ELSEVIER

Journal of Public Economics 92 (2008) 1607–1628

JOURNAL OF PUBLIC ECONOMICS

www.elsevier.com/locate/econbase

Competition and waiting times in hospital markets ☆

Kurt R. Brekke [a,*], Luigi Siciliani [b], Odd Rune Straume [c]

[a] *Department of Economics and Health Economics Bergen, Norwegian School of Economics and Business Administration, Helleveien 30, N-5045 Bergen, Norway*

[b] *Department of Economics and Related Studies, and Centre for Health Economics, University of York, Heslington, York YO10 5DD, UK; and C.E.P.R., 90-98 Goswell Street, London EC1V 7DB, UK*

[c] *Department of Economics and NIPE, University of Minho, Campus de Gualtar, 4710-057 Braga, Portugal; and Health Economics Bergen, Norway*

Received 7 March 2007; received in revised form 2 October 2007; accepted 6 February 2008
Available online 15 February 2008

Abstract

This paper studies the impact of hospital competition on waiting times. We use a Salop-type model, with hospitals that differ in (geographical) location and, potentially, waiting time, and two types of patients: high-benefit patients who choose between neighbouring hospitals (competitive segment), and low-benefit patients who decide whether or not to demand treatment from the closest hospital (monopoly segment). Compared with a benchmark case of monopoly, we find that hospital competition leads to longer waiting times in equilibrium if the competitive segment is sufficiently large. Given a policy regime of hospital competition, the effect of increased competition depends on the parameter of measurement: Lower travelling costs increase waiting times, higher hospital density reduces waiting times, while the effect of a larger competitive segment is ambiguous. We also show that, if the competitive segment is large, hospital competition is socially preferable to monopoly only if the (regulated) treatment price is sufficiently high.
© 2008 Elsevier B.V. All rights reserved.

JEL classification: H42; I11; I18; L13
Keywords: Hospitals; Competition; Waiting times

1. Introduction

Waiting times are a major health policy concern in many OECD countries. Mean waiting times for non-emergency care are above three months in several countries and maximum waiting times can stretch into years. Policymakers often argue that more competition and patient choice can reduce waiting times by encouraging hospitals to compete for patients and revenues (Siciliani and Hurst, 2004, 2005).[1] The mechanisms of how this may work are, however, not very

☆ We thank seminar participants at University of Bergen, Helsinki Centre of Economic Research, University of Porto, New University of Lisbon and Carnegie Mellon University, and two anonymous referees for helpful comments and suggestions.

* Corresponding author.
E-mail addresses: kurt.brekke@nhh.no (K.R. Brekke), ls24@york.ac.uk (L. Siciliani), o.r.straume@eeg.uminho.pt (O.R. Straume).

[1] There are many examples. Norway introduced activity-based funding (DRG-pricing) in 1997 and nation-wide patient choice of hospital in 2001. Both reforms aimed at stimulating competition and reducing waiting times. In the United Kingdom, the policy Payment by Results has been recently introduced, which remunerates hospitals according to a fixed tariff per patient treated. One of the objectives of the policy is to induce hospitals to compete for resources by reducing waiting times. In Denmark patients have had free choice of treatment in any publicly-funded hospital within the county of residence since 1993. In Sweden since 2002 all county councils have introduced free choice among public providers within and between counties.

0047-2727/$ - see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.jpubeco.2008.02.003

clear. Why would hospitals that operate at full capacity and face excessive demand have an incentive to compete for even more patients? The main purpose of this paper is to contribute to the understanding of the relationship between competition and waiting times in hospital markets.

We develop a model of hospital competition within a Salop framework, where hospitals differ in terms of (geographical) location and, possibly, waiting times. We assume that there are two types of patients who differ in expected benefit ("high" and "low") from hospital treatment. Hospitals compete on the segment of demand with high benefit, while they are local monopolists on the demand segment with low benefit. By comparing with a benchmark case of monopoly, we analyse how the introduction of competition in the hospital market affects waiting time and activity in equilibrium. Given a policy regime of hospital competition, we also examine the effects of increasing the *degree of competition*, based on three different measures: (i) patients' travelling costs, (ii) the size of the competitive relative to the monopolistic demand segment, and (iii) hospital density (the number of hospitals). We also derive the socially optimal waiting time and assess the welfare implications of hospital competition.

Most of the existing literature assumes that hospitals are local monopolists (Lindsay and Feigenbaum, 1984; Iversen, 1993, 1997; Martin and Smith, 1999; Olivella, 2002; Barros and Olivella, 2005; see Cullis et al., 2000, for a review of the literature). Two exceptions are Xavier (2003) and Siciliani (2005) who model competition within a Hotelling framework and in a duopoly model with differentiated products, respectively.[2] In these models, competition takes the form of duopoly, with the degree of competition being measured by the substitutability between treatments at the two hospitals, and both find that increased competition (or increased patient choice) leads to *longer* waiting times in equilibrium.

An arguable limitation of both these studies is that the analysis of a potential competition effect is confined to a single competition measure that leaves considerable room for interpretation. Furthermore, the lack of a welfare analysis leaves the more fundamental question of whether hospital competition is desirable in the first place, unanswered.

In the present paper, we complement and extend these studies in several different ways. First, we isolate a pure competition effect by considering monopoly versus competition, something which has not been done in the previous literature on hospital competition and waiting times. Second, the richness of our model allows us to use several different measures of the degree of hospital competition, something that turns out to have a crucial impact with respect to both waiting times and activity levels. Third, we include a welfare analysis where we derive and characterise both the socially optimal waiting time and the optimal treatment price, and analyse under which circumstances hospital competition is socially desirable in a public hospital market. We also deviate from the above mentioned studies by explicitly modelling semi-altruistic health care providers.

We find that *introducing competition*, by allowing previous local monopolies to compete for patients (equivalently, to introduce free patient choice), leads to an increase in equilibrium waiting times (with a corresponding reduction in hospital activity) only if the competitive demand segment is sufficiently large relative to the monopoly segment, and vice versa.[3] Thus, we obtain the previously derived result in the literature as a special case: when the competitive segment tends to one then competition always increases waiting times. Also, given a competition regime, we find that increasing the degree of competition has ambiguous effects on waiting times, depending on the measure of competition. Lower travelling costs for patients increase waiting times, which replicates the result derived by Xavier (2003). In addition, we find that a larger competitive segment has an indeterminate effect, while higher hospital density reduces waiting times.

Furthermore, the relationship between competition and hospital activity is often counter-intuitive. For example, lower travelling costs, which – all else equal – increase demand for hospital treatment, lead in equilibrium to lower hospital activity due to the corresponding increase in waiting time. Similarly, higher hospital density, which – all else equal – reduces demand per hospital, leads in equilibrium to higher per hospital activity due to the corresponding reduction in waiting time.

[2] Another related paper is Dawson et al. (2007) who analyse the impact of introducing patient choice on hospital waiting times. They find that the effect of choice on waiting times depends on the demand elasticities. Their model is, however, very different from ours, as they focus solely on the demand-side, assuming the supply-side to be completely exogenous. Thus, hospital competition is not an issue in their paper at all.

[3] The impact of patient choice on hospital waiting times has received surprisingly little empirical attention. Two notable exceptions are: Dawson et al. (2007) who analyse the impact of the London Patient Choice Project, finding that the project led to shorter (and converging) average waiting times in the London region; Siciliani and Martin (2007) who provide empirical evidence supporting a negative relationship between hospital density and waiting times, for a given level of need.

Regarding social welfare, we show that, if the competitive demand segment is relatively large, hospital competition is socially desirable only if the (regulated) price per treatment is sufficiently high. For a small competitive demand segment, the result is reversed; in this case, competition is desirable only if the treatment price is sufficiently low.

However, the socially optimal waiting time is attainable through optimal price setting, regardless of market regime. We also characterise the socially optimal treatment price and show that whether high-powered incentive schemes substitute or complement competition depends on the measure of competition. Unless the opportunity cost of public funds or altruism is very high, stronger competition through higher hospital density increases the optimal treatment price, while increased competition through lower travelling costs reduces optimal prices.

The rest of the paper is organised as follows. The model is presented in Section 2, while, in Section 3, we derive and characterise the equilibrium waiting time. The effects on waiting time and hospital activity of, first, introducing competition, and, second, increasing the degree of competition, are analysed in Section 4. In Section 5 we derive and characterise both the socially optimal waiting time and the optimal treatment price, and we assess the social desirability of introducing competition in a public hospital market. In Section 6 we extend the analysis in two ways: first, we introduce a copayment; second, we allow for inequality aversion. Finally, Section 7 concludes the paper.

2. Model

The basic model builds on the original formulation of rationing by waiting time of Lindsay and Feigenbaum (1984).[4] The main assumptions are that (i) patients differ in gross valuation of (inherent benefit from) medical treatment (due to, e.g., age, gender, illness severity, opportunity costs); (ii) delay of treatment due to waiting time reduces patients' benefit; and (iii) patients face costs related to obtaining treatment, including any costs related to examinations, referrals and, importantly, travelling. All patients with a non-negative net benefit from medical treatment demand care by joining the waiting list. A higher waiting time lowers patients' benefit, inducing patients with a low valuation (e.g., old patients or patients with mild conditions) to renounce medical treatment.[5,6]

To analyse the impact of competition on waiting times, we need to extend the basic model to more than one single provider. As a consequence, patients are not just deciding whether or not to demand medical treatment, but also *which* provider to demand treatment from. Using the framework of Salop (1979), we consider a market for elective hospital treatment where n hospitals are equidistantly located on a circle with circumference equal to 1. In this market there are two patient types – L and H – differing with respect to the gross valuation of treatment. Both types are uniformly distributed on the circle. A patient demands either one treatment from the most preferred hospital, or no treatment at all. The utility of an H-type patient who is located at x and seeking treatment at hospital i, located at z_i, is given by[7]

$$U^{\mathrm{H}}(x, z_i) = V - kw_i - t|x - z_i|, \tag{1}$$

where V is the gross valuation of (instant) medical treatment for the H-type patient, w_i is the waiting time at hospital i, k is a parameter measuring the (marginal) disutility of delay of treatment, and t is a travelling cost parameter.[8]

[4] See also the theory section in Martin and Smith (1999) and Farnworth (2003).

[5] Lindsay and Feigenbaum (1984) provide empirical evidence showing that higher waiting time ration demand. This is later confirmed by more robust empirical studies by Martin and Smith (1999) and Martin et al. (2007) who show, after controlling for the supply of private beds, that demand for hospital treatment is (weakly) elastic to waiting time.

[6] The presence of private hospitals or clinics offering instant treatment might be an additional reason for waiting time to have a rationing effect on demand as rich patients might opt for the private alternative. A brief analysis of this aspect can be found in Brekke et al. (2007). Besley et al. (1999) provide empirical evidence showing that higher waiting times increase demand for private health insurance in the UK.

[7] In Lindsay and Feigenbaum (1984) the benefit from a treatment received after a wait of w has a present value of $v \cdot e^{-rw}$, where the gross valuation v varies across patients. The extension of their model to n hospitals, forces us to use a linear discount function rather than an exponential one. This assumption makes the model more simple without qualitatively affecting the results. As pointed out by Gravelle and Siciliani (2008), empirical evidence suggests that many individuals do not use exponential discounting of health, but instead use a variety of discounting functions, including hyperbolic discounting. Thus, linear discounting might in many cases be a good approximation.

[8] To make the analysis feasible and focused, we ignore hospital quality of care as a variable. As pointed out by an anonymous referee, there are clear parallels between waiting times and more generally quality of care, as waiting times can be interpreted as a negative form of hospital quality. However, there are also important differences, the main difference being that while increasing quality for the provider is costly, reducing waiting times is not. More precisely, increasing quality increases costs directly and also indirectly through a higher demand. In contrast, reducing waiting times increases demand only.

Equivalently, the utility of a L-type patient who is located at x and seeking treatment at hospital i, located at z_i, is given by

$$U^{\mathrm{L}}(x, z_i) = \upsilon - kw_i - t|x - z_i|, \tag{2}$$

where $V > \upsilon$. Difference in gross valuations across patients can be due to difference in age, gender, illness severity, or simply opportunity costs. For example, old patients with a non-severe condition might have a low valuation of medical treatment.

Travel costs are interpreted broadly and include all costs associated with being far from "home", not just the patients' travelling expenditures. For example, choosing a distant hospital may involve rather high travel and accommodation costs for family and relatives. In addition, distance might also involve non-pecuniary costs to patients due to, for instance, the possibility of less (or no) visits or simply discomfort of being far from home. Finally, for patients living in rural areas, travelling distances might actually be quite long, implying rather high travel expenditures.[9]

The parameter k measures the marginal disutility of waiting ($\partial U/\partial w_i = -k$), and can be thought of reflecting, for instance, illness severity. A high k implies a serious utility loss associated with delay of treatment. Without loss of generality, we normalise the marginal disutility to one, i.e., $k=1$. This implies that we can interpret t as the marginal disutility of travelling relative to waiting. Thus, a low t means that delay in treatment is of relatively more importance to the patient than travelling distance (both measured as disutility in monetary terms).[10]

We concentrate on cases where the H-segment is always covered, while the L-segment is only partially covered. This implies that patients with a high gross valuation (H-types) decide which hospital to demand treatment from, while patients with a low gross valuation (L-types) decide whether or not to join the waiting list of the closest hospital. That is, some L-patients will not seek treatment in equilibrium, as in Lindsay and Feigenbaum (1984). We assume that the H-segment constitutes a share λ of the total number of patients, which is normalised to 1.

Since the distance between hospitals is equal to $1/n$, the H-patient who is indifferent between seeking treatment at hospital i and hospital j is located at x_i^{H}, given by

$$V - tx_i^H - w_i = V - t\left(\frac{1}{n} - x_i^H\right) - w_j,$$

yielding

$$x_i^H = \frac{1}{2t}\left(w_j - w_i + \frac{t}{n}\right). \tag{3}$$

Total demand for hospital i from the H-segment is given by $X_i^{\mathrm{H}} = 2x_i^{\mathrm{H}}$.

L-patients seek treatment only at the nearest hospital, if at all. The L-patient who is indifferent between treatment at hospital i and no treatment is located at x_i^{L}, given by

$$\upsilon - tx_i^{\mathrm{L}} - w_i = 0,$$

yielding

$$x_i^{\mathrm{L}} = \frac{\upsilon - w_i}{t}. \tag{4}$$

Total demand for hospital i from the L-segment is given by $X_i^{\mathrm{L}} = 2x_i^{\mathrm{L}}$. A necessary assumption for waiting time to have a rationing effect on demand is that the L-segment is not fully covered, i.e., $x_i^L < 1/n$, which is the case if and only if $t > (\upsilon - w_i)/n$. We will later derive the conditions for this assumption to hold in equilibrium.

[9] There is strong empirical evidence showing that distance is a major predictor of patients' choice of hospital, see, e.g., Kessler and McClellan (2000) and Tay (2003).

[10] A recent paper by Monstad et al. (2006) measures the marginal substitution rate between waiting time and distance for hip replacements in Norway. They find that patients are not willing to travel far in order to obtain quicker treatment. In light of our model, this suggests that t is quite high.

Total demand facing hospital i from both segments is thus given by

$$X_i^D = \lambda X_i^{\mathrm{H}} + (1-\lambda)X_i^{\mathrm{L}} = \frac{2(1-\lambda)v - w_i(2-\lambda) + \lambda w_i}{t} + \frac{\lambda}{n}, \tag{5}$$

where $\lambda \in (0, 1)$. Notice that $X_i^D \in \left(\frac{\lambda}{n}, \frac{1}{n}\right)$, while total demand is given by $X^{\mathrm{D}} \coloneqq \sum_{i=1}^{\mathrm{n}} X_i^D \in (\lambda, 1)$. To gain a better understanding of the mechanisms of the model, it is useful to see how demand reacts to changes in waiting times at the hospital level. From Eq. (5) we see that

$$\frac{\partial X_i^D}{\partial w_i} = -\frac{2-\lambda}{t} < 0. \tag{6}$$

Notice that lower travelling costs makes it less costly for patients to demand treatment, or to switch between hospitals; this increases the demand responsiveness to changes in waiting times. However, since the demand loss due to increased waiting time is larger in the L-segment, a larger competitive segment (i.e., an increase in λ) will reduce the demand responsiveness to changes in waiting times.

Hospitals are prospectively financed by a public payer offering a lump-sum transfer T and a per-treatment price p. The objective function of hospital i is assumed to be given by

$$\pi_i = T + pX_i^S + \alpha B_i\left(w_i, w_j\right) - C\left(X_i^S\right) - F, \tag{7}$$

where X_i^S is the supply of hospital treatments. Apart from fixed hospital costs, F, the cost of supplying hospital treatments is given by an increasing and strictly convex cost function $C(\cdot)$. The convexity of the cost function captures an important feature in the context of waiting times, namely that hospitals face some capacity constraints.[11] The function $B_i(\cdot)$ gives the benefit of the patients from receiving treatment at hospital i, while the parameter $\alpha \in [0, 1]$ captures the degree of altruism of the provider.[12] More explicitly, the surplus to patients treated at hospital i is given by

$$\begin{aligned} B_i\left(w_i, w_j\right) = {} & 2\lambda \int_0^{\frac{1}{2t}\left(w_j - w_i + \frac{t}{n}\right)} (V - w_i - tx)dx \\ & + 2(1-\lambda) \int_0^{\frac{v - w_i}{t}} (v - w_i - tx)dx, \end{aligned} \tag{8}$$

where the first term is the surplus to H-type patients, and the second term is the surplus to the L-type patients.

Differentiating Eq. (8), we obtain

$$\frac{\partial B_i\left(w_i, w_j\right)}{\partial w_i} = -X_i^D - \frac{\lambda}{t}\left(V - \frac{w_i + w_j}{2} - \frac{t}{2n}\right) < 0. \tag{9}$$

A marginal reduction in the waiting time of hospital i has two effects. First, it reduces the waiting time, and thus increases utility, for all existing patients at hospital i. This is represented by the first term in Eq. (9). Second, it increases demand for treatment at hospital i. At the margin, the increased demand from the L-segment represents a zero utility contribution. However, in the H-segment, there is an inflow of patients with a strictly positive net utility of hospital treatment. This is represented by the second term in Eq. (9). Obviously, the magnitude of this second effect depends on the size of the competitive segment, λ. Notice also that patient surplus at hospital i is a convex function of w_i (implying that the altruistic disutility of waiting ($-\alpha B_i$) is concave in w_i).[13]

[11] A convex variable cost function is supported by evidence suggesting that economies of scale are quite rapidly exhausted in the hospital sector (see, e.g., Ferguson et al., 1999; Folland et al., 2004, for literature surveys).

[12] This formulation is consistent with Ellis and McGuire (1986), Chalkley and Malcomson (1998) and Jack (2005). It is also general. The special case of a profit-maximiser hospital can be obtained by setting $\alpha = 0$.

[13] From Eq. (9) we derive $\frac{\partial^2 B_i\left(w_i, w_j\right)}{\partial w_i^2} = \frac{4-\lambda}{2t} > 0$.

3. Equilibrium waiting times

In deriving the equilibrium, we assume, as is commonly done, that waiting time acts as a re-equilibrating mechanism between demand and supply, i.e., $X^D(w_i, w_j)=X^S$.[14] This implies that it is equivalent whether we maximise the hospital objective function with respect to supply or waiting time. For analytical purposes, we use the latter approach.

Thus, the hospitals simultaneously and independently choose announced waiting times, in order to maximise their objective functions. We assume that the hospitals are not able/allowed to discriminate between different patient types with respect to waiting times. We also assume that hospitals cannot turn down patients seeking treatment. This latter assumption implies that we do not allow for explicit rationing.

Recall that the investment in capacity is captured by the increasing marginal cost assumption. Waiting times have a role precisely because in the absence of waiting times, there would be excess demand. Suppose that waiting times are zero. Then demand is equal to:

$$X_i^D\left(w_i = w_j = 0\right) = \frac{2(1-\lambda)v}{t} + \frac{\lambda}{n}. \tag{10}$$

The optimal supply for provider i is given by X_i^S. If $X_i^S > X_i^D(w_i=0)$, then the optimal waiting time is zero: there is no rationing and the hospital is able to provide instant treatment to all the patients. However, if $X_i^S < X_i^D(w_i=w_j=0)$ there is excess demand equal to $X_i^D(w_i=w_j=0)-X_i^S>0$.

There are two possible interpretations for the equilibrium condition $X^D(w_i, w_j)=X^S$. The first is that at each point in time, providers ration demand and choose a waiting time high enough to bring the market in equilibrium: patients with lowest net benefit are not willing to wait and disappear from the market. The equilibrium waiting time is such that the market always clears: $X_i^S=X_i^D(w_i, w_j)$. This is the standard interpretation in the literature starting from the seminal paper of Lindsay and Feigenbaum (1984), followed more recently by Iversen (1993, 1997), Martin and Smith (1999), Olivella (2002), Barros and Olivella (2005) and Siciliani (2005).

The second possible interpretation is that waiting time adjustment is sluggish. Patients demanding treatment at any point in time t are added to the waiting list. The waiting list increases over time if demand exceeds supply, and vice versa. Before being added to the waiting list, each patient is told that they have to wait a time equal to $w_i(t)$ to clear the current waiting list. If the wait is too long some low-benefit patients will give up the treatment, and are not added to the waiting list. In Appendix A, we show that in the steady state we still obtain the condition $X_i^S=X_i^D(w_i, w_j)$.[15]

Substituting Eq. (5) into Eq. (7) and maximising Eq. (7) with respect to waiting time yields the following first-order condition for hospital i,

$$\frac{\partial \pi_i}{\partial w_i} = \left[p - C'\left(X_i\left(w_i, w_j\right)\right)\right]\frac{\partial X_i\left(w_i, w_j\right)}{\partial w_i} + \alpha\frac{\partial B_i\left(w_i, w_j\right)}{\partial w_i} = 0, \tag{11}$$

which implicitly defines a best response function $w_i(w_j)$. Notice that we have suppressed the superscript on the demand function.[16]

Differentiating Eq. (11), we see that waiting times are strategic complements:[17]

$$\frac{dw_i}{dw_j} = -\frac{\partial^2\pi_i/\partial w_j \partial w_i}{\partial^2\pi_i/\partial w_i^2} = \frac{\left(C''(\cdot)\frac{2-\lambda}{t} - \alpha\right)\frac{\lambda}{t} + \alpha\frac{\lambda}{2t}}{\left(C''(\cdot)\frac{2-\lambda}{t} - \alpha\right)\frac{2-\lambda}{t} - \alpha\frac{\lambda}{2t}} > 0 \tag{12}$$

If, say, firm j increases its waiting time, some (H-type) consumers switch to hospital i, which now faces a higher demand. To meet this increase in demand, hospital i has to increase its supply, but this would increase marginal costs,

[14] See Lindsay and Feigenbaum (1984), Gravelle et al. (2003), Iversen (1993, 1997), Martin and Smith (1999) and Siciliani (2005).

[15] We could potentially analyse also the effect of competition on waiting list as well as waiting time. However, the main focus of policy makers is on waiting times rather than waiting lists (what matters to patients is how long they have to wait, not how many patients are waiting on the list). Therefore, we do not pursue this any further.

[16] The second-order condition is $\partial^2\pi_i/\partial w_i^2 = -\left[\left(C'(\cdot)\frac{2-\lambda}{t} - \alpha\right)\frac{2-\lambda}{t} - \alpha\frac{\lambda}{2t}\right] < 0$, which is always satisfied for a sufficiently convex cost function; also, $\partial^2\pi_i/\partial w_j\partial w_i = \left(C'(\cdot)\frac{2-\lambda}{t} - \alpha\right)\frac{\lambda}{t} + \alpha\frac{\lambda}{2t}$, which is always positive whenever $\partial^2\pi_i/\partial w_i^2<0$.

[17] Applying the second-order condition reported in Footnote 16, it is straightforward to see that both the numerator and the denominator are positive, since $C''(\cdot)\,(2-\lambda)/t>\alpha$ is necessary (but not sufficient) for the second-order condition to be fulfilled.

making the first term in Eq. (11) more positive, implying that $\partial\pi_i/\partial w_i>0$. Since the price is fixed, we see from the first-order condition that the optimal response for hospital i to a higher w_j, is to reduce demand by increasing its waiting time, w_i, until the level where $\partial\pi_i/\partial w_i=0$. Thus, waiting times are strategic complements for competing hospitals.

In a symmetric equilibrium, $w_j=w_i=w^*$. Using Eqs. (5) and (6), the equilibrium waiting time is given by

$$-\frac{(2-\lambda)}{t}[p-C'(X_i(w^*))]=\alpha\left[X_i(w^*)+\frac{\lambda}{t}\left(V-w^*-\frac{t}{2n}\right)\right], \tag{13}$$

where

$$X_i(w^*)=2(1-\lambda)\left(\frac{v-w^*}{t}\right)+\frac{\lambda}{n}, \tag{14}$$

and $w^*=w^*(v, t, \lambda, \alpha, p, n)$.[18] Since the right-hand side of Eq. (13) is positive, the expression in the square brackets on the left hand side of Eq. (13) is negative in an interior solution.[19]

Thus, the equilibrium waiting time is such that the (regulated) price is lower than the marginal treatment cost. In other words, the marginal patient is financially unprofitable to treat for the hospital.

We want to focus on equilibria with strictly positive waiting times. This requires that the cost of treating the last patient who demands treatment at $w=0$ is larger than the treatment price p. This requirement will be met if the supply cost function is sufficiently convex. Furthermore, we restrict attention to interior solutions with a partially covered L-segment in equilibrium, i.e., $x_i^L\in\left(0,\frac{1}{2n}\right)$.

Proposition 1. *Assume that the degree of altruism is sufficiently small. Then there exists an equilibrium waiting time, implicitly defined by* Eq. (13), *which is positive and involves a partially covered L-segment, if* $p\in S:=(\underline{p}, \min\{\bar{p}_1, \bar{p}_2\})$, *where* $\underline{p}$ *and* $\bar{p}_1$ *are implicitly defined by*

$$\underline{p}=C'\left(\frac{\lambda}{n}\right)-\frac{\alpha t}{2-\lambda}\left[\frac{\lambda}{n}+\frac{\lambda}{t}\left(V-w^*\left(\underline{p}\right)-\frac{t}{2n}\right)\right]$$

and

$$\bar{p}_1=C'\left(\frac{1}{n}\right)-\frac{\alpha t}{2-\lambda}\left[\frac{1}{n}+\frac{\lambda}{t}\left(V-w^*\left(\bar{p}_1\right)-\frac{t}{2n}\right)\right],$$

while $\bar{p}_2$ *is given by*

$$\bar{p}_2=C'\left(2(1-\lambda)\frac{v}{t}+\frac{\lambda}{n}\right)-\frac{\alpha t}{2-\lambda}\left[2(1-\lambda)\frac{v}{t}+\frac{\lambda}{2n}+\frac{\lambda}{t}V\right].$$

The equilibrium waiting time is monotonically decreasing in the treatment price p.

All proofs, including the proof of the above proposition, are given in Appendix B.

The inverse relationship between equilibrium waiting times and the treatment price is easily explained. A higher price simply means that the marginal patient becomes less unprofitable to treat, which dampens the incentive to use waiting time as an instrument to shift demand from unprofitable patients towards neighbouring hospitals.

[18] Uniqueness and stability of the equilibrium is confirmed by the positive sign of the Jacobian:

$$\Delta:=\begin{vmatrix}\frac{\partial^2\pi_i}{\partial w_i^2} & \frac{\partial^2\pi_i}{\partial w_j\partial w_i}\\ \frac{\partial^2\pi_j}{\partial w_i\partial w_j} & \frac{\partial^2\pi_j}{\partial w_j^2}\end{vmatrix}=\frac{4}{t}\left(C''(\cdot)\frac{2-\lambda}{t}-\alpha\right)\left[\left(C''(\cdot)\frac{2-\lambda}{t}-\alpha\right)\frac{1-\lambda}{t}-\alpha\frac{\lambda}{2t}\right]>0,$$

where the expression in the square brackets is positive whenever the second-order condition is satisfied.

[19] The term $V-w^*-t/2n$ is the utility to the marginal H-type consumer in equilibrium, which is non-negative due to the market coverage assumption of the competitive segment. Since the right-hand side of Eq. (13) is positive, equilibrium waiting times are such that $p<C'(\cdot)$. The exact conditions for Eqs. (13)–(14) to constitute an interior equilibrium are provided in the Proof of Proposition 1 in Appendix B.

Notice also that, since positive equilibrium waiting times imply that the marginal patient is unprofitable for the hospitals to treat, the equilibrium is "undercutting proof", in the sense that it is never profitable for a hospital to deviate from the equilibrium by reducing waiting times in order to drive neighbouring hospitals out of the market.

4. The impact of competition on waiting times and activity

We will now use the model to analyse if and how competition in hospital markets affects waiting times and hospital activity in equilibrium. The analysis is done in two steps. We start out by considering the effect of *introducing competition* in a hospital market characterised by local monopolies. Subsequently, we consider the effects of different measures to *increase the degree of competition* in a hospital market where there is competition to begin with.

4.1. Introducing competition

Assume that the hospital market described in the previous section consists of local monopolies, where patients are allocated to hospitals purely according to geographical distance. If a patient decides to visit a hospital to undergo treatment, she has to attend the nearest hospital. In our model, this means that hospital i's demand from the H-segment is exogenously given by $X_i^H = \frac{1}{n}$. Total demand for hospital i is thus given by

$$X_i^D(w_i) = \frac{\lambda}{n} + (1-\lambda)\frac{2(v-w_i)}{t}. \tag{15}$$

There is now a demand response to waiting time changes only in the L-segment. Differentiating Eq. (15) with respect to w_i yields

$$\frac{\partial X_i^D(w_i)}{\partial w_i} = -\frac{2(1-\lambda)}{t} < 0. \tag{16}$$

Comparing Eqs. (6) and (16), we see that demand responsiveness is higher with competition.

The surplus to patients treated at hospital i is given by

$$B_i(w_i) = \lambda 2\int_0^{\frac{1}{2n}} (V - w_i - tx)dx + (1-\lambda)2\int_0^{\frac{v-w_i}{t}} (v - w_i - tx)dx, \tag{17}$$

where the first term is the surplus to H-type patients, and the second term is the surplus to the L-type patients. Differentiating Eq. (17), we obtain

$$\frac{\partial B_i(w_i)}{\partial w_i} = -X_i^D(w_i). \tag{18}$$

In the absence of competition, notice how the marginal reduction in patient surplus from waiting is lower in absolute value (cf. Eq. (9)). The reason is that, under monopoly, changing the waiting time has only an effect on inframarginal patients.

Inserting Eq. (15) into the first-order condition, Eq. (11), and applying symmetry, the equilibrium waiting time in a market with local monopolies, w^m, is given by[20]

$$-\frac{2(1-\lambda)}{t}[p - C'(X_i(w^m))] = \alpha X_i(w^m), \quad i = 1,2, \tag{19}$$

[20] The second-order condition is given by $\partial^2\pi_i/\partial w_i^2 = -\left(C''(\cdot)\frac{2(1-\lambda)}{t} - \alpha\right)\frac{2(1-\lambda)}{t} < 0$.

where

$$X_i(w^m) = 2(1-\lambda)\frac{v - w^m}{t} + \frac{\lambda}{n}. \tag{20}$$

Comparing Eqs. (13) and (19) we see that, for $w^* = w^m$, both the left-hand side and the right-hand side of Eq. (19) are smaller than the left-hand side and right-hand side of Eq. (13). This means that $w^m \lesseqgtr w^*$. A closer scrutiny of the two first-order conditions enables us to derive the following result:

Proposition 2. *Introducing competition in a hospital market previously characterised by local monopolies leads to longer (shorter) waiting times and lower (higher) activity in equilibrium if the competitive segment (λ) is sufficiently large (small);*

$$1 - \lambda < (>) \frac{t}{2n(V - v)}.$$

There are two counteracting effects that contribute to this result. First, $\partial X_i/\partial w_i$ increases in absolute value with the introduction of competition (see Eqs. (6) and (16)). In other words, introducing competition means that demand at each hospital becomes more responsive to changes in the waiting time announced by the hospital, and the magnitude of this effect is increasing in λ. This is intuitive, since, without competition, only patients in the L-segment respond to waiting times. So how does the magnitude of $|\partial X_i/\partial w_i|$ affect equilibrium waiting times? Remember that, with a hospital disutility of positive waiting times (due to altruism), the marginal patient is unprofitable to treat. In equilibrium, this financial loss is optimally weighed against the disutility of increasing waiting times. When hospital demand responds to waiting time changes in the competitive demand segment, each hospital gets a stronger incentive to increase the waiting time, since this now becomes an instrument for shifting unprofitable patients to neighbouring hospitals.

However, there is also another effect, related to the altruistic preferences of the hospitals, that works in the opposite direction. Comparing Eqs. (9) and (18) we see that the utility gain of reduced waiting times is higher under hospital competition. With free patient choice, a reduction in waiting times by hospital i attracts patients from neighbouring hospitals who, due to altruism, contribute positively to the hospital objective function. All else equal, this gives the hospitals incentives to reduce waiting times with the introduction of competition.

Thus, the introduction of competition has two different implications: on the one hand, there is competition to avoid treating unprofitable patients, while, on the other hand, there is "altruistic competition" to treat high-benefit patients. Both of these effects get stronger when the relative size of the competitive segment increases. However this relationship is more pronounced for the first effect. The reason is that, since treatment costs are convex, while the altruistic disutility of waiting ($-\alpha B_i$) is concave in w_i, the higher level of demand associated with a larger competitive segment means that competition to avoid treating unprofitable patients becomes a more dominating force as λ increases. Thus, competition leads to longer waiting times in equilibrium if $1 - \lambda < \frac{t}{2n(V-v)}$. Furthermore, we see that an increase in t and/or a reduction of n increase the parameter space for which competition leads to longer waiting times. The reason is that higher travelling costs and/or lower hospital density reduce the (altruistic) utility gain of reducing waiting times under competition, as can be seen from Eq. (9).

It should be noted that the ambiguous nature of the competition effect on equilibrium waiting times is crucially dependent on the way altruism is modelled, where hospitals are (partly) altruistic only toward their own patients. If instead hospitals cared equally about *all* patients in the market, competition would not influence the effect of waiting time changes on the altruistic component in the hospital objective function.[21] In this case, competition would unambiguously increase waiting times. Thus, the first of the two above discussed effects – competition to avoid unprofitable patients – is, in some sense, a more robust effect.[22]

Finally, it is important to notice that introduction of competition does not affect demand *per se*; thus, changes in equilibrium waiting times are driven solely by strategic competition effects.

[21] Under both competition and monopoly, the effect of a waiting time increase on total patient utility is given by $\frac{\partial\left(\sum_{k=1}^{n} B_k\right)}{\partial w_i} = -X_i^D$.

[22] It may also be the case that hospital managers care, to some extent, about all patients, but place a larger altruistic weight on patients at their own hospitals. This intermediate case would weaken the "altruistic competition" effect, without eliminating it completely, increasing the likelihood that competition leads to longer waiting times in equilibrium.

4.2. Increasing the degree of competition

Depending on interpretation, the effect of increased competition (or increased patient choice) on waiting times and activity can work through three different parameters in the model: t, λ and n. First, a reduction in travelling costs, t, will intensify competition between hospitals in the competitive segment of the market. Second, competition will also naturally increase if a larger share of the total market becomes competitive, i.e., if λ increases. One possible (outside-the-model) interpretation is a reduction in fixed costs of undergoing hospital treatment for some patients, implying that a larger share of patients find themselves in the competitive demand segment. Finally, the number of hospitals in the market, n, is a standard measure of the degree of competition. Below we present the comparative statics results with respect to the different competition measures on both waiting time and activity levels, obtained by total differentiation of Eq. (13), applying Cramer's rule.

4.2.1. Lower travelling costs

$$\frac{\partial w^*}{\partial t} = \frac{1}{2}\frac{\left(\frac{2-\lambda}{t}C''(\cdot)-\alpha\right)\frac{\partial X}{\partial t}+\frac{1}{t}\left[(p-C'(\cdot))\frac{2-\lambda}{t}+\alpha\frac{\lambda}{t}(V-w^*)\right]}{\left(C''(\cdot)\frac{2-\lambda}{t}-\alpha\right)\frac{1-\lambda}{t}-\alpha\frac{\lambda}{2t}} < 0, \tag{21}$$

$$\frac{dX(w^*)}{dt} = \frac{\partial X}{\partial t}+\frac{\partial X}{\partial w}\frac{\partial w^*}{\partial t} = \frac{-[(2-\lambda)(p-C'(\cdot))+\alpha\lambda(V-w^*)]+\alpha\lambda(\upsilon-w^*)}{\left(C''(\cdot)\frac{2-\lambda}{t}-\alpha\right)\frac{2(1-\lambda)}{t}-\alpha\frac{\lambda}{t}}\frac{2(1-\lambda)}{t^3} > 0, \tag{22}$$

where $\frac{\partial X}{\partial t} = -\frac{2(1-\lambda)(\upsilon-w)}{t^2} < 0$.[23,24] Lower travelling costs have two different effects on the hospitals' optimal choice of waiting times. First, there is a direct demand effect, as more patients in the L-segment will seek treatment. Each hospital will meet this demand increase by increasing waiting times, and the strength of this response depends on the additional costs of treating more patients relative to the altruistic disutility of longer waiting times. Notice here that a higher level of demand also implies that the utility loss of increasing the waiting time is larger, since there are more patients that need to wait for treatment at hospital i. However, due to the convexity of treatment costs, the net effect is still positive with respect to waiting time. Second, lower travelling costs imply that demand facing each hospital becomes more sensitive to changes in waiting times (see Eq. (6)), which means that it becomes more effective to use waiting times as an instrument to shift unprofitable demand to neighbouring hospitals. Thus, both effects contribute to increased equilibrium waiting times as a result of lower travelling costs.

The effect of lower travelling costs on equilibrium hospital activity is given by the sum of a direct positive demand effect and an indirect negative effect through the increase in equilibrium waiting time. We see from Eq. (22) that the total effect is negative. It is perhaps surprising that lower travelling costs lead to *reduced* activity in equilibrium. This can be explained in the following way: since treatment costs are strictly convex, while the disutility of waiting (due to altruism) is concave in w_i, it is more costly for hospitals to meet increased demand by increasing activity, relative to waiting times. Consequently, the hospitals will meet a demand increase (induced by lower travelling costs) by increasing waiting times until the level where the demand increase is completely offset. However, there is a second effect of lower travelling costs, as explained above. The effect on the responsiveness of demand to waiting times implies that the hospitals have incentives to increase demand even beyond the level where the initial demand increase is nulled out. Thus, a reduction of travelling costs, which initially causes an increase in

[23] Notice that the first-order condition ensures that the expression in the square bracket of the numerator of $\partial w^*/\partial t$ is negative.

[24] $\frac{\partial w^*}{\partial t} = -\frac{\begin{vmatrix}\partial^2\pi_i/\partial w_i\partial t & \partial^2\pi_i/\partial w_i\partial w_j\\ \partial^2\pi_j/\partial w_j\partial t & \partial^2\pi_j/\partial w_j^2\end{vmatrix}}{\Delta}$. Notice that $\partial^2\pi_i/\partial w_i\partial t = \partial^2\pi_j/\partial w_j\partial t$, so that $\frac{\partial w^*}{\partial t} = -\frac{1}{\Delta}\left(\partial^2\pi_i/\partial w_i\partial t\right)\left[\partial^2\pi_j/\partial w_j^2 - \partial^2\pi_i/\partial w_i\partial w_j\right] = -\frac{\partial^2\pi_i/\partial w_i\partial t}{\partial^2\pi_j/\partial w_j^2+\partial^2\pi_i/\partial w_i\partial w_j}$.

demand for hospital treatments, will actually lead to lower activity in equilibrium, due to the equilibrium response in waiting times.

4.2.2. A larger competitive segment

$$\frac{\partial w^*}{\partial \lambda} = \frac{1}{2}\frac{\left(\frac{2-\lambda}{t}C''(\cdot)\right) - \alpha\right)\frac{\partial X}{\partial \lambda} + \frac{p - C'(\cdot)}{t} - \frac{\alpha}{t}\left(V - w^* - \frac{t}{2n}\right)}{\left(C''(\cdot)\frac{2-\lambda}{t} - \alpha\right)\frac{1-\lambda}{t} - \alpha\frac{\lambda}{2t}} \gtrless 0. \tag{23}$$

$$\frac{dX(w^*)}{d\lambda} = \underset{+}{\frac{\partial X}{\partial \lambda}} + \underset{-}{\frac{\partial X}{\partial w}}\underset{-/+}{\frac{\partial w^*}{\partial \lambda}} = \frac{1}{2}\frac{-(p - C'(\cdot))\frac{2(1-\lambda)}{t^2} + \frac{2\alpha}{t^2}\left[(1-\lambda)\left(V - w^* - \frac{t}{2n}\right) + \lambda\left(v - w^* - \frac{t}{2n}\right)\right]}{\left(C''(\cdot)\frac{2-\lambda}{t} - \alpha\right)\frac{1-\lambda}{t} - \alpha\frac{\lambda}{2t}}, \tag{24}$$

where $\frac{\partial X}{\partial \lambda} = 2\left(\frac{1}{2n} - \frac{v-w}{t}\right) > 0$ since, in equilibrium, $x^H = 1/2n$ and $x^L = (v-w)/t$, and, by assumption, $x^L < x^H$.[25]

The first term in the numerator of $\partial w^*/\partial\lambda$ is positive while the second and the third are negative. Notice that even for a low degree of altruism, the effect of λ on waiting time is indeterminate. There are two offsetting effects that contribute to this ambiguity. Since demand is higher from the competitive segment, a higher λ will increase total demand, which – all else equal – contributes to longer waiting times. However, a larger H-segment implies that demand becomes less responsive to changes in waiting times, as seen from Eq. (6). This means that it becomes less effective to use waiting times to shift unprofitable patients to neighbouring hospitals, which – all else equal – reduces equilibrium waiting times. The sum of these two effects is indeterminate.

The effect of a larger competitive segment on equilibrium activity is also indeterminate, although clearly positive for sufficiently low values of λ. The reason is that, for low values of λ, the magnitude of the indirect effect through changes in equilibrium waiting times is relatively low, making the direct demand effect the dominant one. The first term in the numerator of $dX(w^*)/d\lambda$ is always positive. The second term is given by a weighted average of the utility of a H-type patient and a L-type patient when receiving treatment and located at $x = 1/2n$ (by assumption this utility is positive for the H-type and negative for the L-type). This term is consequently also positive if λ is sufficiently low.

4.2.3. Increased hospital density

$$\frac{\partial w^*}{\partial n} = -\frac{1}{2}\frac{\left(C''(\cdot)\frac{2-\lambda}{t} - \alpha\right)\frac{\lambda}{n^2} + \alpha\frac{\lambda}{2n^2}}{\left(C''(\cdot)\frac{2-\lambda}{t} - \alpha\right)\frac{1-\lambda}{t} - \alpha\frac{\lambda}{2t}} < 0 \tag{25}$$

$$\frac{dX(w^*)}{dn} = \underset{-}{\frac{\partial X}{\partial n}} + \underset{-}{\frac{\partial X}{\partial w}}\underset{-}{\frac{\partial w^*}{\partial n}} = \frac{1}{2n^2 t}\frac{\alpha\lambda}{\left(C''(\cdot)\frac{2-\lambda}{t} - \alpha\right)\frac{1-\lambda}{t} - \alpha\frac{\lambda}{2t}} > 0 \tag{26}$$

$$\frac{d[nX(w^*)]}{dn} = X + n\frac{dX}{dn} > 0. \tag{27}$$

Notice that the signs of Eqs. (25) and (26) are determined by applying the second-order condition.[26]

[25] $\frac{\partial w^*}{\partial \lambda} = -\frac{\begin{vmatrix} \partial^2\pi_i/\partial w_i\partial\lambda & \partial^2\pi_i/\partial w_i\partial w_j \\ \partial^2\pi_j/\partial w_j\partial\lambda & \partial^2\pi_j/\partial w_j^2 \end{vmatrix}}{\Delta}$. Notice that $\partial^2\pi_i/\partial w_i\partial\lambda = \partial^2\pi_j/\partial w_j\partial\lambda$, so that $\frac{\partial w^*}{\partial\lambda} = -\frac{1}{\Delta}\left(\partial^2\pi_i/\partial w_i\partial\lambda\right)\left[\partial^2\pi_j/\partial w_j^2 - \partial^2\pi_i/\partial w_i\partial w_j\right] = -\frac{\partial^2\pi_i/\partial w_i\partial\lambda}{\partial^2\pi_j/\partial w_j^2 + \partial^2\pi_i/\partial w_i\partial w_j}$.

[26] $\frac{\partial w^*}{\partial n} = -\frac{\begin{vmatrix} \partial^2\pi_i/\partial w_i\partial n & \partial^2\pi_i/\partial w_i\partial w_j \\ \partial^2\pi_j/\partial w_j\partial n & \partial^2\pi_j/\partial w_j^2 \end{vmatrix}}{\Delta}$. Notice that $\partial^2\pi_i/\partial w_i\partial n = \partial^2\pi_j/\partial w_j\partial n$, so that $\frac{\partial w^*}{\partial n} = -\frac{1}{\Delta}\left(\partial^2\pi_i/\partial w_i\partial n\right)\left[\partial^2\pi_j/\partial w_j^2 - \partial^2\pi_i/\partial w_i w_j\right] = -\frac{\partial^2\pi_i/\partial w_i\partial n}{\partial^2\pi_j/\partial w_j^2 + \partial^2\pi_i/\partial w_i\partial w_j}$.

Increased hospital density unambiguously reduces waiting times in equilibrium. The intuition is quite simple. An increase in n means that – all else equal – each hospital faces a lower demand from the competitive segment. This means, due to the convexity of treatment costs, that the marginal treatment cost (for the last patient) is lower at each hospital. Consequently, the marginal patient becomes less unprofitable to treat and the hospitals will respond by reducing waiting times. Note that increased capacity, in itself, is not enough to reduce waiting times, since the effect on waiting times comes only through the competitive segment, where increased capacity means lower demand for each hospital. This can easily be confirmed by observing that $\partial w^*/\partial n=0$ if $\lambda=0$.

There are two effects – one direct and one indirect – of an increase in n on the equilibrium activity at the hospital level. Increased hospital density in the market means that the number of patients treated per hospital from the competitive segment goes down. However, there is an indirect "spillover" effect from the competitive to the monopoly demand segment. Due to the demand effect in the competitive segment, resulting in shorter waiting times, demand increases from the hospitals' monopoly segments. Eq. (26) shows that the net effect on demand is positive. In this case, the reduction in waiting times fully compensates for the initial drop in demand. Total activity clearly increases with hospital density, given that activity per hospital increases.

The effects of increased hospital competition on waiting times and activity can be summarised as follows:

Proposition 3. *(i) Lower travelling costs increase waiting times and decrease hospital activity.*

(ii) A larger competitive market segment has an indeterminate effect on waiting times and hospital activity. In general, the effect on activity is positive if the competitive segment is sufficiently small.

(iii) Increased hospital density reduces waiting times and increases activity per hospital, as well as total activity in the market.

5. Hospital competition and welfare

Having derived and characterised the equilibrium waiting time, we want to explore the issue of whether competition leads to excessive or suboptimal levels of waiting time from a social welfare perspective. To answer this question, we first need to specify the welfare function. We use the conventional measure of welfare as an unweighted sum of consumers' and producers' surplus. The welfare analysis is conducted at the hospital level; for total welfare just multiply by n.

Since the model is symmetric, the socially optimal waiting time must be uniform across hospitals. Setting $w_i=w_j=w$, the surplus to patients treated at a particular hospital is then given by

$$B(w) = \lambda 2\int_0^{\frac{1}{2n}} (V - w - tx)dx + (1-\lambda)2\int_0^{\frac{v-w}{t}} (v - w - tx)dx, \tag{28}$$

where the first term is the surplus to H-type patients, and the second term is the surplus to the L-type patients. Notice that we are assuming, as we did for the hospitals, that the regulator cannot discriminate between patient types in terms of waiting time. The patient surplus function can be written as

$$B(w) = \frac{\lambda}{n}\left(V - w - \frac{t}{4n}\right) + \frac{(1-\lambda)}{t}(v-w)^2. \tag{29}$$

Not very surprisingly, we see that the consumer surplus is always maximised at zero waiting time.

Writing the social welfare function as the sum of consumers' and producers' surplus net of third-party payments, welfare at the hospital level is given by

$$W(w) = B(w) + T + pX(w) - C(X(w)) - F - (1+\gamma)[pX(w) + T], \tag{30}$$

where $\gamma>0$ is a positive constant denoting the opportunity cost of public funds.[27] Since it is costly for the regulator to fund hospital care, we assume that the lump-sum transfer T is set such that the hospital's participation constraint is

[27] The altruistic component αB is not included in the welfare function as this would lead to double-counting. As argued by Chalkley and Malcomson (1998), "There is a strong case for excluding this benevolent component from social welfare on the grounds that benevolence represents a desire to do what is in the social interest and, as such, should have no role in determining what the social interest is." See also Hammond (1987) for further discussion. Notably, our results will not be qualitatively affected by this in any case.

binding. Adding the (realistic) assumption that the provider also has a limited liability constraint, the transfer is set so that $pX+T=C(X)+F$. The social welfare function then simplifies to

$$W(w) = B(w) - (1+\gamma)[C(X)+F]. \tag{31}$$

5.1. The socially optimal waiting time

The socially optimal waiting time is obtained by maximising welfare with respect to waiting time, yielding the following first-order condition[28]

$$\frac{\partial W}{\partial w} = \frac{\partial B(w)}{\partial w} + (1-\gamma)C'(\cdot)\frac{\partial X(w)}{\partial w} = 0, \tag{32}$$

which states that waiting time is socially optimised at a level where the utility loss to patients from a marginal increase in waiting time is equal to the corresponding reduction of treatment costs.

Using Eqs. (14) and (29), and rearranging Eq. (32), we can write the expression for the socially optimal waiting time, denoted by w^s, as follows:

$$(1+\gamma)C'(X(w^s)) = \frac{-X(w^s)}{\frac{\partial X(w^s)}{\partial w}}, \tag{33}$$

where

$$X(w^s) = 2(1-\lambda)\left(\frac{\upsilon - w^s}{t}\right) + \frac{\lambda}{n}, \tag{34}$$

$$\frac{\partial X(w^s)}{\partial w} = -\frac{2(1-\lambda)}{t}. \tag{35}$$

and $w^s=w^s(\upsilon, t, \lambda, n)$.

Eq. (33) defines an interior solution for the socially optimal waiting time with a partially covered L-segment, i.e., $w^s>0$ and $x^L\in\left(0,\frac{1}{2n}\right)$. Proposition 4 below provides the exact conditions needed to support this equilibrium:

Proposition 4. *There exists a socially optimal waiting time, w^s, implicitly defined by* Eq. (33), *which is strictly positive and involves a partially covered L-segment, if*

$$C'\left(\frac{\lambda}{n}\right) < \frac{t\lambda}{2n(1-\lambda)(1+\gamma)}, \quad \textit{and}$$
$$C'\left(2(1-\lambda)\frac{\upsilon}{t} + \frac{\lambda}{n}\right) > \frac{\upsilon}{1+\gamma} + \frac{t\lambda}{2n(1-\lambda)(1+\gamma)}.$$

We see that a positive socially optimal waiting time with a partially covered L-segment requires that the cost function C is sufficiently convex. The socially optimal waiting time can be characterised by total differentiation, yielding the following comparative statics results:

$$\frac{\partial w^s}{\partial n} = -\frac{\lambda t}{2n^2(1-\lambda)} < 0, \tag{36}$$

[28] The second-order condition is given by

$$\frac{\partial^2 W}{\partial w^2} = -\frac{2(1-\lambda)}{t}(1+\gamma)\left[C''(\cdot)\frac{2(1-\lambda)}{t} - \frac{1}{1+\gamma}\right] < 0$$

Thus, the supply cost function must be sufficiently convex for the condition to be fulfilled, i.e., $C''(\cdot) > \frac{t}{2(1-\lambda)(1+\gamma)}$.

$$\frac{\partial w^s}{\partial t} = -\frac{(v-w)}{t} - \frac{C'(\cdot)2(1-\lambda)/t^2}{\frac{2(1-\lambda)}{t}\left[C''(\cdot)\frac{2(1-\lambda)}{t} - \frac{1}{1+\gamma}\right]} < 0, \tag{37}$$

$$\frac{\partial w^s}{\partial \lambda} = \frac{[t/2n - (v-w)]}{(1-\lambda)} - \frac{C'(\cdot)}{(1-\lambda)\left[C''(\cdot)\frac{2(1-\lambda)}{t} - \frac{1}{1+\gamma}\right]} \gtrless 0, \tag{38}$$

$$\frac{\partial w^s}{\partial \gamma} = \frac{C'(\cdot)}{(1+\gamma)\left[C''(\cdot)\frac{2(1-\lambda)}{t} - \frac{1}{1+\gamma}\right]} > 0, \tag{39}$$

$$\frac{\partial w^s}{\partial \alpha} = 0. \tag{40}$$

The results can be summarised in the following proposition:

Proposition 5. *The socially optimal waiting time is decreasing in hospital density and travelling costs, and increasing in the opportunity cost of public funds. While the size of the competitive demand segment has an indeterminate effect, the degree of altruism has no effect on the socially optimal waiting time.*

Intuitively, an extra provider reduces the demand for each hospital, which reduces the socially optimal waiting time. An increase in travelling costs reduces the demand from low-benefit patients, which reduces the optimal waiting time. Also, higher travelling costs reduce the responsiveness of demand to increases in waiting times, and therefore reduces the effectiveness of waiting times in bringing in equilibrium the demand and supply of treatments, which reduces the optimal waiting times. For both reasons, waiting times decrease in equilibrium when travelling costs go up.

An increase in the proportion of high-benefit patients increases demand as high-benefit patients always demand treatment in equilibrium, while some low-benefit patients do not. This leads to a longer optimal waiting time. On the other hand, a higher proportion of high-benefit patients reduces the responsiveness of demand to increases in waiting times, and therefore reduces the effectiveness of waiting times as a re-equilibrating mechanism. This reduces the optimal waiting times. Consequently, waiting times may increase or decrease.

A higher opportunity cost of public funds naturally increases the socially optimal waiting time, while the degree of altruism has no effect. Since the limited-liability constraint is binding, rather than the participation constraint, the degree of altruism does not influence the socially optimal waiting time.

5.2. The socially optimal treatment price

The socially optimal waiting time can always be implemented by an appropriate choice of p. This price, denoted by p^*, is such that[29]

$$p^* = X(w^*)\frac{t[\lambda + 2(1-\lambda)(1-\alpha)]}{2(1-\lambda)(2-\lambda)} - \gamma C'(X(w^*)) - \frac{\alpha\lambda}{2-\lambda}\left(V - w^* - \frac{t}{2n}\right). \tag{41}$$

Intuitively, the optimal price is higher when the marginal benefit from a reduction in waiting time is higher $\left(-\frac{\partial B}{\partial w} = X\right)$, and it is lower when the degree of altruism α or the opportunity cost of public funds γ is higher. The last term in Eq. (41) takes into account the fact that the marginal benefit from a reduction in waiting time in a competitive setting $\left(-\frac{\partial B(w_i,w_j)}{\partial w_i}\right)$ is higher from the provider's perspective than from the social one $\left(-\frac{\partial B(w^*)}{\partial w^*}\right)$: the larger the difference between the two, the lower is the optimal price.

[29] The optimal price p^* maximises Eq. (31) so that: $[\partial B(w^*)/\partial w^* - (1+\gamma)C'(\cdot)(\partial X(w)/\partial w)]\ \partial w^*/\partial p = 0$, where $\partial B(w^*)/\partial w^* = -X(w^*)$ and $\partial X(w^*)/\partial w^* = -2(1-\gamma)/t$. Comparing the above with Eq. (11), the result is obtained.

The effects of our different competition measures on the optimal price can be derived by total differentiation, yielding the following comparative statics results:

$$\frac{\partial p^*}{\partial n} = \left(\frac{t[\lambda + 2(1-\lambda)(1-\alpha)]}{2(1-\lambda)(2-\lambda)} - \gamma C''(\cdot)\right)\frac{dX(w^*)}{dn} + \frac{\alpha\lambda}{2-\lambda}\left(\frac{dw^*}{dn} - \frac{1}{2n^2}\right) \tag{42}$$

$$\frac{\partial p^*}{\partial t} = \left(\frac{t[\lambda + 2(1-\lambda)(1-\alpha)]}{2(1-\lambda)(2-\lambda)} - \gamma C''(\cdot)\right)\frac{dX(w^*)}{dt} + \frac{\lambda + 2(1-\lambda)(1-\alpha)}{2(1-\lambda)(2-\lambda)}X(w^*) + \frac{\alpha\lambda}{2-\lambda}\left(\frac{dw^*}{dt} + \frac{1}{2n}\right) \tag{43}$$

$$\begin{aligned}\frac{\partial p^*}{\partial \lambda} = &\left(\frac{t[\lambda + 2(1-\lambda)(1-\alpha)]}{2(1-\lambda)(2-\lambda)} - \gamma C''(\cdot)\right)\frac{dX(w^*)}{d\lambda} + 2t\frac{\lambda^2(1-2\alpha) + (\alpha-1)4\lambda + 2(2-\alpha)}{(2(1-\lambda)(2-\lambda))^2}X(w^*) \\ &+\alpha\left(\frac{\lambda}{2-\lambda}\frac{dw^*}{d\lambda} - \frac{1}{2-\lambda}\left(V - w^* - \frac{t}{2n}\right)\right).\end{aligned} \tag{44}$$

The results are summarised as follows:

Proposition 6. *If the degree of altruism or the opportunity cost of public funds is sufficiently low, then a higher hospital density increases the optimal price while lower travelling costs decrease the optimal price. A higher competitive segment has an indeterminate effect on the optimal price.*

Since $dX(w^*)/dn > 0$, a higher hospital density increases activity and increases the social marginal benefit from a reduction in waiting times and therefore increases the optimal price. However a higher activity also increases the marginal cost, which induces a lower price. Furthermore, a higher hospital density reduces waiting times and travelling costs, increasing the marginal benefit from a reduction in waiting time for the semi-altruistic provider, which induces a lower price. Whenever the opportunity cost of public funds or the degree of altruism is sufficiently low the first effect dominates and the optimal price increases.

Since $dX(w^*)/dt > 0$, lower travelling costs reduce activity and reduce the marginal social benefit from a reduction in waiting times and therefore reduces the optimal price. However a lower activity also reduces the marginal cost, which induces a higher price. Furthermore, lower travelling costs imply a more responsive demand, which increases the marginal revenue for the hospital from a reduction in waiting times, inducing a lower optimal price. Finally, lower travelling costs increase waiting times but increase the utility of the patients, so that the marginal benefit from a reduction in waiting time for the semi-altruistic provider can be higher or lower. Whenever the opportunity cost of public funds or the degree of altruism is sufficiently low the optimal price reduces when travelling costs are smaller.

The effect of variations in the competitive segment λ on optimal prices is generally indeterminate. Overall, the analysis in this section suggests that whether higher-powered incentive schemes complements or substitute competition depends on the type of competition. Given that α or γ are not too high, while more competition through a higher hospitals density makes higher-powered incentive schemes more desirable, more competition through lower travelling costs makes higher-powered incentive scheme less desirable.

5.3. Does competition improve social welfare?

Consider the policy choice of monopoly versus competition in the hospital market. Since, for a given waiting time, the patient surplus $B(w)$ is unaffected by this choice of market regime, it is straightforward to see that competition is welfare neutral if the treatment price is set at the level which maximises social welfare, i.e., $p = p^*$. In this case, the effect of competition on equilibrium waiting times will be neutralised by an appropriate adjustment of p, keeping $w^* = w^s$. However, in the general case, where p is not necessarily set at the optimal level,[30] the welfare effect of hospital competition is characterised as follows:

Proposition 7. *Let p^* and p^m be the prices that yield $w^* = w^s$ and $w^m = w^s$, respectively.*

(i) Assume that the competitive demand segment is large; $1 - \lambda < \frac{t}{2n(V-v)}$, implying $w^* > w^m$ and $p^* > p^m$. Then there exists a price $\tilde{p} \in (p^m, p^*)$ such that hospital competition is welfare superior (inferior) if $p > (<) \tilde{p}$.

[30] Indeed, the most frequently used hospital payment system is DRG-pricing, which is close to average cost pricing for specific treatments, and clearly not in line with the optimal pricing rule considered in this section.

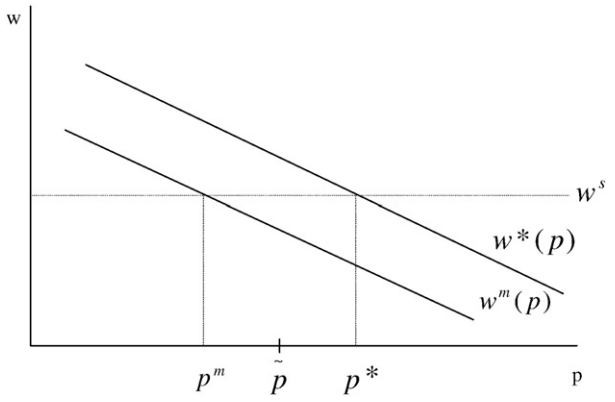

Fig. 1. Welfare effects of competition with a large competitive segment.

(ii) Assume that the competitive demand segment is small; $1-\lambda > \frac{t}{2n(V-v)}$, *implying* $w^* < w^m$ *and* $p^* < p^m$. *Then there exists a price* $\tilde{\tilde{p}} \in (p^*, p^m)$ *such that hospital competition is welfare superior (inferior) if* $p <(>) \tilde{\tilde{p}}$.

Whether or not hospital competition improves social welfare depends here on the characteristics of the reimbursement system (more specifically, the level of prices p) and the relative size of the competitive demand segment (λ). An increase in the treatment price always induces hospitals to increase supply and shorten waiting times. An illustration of the case of $1-\lambda < \frac{t}{2n(V-v)}$ is given in Fig. 1. The following discussion summarises the different cases:

High price. When the price is sufficiently *high*, waiting times in the monopoly equilibrium are shorter than the socially optimal level and activity is excessively high (i.e., the marginal benefit from treating an extra patient is below the marginal cost).

a) If, in addition, the competitive segment λ is sufficiently large so that $w^* > w^m$, then hospital competition increases waiting times towards the optimal level w^s, reducing activity and increasing welfare (see Fig. 1 for $p > p^*$).
b) In contrast, if the competitive segment λ is sufficiently small so that $w^* < w^m$, then hospital competition reduces waiting times even further from the optimal level w^s, increasing activity and reducing welfare.

Low price. The opposite analysis holds if the price is sufficiently *low*. Then waiting times in the monopoly equilibrium are longer than the socially optimal level and activity is excessively low (i.e., the marginal benefit from treating an extra patient is above the marginal cost).

a) If, in addition, the competitive segment λ is sufficiently large so that $w^* > w^m$, then hospital competition increases waiting times further from the optimal level w^s, reducing activity and reducing welfare (see Fig. 1 for $p < p^m$).
b) In contrast, if the competitive segment λ is sufficiently small so that $w^* < w^m$, then hospital competition reduces waiting times towards the optimal level w^s, increasing activity and increasing welfare.

Suppose that we start by a situation where waiting times are excessively high and prices too low. For example, until a few years ago in the UK hospitals were paid with fixed budgets ($p=0$). Similarly, in Norway in 1997 only 30% of the revenues were based on tariffs. In both countries waiting times are considered a major policy concern and are at least perceived as too high. Our analysis suggests that policies that encourage competition will have the expected effect only if the competitive demand segment is sufficiently low. It is only in this case that competition will reduce waiting times, increase activity and increase welfare.

6. Extensions

In this section, we extend our analysis in two directions. First, we allow for a fee, or more precisely, a copayment associated with hospital treatment. Second, we perform a welfare analysis assuming that the regulator might put a stronger weight on poor patients. Redistribution is a relevant concern in hospital treatment.

6.1. Introducing a charge (copayment)

The model so far has assumed that patients receive treatment free of charge. Suppose that on top of receiving a payment p from the government, hospitals can also charge a small fee f. The results obtained in Sections 2–4 are mainly unchanged. Suppose that patients have income m, and that utility is separable between income and benefit from treatment, so that the utility of an H-type (resp. L-type) patient who is located at x and seeking treatment at hospital i, located at z_i, is given by

$$U^{\mathrm{H}}(x,z_i) = V - t|x-z_i| - w_i + u(m-f), \tag{45}$$

$$U^{\mathrm{L}}(x,z_i) = v - t|x-z_i| - w_i + u(m-f). \tag{46}$$

We can show that by redefining

$$V' = V + u(m-f); \quad v' = v + u(m-f); \quad p' = p - f \tag{47}$$

the model is analytically equivalent to the one presented above. More precisely, the equilibrium waiting time is given by Eqs. (13) and (14), where V, v and p are substituted with V', v' and p'. Similarly, for the comparative statics of waiting time with respect to t, λ and n (Section 4.2). Also Proposition 2 regarding the effect of introducing competition remains identical. This is because

$$(V'-v') = V + u(m-f) - v - u(m-f) = (V-v).$$

More insightful is the effect of the introduction of a charge on welfare. The surplus to patients treated at a particular hospital is then given by

$$B(w,f) = \lambda 2\int_0^{\frac{1}{2n}} (V + u(m-f) - w - tx)dx + (1+\lambda)2\int_0^{\frac{v+u(m-f)-w}{t}} (v + u(m-f) - w - tx)dx, \tag{48}$$

where the first term is the surplus to H-type patients, and the second term is the surplus to the L-type patients. Differentiating with respect to the charge f we obtain

$$\frac{\partial B(w,f)}{\partial f} = -u_m(\cdot)X(w,f) < 0.$$

Using the same definition of social welfare as in the previous analysis, welfare at the hospital level is given by

$$W(w) = B(w) + T + (p+f)X(w) - C(X(w)) - F - (1+\gamma)[pX(w) + T]. \tag{49}$$

Assuming that the regulator sets the price so as to induce the socially optimal waiting time, w^s, and assuming, as before, that the lump-sum transfer T is set such that the hospital's limited liability constraint is binding, social welfare is given by

$$W(w^s(f),f) = B(w^s(f),f) - (1+\gamma)[C(X(w^s(f),f)) + F] + f(1+\gamma)X(w^s(f),f). \tag{50}$$

We want to determine the welfare effect of an increase in the charge f. Applying the Envelope Theorem, we ignore the indirect effects of the fee on waiting time, and focus only on the direct effects. Therefore,

$$\frac{\partial W(w^s(f),f)}{\partial f} = -u_m(\cdot)X^s - (1+\gamma)C'(X^s)\frac{\partial X^s}{\partial f} + (1+\gamma)X^s + f(1+\gamma)\frac{\partial X^s}{\partial f}, \tag{51}$$

where

$$X^s := X(w^s(f),f) = \frac{\lambda}{n} + 2(1-\lambda)\left[\frac{v - w^s + u(\cdot)}{t}\right]. \tag{52}$$

From Eq. (51) we can identify four effects: (i) a higher fee reduces consumer surplus, and welfare (first term); (ii) a higher fee reduces demand, which reduces the marginal cost, which increases welfare (second term); (iii) a higher fee reduces the need of distortionary taxation and therefore increases welfare (third term); however (iv) a higher fee reduces demand, which reduces the benefits from lower distortionary taxation (fourth term).

Applying the first-order condition with respect to a socially optimal waiting time, (32), the above expression can be simplified to,

$$\frac{\partial W(w^s(f),f)}{\partial f} = (1+\gamma)X^s(\cdot)\left[1-\epsilon_f^X\right], \tag{53}$$

where

$$\epsilon_f^X := -\frac{\partial X^s}{\partial f}\frac{f}{X^s} = \frac{2n(1-\lambda)fu_m(\cdot)}{t\lambda + 2n(1-\lambda)[v-w^s+u(\cdot)]}. \tag{54}$$

Two results can be straightforwardly obtained, which are summarised in the final proposition of the paper:

Proposition 8. *(i) Introducing a charge is always welfare improving:*

$$\left.\frac{\partial W(w^s(f),f)}{\partial f}\right|_{f=0} > 0;$$

(ii) Increasing the charge is welfare increasing if demand is inelastic: $\epsilon_f^X < 1$.

For hospital care, the empirical evidence normally suggests that the elasticity is well below one. For example, evidence from the Rand Health Insurance Experiment suggests that a higher fee reduces demand by 0.2% (Manning et al., 1987; Newhouse, 1993).

6.2. Inequality aversion

The welfare function used so far is utilitarian as the two types of patients have the same weight. High benefit patients always get treated in our model. Therefore, inequality aversion in this model implies a higher weight for patients with low benefit. Define $\beta > 1$ as the weight given to low-benefit patients. With this reformulation, the patient surplus function becomes:

$$B(w) = \frac{\lambda}{n}\left(V - w - \frac{t}{4n}\right) + \beta\frac{(1-\lambda)}{t}(v-w)^2 \tag{55}$$

with

$$\frac{\partial B}{\partial w} = -\frac{\lambda}{n} - 2\beta\frac{(1-\lambda)}{t}(v-w) < 0. \tag{56}$$

A higher weight β implies that the marginal social cost of waiting is higher, which reduces the optimal social wait:[31]

$$\frac{\partial w}{\partial \beta} = -\frac{(v-w)}{(1+\gamma)\left[C''(\cdot)\frac{2(1-\lambda)}{t} - \frac{\beta}{1+\gamma}\right]} < 0. \tag{57}$$

A lower waiting time, in turn, implies a higher price p if the degree of altruism or the opportunity cost of public funds is sufficiently low. The main results of the analysis are unchanged.

[31] Notice that, with this reformulation of the social welfare function, the second-order condition for an interior solution requires that $C''(\cdot) > \frac{t\beta}{2(1-\lambda)(1+\gamma)}$.

7. Conclusions and policy implications

This study has analysed the impact of hospital competition on waiting times, using a Salop-type model. Our main result is that, compared with a benchmark case of local monopolies, hospital competition reduces waiting times only if the competitive demand segment is sufficiently small. Otherwise, if free choice is relevant for a sufficiently large share of the total patient mass (i.e., if the competitive segment is sufficiently large), then competition *increase*s waiting times. Therefore we suggest that policies that encourage choice and competition in health care markets may not be as successful as policymakers might expect. The intuition for this ambiguous result is that, on the one hand, free patient choice induces hospitals to "compete" to avoid treating unprofitable patients, while, on the other hand, free patient choice also induces semi-altruistic providers to compete to attract high-benefit patients. The first effect dominates when the competitive segment is sufficiently large.

We also find that policies aimed at reducing travelling costs (like reimbursing travel expenses for patients choosing to receive treatment in hospitals outside their catchment area) may surprisingly increase waiting times and reduce overall activity. The reason stems from the fact that reducing travelling costs makes the demand for treatment more elastic, making waiting times a more effective rationing tool.

According to our analysis, policies aimed at increasing hospital density will have the expected effect of reducing waiting times and increasing activity. For example, in countries like Denmark, the UK and Spain, governments have decided to contract out patients to existing private hospitals. This policy can be seen as effectively increasing the density of hospitals by opening the patients from the public waiting list to private providers. Since demand in each hospital is lower and the marginal cost less steep, providers will respond by increasing activity and reducing waiting times.

Many countries increasingly remunerate hospitals according to activity-based funding rules (like DRG pricing in Norway and other European countries or HRG pricing in the UK) where hospitals receive a price for each patient treated. Our analysis suggests that for countries where waiting times are excessively low and prices are too high, hospital competition is socially preferable to monopoly if the competitive demand segment is sufficiently large. In this case, competition will increase waiting time towards the optimal level, reducing activity and increasing welfare.

In contrast, for countries (like perhaps the UK, Finland or Norway) where waiting times are excessively high and prices too low, competition will reduce waiting times, increase activity and increase welfare only if the competitive demand segment is sufficiently small.

Finally, we show that whether higher-powered incentive schemes complements or substitute competition depends on the type of competition. While more competition through a higher hospitals density makes higher-powered incentive schemes more desirable, more competition through lower travelling costs makes higher-powered incentive scheme less desirable.

Appendix A. Adjustment of waiting times to steady state

Denote $y_i(t)$ as the waiting list at time t in hospital i. At any point in time the waiting list increases (decreases) if demand is bigger than supply:

$$\frac{\partial y^i}{\partial t} = X_i^D\big(w_i(t), w_j(t)\big) - X_i^S(t) \tag{A1}$$

The waiting time is given by the number of periods that each patient has to wait before her/his turn arrives, i.e., before all the patients on the current waiting list are treated. Therefore, the waiting time is implicitly defined by (see Siciliani, in press)

$$\int_t^{t+w_i(t)} X_i^S(\tau)d\tau = y_i(t). \tag{A2}$$

Differentiating Eq. (A2) with respect to time, we obtain

$$\left(1 + \frac{\partial w_i}{\partial t}\right)X_i^S(t - w_i) - X_i^S(t) = \frac{\partial y_i}{\partial t}.$$

Substituting $\partial y_i/\partial t$ from Eq. (A1), we obtain the waiting-time dynamics:

$$\frac{\partial w_i}{\partial t} = \frac{X_i^D\big(w_i(t), w_j(t)\big) - X_i^S(t + w_i(t))}{X_i^S(t + w_i(t))}. \tag{A3}$$

The waiting time increases (decreases) if the demand at time t is higher (lower) than the supply at time $t+w$.

We focus on the steady state solution, so that the waiting list, waiting time, supply and demand are constant over time, i.e.,

$$\frac{\partial y_i}{\partial t} = \frac{\partial w_i}{\partial t} = 0,$$

$$w_i(t) = w_i, \quad w_j(t) = w_j,$$

$$X_i^S(t) = X_i^S\big((t + w_i(t)) = X_i^S, \quad X_i^D\big(\big(w_i(t), w_j(t)\big) = X_i^D\big(w_i, w_j\big).$$

In the steady state, $\partial w_i/\partial t=0$, and Eq. (A3) implies $X_i^D(w_i, w_j)=X_i^S$, while the waiting list Eq. (A1) simplifies to $X_i^S w_i=y_i$. This last expression can be written intuitively as $w_i=y_i/X_i^S$. In the steady state, the waiting time is given by the ratio between the waiting list and supply: the waiting time is given by the number of periods to clear the waiting list. In the model derived in the paper, the waiting list y_i does not play any role and it is therefore suppressed.

Appendix B. Proofs

Proof of Proposition 1. We start by confirming the last part of the Proposition. By total differentiation of the first-order conditions, we obtain[32]

$$\frac{\partial w^*}{\partial p} = -\frac{(2-\lambda)/t}{2\left[\big(C''(\cdot)\frac{2-\lambda}{t} - \alpha\big)\frac{(1-\lambda)}{t} - \alpha\frac{\lambda}{2t}\right]} < 0$$

An interior solution with positive equilibrium waiting times requires that the following conditions are met: $w^*>0$ and $x^{\mathrm{L}}\in\big(0, \frac{1}{2n}\big)$. Assume $x^L=0$, which implies $X(w^*) = \frac{\lambda}{n}$. Inserting this into the first-order condition for hospital i, and rearranging, we get

$$p = C'\left(\frac{\lambda}{n}\right) - \frac{\alpha t}{2-\lambda}\left[\frac{\lambda}{n} + \frac{\lambda}{t}\left(V - w^*(p) - \frac{t}{2n}\right)\right]$$

Denote the price that solves this equation by $\underline{p}$. Since $\partial w^*/\partial p<0$ and $\partial x^L/\partial w<0$ we know that $x^L>0$ if $p>\underline{p}$. Now assume $x^{\mathrm{L}} = \frac{1}{2n}$, which implies $X(w^*) = \frac{1}{n}$.

Inserting this into the first-order condition yields

$$p = C'\left(\frac{1}{n}\right) - \frac{\alpha t}{2-\lambda}\left[\frac{1}{n} + \frac{\lambda}{t}\left(V - w^*(p) - \frac{t}{2n}\right)\right].$$

Denote the price that solves this equation by $\overline{p}_1$. Again, since $\partial w^*/\partial p<0$ and $\partial x^L/\partial w<0$ we know that $x^{\mathrm{L}}<\frac{1}{2n}$ if $p<\overline{p}_1$. Finally, assume $w^*=0$, which implies $X(0) = 2(1-\lambda)\frac{v}{t} + \frac{\lambda}{n}$. The first-order condition is then given by

$$p = C'\left(2(1-\lambda)\frac{v}{t} + \frac{\lambda}{n}\right) - \frac{\alpha t}{2-\lambda}\left[2(1-\lambda)\frac{v}{t} + \frac{\lambda}{2n} + \frac{\lambda}{t}V\right]$$

[32] $\frac{\partial w^*}{\partial p} = -\frac{\begin{vmatrix} \partial^2\pi_i/\partial w_i\partial p & \partial^2\pi_i/\partial w_i\partial w_j \\ \partial^2\pi_j/\partial w_j\partial p & \partial^2\pi_j/\partial w_j^2 \end{vmatrix}}{\Delta}$. Notice that $\partial^2\pi_i/\partial w_i\partial p=\partial^2\pi_j/\partial w_j\partial p$, so that $\frac{\partial w^*}{\partial p} = -\frac{1}{\Delta}\big(\partial^2\pi_i/\partial w_i\partial p\big)\Big[\partial^2\pi_j/\partial w_j^2 - \partial^2\pi_i/\partial w_i\partial w_j\Big] = -\frac{\partial^2\pi_i/\partial w_i\partial p}{\partial^2\pi_j/\partial w_j^2+\partial^2\pi_i/\partial w_i\partial w_j}$.

Denote this price by $\overline{p}_2$. By a similar argument as above, $w^*>0$ if $p<\overline{p}_2$. Since $\frac{\lambda}{n}<\min\left\{\frac{1}{n}, 2(1-\lambda)\frac{v}{t}+\frac{\lambda}{n}\right\}$, it is straightforward to see that $\underline{p}<\min\{\overline{p}_1, \overline{p}_2\}$, implying that S is non-empty, if α is sufficiently small. □

Proof of Proposition 2. Subtracting Eq. (13) from Eq. (19) yields

$$\frac{2}{\alpha}[C'(X_i(w^*)) - C'(X_i(w^m))] - 2(w^m - w^*) = \lambda\frac{2(1-\lambda)n(V-v)-t}{n(1-\lambda)(2-\lambda)}.$$

Let us first confirm that the left-hand side (*LHS*) of this equation is monotonic in w^m and w^*. Using Eqs. (5) and (15), we have that $\partial(\text{LHS})/\partial w^* = -\frac{2}{\alpha}C''(X_i)\frac{2-\lambda}{t}+2$ and $\partial(\text{LHS})/\partial w^m = \frac{2}{\alpha}C''(X_i)\frac{2(1-\lambda)}{t}-2$. Applying the second-order conditions, it is straightforward to verify that ∂ (LHS)/$\partial w^*<0$ and ∂(LHS)/$\partial w^m>0$. Since LHS=0 if $w^*=w^m$, it follows that $w^*>(<)w^m$ if the right-hand side of the equation is negative (positive), which is the case if $1-\lambda<(>)\frac{t}{2n(V-v)}$. Since Eqs. (14) and (20) are identical for a given waiting time, $w^m<w^*$ implies that $X_i(w^m)>X_i(w^*)$ and vice versa. □

Proof of Proposition 4. First, $x^L=0$ implies $X(w^s)=\frac{\lambda}{n}$. It follows from Eq. (33) that $C'\left(\frac{\lambda}{n}\right)<\frac{t\lambda}{2n(1-\lambda)(1+\gamma)}$ for $x^L>0$. Second, $x^L=\frac{1}{2n}$ implies $X(w^s)=\frac{1}{n}$. We see from Eq. (33) that $C'\left(\frac{1}{n}\right)>\frac{t}{2n(1-\lambda)(1+\gamma)}$ for $x^L<\frac{1}{2n}$. Third, $w^s=0$ implies $X(0)=2(1-\lambda)\frac{v}{t}+\frac{\lambda}{n}$. From Eq. (33) it is evident that $w^s>0$ requires $C'\left(2(1-\lambda)\frac{v}{t}+\frac{\lambda}{n}\right)>\frac{v}{1+\gamma}+\frac{t\lambda}{2n(1-\lambda)(1+\lambda)}$. Finally, observe that since, by definition, $2(1-\lambda)\frac{v}{t}+\frac{\lambda}{n}\leq\frac{1}{n}$, it follows that $w^s>0$ implies $x^L<\frac{1}{2n}$, making the condition for $x^L<\frac{1}{2n}$ redundant. □

Proof of Proposition 7. We know (Proposition 1) that $\partial w^*/\partial p<0$, and it is straight-forward to show that this also holds under monopoly, i.e., $\partial w^m/\partial p<0$. (i) From Proposition 2 we know that, if $1-\lambda<\frac{t}{2n(V-v)}$, $w^*<w^m$ for all p, implying that $p^m<p^*$. This means that, from a social welfare perspective, waiting time is too long in both regimes if $p<p^m$ and too short in both regimes if $p>p^*$. Since $w^*>w^m$ for all p, it follows that competition is always welfare superior if $p>p^*$, while a monopoly regime is always welfare superior if $p<p^m$. For $p\in(p^m, p^*)$, replacing monopoly with competition means going from a regime with too short waiting to a regime with too long waiting times in equilibrium. Since W is single-peaked in p, there exists a unique price $\tilde{p}\in(p^m, p*)$ such that competition is welfare superior (inferior) if $p>(<)\tilde{p}$. (ii) By the inverse argument we can define an equivalent price $\widetilde{\widetilde{p}}$ for the case of $1-\lambda>\frac{t}{2n(V-v)}$. □

References

Barros, P., Olivella, P., 2005. Waiting lists and patient selection. Journal of Economics and Management Strategy 14, 623–646.

Besley, T., Hall, J., Preston, I., 1999. The demand for private health insurance: do waiting times matter? Journal of Public Economics 72, 155–181.

Brekke, K.R., Siciliani, L., Straume, O.R., 2007. Competition and waiting times in hospital markets. CEPR Discussion Papers, p. 6285.

Chalkley, M., Malcomson, J.M., 1998. Contracting for Health Services when patient demand does not reflect quality. Journal of Health Economics 17, 1–19.

Cullis, P., Jones, J.G., Propper, C., 2000. Waiting and Medical Treatment: Analyses and Policies. Chapter 28. In: Culyer, A.J., Newhouse, J.P. (Eds.), Handbook on Health Economics. Elsevier, Amsterdam.

Dawson, D., Gravelle, H., Jacobs, R., Martin, S., Smith, P.C., 2007. The effects of expanding patient choice of provider on waiting times: evidence from a policy experiment. Health Economics 16, 113–128.

Ellis, R., McGuire, T., 1986. Provider behavior under prospective reimbursement: Cost sharing and supply. Journal of Health Economics 5, 129–151.

Farnworth, M.G., 2003. A game theoretic model of the relationship between prices and waiting times. Journal of Health Economics 22, 47–60.

Ferguson, B., Sheldon, T., Posnett, J. (Eds.), 1999. Concentration and choice in health care. Royal Society of Medicine, London.

Folland, S., Goodman, A.C., Stano, M., 2004. The economics of health and health care. Prentice Hall, Upper Saddle River, NJ.

Gravelle, H., Siciliani, L., 2008. Is waiting-time prioritisation welfare improving? Health Economics 17, 167–184.

Gravelle, H., Smith, P.C., Xavier, A., 2003. Performance signals in the public sector: the case of health care. Oxford Economic Papers 55, 81–103.

Hammond, P., 1987. Altruism. In: Eatwell, J., Milgate, M., Newman, P. (Eds.), The New Palgrave: A Dictionary of Economics. Macmillan, London, pp. 85–87.

Iversen, T., 1993. A theory of hospital waiting lists. Journal of Health Economics 12, 55–71.

Iversen, T., 1997. The effect of private sector on the waiting time in a National Health Service. Journal of Health Economics 16, 381–396.

Jack, W., 2005. Purchasing health care services from providers with unknown altruism. Journal of Health Economics 24, 73–93.

Kessler, D.P., McClellan, M.B., 2000. Is hospital competition socially wasteful? Quarterly Journal of Economics 115, 577–615.

Lindsay, C.M., Feigenbaum, B., 1984. Rationing by waiting lists. American Economic Review 74, 404–417.

Manning, W.G., Newhouse, J.P., Duan, N., Keeler, E.B., Leibowitz, A., 1987. Health Insurance and the demand for medical care: evidence from a randomised experiment. American Economic Review 77, 251–277.

Martin, S., Smith, P.C., 1999. Rationing by waiting lists: an empirical investigation. Journal of Public Economics 71, 141–164.

Martin, S., Rice, N., Jacobs, R., Smith, P.C., 2007. The market for elective surgery: Joint estimation of supply and demand. Journal of Health Economics 26 (2), 263–285.
Monstad, K., Engesæter, L.B., Espehaug, B., 2006. Patients' preferences for choice of hospital. HEB Working Paper No. 05/06. University of Bergen.
Newhouse, J.P., 1993. Free for All? Lessons from the RAND Health Experiment. Harvard University Press, Cambridge, Mass.
Olivella, P., 2002. Shifting public-health-sector waiting lists to the private sector. European Journal of Political Economy 19, 103–132.
Salop, 1979. Monopolistic competition with outside goods. Bell Journal of Economics 10, 141–156.
Siciliani, L., 2005. Does more choice reduce waiting times? Health Economics 14, 17–23.
Siciliani, L., in press. A note on the dynamic interaction between waiting times and waiting lists. Health Economics.
Siciliani, L., Hurst, J., 2004. Explaining waiting times variations for elective surgery across OECD countries. OECD Economic Studies 38, 1–23.
Siciliani, L., Hurst, J., 2005. Tackling excessive waiting times for elective surgery: a comparison of policies in twelve OECD countries. Health Policy 72, 201–215.
Siciliani, L., Martin, S., 2007. An empirical analysis of the impact of choice on waiting times. Health Economics 16, 763–779.
Tay, A., 2003. Assessing competition in hospital care markets: the importance of accounting for quality differentiation. RAND Journal of Economics 34, 786–814.
Xavier, A., 2003. Hospital competition, GP fundholders and waiting times in the UK internal market: the case of elective surgery. International Journal of Health Care Finance and Economics 3, 25–51.