competitive markets.[9] 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 analyzing the effects of improved transparency in markets that are characterized 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.[10]

Finally, our paper complements the multi-task agency literature on the economics of general practice, e.g., Garcia Mariñoso and Jelovac (2003), Malcomson (2004) and González (2006). These papers focus on the dual nature of GP activity, namely, on diagnosing patients and treating or referring them. Optimal payment systems are derived that, at the same time, induce GPs to exert diagnosis effort and give incentives for efficient referral or treatment decisions, i.e., GP treatment for low severity diagnoses and referral for high severity diagnoses.[11] This also refers to the second gain of gatekeeping: the allocation of patients to health care sectors improves since patients more appropriately treated by a GP are screened out through costly diagnosing of all patients. On the other hand, as Malcomson (2004) points out, patients who would not otherwise have been referred, may be referred after being subject to costly diagnosis. Again, health care is used more efficiently. In our paper, GPs are – on the one hand – perfect agents in the sense that they truthfully convey the information about the secondary care market that they have, but – on the other hand – imperfect agents in the sense that diagnosing is noisy. Although we consider diagnosing accuracy to be exogenous, it can, in fact, be seen as a result of an incentive contract like the ones derived in the above cited papers. Instead of analyzing whether or not a patient should be referred to a hospital, we consider that all patients will be referred and concentrate on the improved matching of patients to hospitals through gatekeeping GPs.[12] Although important for the social desirability of gatekeeping, this has not been analyzed before. Moreover, we explicitly model the secondary care sector and introduce imperfect competition, and thereby significantly advance the literature. We demonstrate that the information acquired through gatekeeping affects competition amongst secondary care providers and that this may generate – so far neglected – (adverse) effects of such a system.

---

when prices are regulated. 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. Calem and Rizzo (1995) also analyze horizontal and vertical differentiation of hospitals. However, in contrast to our paper, they neither consider price regulation nor gatekeeping.

[9] See, e.g., Varian (1980), Burdett and Judd (1983), Lommerud and Sørgard (2003) and Schultz (2004, 2005).

[10] Another related paper in this strand of the literature is Baye and Morgan (2001), who analyze 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.

[11] Given the optimal contracts, the question of whether a gatekeeping system dominates free access to secondary care is analyzed. Without going into details here, the results are ambiguous.

[12] In the agency literature cited above, high severity patients finally end up with a specialist as GPs are assumed to be unable to cure these patients. In this sense, our analysis deals with matching of high severity patients to specialists or hospitals.

The remainder of the paper is organized as follows. The basic framework is presented in Section 2. In Section 3, we analyze hospitals' incentives for specialization and quality investments for a given GP attendance rate. In Section 4, we endogenize the GP attendance rate and characterize the corresponding specialization–quality–consultation equilibrium. Section 5 is devoted to welfare effects of gatekeeping and regulation of GP attendance, as well as second-best price regulation. Finally, in Section 6, we provide some concluding remarks.

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 realization 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. Specialization of a hospital – interpreted as a location on S – is denoted x_i, $i=1,\ 2$. Like in Brekke et al. (2006) we make the following assumptions on hospital specializations: $x_1 \in [0,\ \frac{1}{2} - \bar{x}]$ and $x_2 \in [\frac{1}{2} + \bar{x},\ 1]$, where $\bar{x}$ is a (small) positive number. This is done in order to secure existence of pure strategy equilibria throughout the analysis. As specialization is typically difficult to measure we consider x_1 and x_2 non-contractible.[13]

In addition to specialization, there is a second strategic variable used by the hospitals to attract patients, namely the quality of care $q_i \in [\underline{q}, \bar{q}]$, $i=1,\ 2$, where $\underline{q}$ is the minimum quality level allowed by the regulator, and any $q_i < \underline{q}$ can be thought of as malpractice. Apart from malpractice litigation quality is, due to measurement problems, considered not verifiable in a contractual sense. Without loss of generality we assume that $\underline{q} = 0$. Quality costs are assumed to be symmetric and quadratic, kq_i^2, where $k>0$. Placing an upper bound $\bar{q}$ on quality investments is a (crude) way of capturing that it is insurmountably costly to increase quality beyond a certain level.[14] Quality 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.[15] 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.

[13] One may also argue that, although specialization is non-contractible, the regulator is able to prevent that hospitals locate too closely. From the regulator's perspective too close locations may be undesirable since this would imply duplication of fixed costs without the benefit of diversified hospital services.

[14] We can, for instance, think of $\bar{q}$ as the best (state-of-the-art) technology or medical procedure available in the market. Thus, increasing quality above this level is not possible.

[15] 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). Including variable quality costs would obviously imply a more general quality cost structure. However, since prices are fixed in our model, variable quality costs would only weaken the incentives for investing in quality. It can readily be verified that this only complicates the analysis without providing any qualitatively different results. Interested readers may consult Ma and Burgess (1993) for the case of fixed locations or contact the authors for the case of endogenous locations.

The price for one treatment is denoted $p \geq 0$ and is set by some regulatory authority.[16,17] The expected profit of hospital i is given by

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

where D_i is expected demand for hospital i treatment.

Patients derive utility from the quality of hospital care. Furthermore, we assume that a patient's utility is decreasing in the distance between the patient's location (disease) and the location (specialization) of the hospital where (s)he is treated. The utility loss incurred from being treated by a less than perfectly suitable provider of care is referred to as 'mismatch costs'. More specifically, a patient's (ex post) utility when going to hospital i is given by[18]

$$u_i^z = v + q_i - t(z - x_i)^2. \tag{2}$$

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.[19] 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). One may also argue that waiting lists are implicitly modelled as part of the quality decision: a longer list implies lower quality for patients and lower costs for hospitals.[20] The last term measures the mismatch costs incurred – assumed to be quadratically increasing in distance – when treated by hospital $i = 1, 2$. The parameter $t > 0$ determines the importance of mismatch costs relative to the quality of care.

Patients are ex ante uninformed about both their own diagnosis and the qualities and specializations of hospitals. They only know v, the distribution of Z, and that hospital treatment is required, but they cannot observe x_i, q_i and z. For uninformed patients, secondary care is an experience good, and the ex post utility given by (2) can only be learned through actual consumption. However, patients can obtain more information ex ante by consulting a GP before accessing the hospital market. We assume that a GP will convey accurate information about the secondary care market, i.e., hospitals' qualities and specializations, and give the attending patient a diagnosis, i.e., a location on S. This diagnosis is noisy, though. We assume that the GP will provide the correct diagnosis with an exogenous probability $\delta \in (0, 1)$, which we henceforth term 'diagnosing accuracy'. We then make the simplifying assumption that incorrect diagnoses – that occur with probability $(1 - \delta)$ – are uniformly distributed on S.[21] Both the probability of a correct diagnosis and the distribution of incorrect diagnoses are common knowledge. Thus, the GP is a perfect

[16] We mainly think of the price p as a third party payment from the regulator (payer) to the hospitals. Since in our model all individuals are ill and in need for one unit of care, the price p can be interpreted either as a payment per treatment (e.g., DRG-pricing) or per individual (capitation).

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

[18] We could easily include a patient co-payment in the utility function, like, for instance, αp, where $\alpha \in [0, 1]$ is the co-payment rate, or a flat fee $f > 0$. However, this will not affect any of our results, as long as the co-payments are set by the regulator.

[19] 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.

[20] We thank an anonymous referee for suggesting this interpretation.

[21] This assumption eases the presentation of results, while still preserving the relevant features of imperfect diagnosing. It may be more realistic to assume that the densities of incorrect diagnoses are higher in the neighborhood of the true location of a patient. Note, however, that the masspoint at the true location in fact approximates such a density.

agent in the sense that all information is truthfully conveyed to those patients consulting the GP, but an imperfect agent in the sense that diagnosing accuracy is not perfect.[22]

Realistically, there are some individual costs associated with attending a GP to obtain information. To incorporate this, we assume cost heterogeneity with respect to GP consultation, where $y \in [0, 1]$ denotes the cost type of a patient. The associated costs are then 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.[23] There are no other (direct) costs of gatekeeping. To simplify the analysis we assume that patient types are uniformly distributed on the disease space S. As a result, patients are uniformly distributed on the unit square with the disease (or diagnosis) on one axis and cost type on the other. The share of patients who obtain information through GP consultation is denoted by λ. This share is determined either by free choice (voluntary gatekeeping) or by direct regulation (compulsory gatekeeping).

The available regulatory instruments for a social planner are assumed to be λ and p, while hospital quality as well as hospital specialization are not verifiable in a contractual sense.[24] Regarding regulation on λ, it is – in theory – 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 seeking secondary care. Thus, the scope for regulating λ is restricted to setting $\lambda = 1$.

The effect of GP gatekeeping to the market for secondary care is analyzed in a five-stage game:

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 $\lambda = 1$.
2. The hospitals simultaneously decide on their specializations, $x_1 \in [0, \frac{1}{2} - \bar{x}]$ and $x_2 \in [\frac{1}{2} + \bar{x}, 1]$.
3. The hospitals simultaneously set their quality levels $q_1 \in [0, \bar{q}]$ and $q_2 \in [0, \bar{q}]$.
4. Patients choose whether to consult a gatekeeping general practitioner and obtain accurate information about x_i and q_i, and a diagnosis with accuracy $\delta < 1$. If the regulator introduced compulsory gatekeeping at stage 1, there is no choice patients have to make at this stage of the game.
5. All patients choose a hospital for 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 specialization 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

[22] The assumption of perfect GP information on hospital characteristics is made for simplification only. The mechanisms of our model are at work as long as GP consultation leads to more information on hospital characteristics. Obviously, the benefit of a gatekeeping system is lower the poorer GP information.

[23] Without cost heterogeneity either all patients or no patients would approach a gatekeeping GP.

[24] This assumption is appropriate as the quality of care and the degree of specialization are, in general, difficult to measure. To some extent the regulator may be able to control hospital quality and specializations. We capture this by imposing the restrictions that hospitals must provide quality above a minimum threshold, and cannot offer exactly the same services (locate very close to each other).

change of specialization, 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.[25] This sequential structure 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 can easily be justified since introducing a strict gatekeeping system (i.e., setting $\lambda = 1$) 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. (2006) and Nuscheler (2003) there will be an incentive to reoptimize after specializations 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.[26] In any case, since price regulation is not the major focus of the present paper, we will concentrate on the full commitment case.

Although we have a game of imperfect information (the fraction $1 - \lambda$ of the population is uninformed about hospital quality, hospital specialization and about their own disease; the fraction $\lambda(1 - \delta)$ receives accurate quality and specialization information but a wrong diagnosis), subgame perfection is the appropriate solution concept. Note that the standard Nash assumption applies zero conjectural variations, i.e., hospitals optimize their own actions against a given action of the competitor (see, e.g., Bresnahan, 1981). Moreover, there is no collusion amongst hospitals.[27] We solve the game by backward induction, starting with the demand for hospital care. Hospitals then play their sequential specialization–quality game for a given value of λ. This yields reaction functions $x_i^*(\lambda)$ and $q_i^*(\lambda)$ for $i = 1, 2$. This game is analyzed in Section 3.

As hospitals have no means to 'signal' their characteristics, neither specializations nor qualities are observed by patients, although hospitals move before patients decide about whether to consult a GP (and obtain information) or not. Therefore, patients have to decide on GP consultation for given values of the hospitals' strategic variables. So, in a game-theoretic sense, consultation decisions are simultaneous to the specialization–quality game. A reaction function $\lambda^*(x_1, x_2, q_1, q_2)$ results, and the equilibrium of the specialization–quality–consultation subgame is then, as usual, the intersection of the reaction functions where actions are mutually best responses. This subgame is analyzed in Section 4.

The solution of the full game is relegated to Section 5, where social welfare and price regulation is investigated.

3. Hospital specialization and quality

3.1. The demand for secondary care

A share $1 - \lambda$ of the population does not consult a GP, and thus remains uninformed about the actual quality levels and about specializations. Moreover, these patients do not know the exact disease they suffer from. To make a decision about which hospital to approach, patients have to evaluate their expected utility of attending each hospital. As the game is fully symmetric and since

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

[26] The assumption that a regulator can credibly commit to a given price (or, more generally, a given transfer) is extensively applied in the literature; see, e.g., Ma and Burgess (1993), Wolinsky (1997) and Beitia (2003).

[27] Schultz (2005) analyzes the effect of market transparency on tacit price collusion in a Hotelling framework.

hospitals have no means to signal their characteristics, we adopt the standard tie-breaking rule where both hospitals receive half of the uninformed patients, $(1-\lambda)/2$. Any other tie-breaking rule would yield qualitatively similar results. As we concentrate on symmetric equilibria, we also impose symmetry here.

The residual fraction of the population, λ, consults a GP and obtains (perfect) information about hospital characteristics. These patients are responsive to quality investments and specialization decisions, since both strategic variables are observed. Furthermore, the patients consulting a GP receive an imperfect diagnosis. The probability of getting a correct diagnosis is δ and is independent of the disease. If a patient receives a diagnosis z, the probability that he actually suffers from disease z is δ. With the remaining probability, $1-\delta$, z is just a draw from the uniform distribution over the unit interval S. Thus, the expected utility of hospital i treatment, for a patient who has received a diagnosis z, is given by

$$Eu_i^z = v + q_i - \delta t(z - x_i)^2 - (1-\delta)t \int_0^1 (s - x_i)^2 \, \mathrm{d}s \tag{3}$$

Consider $q_1 \in [q_1^l,\ q_1^h]$, where $q_1^l = q_2 + t(x_2 - x_1)(1 - x_2 - x_1) - t\delta(x_2 - x_1)$ and $q_1^h = q_2 + t(x_2 - x_1)(1 - x_2 - x_1) + t\delta(x_2 - x_1)$. Then there exists a unique diagnosis, $\bar{z} \in [0, 1]$, such that a patient who receives this diagnosis is, in expectation, indifferent between the two hospitals. This diagnosis is found by solving $Eu_1^z = Eu_2^z$ for z. If $q_1 < q_1^l$ all patients are – independent of the diagnosis they receive – strictly better off with hospital 2, i.e., hospital 1 gets no demand from informed consumers. If $q_1 > q_1^h$ all patients are strictly better off with hospital 1 treatment. The expected demand for hospital 1 from GP-patients is then given by the expected number of patients who receive a diagnosis $z \le \bar{z}$. Since both true and incorrect diagnoses are uniformly distributed on S, and diagnosing accuracy is the same for all locations, the reported diagnosis is also uniformly distributed on S, implying that the probability of receiving a diagnosis $z \le \bar{z}$ is $\bar{z}$. The share of informed patients that choose hospital 1 is thus given by

$$\bar{z}(q_1, q_2; x_1, x_2) = \begin{cases} 0, & \text{for } q_1 \le q_1^l(q_2) \\ \dfrac{1}{2} + \dfrac{q_1 - q_2}{2t\delta(x_2 - x_1)} - \dfrac{(1 - x_1 - x_2)}{2\delta}, & \text{for } q_1 \in (q_1^l, q_1^h) \\ 1, & \text{for } q_1 \ge q_1^h(q_2) \end{cases} . \tag{4}$$

Overall expected demand for hospital 1 is $D_1 = \lambda\bar{z} + (1-\lambda)/2$. Like $\bar{z}$ the demand function has kinks at q_1^l and q_1^h. Hospital 2 expects to receive the residual demand $D_2 = 1 - D_1 = \lambda(1 - \bar{z}) + (1-\lambda)/2$.

3.2. Quality competition

For given locations and given GP attendance, optimal quality investments are found by inserting demand derived above into the profit function (1) and optimizing with respect to q_i. We assume that $\delta > (1/2) - \bar{x}$. For this case, we show in Appendix A that, if $\bar{q}$ is not too high, a unique pure strategy equilibrium in the quality game exists for all $p > 0$ and $\lambda > 0$, and for all locations $x_1 \in [0, \frac{1}{2} - \bar{x}]$ and $x_2 \in [\frac{1}{2} + \bar{x}, 1]$. This equilibrium is given by

$$q_i^*(\Delta; \lambda, p) = \min\left(\frac{p\lambda}{4tk\delta\Delta}, \bar{q}\right), \qquad i = 1, 2, \tag{5}$$

where $\Delta := x_2 - x_1 \in [2\bar{x}, 1]$. We see that equilibrium quality levels in the interior solution are always symmetric and depend only on the distance between hospitals' locations. This is due to the absence of price competition, where quality investments have a market-expanding effect which, due to the uniform distribution of patients, does not depend on absolute locations. An immediate implication is that optimal specializations will be characterized by some certain distance and not by absolute locations.

Assuming an interior solution, the comparative static results are mostly straightforward. Less product differentiation (lower Δ) will intensify quality competition, i.e., competition is intense when products are close substitutes. Furthermore, patients are more responsive to quality improvements when mismatch costs are small, implying that 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 are the benefits of capturing market 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 as a response to an increase in p.

The degree of information in the market is captured by the two parameters λ and δ. A higher GP attendance (λ) leads to increased quality provision. This is quite intuitive, since more patients obtain information about hospital qualities and thus become responsive to possible quality differences between the hospitals. Improved diagnosing accuracy, on the other hand, has the opposite effect, which might seem a bit surprising at first glance.[28] The underlying mechanism is that lower diagnosing accuracy makes hospital quality a relatively stronger signal for an imperfectly informed patient. If a patient is less certain about his own location, and thus about the expected mismatch costs of attending each hospital, he will attach more weight to hospital quality in making the decision of which hospital to approach for treatment. In other words, improved diagnosing accuracy means that information about hospital specialization becomes more valuable for the patient. All else equal, a higher value of δ thus reduces the degree of competition in the market and leads to lower quality provision in equilibrium.

3.3. Specialization

At this stage of the game hospitals decide on their specialization, taking, for a given λ, the effects on quality competition and demand into account. We look for a symmetric equilibrium in pure strategies. Inserting the optimal quality levels in the *interior* solution into hospital 1's profit function, we obtain the following partial derivative with respect to x_1:

$$\frac{\partial \Pi_1}{\partial x_1} = p\lambda \left(\frac{1}{2\delta} - \frac{p\lambda}{8\Delta^3 k t^2 \delta^2} \right). \tag{6}$$

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, given by[29]

$$x_1^*(\lambda, p) = \frac{1}{2}(1 - \Delta^*) \quad \text{and} \quad x_2^*(\lambda, p) = \frac{1}{2}(1 + \Delta^*), \tag{7}$$

[28] Remember that patients receive perfect information about qualities and specializations, while diagnosis information is imperfect.

[29] It is easily shown that the second-order conditions are met. Moreover, note that symmetry always implies $x_1 + x_2 = 1$.

where

$$\Delta^*(\lambda, p) = \left(\frac{p\lambda}{4t^2k\delta}\right)^{1/3}. \tag{8}$$

In addition, there are two possible corner solutions. If differentiation incentives are very strong, the hospitals will locate at the endpoints, i.e., $\Delta^* = 1$. On the other hand, if the upper bound on quality is sufficiently low, the locations given by (7) and (8) will induce a *corner* solution, $q_i = \bar{q}$, in the ensuing quality game. In this case, the equilibrium in the location game is a corner solution with minimal hospital differentiation, i.e., $\Delta^* = 2\bar{x}$. In the following, we focus on the interior equilibrium given by (7) and (8).[30]

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 (business stealing effect). On the other hand, closer locations imply that quality competition is intensified, as can be seen from Eq. (5).

Consider an increase in the treatment price p. This will strengthen the business stealing effect, since hospitals now receive a higher mark-up on each treatment. However, a price increase also means that quality competition is amplified. From (8), 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.

A similar mechanism determines the relationship between GP attendance and locations. More informed patients will result in stronger quality competition, and hospitals will respond by differentiating more.[31] 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.

Like in the quality game, increased information about the secondary care market through higher GP attendance and improved diagnosing accuracy yield opposite incentives for hospital competition. Since improved diagnosing accuracy *reduces* the intensity of quality competition, hospitals choose to differentiate less.

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 (8) into (5) we obtain the equilibrium quality levels of the game:

$$q^*(\lambda, p) = \left(\frac{p^2\lambda^2}{16tk^2\delta^2}\right)^{1/3}. \tag{9}$$

The following proposition summarizes the comparative statics results:

Proposition 1. *The best responses of the specialization–quality game are both increasing in treatment price and GP consultation, and decreasing in mismatch costs, quality costs and diagnosing accuracy.*

[30] In the specialization equilibrium given by (7) and (8), hospital 1 might also have an incentive to deviate by locating at $(1/2) - \bar{x}$, if such a relocation induces a corner solution in the quality subgame. (Of course, hospital 2 has symmetric incentives.) It can easily be shown that such a deviation is not profitable unless $\bar{q}$ is sufficiently low. We rule out this possibility by assumption.

[31] 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, 2004).

4. GP consultation

In the previous section we derived the equilibrium of the specialization–quality game for a given value of λ, and Eqs. (7)–(9) show the respective best response functions of the hospitals. To solve the game we now have to derive the best response of patients to any given level of Δ and q. This is done by letting patients make the choice of whether or not to consult a GP to obtain more information, based on an assessment of expected benefits and costs.

When deciding whether to approach a (randomly chosen) hospital directly or to consult a GP first, a patient has to weigh the costs of going to a GP against the benefits. As the game is common knowledge, patients know that hospitals provide the same quality. Moreover, the quality received is independent of whether a GP was consulted or not and therefore the consultation decision is independent of qualities. Determining the (individual) benefits of gatekeeping, and thereby the best response $\lambda^*(\Delta)$, simply requires ascertaining the reduction in expected mismatch costs for every degree of product differentiation, Δ, in the market. To simplify the analysis we assume that patients know that the equilibrium will be symmetric, i.e., that hospitals locate equidistantly from the market center, but on opposite sides.[32]

For a given degree of differentiation, Δ, expected mismatch costs for a patient who directly approaches a hospital are

$$M_0 = \frac{t}{2}\int_0^1 \left(z - \frac{1}{2}(1-\Delta)\right)^2 \mathrm{d}z + \frac{t}{2}\int_0^1 \left(z - \frac{1}{2}(1+\Delta)\right)^2 \mathrm{d}z. \tag{10}$$

The first term of Eq. (10) measures the expected mismatch costs when approaching hospital 1 weighted with the probability that this hospital will actually be chosen (which, applying our tie-breaking rule, is 1/2). Expected mismatch costs are calculated over the entire disease space, since patients are unaware of their actual diagnosis. Accordingly, the second term measures the expected mismatch costs when consulting hospital 2, weighted with 1/2. When consulting a GP first, expected mismatch costs are reduced to

$$\begin{aligned} M_{\mathrm{GP}} = {} & t\int_0^{1/2} \left(\delta\left(z - \frac{1}{2}(1-\Delta)\right)^2 + (1-\delta)\int_0^1 \left(s - \frac{1}{2}(1-\Delta)\right)^2 \mathrm{d}s\right) \mathrm{d}z \\ & + t\int_{1/2}^1 \left(\delta\left(z - \frac{1}{2}(1+\Delta)\right)^2 + (1-\delta)\int_0^1 \left(s - \frac{1}{2}(1+\Delta)\right)^2 \mathrm{d}s\right) \mathrm{d}z. \end{aligned} \tag{11}$$

Through GP consultation the patient obtains a diagnosis z and seeks treatment of hospital 1 whenever $z \in [0, \frac{1}{2}]$. The associated expected mismatch costs are given by the first line of Eq. (11). With probability δ the diagnosis z is correct and the corresponding mismatch costs are given by the first term of the integrand (of the outer integral). With the remaining probability $1-\delta$ the diagnosis z is false. The true disease may be at any point of the unit interval and every disease is equally likely. The resulting mismatch costs are given by the second term, i.e., by the inner integral. If $z \in (\frac{1}{2}, 1]$, the patient chooses treatment of hospital 2 and, in expectation, incurs the

[32] This allows us to write the best response function $\lambda^*(\cdot)$ as a function of Δ. Otherwise the benefits of gatekeeping would differ in absolute locations even if relative locations, i.e., Δ, remain unchanged.

second line as mismatch costs. The expected benefit of gatekeeping is thus

$$B{:=}M_0 - M_{\mathrm{GP}} = \frac{t\delta\Delta}{4}. \tag{12}$$

The best response $\lambda^*(\Delta)$ is now obtained by equating the expected benefits of gatekeeping to its actual costs, $t\delta\Delta/4 = ay$, and solving for y, which yields the critical cost type $\tilde{y}$ who is indifferent between consulting a GP for diagnosis (and referral) and approaching a hospital directly. Since the cost type is uniformly distributed on the unit interval the share of patients seeking GP diagnosis is $\int_0^{\tilde{y}} \mathrm{d}y = \tilde{y}$ implying a best response function

$$\lambda^*(\Delta) = \frac{t\delta\Delta}{4a}. \tag{13}$$

The comparative static results can easily explained by the costs and benefits of GP consultation. Of course, the higher consulting cost (a), the lower the share of patients actually attending a GP for consultation. The benefits of gatekeeping are determined by two different factors relating to three different variables or parameters. First, the mismatch costs that, in expectation, can be saved through costly GP consultation are positively related to the degree of horizontal differentiation of services Δ and to the weight t attached to the disease mismatch in the patients' utility function. Thus, the higher $t\Delta$ the more costly, in terms of mismatch costs, to approach the 'wrong' hospital. Second, the savings discussed above are realized with a higher probability the more accurate the diagnosis of the GP, i.e., the larger δ.

Let us now turn to the solution of the game. Eqs. (8) and (13) define the two reaction functions which determine the equilibrium attendance rate and differentiation, λ^* and Δ^*, so that the level of GP attendance is the best response to hospital specializations, and vice versa. Assuming an interior solution for hospital differentiation and GP attendance, the equilibrium values of Δ and λ are found by simultaneously solving (8) and (13), yielding

$$\Delta^*(p) = \frac{1}{4}\left(\frac{p}{tka}\right)^{1/2}, \tag{14}$$

$$\lambda^*(p) = \frac{\delta}{16}\left(\frac{pt}{ka^3}\right)^{1/2}. \tag{15}$$

The corresponding quality levels are obtained by substituting Eq. (15) into (9), yielding

$$q^*(p) = \frac{p}{16ak}. \tag{16}$$

We are now ready to state the comparative static results of the specialization–quality–consultation subgame:

Proposition 2. *The specialization–quality–consultation equilibrium has the following comparative static properties*:

(i) *GP attendance is increasing in treatment price*, *mismatch costs and diagnosing accuracy*, *and decreasing in quality and attendance costs*;
(ii) *hospital differentiation is increasing in treatment price*, *decreasing in mismatch*, *attendance and quality costs*, *and independent of diagnosing accuracy*;
(iii) *hospital quality is increasing in treatment price*, *decreasing in attendance and quality costs*, *and independent of mismatch costs and diagnosing accuracy*.

Several of these effects are quite intuitive. The share of the population attending a GP increases in the mismatch cost, t, as this drives up the benefits of gatekeeping. It also increases in the treatment price. This is an indirect effect stemming from specialization. Price increases boost quality competition and, to dampen this effect, hospitals aim at reducing the substitutability of their services, 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. This reduces the competitive pressure in the hospital market, leading to less differentiation and a lower supply of quality. Equilibrium GP attendance is also increasing in the diagnosing accuracy, δ, since improved accuracy reduces expected mismatch costs and thus increases the benefits of GP gatekeeping. Finally, an increase in the quality cost parameter, k, reduces quality competition and thereby differentiation incentives. This, in turn, reduces the benefits of gatekeeping, leading to a lower GP attendance in equilibrium.

There are also some effects that are less obvious. We see that the mismatch costs parameter t has no effect on equilibrium hospital quality. With exogenous GP attendance, patients were more responsive to quality investments at lower values of t, amplifying quality competition. With endogenous GP attendance, however, this effect is counteracted by the consultation effect. A lower t reduces the benefits of gatekeeping, resulting in lower GP attendance and thus a less competitive market. With linear costs of GP consultation and uniformly distributed consultation cost types these two effects exactly offset. Interestingly, we also see that equilibrium hospital specialization and quality provision are not affected by diagnosing accuracy, δ. For a given level of GP attendance, we know that higher diagnosing accuracy *reduces* the degree of competition in the market, with lower quality provision and less differentiation (direct effect). However, a higher diagnosing accuracy also increases the value of information obtained by attending a GP, leading to higher GP attendance, which, in turn, *increases* the degree of hospital competition (indirect effect). Thus, when the decision of whether or not to attend a GP is taken into account, the indirect effect of improved diagnosing accuracy on hospital competition tends to counteract the direct effect. In our specific model, with linear GP consultation costs, these two effects exactly offset. Obviously, the indirect effect is eliminated if consultation costs are so low that all patients choose to consult a GP before accessing the hospital market. In this case, the equilibrium is a corner solution with $\lambda^* = 1$, where improved diagnosing accuracy dampens competition between secondary care providers.

5. Social welfare

Consider a social planner who aims at maximizing social welfare, defined as the sum of consumers' and producers' surpluses net of any government expenditures.[33] Taking the duopolistic market structure as exogenously given, imposing symmetry, and noting that aggregate GP consultation costs are $a\int_0^{\lambda} s\,\mathrm{d}s = (1/2)a\lambda^2$, expected social welfare is given by

$$W = v + q(1-2kq) - [(1-\lambda)M_0 + \lambda M_{\mathrm{GP}}] - \frac{1}{2}a\lambda^2,$$

or, when substituting for M_0 and M_{GP} from Eqs. (10) and (11),

$$W = v + q(1-2kq) - \frac{t}{12}[1 - 3\Delta(\lambda\delta - \Delta)] - \frac{1}{2}a\lambda^2. \tag{17}$$

[33] If we interpret p as a per treatment or per patient reimbursement from a government agency, we implicitly assume that the third party (i.e., the regulator) is able to raise the necessary funds in a non-distortionary manner.

The interpretation of (17) is straightforward. In addition to the gross utility of hospital treatment (first term), expected social welfare consists of the social net benefit of quality provision (second term) net of expected aggregate mismatch costs (third term) and aggregate GP consultation costs (fourth term).

5.1. Should GP attendance be made compulsory?

Introducing a strict gatekeeping system is equivalent to setting $\lambda = 1$. For illustration, let us first assume that the regulator can decide the optimal rate of GP attendance directly, by choosing any $\lambda^{fb} \in [0, 1]$. From (17), the rate of GP attendance that maximizes social welfare, for a given level of hospital differentiation Δ, is given by

$$\lambda^{fb} = \frac{t\delta\Delta}{4a}. \tag{18}$$

Comparing (18) and (13) we see that, for a given degree of hospital differentiation, private and social incentives for GP attendance coincide. So why should a regulator distort GP consultation? The reason is that, for a given price p, the 'laissez-faire' equilibrium, given by (14), not necessarily produces socially optimal hospital differentiation. Quality provision (16) may also be inefficient. It may therefore be desirable to make GP attendance compulsory, in order to affect both hospital specializations and quality investments in a socially desirable direction, even if this means that GP attendance costs increase beyond the socially (and privately) optimal level.

With *voluntary GP consultation*, expected social welfare is found by inserting the equilibrium expressions for $\Delta^*(p)$, $\lambda^*(p)$ and $q^*(p)$ from (14)–(16) into the welfare function (17), yielding

$$W^*(p) = v - \frac{t}{12} + \frac{(24a - 4p + t\delta^2)p}{512ka^2}. \tag{19}$$

If the regulator enforces *compulsory GP consultation*, expected social welfare is found by setting $\lambda = 1$ in the equilibrium expressions for $\Delta^*(\lambda, p)$ and $q^*(\lambda, p)$ in (8) and (9), and substituting into the welfare function (17), yielding

$$W^*(p)|_{\lambda=1} = v - \frac{t}{12} - \frac{a}{2} + \frac{1}{16}\left[2\left(\frac{2t\delta^2 p}{k}\right)^{1/3} + 3\left(\frac{4p^2}{k^2 t\delta^2}\right)^{1/3} - 4\left(\frac{2p^4}{kt^2\delta^4}\right)^{1/3}\right]. \tag{20}$$

Whether or not an introduction of a strict gatekeeping system is (for a given price) socially desirable is then determined by the sign of the difference between (19) and (20). Unfortunately, it is not feasible to characterize this difference analytically. However, we have done several numerical simulations that produce a clear picture. The details of the simulations are available upon request (see also the supplementary material available in the online version of the paper); here we only summarize the main findings.

It is instructive to express the results with respect to the key parameters t and δ. Our main finding is that enforcing compulsory GP consultation is socially desirable only if t or δ are sufficiently high. Hospital competition is then relatively moderate and insufficient quality and too little hospital differentiation may result. We know that improved patient information in terms of GP consultation (i.e., an increase in λ) increases the degree of competition in the market. The introduction of a strict gatekeeping system thus stimulates competition and may affect both quality and specialization in a socially desirable direction. In contrast, if t or δ are relatively small

hospital competition is already strong. Making GP consultation compulsory might then lead to *excessive competition* with too high quality and too much specialization.

5.2. The treatment price as an additional regulatory instrument

The above result hinges on the assumption that the treatment price is exogenous. We will now relax this assumption and assume that the regulator is able also to use the price as a regulatory instrument in an optimal way. Assuming second-best price regulation, the following result obtains:

Proposition 3. *With second-best price regulation and endogenous GP consultation decisions, there is no scope for direct regulation of GP attendance.*

Proof. Inserting (8) and (9) into (17) yields a welfare function $W(p, \lambda)$. By defining $\hat{p}:=p\lambda$ we can define a new welfare function $\hat{W}(\hat{p}, \lambda):=W(\Delta^*(\hat{p}), q^*(\hat{p}), \lambda)$. Maximizing $W(p, \lambda)$ with respect to p and λ is then equivalent to maximizing $\hat{W}(\hat{p}, \lambda)$ with respect to $\hat{p}$ and λ. Taking the partial derivative with respect to λ yields

$$\frac{\partial \hat{W}(\hat{p}, \lambda)}{\partial \lambda} = -a\lambda + \frac{t\delta}{4}\Delta^*(\hat{p}). \tag{21}$$

By comparing (13) and (21) we see that social and private incentives for GP attendance coincide for every given value of Δ. The regulator can then use $\hat{p}$ to induce the optimal (second-best) levels of q and Δ and let patients choose the socially optimal level of GP attendance themselves. □

When second-best pricing is available, there is no longer any need to use strict gatekeeping as a regulatory mechanism to induce socially more desirable hospital differentiation and quality provision. From (8) and (9), 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 specialization–quality outcome for any given value of λ. If the regulator uses the price instrument to induce second-best differentiation and quality provision, an optimal trade-off between expected mismatch cost reductions and consultation costs will secure the socially optimal level of GP attendance. This is exactly the trade-off that patients make themselves in the described game.

5.3. The second-best optimum

Let us now derive and characterize the second-best price and briefly discuss the efficiency properties of the optimal solution.[34] We focus on an interior solution, where the optimal price is sufficiently low to ensure $q < \bar{q}$.[35] Maximizing (19) with respect to p yields the following second-best treatment price:

$$p^{sb} = 3a + \frac{1}{8}t\delta^2. \tag{22}$$

[34] For a discussion of optimal price regulation under complete information, i.e., where $\lambda\delta = 1$, see Brekke et al. (2006).

[35] It is straightforward to show that the regulator will never induce a corner solution if

$$\frac{(24a + t\delta^2)^2}{819ka^2} + \bar{q}(2k\bar{q} - 1) + \frac{t\bar{x}^2(8a - t\delta^2)}{8a} > 0.$$

The optimal price is increasing in diagnosing accuracy, although δ does not affect quality and specializations in equilibrium. The reason is that higher diagnosing accuracy increases the degree of information about hospital specializations in the market, implying that the socially optimal differentiation is larger. The regulator must then stimulate more differentiation by increasing the treatment price. The effects of the consultation and mismatch cost parameters a and t are more straightforward. An increase in either type of cost dampens hospital competition, leading to less differentiation and lower quality provision, effects that can by counteracted by increasing p.

From the price given in (22), the following equilibrium outcome obtains:

$$\Delta^{\text{sb}} = \frac{1}{4}\left[\frac{1}{k}\left(\frac{3}{t} + \frac{\delta^2}{8a}\right)\right]^{1/2}, \qquad q^{\text{sb}} = \frac{3}{16k} + \frac{t\delta^2}{128ak}$$

$$\text{and} \quad \lambda^{\text{sb}} = \frac{\delta}{64}\left(\frac{2t(24a + t\delta^2)}{a^3 k}\right)^{1/2}. \tag{23}$$

The efficiency properties of the second-best interior solution[36] are summarized as follows:

Proposition 4. *The second-best* (*interior*) *solution of the specialization–quality–consultation game has the following efficiency properties*:

(i) *for* $t\delta^2 < 8a$, *there is too much differentiation given* λ^{sb}, *and too low quality provision*;
(ii) *for* $t\delta^2 = 8a$, *differentiation is first-best given* λ^{sb} *and first-best quality is implemented*;
(iii) *for* $t\delta^2 > 8a$, *there is insufficient differentiation given* λ^{sb}, *and too high quality provision*.

Proof. First-best specialization, $\Delta^{\text{fb}} = \lambda\delta/2$, and first-best quality, $q^{\text{fb}} = (1)/(4k)$, is obtained by partially differentiating Eq. (17) with respect to Δ and q, respectively. From (23), we find that

$$\Delta^{\text{sb}} - \frac{\lambda^{\text{sb}}\delta}{2} = \frac{1}{128a}\sqrt{\frac{48a + 2t\delta^2}{akt}}(8a - t\delta^2) > (<)0, \quad \text{if } t\delta^2 < (>)8a$$

and

$$q^{\text{sb}} - q^{\text{fb}} = \frac{t\delta^2 - 8a}{128ka} < (>)0, \quad \text{if } t\delta^2 < (>)8a. \quad \square$$

The first-best outcome with respect to both hospital differentiation and quality provision is – apart from the knife-edge case (ii) – never achieved. This is not surprising, since the regulator has more policy goals than regulatory instruments.

To see the intuition for the general efficiency characteristics of the second-best equilibrium, consider regimes (i) and (iii), where the benefits of gatekeeping are either low or high compared to its costs. In the former case, GP attendance will be low in equilibrium and social welfare is maximized at a relatively low degree of differentiation. Consequently, the price that yields first-best

[36] It is straightforward to show that second-best pricing yields an interior solution with respect to GP attendance for a subset of the parameter values, defined by $k > \bar{k}$, where

$$\bar{k} := \frac{t\delta^2(24a + t\delta^2)}{2048a^3}.$$

Thus, if $k \leq \bar{k}$ we have a corner solution with $\lambda^{\text{sb}} = 1$, implying a de facto strict gatekeeping regime.

differentiation is not high enough to generate efficient quality provision. Higher quality can then 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. In the latter case, though, GP attendance and thus the first-best level of differentiation are relatively high. The optimal degree of differentiation is then 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.

6. 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). As GPs are usually better informed than patients about the characteristics of the secondary health care market, e.g., about quality and specialization of hospitals, matching of patients to hospitals may be improved by gatekeeping. However, this argument neglects the potential competitive effects in the hospital market. We have presented a model that analyzes the competitive effects of gatekeeping in the presence of hospital non-price competition.

While prices were regulated, we allowed for competition in specialization and quality. We found that when the price is exogenously given, strict gatekeeping may reduce social welfare, especially if mismatch costs and diagnosing accuracy are both sufficiently low. In this case, making it compulsory to attend a GP before receiving secondary care will boost competition to such an extent that excessive hospital specialization and quality occur. This raises doubts about whether gatekeeping improves efficiency. Things change when allowing for second-best price regulation. In this case, we showed that there is no scope for direct regulation of GP attendance, since consultation decisions of patients are the same as what a social planner would implement. Actually, a de facto strict gatekeeping regime arises endogenously if the benefits of gatekeeping are sufficiently high (improved matching outweighs the potentially negative competitive effects) compared to its costs. Finally, we considered the (interior) second-best equilibrium, showing that the solution, in general, will be characterized by inefficient levels of quality and specialization, depending on the relative values of mismatch costs, consultation costs and the diagnosing accuracy. GP consultation, however, was found to be efficient given specializations of hospitals.

The analysis demonstrates that efficiency gains that are usually attributed to GP gatekeeping cannot be taken for granted when the secondary care sector is endogenized and non-price competition amongst providers is considered. 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 specialization so that differentiation increases, which might counteract the positive short run effect.

We have assumed a prospective payment system, where the regulator sets the price per treatment. This payment system corresponds well with DRG-pricing systems, which are extensively used by several countries as a way to reimburse hospitals for their medical treatments. The DRG-price is based on the average cost of treating particular diagnostic groups, where the average cost is derived from reported costs by a representative sample of hospitals. An alternative mechanism is to give hospitals budgets (block grants) based on capitation, which is a payment per individual enrolled in a given health plan (or district). The capitation payment is based on the *expected overall* costs of treating the individuals in the health plan, which depends on the distribution of

diseases, the corresponding probabilities of falling ill, and the treatment costs. In our model there is no uncertainty with respect to health status. Every individual is ill and in need for one unit of care. As a consequence, the regulated price can be interpreted either as a payment per treatment (e.g., DRG-price) or per individual (capitation). As long as the hospitals are not able to affect the payment per patient nor the probability of falling ill, our results can be generalized to also include capitation systems. However, if the hospitals can manipulate the payment per patient by, for instance, upcoding or cream-skimming activities, capitation systems need a separate analysis. This is left for future research.

Acknowledgements

We thank Pierre-Yves Geoffard, Kai A. Konrad, Daniel Krähmer, Kjell Erik Lommerud, Johannes Münster, two anonymous referees, and participants of several workshops, conferences and seminars. The usual caveat applies.

Appendix A. Equilibrium in the quality game

We will show that Eq. (5) is indeed a pure strategy equilibrium of the quality game. What is required is a properly defined upper bound on quality, $\bar{q}$. Moreover, this equilibrium is unique. Uniqueness can be established quite straightforwardly by deriving the best response functions for both hospitals and showing that these functions intersect only once, i.e., at the equilibrium given by (5). Interested readers can contact the authors for details; here we concentrate on giving a rigorous proof of existence.

To establish existence, we need to check that q_1^* and q_2^* are mutually best responses. It is thus sufficient to show that there exists no profitable deviation for hospital 1 from q_1^* given q_2^*, and vice versa. Since the problem at hand is symmetric we restrict attention to hospital 1.

Consider first an *interior solution*, i.e., $q_1^* = q_2^* < \bar{q}$. The profit of hospital 1 is then given by

$$\Pi_1^{\text{int}}(p, \lambda) = \frac{p}{2}\left(1 - \frac{\lambda}{\delta}\left(1 - x_1 - x_2 + \frac{p\lambda}{8t^2 k\delta\Delta^2}\right)\right), \tag{A1}$$

where $\Delta := x_2 - x_1 \in [2\bar{x}, 1]$ and the superscript 'int' refers to interior solutions. Let us now check for possible profitable deviations for hospital 1. First note that one implication from the first-order conditions is that a deviation from q_1^* (given q_2^*) implying $\bar{z} \in (0, 1)$ can never be profitable. In this range the profit function is strictly concave since the demand function is linear in q_1 and the cost function is strictly convex in q_1 (see Eqs. (4) and (1), respectively). Thus, the first-order condition guarantees that q_1^* maximizes hospital 1's profit in that range. Moreover, given q_2^* any quality level q_1 implying $\bar{z} = 1$ cannot be optimal. Like the demand function (4), the profit function (1) has a kink at $q_1 = q_1^h(q_2^*)$. Beyond this quality level demand does not react to quality, but quality costs do make the profit function even more concave in this area. To summarize, the first-order condition and the strict concavity of the profit function on the set $\{q_1 | 0 < \bar{z}(q_1, q_2^*) \leq 1\}$ guarantee that q_1^* is a best response to q_2^* on that set. When checking for profitable deviations of hospital 1 we can therefore direct attention to situations where $\bar{z} = 0$. However, when hospital 1 receives no demand from informed consumers, it maximizes its profit by not investing in quality at all. The optimal deviation thus implies $q_1 = 0$ and a payoff

$$\hat{\Pi}_1(p, \lambda) = \frac{(1 - \lambda)p}{2}. \tag{A2}$$

Deviation is not profitable if $\Phi_1^{\text{int}}(p,\lambda) := \Pi_1^{\text{int}}(p,\lambda) - \hat{\Pi}_1(p,\lambda) \geq 0$, where Φ_1^{int} can be expressed as

$$\Phi_1^{\text{int}}(p,\lambda) = \frac{p\lambda}{2}\left[1 - \frac{1}{\delta}\left(1 - x_1 - x_2 + \frac{p\lambda}{8t^2k\delta\Delta^2}\right)\right]. \tag{A3}$$

Let us now study the properties of Φ_1^{int}. Since $p=0$ yields $\Pi_1^{\text{int}}(0,\lambda) = \hat{\Pi}_1(0,\lambda) = 0$, we have $\Phi_1^{\text{int}}(0,\lambda) = 0$. So, not surprisingly, there is no incentive to deviate when the price is zero. The first and second-order partial derivatives with respect to p are

$$\frac{\partial\Phi_1^{\text{int}}(p,\lambda)}{\partial p} = \frac{\lambda}{2\delta}\left(x_1 + x_2 + \delta - 1 - \frac{p\lambda}{4k\delta t^2\Delta^2}\right), \tag{A4}$$

$$\frac{\partial^2\Phi_1^{\text{int}}(p,\lambda)}{\partial p^2} = -\frac{\lambda^2}{8kt^2\delta^2\Delta^2} < 0, \tag{A5}$$

which confirms that Φ_1^{int} is concave in p. For a deviation to be unprofitable for small values of p we need to have

$$\lim_{p\to 0}\frac{\partial\Phi_1^{\text{int}}(p,\lambda)}{\partial p} = \frac{\lambda}{2\delta}(x_1 + x_2 + \delta - 1) > 0. \tag{A6}$$

It is easy to show that inequality (A6) holds for all $x_1 \in [0, \frac{1}{2} - \bar{x}]$ and $x_2 \in [\frac{1}{2} + \bar{x}, 1]$ if $\delta > (1/2) - \bar{x}$. Thus, there exists a sufficiently low value of p such that deviation is not profitable provided that $\lambda > 0$ and $\delta > (1/2) - \bar{x}$.[37]

When there is no upper bound on quality we know that $\Phi_1^{\text{int}} \to -\infty$ if $p \to \infty$. Together with the concavity of Φ_1^{int}, shown in Eq. (A5), and the assumptions made above, we know that Φ_1^{int} has exactly two roots, one at $p=0$ and one at some $\hat{p} \in (0,\infty)$, where $\hat{p}$ depends on locations, i.e., $\hat{p} = \hat{p}(x_1, x_2)$. Thus, for sufficiently high prices, i.e., for $p > \hat{p}(x_1, x_2)$, deviation to $q_1 = 0$ is profitable. We now define an upper bound on quality, $\bar{q}$, such that this will never happen.

Define $\hat{p}(x_1, x_2)$ implicitly as the positive solution to $\Phi_1^{\text{int}}(p, 1) = 0$. From (A3), this is given by

$$\hat{p}(x_1, x_2) = 8kt^2\Delta^2\delta(\delta + x_1 + x_2 - 1).$$

Note that this critical price is, given the assumptions made so far and given the properties of Φ_1^{int}, always well defined. Moreover, since Eq. (5) establishes the monotonicity of $q_i^*(p,\lambda)$ in p and λ, the critical price implies a critical quality level $\hat{q}(x_1, x_2) := q_i^*(\hat{p}(x_1, x_2), 1)$, $i = 1, 2$. Now assume that $p \leq \hat{p}(x_1, x_2)$; then the candidate equilibrium (q_1^*, q_2^*) has quality levels below $\hat{q}(x_1, x_2)$ and there is no incentive to deviate. If $p > \hat{p}(x_1, x_2)$ there is no incentive to deviate if (i) the upper bound on quality $\bar{q}$ is not above $\hat{q}(x_1, x_2)$ and if (ii) the hospital has no incentive to deviate from $\bar{q}$. Let

$$\bar{p} := \min\left\{\hat{p}(x_1, x_2) | x_1 \in \left[0, \tfrac{1}{2} - \bar{x}\right], x_2 \in \left[\tfrac{1}{2} + \bar{x}, 1\right]\right\}$$

Then the definition

$$\bar{q} := q_1^*(\bar{p}, 1)$$

[37] Note that we have not made use of symmetry in locations. But consider locations were symmetric, then $x_1 + x_2 = 1$, implying that (A6) would be satisfied for any diagnosing accuracy $\delta > 0$.

guarantees that hospital 1 never has an incentive to deviate from the candidate equilibrium—independent of locations.[38] Due to symmetry, the exact same argument applies of course for hospital 2.

Now consider the *corner solution*. Due to the first-order conditions, where quality incentives depend on relative, but not absolute, locations, the corner solution must also be symmetric; $q_1^*, q_2^* = \bar{q}$. In this case, the profit of hospital 1 is given by

$$\Pi_1^{\text{cor}}(p, \lambda, \bar{q}) = \frac{p}{2}\left(1 - \frac{\lambda(1 - x_1 - x_2)}{\delta}\right) - k\bar{q}^2, \tag{A7}$$

where the superscript 'cor' refers to corner solutions. Optimal deviation profits are still given by (A2), implying

$$\Phi_1^{\text{cor}}(p, \lambda, \bar{q}) = \frac{\lambda p}{2}\left(1 - \frac{1 - x_1 - x_2}{\delta}\right) - k\bar{q}^2. \tag{A8}$$

Since Φ_1^{cor} is necessarily – by the definition $\bar{q}{:=}q_1^*(\bar{p}, 1)$ – non-negative at $\bar{q}$, the properties $(\partial\Phi_1^{\text{cor}}(p, \lambda, \bar{q}))/(\partial p) > 0$ and $(\partial\Phi_1^{\text{cor}}(p, \lambda, \bar{q}))/(\partial\lambda) > 0$ guarantee that no profitable deviation exists.

To summarize, a symmetric pure strategy equilibrium exists for all $\delta > (1/2) - \bar{x}$, p and λ, and all locations $x_1 \in [0, \frac{1}{2} - \bar{x}]$ and $x_2 \in [\frac{1}{2} + \bar{x}, 1]$. This equilibrium is given by (5).

Appendix B. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jhealeco.2006.04.004.

References

Arrow, K., 1963. Uncertainty and the welfare economics of medical care. American Economic Review 53, 941–973.

Barros, P.P., 1998. The black box of health care expenditure growth determinants. Health Economics 7, 533–544.

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

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

Beitia, A., 2003. Hospital quality choice and market structure in a regulated duopoly. Journal of Health Economics 22, 1011–1036.

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

Brekke, K.R., Nuscheler, R., Straume, O.R., 2006. Quality and location choices under price regulation. Journal of Economics & Management Strategy 15, 207–227.

Bresnahan, T., 1981. Duopoly models with consistent conjectures. American Economic Review 71, 934–945.

Burdett, K., Judd, K.L., 1983. Equilibrium price dispersion. Econometrica 51, 955–969.

[38] If $\bar{x}$ is sufficiently small, $\hat{p}(x_1, x_2)$ is minimized when the hospitals locate as close as possible. Then,

$$\bar{p} = 32k(t\delta\bar{x})^2,$$

which implies

$$\bar{q} = 4t\delta\bar{x}.$$

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

D'Aspremont, C., Gabszewicz, J.J., Thisse, J.-F., 1979. On Hotelling's 'stability in competition'. Econometrica 47, 1145–1150.

Dranove, D., Kessler, D., McClelland, M., Satterthwaite, M., 2003. Is more information better? The effects of "report cards"on health care providers. Journal of Political Economy 111, 555–588.

Economides, N., 1989. Quality variations and maximal variety differentiation. Regional Science and Urban Economics 19, 21–29.

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

Ferris, T.G., Chang, Y., Blumenthal, D., Pearson, S.D., 2001. Leaving gatekeeping behind—effects of opening access to specialists for adults in a Health Maintenance Organization. New England Journal of Medicine 345, 1312–1317.

Garcia Mariñoso, B., Jelovac, I., 2003. GPs' payment contracts and their referral practice. Journal of Health Economics 22, 617–635.

González, P., 2006. The Gatekeeping Role of General Practitioners. Does Patients' Information Matter? Working Paper ECON 06.09, Universidad Pablo de Olavide.

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

Gravelle, H., Masiero, G., 2000. Quality incentives in a regulated market with imperfect information and switching costs: capitation in general practice. Journal of Health Economics 19, 1067–1088.

Hotelling, H., 1929. Stability in competition. Economic Journal 39, 41–57.

Lommerud, K.E., Sørgard, L., 2003. Entry in telecommunication: customer loyalty, price sensitivity and access prices. Information Economics and Policy 15, 55–72.

Lyon, T.P., 1999. Quality competition, insurance, and consumer choice in health care markets. Journal of Economics & Management Strategy 8, 545–580.

Ma, C.-t.A., Burgess Jr., J.F., 1993. Quality competition, welfare, and regulation. Journal of Economics 58, 153–173.

Malcomson, J.M., 2004. Health service gatekeepers. RAND Journal of Economics 35, 401–421.

McGuire, T.G., 2000. Physician agency. In: Culyer, A.J., Newhouse, J.P. (Eds.), Handbook of Health Economics, vol. 1A. North-Holland, Amsterdam, pp. 462–535.

Nuscheler, R., 2003. Physician reimbursement, time-consistency and the quality of care. Journal of Institutional and Theoretical Economics 159, 302–322.

Salop, S.C., 1979. Monopolistic competition with outside goods. Bell Journal of Economics 10, 141–156.

Schultz, C., 2004. Market transparency and product differentiation. Economics Letters 83, 173–178.

Schultz, C., 2005. Transparency on the consumer side and tacit collusion. European Economic Review 49, 279–297.

Scott, A., 2000. Economics of general practice. In: Culyer, A.J., Newhouse, J.P. (Eds.), Handbook of Health Economics, vol. 1B. North-Holland, Amsterdam, pp. 1175–1200.

Varian, H., 1980. A model of sales. American Economic Review 70, 651–659.

Wolinsky, A., 1997. Regulation of duopoly: managed competition vs. regulated monopolies. Journal of Economics & Management Strategy 8, 821–847.