Sá, Luís; Siciliani, Luigi; Straume, Odd Rune

Working Paper

Dynamic hospital competition under rationing by waiting times

CESifo Working Paper, No. 7661

Provided in Cooperation with:
Ifo Institute – Leibniz Institute for Economic Research at the University of Munich

Suggested Citation: Sá, Luís; Siciliani, Luigi; Straume, Odd Rune (2019) : Dynamic hospital competition under rationing by waiting times, CESifo Working Paper, No. 7661, Center for Economic Studies and ifo Institute (CESifo), Munich

This Version is available at:
https://hdl.handle.net/10419/201887

CESifo WORKING PAPERS

7661
2019

May 2019

Dynamic hospital competition under rationing by waiting times

Luís Sá, Luigi Siciliani, Odd Rune Straume

CESifo

Impressum:

CESifo Working Papers
ISSN 2364-1428 (electronic version)
Publisher and distributor: Munich Society for the Promotion of Economic Research - CESifo GmbH
The international platform of Ludwigs-Maximilians University's Center for Economic Studies and the ifo Institute
Poschingerstr. 5, 81679 Munich, Germany
Telephone +49 (0)89 2180-2740, Telefax +49 (0)89 2180-17845, email office@cesifo.de
Editor: Clemens Fuest
www.cesifo-group.org/wp

An electronic version of the paper may be downloaded
· from the SSRN website: www.SSRN.com
· from the RePEc website: www.RePEc.org
· from the CESifo website: www.CESifo-group.org/wp

CESifo *Working Paper No. 7661*
Category 11: Industrial Organisation

Dynamic hospital competition under rationing by waiting times

Abstract

We develop a dynamic model of hospital competition where (i) waiting times increase if demand exceeds supply; (ii) patients choose a hospital based in part on waiting times; and (iii) hospitals incur waiting time penalties. We show that, whereas policies based on penalties will lead to lower waiting times, policies that promote patient choice will instead lead to higher waiting times. These results are robust to different game-theoretic solution concepts, designs of the hospital penalty structure, and patient utility specifications. Furthermore, waiting time penalties are likely to be more effective in reducing waiting times if they are designed with a linear penalty structure, but the counterproductive effect of patient choice policies is smaller when penalties are convex. These conclusions are partly derived by calibration of our model based on waiting times and elasticities observed in the English NHS for a common treatment (cataract surgery).

JEL-Codes: C730, H420, I110, I180, L420.

Keywords: hospital competition, waiting times, patient choice, differential games.

*Luís Sá**
Department of Economics/NIPE
University of Minho, Campus de Gualtar
Portugal - 4710-057 Braga
luis.sa@uminho.pt

Luigi Siciliani
Department of Economics and Related Studies
University of York
United Kingdom - Heslington, York YO10 5DD
luigi.siciliani@york.ac.uk

Odd Rune Straume
Department of Economics/NIPE
University of Minho, Campus de Gualtar
Portugal - 4710-057 Braga
o.r.straume@eeg.uminho.pt

*corresponding author

May 2019
We thank two anonymous referees for valuable suggestions. Sá and Straume acknowledge funding from COMPETE (ref. no. POCI-01-0145-FEDER-006683), with the FCT/MEC's (Fundação para a Ciência e Tecnologia, I.P.) Financial support through national funding and by the ERDF through the Operational Programme on Competitiveness and Internationalization - COMPETE 2020 under the PT2020 Partnership Agreement. Sá thanks FCT/MEC for the PhD Studentship SFRH/BD/129073/2017, financed through national funding and by the ESF.

1 Introduction

Waiting times for non-emergency (elective) treatments are a key health policy concern across OECD countries, such as Australia, Canada, Ireland, Finland, Norway, Portugal, and the United Kingdom. Mean waiting times range between 50 and 150 days across countries for common procedures such as cataract surgery, hip and knee replacement, hernia, hysterectomy, and prostatectomy (Siciliani et al., 2014). Although some countries like Finland and the UK have had successes in 2000-2005 in reducing waiting times from high levels (e.g., more than 150 days on average for hip and knee replacement), waiting times have stalled in most countries since the financial crisis and have slowly started to rise again in some countries. In countries like Chile, Poland, and Estonia, waiting times for hip and knee procedures are still above one year (OECD, 2017).

Waiting times are a major source of dissatisfaction for patients since they postpone health benefits, may worsen symptoms, deteriorate patients' conditions, and lead to worse clinical outcomes. In response to the dissatisfaction that they generate, governments have taken a variety of measures to reduce waiting times. Many OECD countries have adopted some form of maximum waiting time guarantees (Siciliani, Moran, and Borowitz, 2013). However, the design and implementation of these guarantees can differ significantly across countries.

Two common approaches are to link maximum wait guarantees either to penalties or to competition (and patient choice) policies. The first approach was followed by Finland and England, which combined maximum waiting times with sanctions for failure to fulfil the guarantee. Targets with penalties were introduced in England in 2000-05 with political oversight from the Prime Ministerial Delivery Unit and the Health Care Commission. Senior health administrators risked losing their jobs if targets were not met. As a result, the proportion of patients waiting over six months was reduced by 6-9 percentage points (Propper, Sutton, et al., 2008). In 2010, maximum wait guarantees became a patient entitlement codified into the NHS Constitution, establishing a patient right to a maximum of 18 weeks from GP referral to treatment. In Finland, waiting time guarantees were combined with targets as part of the Health Care Guarantee in 2005, subsequently included in the 2010 Health Care Act. A National Supervisory Agency supervised the implementation of the guarantee through targets and penalised municipalities failing to comply. The number of patients waiting over six months was reduced from 12.6 per $1,000$ population in 2002 to 6.6 per $1,000$ in 2005 (Siciliani, Moran, and Borowitz, 2013).

The second approach involves combining maximum waiting time guarantees with patient choice

and competition policies. For example, in Denmark, if the hospital foresees that the maximum waiting time guarantee will not be fulfilled, the patient can choose another public or private hospital. In Portugal, when a patient on the waiting list reaches 75% of the maximum guaranteed time, a voucher that allows the patient to seek treatment at any other provider, including private sector providers, is issued. In several countries, like England and Norway, patients are free to choose any provider within the country (Siciliani et al., 2017).

From an economics perspective, waiting times act as a non-price rationing device to bring into equilibrium the demand for and the supply of health care in publicly-funded health systems. Many countries with a National Health Service or public health insurance combine the absence of co-payments with the presence of capacity constraints. As a result, an excess demand arises, which translates into a waiting list. One way to bring the demand for and the supply of treatments into equilibrium is to rely on waiting times. As argued by Lindsay and Feigenbaum (1984), Martin and Smith (1999), and Iversen (1993, 1997), waiting times tend to discourage demand if patients give up the treatment or opt for treatment in the private sector. Waiting times may also influence positively the supply of health services if altruistic providers exert greater effort and treat more patients when waiting times are higher.

In the present study, we investigate whether competition and patient choice policies play a useful role in reducing waiting times, and the extent to which such a role is altered in the presence of penalties for providers with long waits. Our model is dynamic to capture a key feature of the waiting time phenomenon. Waiting times tend to increase when demand for treatment is higher than the supply of treatment so that new patients are added to the waiting list. Similarly, waiting times tend to reduce when more patients are removed from the waiting list than those added. A second feature of our model is that hospitals compete for patients, with hospitals with lower waiting times attracting more patients.

The combination of a dynamic approach with strategic interactions across providers calls for a differential-game approach. Although we solve the model for both open-loop and closed-loop decision rules (Dockner, 2000), our main analysis is based on the arguably more realistic feedback (closed-loop) solution, where hospitals can observe (and react to) waiting times at each point in time, implying that supply decisions can be continuously revised based on the evolution of waiting times. Under open-loop decision rules, hospitals compute their optimal supply paths at the beginning of the game and are restricted to follow such plans thereafter. It seems plausible that

hospitals can adjust supply over time in response to the dynamics of waiting times (own and those of rival hospitals).

To model the demand for healthcare faced by each provider, we use a Hotelling approach with two hospitals located at each endpoint of the unit line segment. We adopt a general specification, which allows for two types of patients who differ in the valuation of their outside option (e.g., to seek treatment in the private sector or to forego treatment altogether), which in turn implies different net benefits, 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.

Our main aim is to investigate the effect of policies that facilitate *patient choice*, commonly interpreted as policies that stimulate competition, and how such policies interact with policies based on waiting time penalties. Within our analytical framework, patient choice policies are modelled as a reduction in patients' transportation costs, which makes each hospital's demand more responsive to changes in waiting times and is a standard competition measure in spatial competition models. The effect of such policies is studied in contexts where waiting time penalties are either *linear* in waiting times or *convex* in waiting times, with the marginal penalty increasing with waiting.

We obtain several policy relevant findings. Importantly, we find that policies to increase patient choice lead to *higher* steady-state waiting times as long as hospitals suffer a disutility from positive waiting times. Increased patient choice makes demand more responsive to changes in waiting times, which implies that a unilateral reduction in waiting time at one hospital will lead to a larger demand increase for this hospital. This implies, in turn, that it becomes more difficult for each hospital to reduce waiting times through a unilateral increase in the supply of treatments. In other words, patient choice policies reduce the effectiveness of treatment supply as an instrument to reduce waiting times. The policy implication of this result is that patient choice policies are counterproductive, in terms of reducing waiting times, in the presence of waiting time penalties. Moreover, higher waiting penalties make patient choice policies even more counterproductive. We also show that a combined policy of more patient choice and higher waiting time penalties will lead to higher waiting times if the waiting time penalty is sufficiently high to begin with.

The above described results are derived analytically for the case of constant marginal provider disutility of waiting time, for example because of linear waiting time penalties. For the case of convex waiting time penalties, a closed-form solution cannot be obtained, and our results are

therefore numerically derived. To make the results more salient, we calibrate our model based on waiting times observed in the English NHS for a common treatment (cataract surgery). The calibration is also informed by demand elasticities which have been estimated in the empirical literature (Martin and Smith, 1999; Sivey, 2012).

The calibration output shows that our main result, that patient choice policies lead to higher waiting times, also carries over to the case of convex waiting time penalties. This comes as no surprise, the intuition behind this result does not rely on the shape of the provider disutility function but rather on the responsiveness of demand to waiting times. Not only is this result robust to the design of the waiting time penalty structure, it holds under a fairly general patient utility specification and is independent of the choice of game-theoretic solution concept, as it arises also under open-loop decision rules.

However, under closed-loop rules (where hospitals can observe and react to waiting times at each point in time), convex waiting time penalties introduce an additional strategic effect by creating *dynamic strategic substitutability* in supply. This implies that lower treatment supply by one hospital will be optimally met by increased supply by the competing hospital, which dampens the initial increase in waiting time caused by the supply reduction. This strategic substitutability gives each hospital an incentive to reduce its supply in order to 'free-ride' on the subsequent supply increase by the other hospital. The policy implication of this result is that, all else equal, waiting time penalties are likely to be more effective in reducing waiting times if they are designed with a linear penalty structure. On the other hand, we also show that the counterproductive effect of patient choice policies is smaller when penalties are convex instead of linear, which gives rise to yet another inherent conflict between these two policies. Waiting time penalties are more effective if they are linear, but linear penalties make patient choice policies more counterproductive.

The rest of the study is organised as follows. In the next section, we present a brief overview of the literature and explain how we contribute to it. In Section 3, we present the model, whereas the main analysis, based on the closed-loop solution, is given in Section 4. Section 5 considers patient welfare. Section 6 examines the robustness of our main result to non-linear patient utility in waiting time and distance. Finally, Section 7 provides concluding remarks, including a discussion of how our main results relate to the empirical literature on patient choice and waiting times.

2 Related literature

Our study brings together two different strands of the theoretical literature. The first is the literature that investigates the role of waiting times in the health sector. As mentioned above, the idea that waiting times may help bringing the supply and the demand for healthcare into equilibrium goes back to Lindsay and Feigenbaum (1984) and Iversen (1993). Iversen (1997) also investigates whether allowing patients to be treated in the private sector will reduce waiting times in the public sector and shows that the answer depends on the demand elasticity for public treatment with respect to waiting time. Demand and supply responsiveness to waiting times are estimated by Martin and Smith (1999) using English data, and they find that demand is generally inelastic (with an elasticity of about -0.1).

There are also normative analyses in this strand of the literature. Hoel and Sæther (2003) show that concerns for equity can make it optimal to have a mixed system of public and private provision with a positive waiting time in the public sector, though Marchand and Schroyen (2005) find, through a calibration exercise, that the welfare gains of a mixed system might be quite low. Gravelle and Siciliani (2008a, 2008c) investigate the scope for waiting time prioritisation policies across and within treatments and find that prioritisation is generally welfare improving even in a setting where the provider can only observe some dimensions of patient benefit. Gravelle and Siciliani (2008b) also show that rationing by copay tends to be welfare improving relative to rationing by waiting. All the above studies use a static approach assuming that demand and supply adjust instantaneously to reach equilibrium. One exception is Siciliani (2006), who investigates the behaviour of a monopolist in a dynamic set-up. We model waiting time dynamics in a similar way but critically allow for strategic interactions across providers to investigate the role of patient choice and competition.

The second strand of the literature relates to hospital competition with fixed prices. Though most of this literature consists of studies using a static framework, there is a limited but growing literature that models hospital competition in a dynamic framework. It focuses, however, on incentives for quality provision rather than on waiting times.[1] Brekke et al. (2010, 2012) find that, if quality is modelled as a stock variable which increases if quality investments are higher than its depreciation, or if demand is sluggish so that an increase in quality only partially translates into an increase in demand, then quality is higher under the open-loop solution if hospitals face increasing

[1] See Brekke et al. (2014) for a review of the theoretical literature on hospital competition under regulated prices.

marginal treatment costs. Equilibrium quality instead coincide under the two solution concepts if marginal treatment costs are constant. Siciliani, Straume, and Cellini (2013) suggest that these results can be overturned in the presence of altruistic preferences, so that quality is higher under the closed-loop solution.

Our modelling of waiting times differs analytically from these previous contributions because the state variable (i.e., waiting time) of the rival enters the dynamic constraint of the maximisation problem of each provider. This is not the case when quality is modelled as a stock (as in Brekke et al., 2010) because neither the state nor control variable of the rival provider enters the quality stock function. It is also not the case when demand is modelled as sluggish (as in Brekke et al., 2012) because demand depends on the control variable of the rival, not the state variable. Thus, because of these fundamental differences in the dynamic nature of the problems, the results from models of dynamic quality competition do not automatically carry over to the case of waiting times. In other words, if we want to study the effects of patient choice and competition on waiting times in a dynamic context, we cannot simply interpret waiting time as 'negative quality' and apply the results from the above mentioned studies of dynamic quality competition.

As previously mentioned, in the main bulk of the theoretical literature on hospital competition, the theoretical framework is a static one. To our knowledge, Brekke et al. (2008) were the first to deal with waiting times. Similarly to the present study, they identify a potentially positive relationship between patient choice and equilibrium waiting times. However, the underlying mechanisms are very different. In the static model (Brekke et al., 2008), hospitals choose waiting times to influence demand and in turn revenues. Increased competition (patient choice) makes demand more responsive to changes in waiting time, which then becomes a more effective tool for each hospital to steer demand in the desired direction. If hospitals are semi-altruistic, the equilibrium is such that price is below marginal cost (for the marginal treatment). Hospitals might therefore have an incentive to *reduce* demand, and waiting times become a more powerful tool to achieve this when patient choice increases, paving the way for a positive relationship between patient choice and equilibrium waiting times.

In the present dynamic approach, more competition also makes demand more responsive to waiting times, but then the similarities end. Hospitals choose treatment supply but cannot directly control waiting times. The supply decision is instead used as an instrument to affect waiting times, and this instrument becomes less effective with increased patient choice. This is why more

competition leads to higher waiting times in our dynamic setting, and the underlying mechanism is not related to price being below marginal cost in equilibrium, although this feature is also present here. Thus, the present study is not just a dynamic version of Brekke et al. (2008), in the sense that the results rely on the same mechanisms placed in a dynamic context. Rather, placing the analysis in a dynamic framework allows us to uncover new mechanisms that are uniquely related to the dynamic process that generates changes in waiting times. In this sense, the present dynamic analysis complements and reinforces the previous results based on a static framework.

More recently, Chen et al. (2016) developed a two-period signalling model in which they analyse the effect of waiting time report cards (i.e., the public reporting of waiting times) on the supply decisions and waiting times of two hospitals. Waiting times report cards increase competition in the market by providing patients with information and, hence, making demand responsive to waiting times. This generally gives hospitals incentives to increase their service rates (supply) up to the point where the marginal revenue equals the marginal cost, causing waiting times to fall in equilibrium. However, if the exogenous hospital qualities differ and are unknown to some patients, an incentive to use long waiting times as a signal for treatment quality arises for the high-quality hospital. Chen et al. (2016) show that the competitive effect (to attract patients) induced by waiting time report cards outweighs the signalling effect, so that both hospitals' waiting times are shorter than when there are no report cards, thus establishing a negative link between increased competition and waiting times (regardless of whether hospital qualities differ or are identical, which is the case that is equivalent to our analysis).

Their model shares with ours the feature that hospitals may only affect waiting times indirectly through supply but, crucially, assumes that hospitals face no form of disutility of waiting time. In the present analysis, increased supply is used not only to increase revenues but also to reduce waiting times and, hence, the disutility thereof. Increased supply reduces waiting times, which, in turn, attracts patients and thus dampens the initial decrease in waiting times. This demand response is stronger the greater is the degree of patient choice in the market. Higher demand responsiveness weakens the incentive hospitals have to increase supply and this is why the negative relationship between increased competition (patient choice) and waiting times fails to arise in the presence of hospital disutility of waiting time.

3 The Model

Consider a duopolistic health care market in which hospitals, indexed by i and j, are located at each endpoint of the unit line segment $[0, 1]$. There are N potential patients uniformly distributed on the line segment. In every period t, each of these patients may benefit from treatment at either of the two hospitals. In order to consume one unit of treatment, patients bear no out-of-pocket expenditures at the hospital but face expenses (or disutility) in the form of travelling costs. Furthermore, patients are required to join a waiting list and therefore suffer a disutility of waiting.

There are two types of patients, differing with respect to the value of their outside option (i.e., the utility of not being treated by either of the two hospitals). Whereas a share β of the patients are assumed to have no valuable outside option, the remaining share $(1-\beta)$ have a strictly positive outside option $k > 0$. For simplicity, we assume that these shares are constant along the line segment. The difference between these two patient types can be attributed either to a difference in illness severity, which creates a difference in the utility of being untreated, or to a difference in the ability to seek treatment elsewhere (e.g., in a private market or abroad), for example, due to differences in income or wealth.

Both types of patients make utility-maximising treatment consumption decisions, taking into account travelling costs as well as the length of time between the moment they join the waiting list and that when treatment is supplied (i.e., the waiting time). The utility in period t of a patient with no valuable outside option, who is located at $x \in [0, 1]$ and chooses Hospital i, located at z_i, is given by

$$u(x, z_i, t) = v - w_i(t) - \tau|x - z_i|, \tag{1}$$

where v is the gross valuation of treatment, $w_i(t)$ is the waiting time at Hospital i in period t, and τ is the marginal disutility of travelling. The marginal disutility of waiting is normalised to one, which allows τ to be interpreted as the marginal disutility of travelling relative to waiting. The equivalent utility in period t of a patient with a strictly positive outside option is

$$u(x, z_i, t) = v - k - w_i(t) - \tau|x - z_i|. \tag{2}$$

For patients with a positive outside option, we assume that k is sufficiently high such that some of these patients will strictly prefer the outside option to being treated by any of the two hospitals

in the market. This implies that the relevant choice for each of these patients is between seeking treatment at the most preferred hospital or exercising the outside option. We will refer to this as the *monopolistic segment* of the market. For all the patients without a valuable outside option, we assume that utility is maximised by seeking treatment at one of the hospitals. These patients therefore constitute the *competitive segment* of the market. By concentrating on cases where the competitive segment is fully covered, whereas the monopolistic segment is only partially covered, we ensure that total demand is elastic with respect to waiting times, implying that waiting times have a rationing effect on demand.

3.1 Demand for hospital treatment

In the *competitive* segment, the patient who is indifferent between seeking treatment at Hospital i and Hospital j is located at $x_C(t)$, implicitly given by

$$v - w_i(t) - \tau x_C = v - w_j(t) - \tau(1 - x_C), \tag{3}$$

yielding

$$x_C(t) = \frac{1}{2} + \frac{w_j(t) - w_i(t)}{2\tau}. \tag{4}$$

In the *monopolistic* segment, the patient who is indifferent between demanding treatment at Hospital i and consuming his or her outside option is located at $x_M^i(t)$, implicitly given by

$$v - w_i(t) - \tau x_M^i = k, \tag{5}$$

yielding

$$x_M^i(t) = \frac{v - k - w_i(t)}{\tau}. \tag{6}$$

A similar expression can be obtained for Hospital j: $x_M^j(t) = (v - k - w_j(t))/\tau$.

With a total mass N of patients in the market, demand faced by Hospitals i and j is a weighted sum of demand from the competitive and the monopolistic segments and is respectively given by

$$D_i(w_i(t), w_j(t)) = N[\beta x_C(t) + (1 - \beta)x_M^i(t)] \tag{7}$$

and

$$D_j(w_i(t), w_j(t)) = N[\beta(1 - x_C(t)) + (1 - \beta)x_M^j(t)]. \tag{8}$$

3.2 Hospital objectives and treatment supply

In each period t, Hospital i treats $S_i(t)$ patients. Hospitals are financed by a third-payer (e.g., a regulator or insurer) that offers a prospective payment p for each unit of treatment supplied and a lump-sum transfer T. The instantaneous objective function of Hospital i is assumed to be

$$\Pi_i(t) = T + pS_i(t) - C(S_i(t)) - \Phi(w_i(t)). \tag{9}$$

The cost of supplying hospital treatments is given by an increasing and strictly convex cost function $C(S_i(t)) = \frac{\gamma}{2}S_i(t)^2$, with $\gamma > 0$. The convexity of the cost function captures an important feature in the context of waiting times, namely that hospitals face capacity constraints.[2] The function $\Phi(w_i(t))$ captures the provider disutility of having positive waiting times. The disutility of waiting time is monetary if the hospital faces penalties levied by the regulator or reductions in funding. Alternatively, it is non-monetary if the hospital takes into the account the reputational damage of reporting long waiting times, or if the hospital is subject to a more stringent monitoring regime by the regulator. We assume that the disutility of waiting time takes the linear-quadratic form

$$\Phi(w_i(t)) = \alpha_1 w_i(t) + \frac{\alpha_2}{2}w_i(t)^2, \tag{10}$$

with $\alpha_1 \geq 0$ and $\alpha_2 \geq 0$. Whether waiting times penalties have a linear or non-linear effect on hospital utility depends on the institutional context. In settings where hospital managers can lose their jobs when waiting times become very long, penalties are arguably non-linear, with the marginal penalty increasing with waiting. This may also be the case in health systems where health regulators have mechanisms that escalate from warning messages to agreeing and monitoring action plans with the providers. Other health systems may instead gradually penalise hospitals with longer wait through a proportionate reduction in revenues.

Hospital targets are set for broad areas of care, typically all elective (non-emergency) care. Only in recent years some more stringent maximum waiting times have been specified for prioritised areas

[2] A strictly convex treatment cost function captures the case of *smooth* capacity constraints, where capacity can be increased, but only at an increasing marginal cost.

of care, such as cancer patients or certain cardiac surgeries (Siciliani, Moran, and Borowitz, 2013). Although our model is specified for a specific treatment which is reimbursed with DRG price p, any increase in supply for a specific treatment will contribute to reduce waiting times and help to satisfy the targets across all elective care. In Section 4.3, we calibrate the model for a specific treatment, cataract surgery. We choose this procedure because it has high volume and is correlated with waiting times for other high-volume procedures (such as hip and knee replacement; Siciliani et el., 2014). It has also similar demand elasticity to waiting across all elective care (Martin and Smith, 1999; Sivey, 2012).

Waiting times evolve dynamically over time according to

$$\frac{dw_i(t)}{dt} = \dot{w}_i(t) = \theta[D_i(w_i(t), w_j(t)) - S_i(t)] \tag{11}$$

and

$$\frac{dw_j(t)}{dt} = \dot{w}_j(t) = \theta[D_j(w_i(t), w_j(t)) - S_j(t)], \tag{12}$$

where $\theta > 0$ relates changes in waiting times to the difference between the demand faced by each hospital and its activity (i.e., changes in the waiting list). Under this formulation, waiting times increase when current demand exceeds current supply and vice versa, and the speed at which waiting times respond to changes in demand or supply is given by θ.

We are implicitly assuming that the waiting time at each hospital is positive in every period. The hospital objective function depends on the hospital's supply decision, which is given by the number of treatments performed by Hospital i in period t, $S_i(t)$. The objective function does not instead depend directly on demand, which is given by the number of patients added to Hospital i's waiting list in period t, $D_i(w_i(t), w_j(t))$. If $S_i(t) < D_i(w_i(t), w_j(t))$, there is a net increase in the waiting list and the (expected or average) waiting time increases. On the other hand, if $S_i(t) > D_i(w_i(t), w_j(t))$, there is a net reduction in the waiting list and the waiting time therefore falls. In either case, as long as the waiting list is not emptied, the number of treatments performed in period t is given by the hospital's supply of treatments. Demand for treatments only affects the actual number of treatments indirectly through waiting times, which in turn affect each hospital's optimal supply decisions, as we will show later.

We assume that the hospitals maximise their payoffs over an infinite time horizon and have a

common constant discount rate, ρ. Formally, the maximisation problem of Hospital i is given by

$$\begin{aligned}
\max_{S_i(t)\in\mathbb{R}_0^+} \quad & \int_0^\infty e^{-\rho t}\Pi_i(t)dt \\
\text{subject to} \quad & \dot{w}_i(t) = \theta[D_i(w_i(t), w_j(t)) - S_i(t)], \\
& \dot{w}_j(t) = \theta[D_j(w_i(t), w_j(t)) - S_j(t)], \\
& w_i(0) = w_{i0} > 0, \\
& w_j(0) = w_{j0} > 0.
\end{aligned}$$

Although, in reality, hospitals do not plan their activity over an infinite time horizon, we argue that this is a reasonable approximation if hospitals are regarded as lasting institutions. Managerial and medical structures are periodically replaced, but the hospital's *mission*—to provide care given its production technology and the regulatory scheme it faces—is likely to remain the same over long periods of time. This is likely if hierarchies are substituted by others with similar objective functions.

3.3 Solution concepts

There are two main solution concepts established by the differential-game literature (see Dockner et al., 2000). Under the *open-loop* solution, hospitals either compute their optimal supply paths at the beginning of the game and are restricted to follow such plans thereafter, or they may observe the state of the world (i.e., waiting times) only at $t = 0$ and cannot therefore condition their actions (i.e., supply) on these observations thereafter. In both cases, strategies are time-profiles that specify the supply to be provided at each point in time.

If, besides current time, hospitals observe waiting times in every period and factor them in their decision making, a *closed-loop* solution arises. Under this solution concept, Hospital i's supply is a function of the contemporaneous waiting times in each t. While the closed-loop solution is informationally more demanding, it involves weaker commitment since hospitals are allowed to adjust supply as waiting times evolve.

The appropriateness of each solution concept depends on the assumptions regarding the players' information set as well as commitment requirements. The open-loop solution implies that hospitals have no information concerning waiting times once the game starts or are committed to the supply plans computed at the beginning of the game, which might be considered an excessively stringent

assumption. Due to regulatory requirements, hospitals periodically collect and report data on waiting times, upon which their activity may be conditioned.[3] Moreover, a setting in which hospitals adjust activity according to waiting times is more realistic and relevant for policy-making.[4] Thus, although the closed-loop solution is computationally much more demanding, it is based on a set of assumptions that are arguably more realistic and we will therefore conduct our main analysis under the assumption that hospital behaviour is characterised by closed-loop decision rules.

4 Treatment supply and waiting times in the closed-loop solution

Suppose that hospitals are able to observe the evolution of waiting times and make supply decisions dependent on current waiting times. When solving for the closed-loop solution, we restrict attention to Markovian stationary strategies, whereby the controls (i.e., supply decisions) at time t depend only on the current values of the states (i.e, the waiting times), which summarise the history of the game. We also focus on a symmetric equilibrium with non-negative waiting times and a partially covered monopolistic segment.

We will present our results distinguishing between two different cases, namely *constant* and *increasing* marginal provider disutility of waiting time. As mentioned above, which case is more plausible depends on the institutional context and this may differ across countries or even within a country at different points in time. For example, one could argue that in England in 2000-2005 the marginal disutility was increasing in waiting times when senior health administrators risked losing their jobs if targets were not met. This would be the case if small deviations from the target would only lead to additional monitoring from the regulator, but a large deviation from the target would culminate into the hospital CEO being dismissed. In contrast, the marginal disutility of waiting time could be constant if deviations from a target led to a proportionate reduction in hospital income, which was implemented later in England. Therefore, both scenarios are important from a policy perspective. We discuss them in turn, starting with the case of constant marginal disutility, which allows us to obtain closed-form solutions for equilibrium supply and waiting times.

[3] See Siciliani, Moran, and Borowitz (2013) for a description of waiting times regulatory arrangements and policies across OECD countries.

[4] This need not be the case of other analyses of hospital behaviour. The case of quality competition as analysed in, for example, Brekke et al. (2010) provides a setting in which the open-loop solution might be, at least, as appropriate. If hospitals devise investment plans that ought to be followed for long periods of time, meaning that their discretion is strongly restricted, their actions (investment decisions) are *as if* they are not conditional on the state of the world (the stock of quality).

4.1 Constant marginal provider disutility of waiting time

Suppose that the disutility of waiting time is given by (10) with $\alpha_1 > 0$ and $\alpha_2 = 0$. In this case, it can be shown (see Appendix A) that the optimal supply rule for each hospital at time t is equal to the steady-state supply, S^{CL}, and given by

$$S_i(t) = S_j(t) = S^{CL} = \frac{p}{\gamma} + \frac{2\theta\tau\alpha_1}{\gamma\phi}, \tag{13}$$

where

$$\phi = \theta(2-\beta)N + 2\tau\rho - \frac{(\theta\beta N)^2}{\theta(2-\beta)N + 2\tau\rho} \in (0,1). \tag{14}$$

In other words, the optimal supply rule is independent of waiting times. We thus obtain the following result:

Proposition 1 *If the marginal provider disutility of waiting time is constant, the equilibrium is characterised by constant supply of treatments over time.*

This result is explained by the lack of strategic interaction between the hospitals when waiting time disutility is linear in waiting times. A unilateral increase in supply by Hospital i leads to an initial reduction in waiting times at this hospital. This will shift demand from the rival hospital and therefore will also reduce the waiting time at Hospital j. However, if $\alpha_2 = 0$, the reduction in waiting time at Hospital j does not affect the hospital's marginal disutility of waiting time, so that the hospital will not respond by changing its supply.[5]

The intuition behind each hospital's optimal supply rule is perhaps easier gained by re-writing (13) as

$$p + \frac{2\theta\tau\alpha_1}{\phi} = \gamma S_i. \tag{15}$$

On the one hand, a marginal increase in supply (i) generates more revenues and (ii) reduces the waiting time and its associated disutility. These two elements of the marginal benefit of supply are given by the two terms on the left-hand side of (15). On the other hand, increasing supply is costly, with the marginal cost of supply given by the right-hand side of (15). Each hospital offers a supply of treatments such that the marginal benefit is exactly offset by the marginal cost. This trade-off is key to understanding the main intuition behind most of our subsequently derived results.

[5] When $\alpha_2 = 0$, our differential game belongs to the class of the so-called linear-state games, which is characterised by the coincidence between the time path of controls and states under the open- and closed-loop solution concepts. The calibration in Section 4.3 illustrates this general result.

It also follows directly from (15) that, in an interior-solution equilibrium, each hospital operates at a level where the price-cost margin is negative, implying that the marginal patient is unprofitable to treat.[6] This is a result of the disutility of waiting time, which gives each hospital an incentive to expand supply beyond the level where the price is equal to marginal treatment costs.

The corresponding steady-state waiting time is given by[7]

$$w^{CL} = \frac{\tau}{(1-\beta)N}\left\{N\left[\frac{\beta}{2} + (1-\beta)\left(\frac{v-k}{\tau}\right)\right] - \frac{p}{\gamma} - \frac{2\theta\tau\alpha_1}{\gamma\phi}\right\}. \tag{16}$$

We can see directly from (16) that the steady-state waiting time is decreasing in p and α_1, which is very intuitive. A higher price (p) makes the marginal patient more profitable (or less unprofitable) to treat, whereas a higher waiting time penalty (α_1) increases the disutility of waiting time. In both cases, the hospitals have stronger incentives to increase supply and equilibrium waiting times will therefore go down.

4.1.1 Patient choice and waiting times

How does the degree of patient choice affect steady-state supply and waiting times? In our framework, the degree of patient choice can be inversely measured by the parameter τ, which is a standard (inverse) measure of competition intensity in the hospital competition literature that is based on models of spatial competition. A reduction in τ makes demand more responsive to changes in waiting times, thus reflecting a higher degree of patient choice.

The effect of a marginal change in τ on the steady-state waiting time and supply can be expressed as

$$\frac{\partial w^{CL}}{\partial \tau} = -\frac{(1-\beta)x_M^{CL} + \frac{\tau}{N}\frac{\partial S^{CL}}{\partial \tau}}{1-\beta} < 0, \tag{17}$$

where $x_M^{CL} = \left(v - k - w^{CL}\right)/\tau$ is the location of the indifferent patient in the monopolistic segment, and

$$\frac{\partial S^{CL}}{\partial \tau} = N\theta^2\alpha_1\frac{(1-\beta)[N\theta(2-\beta)+4\tau\rho]\theta N + (2-\beta)(\tau\rho)^2}{2\gamma(N\theta+\tau\rho)^2[N(1-\beta)\theta+\tau\rho]^2} > 0, \tag{18}$$

allowing us to establish the following result:

Proposition 2 *If the marginal provider disutility of waiting time is constant, a higher degree of*

[6] Notice that, when treatment costs are strictly convex, a negative price-cost margin for the *marginal* patient does not imply that the price-cost margin is negative for the *average* patient.

[7] In Appendix A, we show that a sufficiently large γ ensures that the steady-state is characterised by non-negative waiting times and a partially covered monopolistic segment.

patient choice leads to lower treatment supply and higher waiting times in the steady-state.

The negative relationship between τ and w^{CL} is a consequence of two effects that work in the same direction. First, there is a direct demand effect. A reduction in τ increases total demand (and therefore demand for each hospital) since a larger number of patients in the monopolistic segment chooses to opt for treatment (at the nearest hospital). A higher demand directly increases the waiting time at each hospital. This effect is given by the first term in the numerator of (17), and the size of this effect is smaller the larger the relative size of the competitive segment, β.

The second effect is related to how τ affects the demand responsiveness to waiting times in the competitive segment of demand, and is thus more directly related to the patient choice interpretation of the parameter τ. This is an indirect effect that works through changes in each hospital's incentive to affect waiting times through its treatment supply decision. Each hospital can lower its waiting time by increasing the supply of treatments, and the effect of a unilateral increase in treatment supply on the waiting time is given by a direct and an indirect (feedback) effect. For a given demand, an increase in treatment supply will reduce the waiting time. However, a lower waiting time will increase demand and therefore dampen the initial reduction in the waiting time. Crucially, the strength of this feedback effect depends on how strongly demand responds to waiting time changes. A lower τ makes demand more responsive to changes in waiting times, which increases the feedback effect and therefore makes treatment supply a less effective instrument to reduce waiting times. Consequently, this *reduces* the marginal benefit of treatment supply and gives each hospital an incentive to reduce the supply of treatments. This effect is captured by the second term in the numerator of (17).

Notice that the effect of a reduction in τ on steady-state supply does not depend on the direct demand effect, only on the indirect effect through demand responsiveness. Consider the special case of no waiting time disutility, $\alpha_1 = 0$. In this case, the second effect vanishes, since the hospitals have no incentives to adjust supply in order to affect waiting times. A reduction in τ will not affect the hospitals' supply decisions and waiting times increase only because of higher demand (i.e., waiting times increase only through the first of the two above mentioned effects). Thus, it is the presence of waiting time disutility ($\alpha_1 > 0$) that causes a negative relationship between patient choice and treatment supply. This has potentially interesting policy implications which we will explore in the following.

4.1.2 Combining patient choice policies with waiting time penalties

Suppose that policymakers aim at reducing hospital waiting times. Two commonly suggested policy options is to either directly target the perceived problem by introducing (or increasing) waiting time penalties, or to stimulate patient choice (e.g., by public reporting of waiting times) with the aim of achieving lower waiting times through increased intensity of competition between the hospitals. In our model, as Proposition 1 shows, only the former policy works, whereas the latter policy is counterproductive. Moreover, the former policy makes the latter policy more counterproductive. All else equal, the larger the waiting time penalties, the larger is the increase in steady-state waiting times as a result of more patient choice.

Many countries have introduced both choice policies and waiting time penalties. While our analysis shows that these two policies have counteracting effects on treatment supply and waiting times, it remains to show what determines the direction of the overall effect in a context where the two policies are combined. Consider therefore a policy package consisting of a marginal increase in the degree of patient choice combined with a marginal increase in the waiting time penalty. The resulting effect on steady-state waiting times is given by

$$\frac{\partial w^{CL}}{\partial \alpha_1} - \frac{\partial w^{CL}}{\partial \tau} = \frac{1}{N\left(1-\beta\right)}\left[\left(1-\beta\right)Nx_M^{CL} + \tau\left(\frac{\partial S^{CL}}{\partial \tau} - \frac{2\theta\tau}{\gamma\phi}\right)\right]. \tag{19}$$

If we exclude the demand effect of lower travelling costs, thus focusing exclusively on the patient choice interpretation of τ, the overall effect of this dual policy on waiting times is given by the sign of the second term in the square brackets of on the right-hand side of (19). It can easily be shown that the sign of this effect is positive, implying higher waiting times, if

$$\alpha_1 > \frac{\left(N\theta+\tau\rho\right)\left(\left(1-\beta\right)N\theta+\tau\rho\right)\left(\left(2-\beta\right)N\theta+2\tau\rho\right)\tau}{N\theta\left(N\theta\left(1-\beta\right)\left(\left(2-\beta\right)N\theta+4\tau\rho\right)+\left(2-\beta\right)\tau^2\rho^2\right)}. \tag{20}$$

Thus, a combined policy of increased patient choice and higher waiting time penalties is more likely to yield higher waiting times (and lower treatment supply) if the disutility of waiting time is sufficiently high to begin with. The reason is that the marginal effect of a higher waiting time penalty on waiting times is constant (as we can see from (16)), whereas the marginal effect of increased patient choice on waiting times is increasing in the disutility of waiting times. Consequently, the counterproductive effect of increased patient choice dominates for sufficiently high values of α_1.

It can also be shown that, unless β is very close to 1, the right-hand side of (20) is decreasing in θ and approaches τ as $\theta \to \infty$. This implies that the scope for a waiting time increase as a result of the combined policy is larger the faster waiting times adjust to changes in supply.

If we interpret the waiting time disutility as reflecting only waiting time penalties, we can summarise the above policy analysis as follows:

Proposition 3 *Suppose that waiting time penalties are linear in waiting times. In this case, (i) the counterproductive effect of patient choice policies on treatment supply and waiting times is larger the higher the waiting time penalty. Furthermore, (ii) a combined policy of increased patient choice and higher waiting time penalties has an ambiguous effect on treatment supply and waiting times, but is more likely to be counterproductive the higher the initial waiting time penalty.*

4.2 Increasing marginal provider disutility of waiting time

Suppose that the disutility of waiting time is given by (10) with $\alpha_1 > 0$ and $\alpha_2 > 0$. In this scenario, a closed-form solution of supply and waiting times cannot be obtained. Our game belongs to the class of linear-quadratic differential games wherein the state variables enter the objective function quadratically, while they enter the dynamic constraints linearly. Although the closed-loop solution of linear-quadratic games may generally be computed analytically, this is not always assured. This is the case of our model whose particular structure features both state variables entering the dynamic constraints and has algebraic properties that limit the tractability of its closed-loop solution.

We are, however, able to solve for the solution numerically. To make the analysis more salient and policy relevant, we take this constraint as an opportunity to calibrate the model based on real data and available empirical evidence. The rest of this subsection characterises some general features of the solution, and the next one provides the calibration of the closed-loop solution.

Proposition 4 *If the marginal provider disutility of waiting time is increasing, the optimal closed-loop supply rule for Hospital i is given by:*

$$S_i(w_i, w_j, t) = \frac{p - \theta(\omega_1 + \omega_3 w_i(t) + \omega_5 w_j(t))}{\gamma}, \tag{21}$$

where $\omega_3 < 0$ is required by the concavity of the value function and $\omega_5 \in \Omega$.

See Appendix A for the definition Ω and proof of Proposition 4.

In contrast to the case of constant marginal disutility of waiting time, a dynamic strategic interaction is present when the marginal disutility is increasing. This implies that the equilibrium supply of Hospital i at time t depends both on own waiting time, $w_i(t)$, and the waiting time at Hospital j, $w_j(t)$. Considering first the relationship between optimal treatment supply and own waiting time, $\omega_3 < 0$ implies that an increase in the waiting time of Hospital i increases the hospital's optimal treatment supply. The reason is that a longer waiting time increases the hospital's marginal disutility of waiting time and therefore increases the marginal benefit of supply.

The relationship between the treatment supply at Hospital i and the waiting time at Hospital j is determined be the sign of ω_5. Although it is not possible to unambiguously determine the sign of ω_5 analytically (see Appendix A), our calibration results provided in the next subsection show that ω_5 is negative for all the parameter configurations considered. If ω_5 is negative, then hospitals' supply decisions are characterised by strategic *substitutability*, $\partial S_i(w_i, w_j)/\partial w_j > 0$, for which we provide the following intuition. Consider a unilateral increase in supply by Hospital i. This leads to lower waiting times at Hospital i, which in turn shifts demand from Hospital j to Hospital i, causing a reduction in waiting times also at Hospital j. A lower waiting time at Hospital j reduces its marginal disutility of waiting time, and thus its marginal benefit of supply. Hospital j will therefore optimally respond by reducing its supply of treatments. In other words, a supply increase by Hospital i triggers a supply decrease by Hospital j.

The above described strategic interaction has important implications for the supply incentives of each hospital. Consider once more a unilateral increase in supply by Hospital i, which leads to an initial reduction in waiting time at this hospital. However, because of strategic substitutability, Hospital j will respond by reducing its supply, as explained above. The subsequent increase in waiting time at Hospital j shifts some demand towards Hospital i, thereby dampening the initial reduction in the waiting time caused by the supply increase of Hospital i. Thus, dynamic strategic substitutability lowers the marginal benefit of treatment supply, giving each hospital an incentive to reduce its own supply in order to 'free-ride' on the subsequent supply increase of the rival hospital.

In Appendix A, we also show that, if the initial waiting times are the same in both hospitals or if the average initial waiting time equals the steady-state waiting time, then waiting times, supply and demand in both segments of the market converge *monotonically* to the steady-state. In this case, if the condition $|\omega_3| > |\omega_5|$ holds, the equilibrium path to the steady-state is characterised by periods of increasing (decreasing) hospital activity and increasing (decreasing) waiting time, which

is in line with Siciliani (2006) in a monopoly setting. Notice that $|\omega_3| > |\omega_5|$ implies that the own waiting time effect on hospital activity is larger than the effect of the waiting time of the competing hospital, which is both intuitive and confirmed by our calibration exercise below.[8]

However, *non-monotonic* convergence may also arise. In Appendix A we show that, if the average initial waiting time is above (below) the steady-state waiting time, the hospital with the shortest (longest) initial waiting time might experience a non-monotonic convergence along the equilibrium path, with the waiting time first increasing (decreasing) before decreasing (increasing) towards the steady-state. One policy implication is that short-run provider performance on waiting times may not be representative of its long-run one.

4.3 Calibration

We calibrate the model using data from the English NHS on cataract surgery, which is a common non-emergency procedure across OECD countries (Siciliani et al., 2014). Our two key variables in the model are the steady-state waiting time and supply.

Waiting time data for cataract surgery is obtained from the Hospital Episode Statistics published by NHS Digital. Table 1 reports the mean and median waiting times (in days) for a cataract procedure provided either by NHS hospitals or the independent sector (private hospitals treating publicly-funded patients).[9]

Table 1. Evolution of waiting times for cataract procedures in the English NHS

Financial year	2011-12	2012-13	2013-14	2014-15	2015-16	2016-17
Mean waiting time	66	67	71	70	70	70
Median waiting time	59	60	63	62	59	58

Waiting times have remained relatively stable in recent years. They coincide with a period in which NHS England (the main regulator) did not specify performance standards for non-emergency care (The King's Fund, 2017). We interpret this as a regime where no significant penalties have been imposed on providers with longer waits. Within our model this corresponds to the special case when there is no hospital disutility of waiting time ($\alpha_1 = \alpha_2 = 0$). We therefore use the data in Table 1

[8] Additionally, it follows from equations (A.21) and (A.22) in Appendix A that $|\omega_3| > |\omega_5|$ is a sufficient (but not necessary) condition for convergence to be verified.

[9] Healthcare Resource Group (HRG) code BZ02Z, *Phacoemulsification Cataract Extraction and Lens Implant*, in the HRG4 classification system. In 2011-12, episodes were grouped according to the HRG3.5 version, and the corresponding HRG code is B13.

as a measure of waiting times in a steady-state with no penalties, which we denote by superscript s. To make the analysis consistent with the study of Propper et al. (2010), we employ the mean waiting time, measured in months, and focus on the financial year 2016-17, giving $w^s = 2.3$.

According to the National Schedule of Reference Costs from NHS Improvement, 234 NHS providers performed 286,596 cataract procedures in the same year.[10] This gives a monthly average of approximately 100 procedures per provider, so that $S^s = D^s = 100$.

On the *supply* side, two key parameters are the tariff for a cataract surgery (the DRG-type price) and the marginal cost of treatment. From the National Schedule of Reference Costs, the national tariff in 2016-17 for a cataract procedure was 731£. We set $p = 731$. Given that the first-order condition $S^s = p/\gamma$ has to hold (when $\alpha_1 = \alpha_2 = 0$), we recover the parameter related to the marginal cost of treatment, $\gamma = 7.31$.

On the *demand* side, the key parameters are the potential demand, the size of the competitive segment, the demand responsiveness, the gross valuation of treatment, and the value of the outside option. These parameters are less easy to obtain but we infer them in the following way. According to OECD (2018), 10.5% of the UK population was covered by private health insurance in 2015. We assume that patients with private insurance opt for private treatment and that publicly-funded cataract procedures account for about 90% of the market.[11] Given that the steady-state supply in each hospital is $S^s = 100$, potential demand across the two hospitals is then given by $N = 222$.

Sivey (2012) estimates a demand elasticity for cataract surgery across NHS providers that is approximately -0.1. The waiting time elasticity of demand evaluated at the steady-state values and $N = 222$ gives

$$\frac{\partial D_i(w_i(t), w_j(t))}{\partial w_i(t)} \frac{w^s}{D^s} = -\frac{N(2-\beta)}{2\tau} \frac{w^s}{D^s} = -\frac{222(2-\beta)}{2\tau} \frac{2.3}{100} = -0.1. \tag{22}$$

We do not know how large is the competitive segment. In order to account for patient heterogeneity, we conduct the analysis for three different values, $\beta = \{0.2, 0.5, 0.8\}$. We start by assuming $\beta = 0.2$, so that the competitive segment accounts for 20% of potential demand and is therefore relatively small, and then check how the results differ when it is 50% and 80% (relatively large). If $\beta = 0.2$, then, from (22), the demand elasticity implies that $\tau = 45.954$. Moreover, from the demand

[10] The National Schedule of Reference Costs is detailed according to the HRG4+ classification system, which presents a more thorough description of cataract episodes than the HRG4. Focusing on *Phacoemulsification Cataract Extraction and Lens Implant*, the HRGs considered are BZ34A, BZ34B, and BZ34C in HRG4+.

[11] This is an approximation since some patients without private insurance may also obtain private care if they pay out of pocket and some with private insurance may not seek private care if they face co-payments.

equation evaluated at the steady-state,

$$D^s = N\left[\frac{\beta}{2} + (1-\beta)\left(\frac{v-k-w^s}{\tau}\right)\right], \tag{23}$$

we can recover the difference between the gross valuation of treatment and the value of the outside option: $v - k = 22.4308$. If $\beta = 0.5$, then, from (22), we obtain $\tau = 38.295$ and, from (23), we obtain $v - k = 17.653$. If $\beta = 0.8$, then, from (22), we obtain $\tau = 30.636$ and, from (23), we obtain $v - k = 10.028$. We have thus recovered the demand-side parameters for $\beta = \{0.2, 0.5, 0.8\}$.

We adopt a discount factor of 0.95 per year and take each period t as one month. The monthly discount rate is therefore $\rho = 0.004$ (computed from $e^{-12\rho} = 0.95$).

In the steady-state, it takes one month for Hospital i to treat 100 patients. This implies that, if 10 additional patients are added to the list, the waiting time will increase by 0.1 months (about 3 days). More formally, from the dynamic constraint, $\Delta w^s \approx \theta\Delta(D^s - S^s)$, which gives $\theta = \frac{\Delta w^s}{\Delta(D^s - S^s)} = \frac{0.1}{10} = 0.01$ in the neighbourhood of the steady-state.

We are interested in understanding provider behaviour in the presence of penalties. We therefore need to identify plausible values for α_1 and α_2 under a penalty regime. In order to do this, we make use of the open-loop solution, for which we can derive a closed-form solution for the steady-state waiting time when $\alpha_2 > 0$ (see Appendix B). We denote variables in the open-loop steady-state by the superscript OL. Propper et al. (2010) find that the introduction of waiting time penalties in the English NHS in 2000-05 reduced the mean waiting time by 13 days (i.e., 0.43 months). Although this estimate refers to an earlier period, it provides us with a plausible order of magnitude if such penalties were re-introduced in 2016-17. We then use this figure to compute the difference between the steady-state waiting time in the model with no disutility of waiting time and the open-loop steady-state waiting time, which is given by

$$w^s - w^{OL} = 2.3 - \frac{\gamma\phi\tau}{(1-\beta)\gamma\phi N + 2\theta\tau^2\alpha_2}\left\{N\left[\frac{\beta}{2} + (1-\beta)\left(\frac{v-k}{\tau}\right)\right] - \frac{p}{\gamma} - \frac{2\theta\tau\alpha_1}{\gamma\phi}\right\} = 0.43. \tag{24}$$

Inserting the above described parameter values when $\beta = 0.2$, the solution to (24) has one degree of freedom and is given by

$$\alpha_2 = 30.5274 - 0.53486\alpha_1. \tag{25}$$

All α_1 and α_2 that satisfy (25) yield a reduction of 0.43 months in the open-loop steady-state

waiting time compared to the case with no disutility of waiting time. We consider three disutility structures: (i) linear disutility ($\alpha_2 = 0$), yielding $\alpha_1 = 57.0826$; (ii) quadratic disutility ($\alpha_1 = 0$), yielding $\alpha_2 = 30.5274$; and (iii) an intermediate case in which $\alpha_1 = \frac{57.0826}{2}$ and $\alpha_2 = \frac{30.5274}{2}$.

We insert all parameter values and solve the system (A.6)-(A.8) in Appendix A to yield ω_1, ω_3, and ω_5, which are plugged into (A.27) to obtain the closed-loop steady-state waiting time. With ω_1, ω_3, ω_5, and w^{CL}, we use (21) to retrieve the closed-loop steady-state supply. For the open-loop steady-state waiting time and supply, we insert the parameter values into equations (B.6) and (B.8) in Appendix B.

The same steps were then repeated for $\beta = 0.5$ and $\beta = 0.8$.

4.3.1 Linear versus convex waiting time disutility

The results generated by the above described calibration procedure are summarised in Table 2.

Table 2. Calibration results for a waiting time elasticity of demand of -0.1

β	α_1	α_2	w^{OL}	w^{CL}	S^{OL}	S^{CL}	ω_3	ω_5
0.2	57.0862	0	1.8700	1.8700	101.6620	101.6620	0	0
0.2	28.5431	15.2637	1.8700	1.8703	101.6620	101.6609	−164.6061	−8.3753
0.2	0	30.5274	1.8700	1.8705	101.6620	101.6600	−321.6537	−15.3715
0.5	39.2269	0	1.8700	1.8700	101.2464	101.2464	0	0
0.5	19.6143	10.4885	1.8700	1.8720	101.2464	101.2402	−119.1899	−19.2189
0.5	0	20.9769	1.8700	1.8734	101.2464	101.2353	−233.5920	−36.3039
0.8	13.5675	0	1.8700	1.8700	100.6232	100.6232	0	0
0.8	6.7837	3.6277	1.8700	1.8755	100.6232	100.6147	−52.3298	−20.3702
0.8	0	7.2553	1.8700	1.8795	100.6232	100.6077	−102.5480	−38.9404

Our calibration results confirm that, as explained in Section 4.2, the dynamic interaction introduced by increasing marginal disutility of waiting time leads to longer steady-state waiting times. As the waiting time disutility becomes more convex (i.e., more weight is placed on the quadratic term), the longer is the waiting time and the lower is supply in the closed-loop steady-state. The reason is simply that a more convex disutility function increases the magnitude of each hospital's supply response to changes in the waiting time, which reinforces each hospital's incentive to reduce supply in order to provoke a supply increase by the rival hospital, which in turn benefits the former

hospital in the form of a lower waiting time. This result has potentially interesting policy implications, as it suggests that *linear penalties* are more effective in reducing waiting times, all else equal. Notice also that the importance of the design of the penalty structure is larger for higher values of the competitive segment, β. This is intuitive, since the strategic substitutability relies on the existence of a competitive segment, wherein changes in the waiting time at one hospital affect demand faced by the rival hospital. Thus, a larger relative size of the competitive segment will magnify the effects of strategic substitutability.

Besides confirming that they coincide when $\alpha_2 = 0$, another key insight from Table 2 is that the difference in waiting times under the open- and closed-loop solutions is very small (less than 1%) when $\alpha_2 > 0$. This suggests that, even with non-linear penalties, the less computationally demanding open-loop solution offers a close approximation of the closed-loop one.

4.3.2 Higher waiting time elasticity of demand

One may worry that the results from Table 2 are due to the low demand elasticity. We therefore extend the analysis under the assumption that the waiting time elasticity is higher. We consider two additional cases. First, we assume that the elasticity is -0.2, twice as large, which is the highest that has been reported in studies for England (see Iversen and Siciliani (2011) for an overview). Second, we assume that the elasticity is -1. This is an upper bound. There is only one study from Australia which provides such a large estimate (Stavrunova and Yerokhin, 2011), and this is consistent with the features of the Australian health system where more than half of the population is treated privately. Tables 3 and 4 provide the results for waiting time elasticities of demand of -0.2 and -1, respectively, and they are derived following the steps detailed above. We see that an increase in the waiting time elasticity of demand reinforces the relative effectiveness of linear (as opposed to convex) waiting time penalties. Still, the quantitative difference between steady-state

waiting times in the open- and closed-loop solutions remains small.

Table 3. Calibration results for a waiting time elasticity of demand of −0.2

β	α_1	α_2	w^{OL}	w^{CL}	S^{OL}	S^{CL}	ω_3	ω_5
0.2	218.4948	0	1.8700	1.8700	103.3237	103.3237	0	0
0.2	109.2474	58.4211	1.8700	1.8703	103.3237	103.3212	−322.1649	−16.7563
0.2	0	116.8421	1.8700	1.8705	103.3237	103.3193	−629.5163	−31.3129
0.5	148.9097	0	1.8700	1.8700	102.4928	102.4928	0	0
0.5	74.4548	39.8154	1.8700	1.8722	102.4928	102.4791	−231.9189	−38.3288
0.5	0	79.6308	1.8700	1.8738	102.4928	102.4683	−454.4837	−72.3946
0.8	49.2070	0	1.8700	1.8700	101.2464	101.2464	0	0
0.8	24.6037	13.1571	1.8700	1.8762	101.2464	101.2272	−98.9201	−39.8422
0.8	0	26.3142	1.8700	1.8807	101.2464	101.2115	−193.7462	−76.1410

Table 4. Calibration results for a waiting time elasticity of demand of −1

β	α_1	α_2	w^{OL}	w^{CL}	S^{OL}	S^{CL}	ω_3	ω_5
0.2	5265.7273	0	1.8700	1.8700	116.6184	116.6184	0	0
0.2	2632.8636	1407.9485	1.8700	1.8703	116.6184	116.6049	−1581.6013	−83.7851
0.2	0	2815.8969	1.8700	1.8706	116.6184	116.5947	−3090.4383	−156.5665
0.5	3561.6215	0	1.8700	1.8700	112.4638	112.4638	0	0
0.5	1788.8108	952.3052	1.8700	1.8724	112.4638	112.3893	−1132.2781	−190.9560
0.5	0	1904.6104	1.8700	1.8741	112.4638	112.3307	−2218.7774	−360.6664
0.8	1126.6109	0	1.8700	1.8700	106.2319	106.2319	0	0
0.8	563.3055	301.2329	1.8700	1.8769	106.2319	106.1257	−469.2457	−194.4604
0.8	0	602.4657	1.8700	1.8818	106.2319	106.0397	−918.7196	−371.5998

4.3.3 Higher waiting times and hospital heterogeneity

In this section, we investigate whether our calibration results are robust to providers with longer waiting times. We simulate scenarios in which the baseline waiting time is 50% higher (i.e., $w^s = 3.45$). This is in line with Sivey (2012), who finds that the standard deviation of waiting times for cataract patients is about half of the mean wait.

Since long waiting times may be observed both at hospitals with high and low volumes, we

recalibrate the model with the higher baseline waiting time ($w^s = 3.45$) and set steady-state supply respectively at $S^s = 300$ (high volume) and $S^s = 50$ (low volume) in Tables 5 and 6. This is in line with HES data that reveal significant dispersion in hospital volumes even at the upper tail of the waiting times distribution across all procedures.[12]

By repeating the steps outlined at the beginning of Section 4.3, we obtain the results in Tables 5 and 6, which show that the effect of linear versus convex penalties is qualitatively similar to our previously derived results (in Tables 2-4). And again, the waiting times under the open-loop solution are very similar to those under closed-loop solution.

Table 5. Calibration results for larger hospitals and a higher baseline waiting time

β	α_1	α_2	w^{OL}	w^{CL}	S^{OL}	S^{CL}	ω_3	ω_5
0.2	79.3777	0	3.0200	3.0200	303.3237	303.3237	0	0
0.2	39.6889	13.1420	3.0200	3.0202	303.3237	303.3224	−209.9135	−10.6351
0.2	0	26.2840	3.0200	3.0203	303.3237	303.3214	−413.6013	−20.3550
0.5	54.9493	0	3.0200	3.0200	302.4928	302.4928	0	0
0.5	27.4747	9.0976	3.0200	3.0212	302.4928	302.4860	−152.6061	−24.3845
0.5	0	18.1951	3.0200	3.0222	302.4928	302.4803	−301.2599	−47.0010
0.8	19.7392	0	3.0200	3.0200	301.2464	301.2464	0	0
0.8	9.8696	3.2681	3.0200	3.0232	301.2464	301.2374	−68.5619	−26.0926
0.8	0	6.5362	3.0200	3.0258	301.2464	301.2294	−135.3564	−50.6886

[12] In 2016-17, the standard deviation of finished consultant episodes for hospitals above the 90^{th} percentile of the waiting times distribution was over three times larger than the mean.

Table 6. Calibration results for smaller hospitals and a higher baseline waiting time

β	α_1	α_2	w^{OL}	w^{CL}	S^{OL}	S^{CL}	ω_3	ω_5
0.2	13.2296	0	3.0200	3.0200	50.5539	50.5539	0	0
0.2	6.6148	2.1903	3.0200	3.0202	50.5539	50.5537	−34.9856	−1.7725
0.2	0	4.3807	3.0200	3.0203	50.5539	50.5536	−68.9335	−3.3925
0.5	9.1582	0	3.0200	3.0200	50.4155	50.4155	0	0
0.5	4.5791	1.5163	3.0200	3.0212	50.4155	50.4143	−25.4343	−4.0641
0.5	0	3.0325	3.0200	3.0222	50.4155	50.4134	−50.2100	−7.8335
0.8	3.2899	0	3.0200	3.0200	50.2077	50.2077	0	0
0.8	1.6449	0.5447	3.0200	3.0232	50.2077	50.2062	−11.4270	−4.3488
0.8	0	1.0894	3.0200	3.0258	50.2077	50.2049	−22.5594	−8.4481

4.3.4 Patient choice and waiting times

One of our main aims is to analyse the relationship between patient choice and waiting times. In line with the analysis in Section 4.1.1, we therefore conduct comparative statics with respect to the patient choice parameter τ. The fourth and fifth columns of Table 7 show the effects (on steady-state waiting times and supply) of a 10% reduction in τ, with all other parameters kept unchanged from our main calibration analysis, which implies that the results displayed in Table 2 serve as a reference point of comparison. In the last two columns of Table 7, we report the equivalent effects of a combined policy package, where a 10% reduction in τ is accompanied by a 10% increase in waiting time penalties (equivalent to the analysis in Section 4.1.2).

In qualitative terms, the effects of increased patient choice on steady-state waiting times and supply, as shown in the fourth and fifth columns of Table 7, confirm that the result stated in Proposition 2 generalises beyond the case of constant marginal disutility of waiting time. Regardless of the shape of the hospitals' waiting time disutility function, a reduction in τ leads to higher

steady-state waiting times.[13]

Table 7. Steady-state effects of policy reforms

			Patient choice[1]		Joint policy[2]	
β	α_1	α_2	$\Delta\%w^{CL}$	$\Delta\%S^{CL}$	$\Delta\%w^{CL}$	$\Delta\%S^{CL}$
0.2	57.0862	0	111.86	−0.15	109.98	0
0.2	28.5431	15.2637	102.24	0.61	99.67	0.82
0.2	0	30.5274	94.15	1.25	91.14	1.49
0.5	39.2269	0	86.27	−0.11	84.39	0
0.5	19.6143	10.4885	78.76	0.33	76.41	0.47
0.5	0	20.9769	72.52	0.70	69.87	0.85
0.8	13.5675	0	45.34	−0.05	43.45	0.01
0.8	6.7837	3.6277	41.25	0.07	39.23	0.12
0.8	0	7.2553	37.93	0.17	35.85	0.23

[1]10% reduction in τ

[2]10% reduction in τ and 10% increase in α_1 and/or α_2

However, even if more patient choice increases steady-state waiting times for all parameter configurations considered in Table 7, there is a clear pattern showing that this effect is quantitatively smaller if the waiting time disutility is more convex. The reason is that a reduction in τ has two counteracting effects on steady-state supply when $\alpha_2 > 0$. On the one hand, a lower τ makes treatment supply a less effective instrument to reduce waiting times, as previously explained, which gives each hospital an incentive to reduce their supply. On the other hand, a lower τ also increases demand (from the monopolistic segment), which—all else equal—leads to higher waiting times. If the disutility of waiting time is strictly convex (i.e., if $\alpha_2 > 0$), such increase in waiting time increases the marginal disutility of waiting time and therefore increases the marginal benefit of supply. In other words, with a strictly convex waiting time disutility function, the waiting time increase due to increased patient choice is partly dampened by the hospitals' incentives to increase supply in response to higher waiting times. Indeed, the fifth column in Table 7 shows that steady-state supply increases for the parameter configurations with $\alpha_2 > 0$.

[13]In the open-loop solution, for which a closed-form solution can be derived also in the case of increasing marginal waiting time disutility (see Appendix B), it is also easily shown that a reduction in τ leads to higher steady-state waiting times for all parameter values that are compatible with equilibrium existence.

This illustrates another aspect of the inherent conflict between waiting time penalties and patient choice policies, as previously discussed in Section 4.1.2. On the one hand, waiting time penalties are more effective in reducing waiting times when they are designed as linear penalties (as shown by Tables 2-6). On the other hand, the counterproductive effect of patient choice policies on waiting times is larger when penalties are linear (as shown by Table 7).

The last two columns of Table 7 show the effects of a policy package where the increased in patient choice is combined with a (10%) increases in waiting time penalties. Not surprisingly, this dampens the increase in waiting times induced by more patient choice. However, we see that the patient choice effect clearly dominates, implying that such a policy package leads to an overall increase in steady-state waiting times.

5 Patient Welfare

In this section, we briefly investigate the effect of choice policies on overall patient welfare. In the symmetric steady-state equilibrium, overall patient welfare, denoted by U, is given by the sum of patients' utility

$$U = 2N\beta \int_0^{\frac{1}{2}} (v - w^{CL} - \tau x)dx + 2N(1-\beta) \int_0^{x_M^{CL}} (v - k - w^{CL} - \tau x)dx, \tag{26}$$

and the effect of *lower* travelling costs is

$$\frac{\partial U}{\partial \tau} = -2D^{CL}\frac{\partial w^{CL}}{\partial \tau} - N\left[\frac{\beta}{4} + (1-\beta)(x_M^{CL})^2\right]. \tag{27}$$

The first term is negative and captures the utility loss due to longer waiting times endured by all patients. The second term is positive and captures the utility increase from lower travelling costs, which we interpret more broadly as simpler access to health care. Note that there is a third term since an increase in waiting times reduces demand at the margin, but given that the marginal patient is indifferent between treatment and no treatment, this has no effect on welfare. Therefore, the effect of choice policies on overall welfare is indeterminate and is positive only if the direct effect of easier access overcomes the utility loss from longer waiting times.

The above approach takes a utilitarian perspective. Suppose that a health authority or regulator (a Ministry of Health) is only interested in the *health* component of patient welfare (Gravelle and

Siciliani, 2008c). This approach has been sometimes referred as the extra-welfarist approach since it ignores non-health components which affect patient utility. Aggregate health patient benefit, denoted B, at the symmetric steady-state, is

$$B = 2N\beta \int_0^{\frac{1}{2}} (v - w^{CL})dx + 2N(1-\beta) \int_0^{x_M^{CL}} (v - k - w^{CL})dx, \tag{28}$$

and the effect of lower travelling costs is

$$\frac{\partial B^W}{\partial \tau} = -2D^{CL}\frac{\partial w^{CL}}{\partial \tau} + 2(v - k - w^{CL})\frac{\partial S^{CL}}{\partial \tau}. \tag{29}$$

If providers' penalties are linear in waiting times, patient choice policies increase waiting times for each patient and reduce supply with fewer patients gaining a health benefit from treatment, thus unambiguously reducing aggregate health benefits.

If providers' penalties are non-linear in waiting times, choice policies simultaneously increase waiting times and supply. Therefore, the effect on aggregate health benefit is in principle ambiguous. However, our calibration exercise shows that the supply effect is a second-order effect and that patient choice reduces aggregate health benefit also when $\alpha_2 > 0$. In more detail, Table 8 reports the percent change in B and U induced by a 10% reduction in τ, which is computed using the welfare values associated with Tables 2 and 7.

Table 8. Steady-state effects of a 10% reduction in τ on patient welfare

β	α_1	α_2	$\Delta\%w^{CL}$	$\Delta\%S^{CL}$	$\Delta\%U$	$\Delta\%B$
0.2	57.0862	0	111.86	−0.15	−9.81	−10.04
0.2	28.5431	15.2637	102.24	0.61	−8.10	−8.52
0.2	0	30.5274	94.15	1.25	−6.66	−7.23
0.5	39.2269	0	86.27	−0.11	−8.26	−9.22
0.5	19.6143	10.4885	78.76	0.33	−6.70	−8.08
0.5	0	20.9769	72.52	0.70	−5.38	−7.12
0.8	13.5675	0	45.34	−0.05	−1.71	−5.88
0.8	6.7837	3.6277	41.25	0.07	−0.69	−5.31
0.8	0	7.2553	37.93	0.17	0.15	−4.84

6 Robustness

In order to facilitate analytical tractability, our model has a linear-quadratic structure. One implication is that patient (dis)utility is assumed to be linear in waiting times, and travelling costs are linear in distance. Here we will briefly evaluate whether our main result—that more patient choice leads to increased waiting times—is robust to a relaxation of these assumptions. Unfortunately, it is only possible to perform these robustness checks in the context of the open-loop solution. However, our previous analysis has shown that the open-loop solution is a very close approximation of the closed-loop solution in our setting. The two solutions coincide if $\alpha_2 = 0$, and our calibration results show that the two solutions concepts produce quantitatively almost identical results if $\alpha_2 > 0$. More importantly, the positive relationship between patient choice and waiting times does not depend on the choice of the solution concept.

6.1 Non-linear patient disutility of waiting

Suppose that, in the patient utility functions (1) and (2), we replace w_i with a strictly increasing function $f(w_i)$. Total demand for Hospital i is then given by

$$D_i(w_i, w_j) = N\left\{\beta\left[\frac{1}{2} + \frac{f(w_j) - f(w_i)}{2\tau}\right] + (1-\beta)\left[\frac{v - k - f(w_i)}{\tau}\right]\right\}. \tag{30}$$

Let w^{OL} be the steady-state waiting time in the open-loop solution. In Appendix B, we show that this solution exists if $f(\cdot)$ is either concave or convex with a sufficiently low degree of convexity. Furthermore, we also show that, under the conditions of equilibrium existence, $\partial w^{OL}/\partial\tau < 0$. Thus:

Proposition 5 *Regardless of whether patient utility is concave or convex in waiting time, the steady-state waiting time in the open-loop solution, if it exists, is increasing in the degree of patient choice.*

This result is not surprising, given the intuition behind the previously derived positive relationship between patient choice and steady-state waiting times, which is related to the responsiveness of demand to changes in waiting times. As long as increased patient choice makes demand more responsive to changes in waiting times, it becomes more difficult for each hospital to curb waiting times by unilaterally increasing supply, which in turn leads to longer steady-state waiting times

at both hospitals. This mechanism only requires that patient utility decreases with longer waiting times; it does *not* depend on whether patient utility decreases at a faster or slower rate when waiting times increase. Thus, we conjecture that the result stated in Proposition 5 also hold in a closed-loop setting.

6.2 Non-linear patient disutility of travelling

Consider next the case where, in the patient utility functions (1) and (2), we replace $|x - z_i|$ with a strictly increasing function $g(|x - z_i|)$. This generalisation prevents a closed-form derivation of demand. However, by the Implicit Function Theorem, we can derive the demand responsiveness to waiting time as

$$\frac{\partial D_i(w_i(t), w_j(t))}{\partial w_i(t)} = -\frac{N}{\tau}\left(\frac{\beta}{\tau[g'(x_C(t)) + g'(1 - x_C(t))]} + \frac{(1-\beta)}{g'(x_M^i(t)))}\right) < 0 \tag{31}$$

and

$$\frac{\partial D_i(w_i(t), w_j(t))}{\partial w_j(t)} = \frac{N\beta}{\tau[g'(x_C(t)) + g'(1 - x_C(t))]} > 0. \tag{32}$$

Still using τ as an inverse measure of the degree of patient choice, we derive (see Appendix B) the following result:

Proposition 6 *(i) The steady-state waiting time in the open-loop solution is increasing in the degree of patient choice if the patient disutility of travelling is either concave or not strongly convex in travelling distance. (ii) In the case of constant marginal provider disutility of waiting time, the open-loop steady-state waiting time is increasing in the degree of patient choice if it exists.*

Thus, unless patient utility is strongly convex in travelling distance, our main result holds also in the case of non-linear patient disutility of travelling. And it always holds in the case of linear waiting time penalties, given that the open-loop solution exists. The general condition stated in Proposition 6 covers, for example, the empirical specification of Sivey (2012), who assumes that the utility of English cataract patients is a function of the natural log of travel time.

7 Concluding remarks

We have investigated whether increased competition through patient choice policies play a useful role in reducing waiting times and the extent to which such a role is altered in the presence

of penalties for providers with long waits. Our main results suggest, perhaps surprisingly, that increased patient choice leads to *higher* waiting times and that patient choice policies are therefore *counterproductive* in this respect. Furthermore, in the presence of waiting time penalties, we have shown that larger penalties make patient choice policies even more counterproductive.

The counterproductive effect of patient choice policies follows from the fact that increased patient choice makes each hospital's demand more responsive to changes in waiting times, which in turn makes it harder for each hospital to reduce waiting times by unilaterally increasing supply. In other words, increased patient choice makes each hospital's supply decision a less effective instrument to reduce waiting times, thereby leading to higher waiting times in equilibrium. This is a highly robust result which, in qualitative terms, does not depend on the choice of game-theoretic solution concept (closed-loop versus open-loop), nor on the design of the waiting time penalty structure (linear versus convex penalties). We have also shown that this result is robust to a fairly general patient utility specification. The result holds when patients' disutility of waiting is non-linear, and it also holds when patients' disutility of travelling is non-linear (though not too strongly convex).

While our main result might perhaps appear counterintuitive, it is consistent with a recent empirical study which shows that the introduction of patient choice policies in England since 2006 led to an increase in waiting times for hip and knee replacement (with one additional rival increasing waiting times by about 3-4%) and had no effect on waiting times for coronary bypass (Moscelli et al., 2019) or the proportion of patients waiting more than three months (Gaynor et al., 2013, footnote 16). Our results are also in line with an earlier study which showed that, for hip and knee replacement, hospitals facing more competition had higher readmissions (Moscelli et al., 2018a). Therefore, it appears that waiting times and quality worsened for some elective treatments, despite the improvements found for heart attack mortality and overall mortality (Cooper et al., 2011; Gaynor et al., 2013), and for hip fracture mortality (Moscelli et al., 2018b).

Our findings are instead in contrast with the older study by Propper, Burgess, and Gossage (2008), which found that competition in the late nineties reduced waiting times in England. However, this result was obtained in a different institutional setting than the one covered in our study. Patients had no or very limited choice. Hospitals prices were not fixed, but negotiated between health authorities and providers. Clinical quality measures were not available to the funders so that providers competed for funding from health authorities based on prices and waiting times.

As mentioned in the Introduction, countries like Denmark and Portugal have introduced patient choice policies. Although there is no evaluation study, in Denmark, waiting times reduced to some extent following the introduction of patient choice (and other) policies. These however can be explained by an expansion in capacity since the use of private providers to treat publicly-funded patients increased from 2 to 4% (Siciliani et al., 2013). Moreover, in Denmark, hospitals did not face any direct penalties for longer waiting times. In Portugal, preliminary evidence from 2016-2017 suggests that following the introduction of choice policies, median waiting time for first outpatient consultation increased in five specialties and reduced in two specialties (Simões et al., 2017). This suggests that choice policies did not have the intended effect of stimulating higher supply.

In summary, our model and analysis suggest that although policies based on provider penalties will have the intended effect in reducing waiting times, policies which stimulate patient choice and competition will not.

References

[1] Brekke, K. R., Cellini, R., Siciliani, L., and Straume, O. R. (2010). Competition and quality in health care markets: A differential-game approach. Journal of Health Economics, 29(4), 508–523.

[2] Brekke, K. R., Cellini, R., Siciliani, L., and Straume, O. R. (2012). Competition in regulated markets with sluggish beliefs about quality. Journal of Economics & Management Strategy, 21(1), 131–178.

[3] Brekke, K. R., Gravelle, H., Siciliani, L., and Straume, O. R. (2014). Patient choice, mobility and competition among health care providers. In R. Levaggi and M. Montefiori (Eds.), Health care provision and patient mobility. Developments in health economics and public policy (Vol. 12).

[4] Brekke, K. R., Siciliani, L., and Straume, O. R. (2008). Competition and waiting times in hospital markets. Journal of Public Economics, 92 (7), 1607–1628.

[5] Chen, Y., Meinecke, J., and Sivey, P. (2016) A theory of waiting time reporting and quality signaling. Health Economics, 25(11), 1355–1371.

[6] Cooper, Z., Gibbons, S., Jones, S., McGuire, A. (2011). Does hospital competition save lives? evidence from the English NHS patient choice reforms. The Economic Journal, 121(554), 228–260.

[7] Dockner, E. J., Jorgensen, S., Long, N. V., and Sorger, G. (2000). Differential games in economics and management science. Cambridge: Cambridge University Press.

[8] Gaynor, M., Moreno-Serra, R., Propper, C. (2013). Death by market power: Reform, competition, and patient outcomes in the National Health Service. American Economic Journal: Economic Policy, 5, 134–166.

[9] Gravelle, H., and Siciliani, L. (2008a). Is waiting-time prioritisation welfare improving? Health Economics, 17(2), 167–184.

[10] Gravelle, H., and Siciliani, L. (2008b). Optimal quality, waits and charges in health insurance. Journal of Health Economics, 27(3), 663–674.

[11] Gravelle, H., and Siciliani, L. (2008c). Ramsey waits: Allocating public health service resources when there is rationing by waiting. Journal of Health Economics, 27(5), 1143–1154.

[12] Hoel, M., and Sæther, E. M. (2003). Public health care with waiting time: The role of supplementary private health care. Journal of Health Economics, 22(4), 599–616.

[13] Iversen, T. (1993). A theory of hospital waiting lists. Journal of Health Economics, 12(1), 55–71.

[14] Iversen, T. (1997). The effect of a private sector on the waiting time in a national health service. Journal of Health Economics, 16(4), 381–396.

[15] Iversen, T., and Siciliani, L. (2011). Non-price rationing and waiting times. In S. Glied and P. C. Smith (Eds.), The Oxford handbook of health economics. Oxford: Oxford University Press.

[16] Lindsay, C. M., and Feigenbaum, B. (1984). Rationing by waiting lists. The American Economic Review, 74(3), 404–417.

[17] Marchand, M., and Schroyen, F. (2005). Can a mixed health care system be desirable on equity grounds? The Scandinavian Journal of Economics, 107(1), 1–23.

[18] Martin, S., and Smith, P. C. (1999). Rationing by waiting lists: An empirical investigation. Journal of Public Economics, 71(1), 141–164.

[19] Moscelli, G., Gravelle, H., Siciliani, L. (2019). The Effect of Hospital Choice and Competition on Waiting Times and Inequalities in Waiting Times, mimeo, University of York.

[20] Moscelli, G., Gravelle, H., Siciliani, L. (2018a). Effects of Market Structure and Patient Choice on Hospital Quality for Planned Patients (No. 1118). School of Economics, University of Surrey.

[21] Moscelli, G., Gravelle, H., Siciliani, L., Santos, L., (2018b) Heterogeneous effects of patient choice and hospital competition on mortality, Social Science and Medicine, 216, 50-58.

[22] OECD. (2017). Health at a glance 2017: OECD indicators. Paris: OECD Publishing.

[23] OECD. (2018). OECD health statistics 2018 [Data set]. Retrieved from https://stats.oecd.org/index.aspx?DataSetCode=HEALTH STAT

[24] Propper, C., Burgess, S., Gossage, D. (2008). Competition and quality: evidence from the NHS internal market 1991–1999. Economic Journal, 118(525), 138–170.

[25] Propper, C., Sutton, M., Whitnall, C., and Windmeijer, F. (2008). Did targets and terror reduce waiting times in England for hospital care? The B.E. Journal of Economic Analysis & Policy, 8 (2).

[26] Propper, C., Sutton, M., Whitnall, C., and Windmeijer, F. (2010). Incentives and targets in hospital care: Evidence from a natural experiment. Journal of Public Economics, 94(3), 318–335.

[27] Siciliani, L. (2006). A dynamic model of supply of elective surgery in the presence of waiting times and waiting lists. Journal of Health Economics, 25(5), 891–907.

[28] Siciliani, L., Chalkley, M., and Gravelle, H. (2017). Policies towards hospital and GP competition in five European countries. Health Policy, 121(2), 103–110.

[29] Siciliani, L., Moran, V., and Borowitz, M. (Eds.). (2013). Waiting time policies in the health sector: What works? Paris: OECD Health Policy Studies, OECD Publishing.

[30] Siciliani, L., Moran, V., and Borowitz, M. (2014). Measuring and comparing health care waiting times in OECD countries. Health Policy, 118(3), 292–303.

[31] Siciliani, L., Straume, O. R., and Cellini, R. (2013). Quality competition with motivated providers and sluggish demand. Journal of Economic Dynamics and Control , 37 (10), 2041–2061.

[32] Simões, J., Augusto, G.F., Fronteira, I. (2017). Introduction of freedom of choice for hospital outpatient care in Portugal: Implications and results of the 2016 reform, Health Policy, 121, 1203–1207.

[33] Sivey, P. (2012). The effect of waiting time and distance on hospital choice for English cataract patients. Health Economics, 21(4), 444–456.

[34] Stavrunova, O., and Yerokhin, O. (2011). An equilibrium model of waiting times for elective surgery in NSW public hospitals. Economic Record, 87(278), 384–398.

[35] The King's Fund. (2017). What is happening to waiting times in the NHS? [Article]. Retrieved from https://www.kingsfund.org.uk/publications/articles/nhs-waiting-times

Appendix A: Closed-loop solution

Given the linear-quadratic structure of our model, we conjecture that the value function for Hospital i takes the form:

$$V^i(w_i, w_j) = \omega_0 + \omega_1 w_i + \omega_2 w_j + \frac{\omega_3}{2} w_i^2 + \frac{\omega_4}{2} w_j^2 + \omega_5 w_i w_j. \tag{A.1}$$

This value function must satisfy the Hamilton-Jacobi-Bellman (HJB) equation for Hospital i, which is given by[14]

$$\rho V^i(w_i, w_j) = \max \left\{ T + pS_i - \frac{\gamma}{2} S_i^2 - \alpha_1 w_i - \frac{\alpha_2}{2} w_i^2 + \theta \frac{\partial V^i}{\partial w_i}(D_i - S_i) + \theta \frac{\partial V^i}{\partial w_j}(D_j - S_j) \right\}. \tag{A.2}$$

Maximisation of the right-hand side of the HJB equations yields:

$$S_i(w_i, w_j) = \frac{p - \theta(\omega_1 + \omega_3 w_i + \omega_5 w_j)}{\gamma}. \tag{A.3}$$

Substituting Hospital i's supply rule and the analogous supply rule for Hospital j into the HJB

[14] To save notation, we omit the time index t in all subsequent expressions.

equation, together with (7)-(8), we obtain:

$$\begin{aligned}\rho V^i(w_i,w_j) &= T + p\left[\frac{p-\theta(\omega_1+\omega_3 w_i+\omega_5 w_j)}{\gamma}\right] - \frac{\gamma}{2}\left[\frac{p-\theta(\omega_1+\omega_3 w_i+\omega_5 w_j)}{\gamma}\right]^2 - \alpha_1 w_i - \frac{\alpha}{2}w_i^2 \\ &+\theta(\omega_1+\omega_3 w_i+\omega_5 w_j)\left[\beta\left(\frac{1}{2}+\frac{w_j-w_i}{2\tau}\right)N + (1-\beta)\left(\frac{v-k-w_i}{\tau}\right)N - \frac{p-\theta(\omega_1+\omega_3 w_i+\omega_5 w_j)}{\gamma}\right] \\ &+\theta(\omega_2+\omega_4 w_j+\omega_5 w_i)\left[\beta\left(\frac{1}{2}+\frac{w_i-w_j}{2\tau}\right)N + (1-\beta)\left(\frac{v-k-w_j}{\tau}\right)N - \frac{p-\theta(\omega_1+\omega_3 w_j+\omega_5 w_i)}{\gamma}\right],\end{aligned} \tag{A.4}$$

which can be rewritten as

$$\begin{aligned}&\left\{T + \frac{p^2}{2\gamma} + \sigma(\omega_1+\omega_2) + \frac{\theta^2}{2\gamma}\omega_1^2 + \frac{\theta^2}{\gamma}\omega_1\omega_2 - \rho\omega_0\right\} \\ &+ w_i\left\{-\left[\rho + \frac{\theta(2-\beta)N}{2\tau}\right]\omega_1 + \frac{\theta\beta N}{2\tau}\omega_2 + \sigma(\omega_3+\omega_5) + \frac{\theta^2}{\gamma}\omega_1\omega_3 + \frac{\theta^2}{\gamma}\omega_1\omega_5 + \frac{\theta^2}{\gamma}\omega_2\omega_5 - \alpha_1\right\} \\ &+ w_j\left\{\frac{\theta\beta N}{2\tau}\omega_1 - \left[\rho + \frac{\theta(2-\beta)N}{2\tau}\right]\omega_2 + \sigma(\omega_4+\omega_5) + \frac{\theta^2}{\gamma}\omega_1\omega_4 + \frac{\theta^2}{\gamma}\omega_1\omega_5 + \frac{\theta^2}{\gamma}\omega_2\omega_3\right\} \\ &+ w_i^2\left\{-\left[\frac{\rho}{2} + \frac{\theta(2-\beta)N}{2\tau}\right]\omega_3 + \frac{\theta^2}{2\gamma}\omega_3^2 + \frac{\theta\beta N}{2\tau}\omega_5 + \frac{\theta^2}{\gamma}\omega_5^2 - \frac{\alpha_2}{2}\right\} \\ &+ w_j^2\left\{-\left[\frac{\rho}{2} + \frac{\theta(2-\beta)N}{2\tau}\right]\omega_4 + \frac{\theta^2}{\gamma}\omega_3\omega_4 + \frac{\theta\beta N}{2\tau}\omega_5 + \frac{\theta^2}{2\gamma}\omega_5^2\right\} \\ &+ w_i w_j\left\{\frac{\theta\beta N}{2\tau}(\omega_3+\omega_4) - \left[\rho + \frac{\theta(2-\beta)N}{\tau}\right]\omega_5 + \frac{2\theta^2}{\gamma}\omega_3\omega_5 + \frac{\theta^2}{\gamma}\omega_4\omega_5\right\} = 0,\end{aligned} \tag{A.5}$$

where $\sigma = \frac{\theta\beta N}{2} + \theta(1-\beta)\left(\frac{v-k}{\tau}\right)N - \frac{\theta}{\gamma}p$.

For the equality to hold, the terms in curly brackets in the above equation have to be equal to zero. Since the last three terms depend only on ω_3, ω_4, and ω_5, we focus on the system of three equations and three unknowns given by:

$$-\left[\frac{\rho}{2} + \frac{\theta(2-\beta)N}{2\tau}\right]\omega_3 + \frac{\theta^2}{2\gamma}\omega_3^2 + \frac{\theta\beta N}{2\tau}\omega_5 + \frac{\theta^2}{\gamma}\omega_5^2 - \frac{\alpha_2}{2} = 0, \tag{A.6}$$

$$-\left[\frac{\rho}{2} + \frac{\theta(2-\beta)N}{2\tau}\right]\omega_4 + \frac{\theta^2}{\gamma}\omega_3\omega_4 + \frac{\theta\beta N}{2\tau}\omega_5 + \frac{\theta^2}{2\gamma}\omega_5^2 = 0, \tag{A.7}$$

$$\frac{\theta\beta N}{2\tau}(\omega_3+\omega_4) - \left[\rho + \frac{\theta(2-\beta)N}{\tau}\right]\omega_5 + \frac{2\theta^2}{\gamma}\omega_3\omega_5 + \frac{\theta^2}{\gamma}\omega_4\omega_5 = 0. \tag{A.8}$$

A.1 Constant marginal waiting time disutility

Consider first the closed-loop solution under constant marginal waiting time disutility. When $\alpha_2 = 0$, the system of equations (A.6)-(A.8) has a single candidate solution for which the value function is not convex with respect to w_i. The remaining five candidates have $\omega_3 > 0$ and cannot therefore constitute a solution the hospital's maximisation problem. The solution that yields a linear—hence, concave—value function with respect to w_i is $\omega_3 = \omega_4 = \omega_5 = 0$. This linearity of the value function with respect to waiting times is not surprising given the linear structure of the game when $\alpha_2 = 0$. Setting $\omega_3 = \omega_5 = 0$ in (A.3), Hospital i's optimal supply rule becomes

$$S_i(w_i, w_j) = \frac{p - \theta\omega_1}{\gamma}, \tag{A.9}$$

implying that supply is constant, and thus independent of waiting times, in each t.

With $\omega_3 = \omega_4 = \omega_5 = 0$, (A.5) simplifies to:

$$\begin{aligned}\left\{T + \frac{p^2}{2\gamma} + \sigma(\omega_1 + \omega_2) + \frac{\theta^2}{2\gamma}\omega_1^2 + \frac{\theta^2}{\gamma}\omega_1\omega_2 - \rho\omega_0\right\} \\ + w_i\left\{-\left[\rho + \frac{\theta(2-\beta)N}{2\tau}\right]\omega_1 + \frac{\theta\beta N}{2\tau}\omega_2 - \alpha_1\right\} \\ + w_j\left\{\frac{\theta\beta N}{2\tau}\omega_1 - \left[\rho + \frac{\theta(2-\beta)N}{2\tau}\right]\omega_2\right\} = 0. \end{aligned} \tag{A.10}$$

Since the last two terms depend only on ω_1 and ω_2, we focus on the 2×2 system and solve for ω_1. The solution is given by

$$\omega_1 = -\frac{\tau\alpha_1\left[2\rho\tau + \theta(2-\beta)N\right]}{2\left[\rho\tau + \theta(1-\beta)N\right]\left[\rho\tau + \theta N\right]} = -\frac{2\tau\alpha_1}{\phi}. \tag{A.11}$$

Inserting the expression for ω_1 into the optimal supply rule for hospitals i and j yields $S_i = S_j = S^{CL}$ as given by (13) in Section 4. Using this result, the closed-loop steady-state waiting time is derived from the equations of motion (11)-(12), with $\dot{w}_i(t) = \dot{w}_j(t) = 0$. Simple algebra shows that $w_i = w_j = w^{CL}$ as given by (16) in Section 4.

From (16), the steady-state waiting time is positive if and only if $p \leq \overline{p}$, given by

$$\overline{p} = \gamma N\left[\frac{\beta}{2} + (1-\beta)\left(\frac{v-k}{\tau}\right)\right] - \frac{2\theta\tau\alpha_1}{\phi}. \tag{A.12}$$

Furthermore, in order to have a partially covered monopolistic segment in the steady-state, the following condition must be satisfied:

$$0 < \frac{v - k - w^{CL}}{\tau} < \frac{1}{2}. \tag{A.13}$$

The lower bound is satisfied if $p > \underline{p}$, given by

$$\underline{p} = \frac{\beta\gamma N}{2} - \frac{2\theta\tau\alpha_1}{\phi}, \tag{A.14}$$

whereas the upper bound is satisfied if $p < \frac{\gamma N}{2} - \frac{2\theta\tau\alpha_1}{\phi}$, which always holds if $p < \overline{p}$. Thus, an interior-solution equilibrium (i.e., positive waiting times with a partially covered monopolistic segment) requires $p \in \mathcal{P} = (\max\{0, \underline{p}\}, \overline{p})$. Since $\overline{p} > \underline{p}$ for $\beta \in (0, 1)$, $\mathcal{P}$ is non-empty if $\overline{p} > 0$, which requires that γ is sufficiently large.

A.2 Increasing marginal disutility of waiting time

When $\alpha_2 > 0$, the solution to (A.6)-(A.8) depends on the root of a sixth degree polynomial, precluding the computation of an analytical solution. Assume, for now, that a solution exists and that it is such that (21) in Proposition 4 constitutes a Markov Perfect Nash Equilibrium.

From (A.6), two candidate solutions for ω_3 (as functions of ω_5) ensue:

$$\omega_3 = \frac{\gamma}{\theta^2}\left\{\left[\frac{\rho}{2} + \frac{\theta(2-\beta)N}{2\tau}\right] \pm \sqrt{\left[\frac{\rho}{2} + \frac{\theta(2-\beta)N}{2\tau}\right]^2 - \frac{2\theta^2}{\gamma}\left[\frac{\theta^2}{\gamma}\omega_5^2 + \frac{\theta\beta N}{2\tau}\omega_5 - \frac{\alpha_2}{2}\right]}\right\}. \tag{A.15}$$

A solution to Hospital i's maximisation problem is attained if the value function is concave with respect to w_i, which requires $\omega_3 < 0$. The greater root (unambiguously positive) is therefore ruled out. For the smaller root to be negative, the second term under the square-root must be positive, which is true for $\omega_5 \in (\underline{\omega_5}, \overline{\omega_5})$, with

$$\underline{\omega_5} = -\frac{\gamma}{2\theta^2}\left[\frac{\theta\beta N}{2\tau} + \sqrt{\left(\frac{\theta\beta N}{2\tau}\right)^2 + \frac{2\theta^2\alpha_2}{\gamma}}\right] < 0, \tag{A.16}$$

$$\overline{\omega_5} = -\frac{\gamma}{2\theta^2}\left[\frac{\theta\beta N}{2\tau} - \sqrt{\left(\frac{\theta\beta N}{2\tau}\right)^2 + \frac{2\theta^2\alpha_2}{\gamma}}\right] > 0. \tag{A.17}$$

Additionally, in order for (21) to be a Markov Perfect Nash Equilibrium, the value function must be bounded from above. A necessary and sufficient condition for this requirement to hold is that waiting times converge in equilibrium. Inserting (7), (8), (21), and the analogous supply rule for Hospital j into (11)-(12) yields the following system of differential equations:

$$\frac{\dot{w}_i}{\theta} = \left[-\frac{(2-\beta)N}{2\tau} + \frac{\theta}{\gamma}\omega_3\right] w_i + \left[\frac{\beta N}{2\tau} + \frac{\theta}{\gamma}\omega_5\right] w_j + N\left[\frac{\beta}{2} + (1-\beta)\left(\frac{v-k}{\tau}\right)\right] - \left(\frac{p-\theta\omega_1}{\gamma}\right), \tag{A.18}$$

$$\frac{\dot{w}_j}{\theta} = \left[\frac{\beta N}{2\tau} + \frac{\theta}{\gamma}\omega_5\right] w_i + \left[-\frac{(2-\beta)N}{2\tau} + \frac{\theta}{\gamma}\omega_3\right] w_j + N\left[\frac{\beta}{2} + (1-\beta)\left(\frac{v-k}{\tau}\right)\right] - \left(\frac{p-\theta\omega_1}{\gamma}\right). \tag{A.19}$$

The Jacobian of (A.18)-(A.19) is

$$J^{CL} = \theta \begin{bmatrix} -\frac{(2-\beta)N}{2\tau} + \frac{\theta}{\gamma}\omega_3 & \frac{\beta N}{2\tau} + \frac{\theta}{\gamma}\omega_5 \\ \frac{\beta N}{2\tau} + \frac{\theta}{\gamma}\omega_5 & -\frac{(2-\beta)N}{2\tau} + \frac{\theta}{\gamma}\omega_3 \end{bmatrix} \tag{A.20}$$

and its eigenvalues are

$$s_1 = \theta\left[-\frac{N}{\tau} + \frac{\theta}{\gamma}(\omega_3 - \omega_5)\right], \tag{A.21}$$

$$s_2 = \theta\left[-\frac{(1-\beta)N}{\tau} + \frac{\theta}{\gamma}(\omega_3 + \omega_5)\right]. \tag{A.22}$$

A sufficient condition for waiting times to converge is that both eigenvalues are negative. Then, $s_1 < 0$ if $\omega_5 > -\frac{\gamma N}{\theta\tau} + \omega_3$ and $s_2 < 0$ if $\omega_5 < \frac{\gamma(1-\beta)N}{\theta\tau} - \omega_3$.

Using the expression for ω_3 as a function of ω_5, (A.15), the necessary condition $s_1 < 0 \wedge s_2 < 0 \wedge \omega_3 < 0$ is satisfied if $\omega_5 \in \Omega = \left(\max\{\underline{\omega_5}, \underline{\omega_5}'\}, \min\{\overline{\omega_5}, \overline{\omega_5}'\}\right)$, where

$$\underline{\omega_5}' = \frac{\gamma}{6\theta^2}\left[\rho - \frac{2\theta\beta N}{\tau} - \sqrt{\left(\rho - \frac{2\theta\beta N}{\tau}\right)^2 + \frac{12\theta^2}{\gamma}\left[\frac{\gamma N}{\theta\tau}\left(\rho + \frac{\theta(1-\beta)N}{\tau}\right) + \alpha_2\right]}\right] < 0, \tag{A.23}$$

$$\overline{\omega_5}' = \frac{\gamma}{6\theta^2}\left[-\left(\rho + \frac{2\theta\beta N}{\tau}\right) + \sqrt{\left(\rho + \frac{2\theta\beta N}{\tau}\right)^2 + \frac{12\theta^2}{\gamma}\left[\frac{\gamma(1-\beta)N}{\theta\tau}\left(\rho + \frac{\theta N}{\tau}\right) + \alpha_2\right]}\right] > 0. \tag{A.24}$$

Thus, provided that a solution to (A.6)-(A.8) exists, it constitutes a Markov Perfect Nash Equilibrium (or closed-loop equilibrium) if $\omega_5 \in \Omega$. Finally, an equilibrium with $\omega_5 = 0$ is ruled out by inspection of (A.6)-(A.8).

The eigenvalues given by (A.21)-(A.22) also provide confirmation that the supply rules derived

in the previous subsection, under constant marginal disutility of waiting time, constitute a Markov Perfect Nash Equilibrium. It is straightforward to see from (A.21) and (A.22) that $s_1 < 0$ and $s_2 < 0$ when $\omega_3 = \omega_5 = 0$.

A.2.1 Transitional dynamics

In order to analyse the convergence to the steady-state in the closed-loop solution, we turn to its open-loop representation. That is, we derive time-profiles of the waiting time, supply, and demand from the optimal closed-loop supply rule. Let the superscript CL denote the closed-loop steady-state. The eigenvalues governing the system of differential equations (A.18)-(A.19), s_1 and s_2, are respectively associated with the eigenvectors $\nu_1 = c_1 \begin{bmatrix} 1 & -1 \end{bmatrix}^T$ and $\nu_2 = c_2 \begin{bmatrix} 1 & 1 \end{bmatrix}^T$, with $c_1, c_2 \in \mathbb{R}$. Setting $c_1 = c_2 = 1$, the solution of the system of differential equations (A.18)-(A.19) takes the form:

$$w_i(t) = C_1 e^{s_1 t} + C_2 e^{s_2 t} + w^{CL}, \tag{A.25}$$

$$w_j(t) = -C_1 e^{s_1 t} + C_2 e^{s_2 t} + w^{CL}, \tag{A.26}$$

where C_1 and C_2 are arbitrary constants. The closed-loop steady-state waiting time w^{CL} is retrieved by setting $\dot{w}_i = \dot{w}_j = 0$ in (A.18)-(A.19) and solving for w_i and w_j. This yields

$$w^{CL} = \frac{N\left[\frac{\beta}{2} + (1-\beta)\left(\frac{v-k}{\tau}\right)\right] - \left(\frac{p-\theta\omega_1}{\gamma}\right)}{\frac{(1-\beta)N}{\tau} - \frac{\theta}{\gamma}(\omega_3 + \omega_5)}. \tag{A.27}$$

Inserting the initial conditions $w_i(0) = w_{0i}$ and $w_j(0) = w_{0j}$ into (A.25)-(A.26) and solving for C_1 and C_2 gives $C_1 = \frac{w_{0i} - w_{0j}}{2}$ and $C_2 = \frac{w_{0i} + w_{0j}}{2} - w^{CL}$. Then, waiting times at Hospital i converge to the steady-state according to:

$$w_i(t) = \left(\frac{w_{0i} - w_{0j}}{2}\right) e^{s_1 t} + \left(\frac{w_{0i} + w_{0j}}{2} - w^{CL}\right) e^{s_2 t} + w^{CL}. \tag{A.28}$$

Consider, now, the dynamics of supply and demand. Inserting (A.28) and the analogous equation for $w_j(t)$ into (21) yields:

$$S_i(t) = \frac{\theta}{\gamma}\left[(\omega_5 - \omega_3)\left(\frac{w_{0i} - w_{0j}}{2}\right) e^{s_1 t} - (\omega_3 + \omega_5)\left(\frac{w_{0i} + w_{0j}}{2} - w^{CL}\right) e^{s_2 t}\right] + \frac{p - \theta[\omega_1 + (\omega_3 + \omega_5) w^{CL}]}{\gamma}. \tag{A.29}$$

Using (7), (A.28), and the analogous equation for $w_j(t)$, the dynamics of demand faced by Hospital i in the competitive and monopolistic segments are respectively given by

$$D_C^i(t) = \beta N \left[\frac{1}{2} + \left(\frac{w_{0j} - w_{0i}}{2\tau}\right) e^{s_1 t}\right] \tag{A.30}$$

and

$$D_M^i(t) = \frac{(1-\beta)N}{\tau} \left[v - k - w^{CL} + \left(\frac{w_{0j} - w_{0i}}{2}\right) e^{s_1 t} + \left(w^{CL} - \frac{w_{0i} + w_{0j}}{2}\right) e^{s_2 t}\right]. \tag{A.31}$$

If $w_{0i} = w_{0j}$, it follows from equations (A.28)-(A.31) that the dynamics of waiting times, supply, and demand are uniquely governed by s_2, and convergence is thus monotonic. By the same token, if the initial waiting times differ but their average equals the steady-state waiting time w^{CL}, dynamics are uniquely governed by s_1, and convergence is monotonic as well in this case. Note, additionally, that demand in the competitive segment always converges monotonically to $\beta N/2$.

For the transitional dynamics in the closed-loop solution under constant marginal disutility of waiting time, simply set $\omega_3 = \omega_5 = 0$ in equations (A.28)-(A.31). Constant hospital activity over time for $\alpha_2 = 0$ is then confirmed by (A.29).

A.2.2 Non-monotonic convergence

Equations (A.28)-(A.31) show that convergence to the steady-state depends on two, possible opposing, forces. It depends on whether a hospital's initial waiting time is longer than that of the rival, and whether the average initial waiting time in the market differs from the steady-state waiting time. When these two conditions hold, the possibility of non-monotonic convergence arises. To see why non-monotonic convergence might occur, consider the equilibrium dynamics of waiting times described in (A.28). If the average initial waiting time is above (below) the steady-state, the first two terms have opposite signs for the hospital with the shorter (longer) waiting time. In both cases, whether or not non-monotonic convergence emerges depends on the relative size and speed of convergence (to zero) of each of those terms.

Differentiating (A.28) with respect to time and equating to zero yields a single critical point given by

$$t^* = \left(\frac{1}{s_1 - s_2}\right) ln \left[-\frac{s_2}{s_1}\left(\frac{w_{0i} + w_{0j} - 2w^{CL}}{w_{0i} - w_{0j}}\right)\right], \tag{A.32}$$

where s_1 and s_2 are given by (A.21) and (A.22), respectively. Convergence is non-monotonic for Hospital i if and only if $t^* \in \mathbb{R}^+$. With $s_1, s_2 < 0$, the first factor in (A.32) is negative if $|s_1| > |s_2|$. Thus, $t^* \in \mathbb{R}^+$ if and only if the second factor in (A.32) is defined and is negative, which requires that the expression in the square brackets lies between 0 and 1. It is possible to derive some easily interpretable conditions for this expression to be positive. Since $-\frac{s_2}{s_1} < 0$, we must have $\frac{w_{0i}+w_{0j}-2w^{CL}}{w_{0i}-w_{0j}} < 0$. Two cases then arise:

1. If the average initial waiting time is above the steady-state waiting time, the numerator is positive, and $\frac{w_{0i}+w_{0j}-2w^{CL}}{w_{0i}-w_{0j}}$ is negative only if Hospital i has an initial waiting time below that of Hospital j.

2. If the average initial waiting time is below the steady-state waiting time, the numerator is negative, and $\frac{w_{0i}+w_{0j}-2w^{CL}}{w_{0i}-w_{0j}}$ is negative only if Hospital i has an initial waiting time above that of Hospital j.

Therefore, when the average initial waiting time is above (below) the steady-state waiting time, it is the hospital with the shortest (longest) waiting time that exhibits non-monotonic convergence, provided that $|s_1| > |s_2|$ and $-\frac{s_2}{s_1}\left(\frac{w_{0i}+w_{0j}-2w^{CL}}{w_{0i}-w_{0j}}\right) \in (0,1)$.

To conclude the proof, we consider the shape of (A.28). Evaluating its second-order derivative with respect to t at t^* yields the following results:

1. If $(w_{0i} + w_{0j} > 2w^{CL}) \wedge (w_{0i} < w_{0j})$, then $w_i''(t^*) < 0$ simplifies to:

$$\left(\frac{s_1}{s_2}\right)^2 e^{(s_1-s_2)t^*}(w_{0i} - w_{0j}) < -(w_{0i} + w_{0j} - 2w^{CL}). \tag{A.33}$$

   Diving both sides by $(w_{0i}-w_{0j})$ reverses the inequality sign. Then, using (A.32), the inequality becomes $\frac{s_1}{s_2} > 1$, which is true.

2. If $(w_{0i} + w_{0j} < 2w^{CL}) \wedge (w_{0i} > w_{0j})$, then $w_i''(t^*) > 0$ simplifies to:

$$\left(\frac{s_1}{s_2}\right)^2 e^{(s_1-s_2)t^*}(w_{0i} - w_{0j}) > -(w_{0i} + w_{0j} - 2w^{CL}). \tag{A.34}$$

   Diving both sides by $(w_{0i} - w_{0j})$ does not reverse the inequality sign. Then, using (A.32), the inequality becomes $\frac{s_1}{s_2} > 1$, which is true.

Hence, if $|s_1| > |s_2|$, $-\frac{s_2}{s_1}\left(\frac{w_{0i}+w_{0j}-2w^{CL}}{w_{0i}-w_{0j}}\right) \in (0,1)$, and the average initial waiting time is above (below) the steady-state waiting time, the dynamics of the waiting time at the hospital with the shortest (longest) initial wait has a unique maximum (minimum). This implies that the waiting time at the hospital with the shortest (longest) initial wait first increases (decreases) before decreasing (increasing) towards the steady-state.

Appendix B: The open-loop solution

Let $\mu_i(t)$ and $\lambda_i(t)$ denote, respectively, the costate variables associated with the dynamic equations of $w_i(t)$ and $w_j(t)$, given by (11) and (12), respectively, for Hospital i. That is, $\mu_i(t)$ is associated with Hospital i's waiting time and $\lambda_i(t)$ with that of the rival. The current-value Hamiltonian is

$$H_i = T + pS_i(t) - \frac{\gamma}{2}S_i(t)^2 - \alpha_1 w_i(t) - \frac{\alpha_2}{2}w_i(t)^2 + \mu_i(t)\theta[D_i(w_i(t), w_j(t)) - S_i(t)] + \lambda_i(t)\theta[D_j(w_i(t), w_j(t)) - S_j(t)]. \quad \text{(B.1)}$$

Candidates for optimal supply path $S_i(t)$ and costate trajectories $\mu_i(t)$ and $\lambda_i(t)$ must satisfy $\partial H_i/\partial S_i(t) = 0$, $\dot{\mu}_i(t) = \rho\mu_i(t) - \partial H_i/\partial w_i(t)$, and $\dot{\lambda}_i(t) = \rho\lambda_i(t) - \partial H_i/\partial w_j(t)$. More extensively:

$$p - \gamma S_i(t) = \theta\mu_i(t), \quad \text{(B.2)}$$

$$\dot{\mu}_i(t) = \left[\rho + \frac{\theta(2-\beta)N}{2\tau}\right]\mu_i(t) - \frac{\theta\beta N}{2\tau}\lambda_i(t) + \alpha_1 + \alpha_2 w_i(t), \quad \text{(B.3)}$$

and

$$\dot{\lambda}_i(t) = \left[\rho + \frac{\theta(2-\beta)N}{2\tau}\right]\lambda_i(t) - \frac{\theta\beta N}{2\tau}\mu_i(t). \quad \text{(B.4)}$$

The solution must also satisfy the transversality conditions

$$\lim_{t\to\infty} e^{-\rho t}\mu_i(t)w_i(t) = 0 \quad \text{and} \quad \lim_{t\to\infty} e^{-\rho t}\lambda_i(t)w_j(t) = 0. \quad \text{(B.5)}$$

Optimality is established by concavity of the current-value Hamiltonian with respect to $S_i(t)$ and $w_i(t)$. Inserting the definition of demand (7) and the optimality condition for supply (B.2) into the

dynamic constraint (11) yields

$$\dot{w}_i(t) = \theta N \left[\beta \left(\frac{1}{2} + \frac{w_j(t) - w_i(t)}{2\tau} \right) + (1-\beta) \left(\frac{v - k - w_i(t)}{\tau} \right) \right] - \theta \left(\frac{p - \theta \mu_i(t)}{\gamma} \right). \tag{B.5}$$

Let the superscript OL denote the symmetric open-loop steady-state in which $w_i(t) = w_j(t) = w^{OL}$, $\mu_i(t) = \mu_j(t) = \mu^{OL}$, and $S_i(t) = S_j(t) = S^{OL}$. Setting $\dot{w}(t) = \dot{\mu}(t) = \dot{\lambda}(t) = 0$ in equations (B.3), (B.4), and (B.5) and solving for the steady-state waiting time and costate variable gives

$$w^{OL} = \frac{\gamma\phi\tau}{(1-\beta)\gamma\phi N + 2\theta\tau^2\alpha_2} \left\{ N \left[\frac{\beta}{2} + (1-\beta) \left(\frac{v-k}{\tau} \right) \right] - \frac{p}{\gamma} - \frac{2\theta\tau\alpha_1}{\gamma\phi} \right\} \tag{B.6}$$

and

$$\mu^{OL} = -\frac{2\tau}{\phi}(\alpha_1 + \alpha_2 w^{OL}), \tag{B.7}$$

where ϕ is defined by (14) in Section 4. The corresponding steady-state supply is

$$S^{OL} = \frac{p}{\gamma} + \frac{2\theta\tau\left(\alpha_1 + \alpha_2 w^{OL}\right)}{\gamma\phi}. \tag{B.8}$$

It can be shown (calculations available upon request) that this equilibrium is stable in the saddle sense, and that the steady-state is characterised by non-negative waiting times and a partially covered monopolistic segment if the cost parameter γ is sufficiently large.

From (B.6) we derive

$$\frac{\partial w^{OL}}{\partial \tau} = -\frac{(1-\beta)x_M^{OL} + \frac{\tau}{N}\frac{\partial S^{OL}}{\partial \tau}}{1 - \beta + \frac{2\theta\tau^2\alpha_2}{\gamma\phi N}} < 0, \tag{B.9}$$

where

$$\frac{\partial S^{OL}}{\partial \tau} = N\theta^2(\alpha_1 + \alpha_2 w^{OL}) \frac{(1-\beta)[N\theta(2-\beta) + 4\tau\rho]\theta N + (2-\beta)(\tau\rho)^2}{2\gamma(N\theta + \tau\rho)^2[N(1-\beta)\theta + \tau\rho]^2} > 0 \tag{B.10}$$

is the marginal effect of τ on steady-state supply for a *given* waiting time. Thus, regardless of whether the marginal provider disutility of waiting time is constant or increasing, more patient choice leads to higher steady-state waiting times.

B.1 Non-linear patient disutility of waiting

Suppose that hospital demand is given by (30) in Section 6.1. Defining the Hamiltonian as before, the optimality conditions in the symmetric steady-state are now given by

$$p - \gamma S^{OL} = \theta \mu^{OL}, \tag{B.11}$$

$$\left[\rho - \theta \frac{\partial D_i(w^{OL})}{\partial w_i}\right] \mu^{OL} - \theta \frac{\partial D_j(w^{OL})}{\partial w_i} \lambda^{OL} + \alpha_1 + \alpha_2 w^{OL} = 0, \tag{B.12}$$

and

$$\left[\rho - \theta \frac{\partial D_j(w^{OL})}{\partial w_j}\right] \lambda^{OL} - \theta \frac{\partial D_i(w^{OL})}{\partial w_j} \mu^{OL} = 0. \tag{B.13}$$

Using (30) and (B.12)-(B.13) to solve for μ^{OL} and λ^{OL}, we obtain

$$\mu^{OL} = -\frac{\tau}{2} \frac{2\rho\tau + \theta(2-\beta)N\frac{\partial f(w^{OL})}{\partial w}}{\left[\rho\tau + \theta(1-\beta)N\frac{\partial f(w^{OL})}{\partial w}\right]\left[\rho\tau + \theta N\frac{\partial f(w^{OL})}{\partial w}\right]} (\alpha_1 + \alpha_2 w^{OL}) < 0 \tag{B.14}$$

and

$$\lambda^{OL} = \left[\frac{\theta\beta N\frac{\partial f(w^{OL})}{\partial w}}{2\rho\tau + \theta(2-\beta)N\frac{\partial f(w^{OL})}{\partial w}}\right] \mu^{OL} < 0. \tag{B.15}$$

Using the dynamic constraint, (30), and (B.11), the steady-state waiting time is then implicitly defined by

$$N\left[\frac{\beta}{2} + (1-\beta)\left(\frac{v - k - f(w^{OL})}{\tau}\right)\right] - \frac{p - \theta\mu^{OL}}{\gamma} = 0. \tag{B.16}$$

Existence requires that the second-order conditions of the hospitals' maximisation problem are satisfied. These are given by $\partial^2 H_i/\partial S_i^2 \leq 0$, $\partial^2 H_i/\partial w_i^2 \leq 0$, and $(\partial^2 H_i/\partial S_i^2)(\partial^2 H_i/\partial w_i^2) - \partial^2 H_i/\partial S_i \partial w_i \geq 0$. Since $\partial^2 H_i/\partial S_i^2 = -\gamma$ and $\partial^2 H_i/\partial S_i \partial w_i = 0$, concavity of the Hamiltonian requires that

$$\frac{\partial^2 H_i}{\partial w_i^2} = -\alpha_2 - \left[\frac{\theta(2-\beta)N}{2\tau}\mu_i - \frac{\theta\beta N}{2\tau}\lambda_i\right] \frac{\partial^2 f}{\partial w_i^2} \leq 0. \tag{B.17}$$

Evaluated at the steady-state, this expression becomes

$$-\alpha_2 + \frac{\left[\rho\tau(2-\beta) + 2\theta(1-\beta)N\frac{\partial f(w^{OL})}{\partial w}\right]\theta N(\alpha_1 + \alpha_2 w^{OL})}{2\left[\rho\tau + \theta(1-\beta)N\frac{\partial f(w^{OL})}{\partial w}\right]\left[\rho\tau + \theta N\frac{\partial f(w^{OL})}{\partial w}\right]} \frac{\partial^2 f(w^{OL})}{\partial w^2} \leq 0. \tag{B.18}$$

If $\partial^2 f\left(w^{OL}\right)/\partial w^2 \leq 0$, the second-order conditions are always satisfied, whereas, if $\partial^2 f\left(w^{OL}\right)/\partial w^2 > 0$, the second-order conditions are satisfied if $\alpha_2 > 0$ and the degree of convexity of f is sufficiently

small. More specifically, the second-order conditions are satisfied if

$$\frac{\partial^2 f(w^{OL})}{\partial w^2} \leq \frac{2\left[\rho\tau + \theta(1-\beta)N\frac{\partial f(w^{OL})}{\partial w}\right]\left[\rho\tau + \theta N\frac{\partial f(w^{OL})}{\partial w}\right]}{\rho\tau(2-\beta) + 2\theta(1-\beta)N\frac{\partial f(w^{OL})}{\partial w}}\frac{\alpha_2}{\theta N(\alpha_1 + \alpha_2 w^{OL})}. \tag{B.19}$$

Implicitly differentiating (B.16) with respect to w^{OL} and τ yields

$$\frac{\partial w^{OL}}{\partial \tau} = -\frac{(1-\beta)x_M^{OL} - \frac{\tau\theta}{N\gamma}\frac{\partial \mu^{OL}}{\partial \tau}}{(1-\beta)\frac{\partial f(w^{OL})}{\partial w} - \frac{\tau\theta}{N\gamma}\frac{\partial \mu^{OL}}{\partial w^{OL}}} < 0, \tag{B.20}$$

where $x_M^{OL} = (v - k - f(w^{OL}))/\tau > 0$ is the location on the indifferent patient in the monopolistic segment, and where

$$\frac{\partial \mu^{OL}}{\partial \tau} = -\frac{\partial f(w^{OL})}{\partial w}\frac{\theta N\Gamma(w^{OL})(\alpha_1 + \alpha_2 w^{OL})}{2\left[\rho\tau + \theta(1-\beta)N\frac{\partial f(w^{OL})}{\partial w}\right]^2\left[\rho\tau + \theta N\frac{\partial f(w^{OL})}{\partial w}\right]^2} < 0, \tag{B.21}$$

$$\begin{aligned}\frac{\partial \mu^{OL}}{\partial w^{OL}} = &-\frac{\tau}{2}\frac{\left[2\rho\tau + \theta(2-\beta)N\frac{\partial f(w^{OL})}{\partial w}\right]\alpha_2}{\left[\rho\tau + \theta(1-\beta)N\frac{\partial f(w^{OL})}{\partial w}\right]\left[\rho\tau + \theta N\frac{\partial f(w^{OL})}{\partial w}\right]} \\ &+ \frac{\partial^2 f(w^{OL})}{\partial w^2}\frac{\tau\theta N\Gamma(w^{OL})(\alpha_1 + \alpha_2 w^{OL})}{2\left[\rho\tau + \theta(1-\beta)N\frac{\partial f(w^{OL})}{\partial w}\right]^2\left[\rho\tau + \theta N\frac{\partial f(w^{OL})}{\partial w}\right]^2} \leq 0,\end{aligned} \tag{B.22}$$

and

$$\Gamma(w^{OL}) = (\rho\tau)^2(2-\beta) + 4\rho\tau\theta(1-\beta)N\frac{\partial f(w^{OL})}{\partial w} + \theta^2(1-\beta)(2-\beta)N^2\left(\frac{\partial f(w^{OL})}{\partial w}\right)^2 > 0. \tag{B.23}$$

To show that (B.22) is always non-positive in the steady-state equilibrium, notice that the right-hand side of (B.22) is increasing in $\partial^2 f\left(w^{OL}\right)/\partial w^2$, while the second-order conditions dictate that $\partial^2 f\left(w^{OL}\right)/\partial w^2$ must be sufficiently low (cf. (B.19)). Replacing $\partial^2 f\left(w^{OL}\right)/\partial w^2$ in equation (B.22) with the right-hand side of (B.19), which is the maximum value of $\partial^2 f\left(w^{OL}\right)/\partial w^2$ that still ensures equilibrium existence, yields

$$\frac{\partial \mu^{OL}}{\partial w^{OL}} = -\frac{\rho\theta(\tau\beta)^2 N\frac{\partial f(w^{OL})}{\partial w}\alpha_2}{2\left[\rho\tau + \theta(1-\beta)N\frac{\partial f(w^{OL})}{\partial w}\right]\left[\rho\tau + \theta N\frac{\partial f(w^{OL})}{\partial w}\right]\left[\rho\tau(2-\beta) + 2\theta(1-\beta)N\frac{\partial f(w^{OL})}{\partial w}\right]} \leq 0. \tag{B.24}$$

This implies that $\partial \mu^{OL}/\partial w^{OL} \leq 0$, and thus $\partial w^{OL}/\partial \tau < 0$, for every specification of $f(w)$ that is

compatible with equilibrium existence under open-loop rules.

B.2 Non-linear patient disutility of travelling

Suppose the patient utility function is redefined as indicated in Section 6.2. The optimality conditions, evaluated at the symmetric steady-state, are the given by (B.11) and

$$\left[\rho + \frac{\theta\beta N}{2\tau g'\left(\frac{1}{2}\right)} + \frac{\theta(1-\beta)N}{\tau g'\left(x_M^{OL}\right)}\right]\mu^{OL} - \frac{\theta\beta N}{2\tau g'\left(\frac{1}{2}\right)}\lambda^{OL} + \alpha_1 + \alpha_2 w^{OL} = 0, \tag{B.25}$$

and

$$\left[\rho + \frac{\theta\beta N}{2\tau g'\left(\frac{1}{2}\right) + \frac{\theta(1-\beta)N}{\tau g'\left(x_M^{OL}\right)}}\right]\lambda^{OL} - \frac{\theta\beta N}{2\tau g'\left(\frac{1}{2}\right)}\mu^{OL}. \tag{B.26}$$

Using (B.25) and (B.26) to solve for μ^{OL} and λ^{OL}, we obtain:

$$\mu^{OL} = -\frac{\tau}{2}\frac{2g'\left(\frac{1}{2}\right)\rho\tau + \theta\beta N + \frac{2g'\left(\frac{1}{2}\right)\theta(1-\beta)N}{g'\left(x_M^{OL}\right)}}{\left[\rho\tau + \frac{\theta(1-\beta)N}{g'\left(x_M^{OL}\right)}\right]\left[g'\left(\frac{1}{2}\right)\rho\tau + \theta\beta N + \frac{g'\left(\frac{1}{2}\right)\theta(1-\beta)N}{g'\left(x_M^{OL}\right)}\right]}(\alpha_1 + \alpha_2 w^{OL}) < 0. \tag{B.27}$$

and

$$\lambda^{OL} = \left[\frac{\theta\beta N}{2g'\left(\frac{1}{2}\right)\rho\tau + \theta\beta N + \frac{2g'\left(\frac{1}{2}\right)\theta(1-\beta)N}{g'\left(x_M^{OL}\right)}}\right]\mu^{OL} < 0. \tag{B.28}$$

Using the dynamic constraint and (B.11), the steady-state waiting time is implicitly defined by

$$N\left[\frac{\beta}{2} + (1-\beta)x_M^{OL}\right] - \frac{p - \theta\mu^{OL}}{\gamma} = 0. \tag{B.29}$$

Existence requires that the second-order conditions of the hospitals' maximisation problem are satisfied. These are given by $\partial^2 H_i/\partial S_i^2 \leq 0$, $\partial^2 H_i/\partial w_i^2 \leq 0$, and $(\partial^2 H_i/\partial S_i^2)(\partial^2 H_i/\partial w_i^2) - \partial^2 H_i/\partial S_i \partial w_i \geq 0$. Since $\partial^2 H_i/\partial S_i^2 = -\gamma$ and $\partial^2 H_i/\partial S_i \partial w_i = 0$, concavity of the Hamiltonian requires that

$$\frac{\partial^2 H_i}{\partial w_i^2} = -\alpha_2 + \left[\frac{\theta(1-\beta)N}{\tau}\frac{g''(x_M^i)}{[g''(x_M^i)]^2}\frac{\partial x_M^i}{\partial w_i}\right]\mu_i \leq 0. \tag{B.30}$$

Evaluated at the steady-state, this expression becomes

$$-\alpha_2 - \left[\frac{\theta(1-\beta)N}{\tau^2}\frac{g''\left(x_M^{OL}\right)}{[g'\left(x_M^{OL}\right)]^3}\right]\mu^{OL} \leq 0. \tag{B.31}$$

If $g''\left(x_M^{OL}\right) \leq 0$, the second-order conditions are always satisfied, whereas, if $g''\left(x_M^{OL}\right) > 0$, the second-order conditions are satisfied if $\alpha_2 > 0$ and the degree of convexity of g is sufficiently small.

Implicitly differentiating (B.29) with respect to w^{OL} and τ yields

$$\frac{\partial w^{OL}}{\partial \tau} = -\frac{N(1-\beta)\frac{\partial x_M^{OL}}{\partial \tau} + \frac{\theta}{\gamma}\frac{\partial \mu^{OL}}{\partial \tau}}{N(1-\beta)\frac{\partial x_M^{OL}}{\partial w^{OL}} + \frac{\theta}{\gamma}\frac{\partial \mu^{OL}}{\partial w^{OL}}} \tag{B.32}$$

where

$$\frac{\partial x_M^{OL}}{\partial \tau} = -\frac{g(x_M^{OL})}{\tau g'\left(x_M^{OL}\right)} < 0, \tag{B.33}$$

$$\frac{\partial x_M^{OL}}{\partial w^{OL}} = -\frac{1}{\tau g'\left(x_M^{OL}\right)} < 0, \tag{B.34}$$

$$\frac{\partial \mu^{OL}}{\partial \tau} = -\frac{\left[\Delta_1 - \Delta_2 \frac{g'\left(x_M^{OL}\right) g''\left(x_M^{OL}\right)}{[g'\left(x_M^{OL}\right)]^2}\right](\alpha_1 + \alpha_2 w^{OL})}{\left[\rho\tau + \frac{\theta(1-\beta)N}{g'\left(x_M^{OL}\right)}\right]^2 \left[g'\left(\frac{1}{2}\right)\rho\tau + \theta\beta N + \frac{g'\left(\frac{1}{2}\right)\theta(1-\beta)N}{g'\left(x_M^{OL}\right)}\right]^2}, \tag{B.35}$$

$$\Delta_1 = \left[4g'\left(\frac{1}{2}\right)\rho\tau + \theta\beta N + \frac{2g'\left(\frac{1}{2}\right)\theta(1-\beta)N}{g'\left(x_M^{OL}\right)}\right]\left[g'\left(\frac{1}{2}\right)\left[\frac{\theta(1-\beta)N}{g'\left(x_M^{OL}\right)}\right]^2 + \frac{(\theta N)^2\beta(1-\beta)}{g'\left(x_M^{OL}\right)}\right] + \frac{2\left[g'\left(\frac{1}{2}\right)\rho\tau\right]^2\theta(1-\beta)N}{g'\left(x_M^{OL}\right)} + g'\left(\frac{1}{2}\right)(\rho\tau)^2\theta N\beta > 0, \tag{B.36}$$

$$\Delta_2 = g'\left(\frac{1}{2}\right)\left[4g'\left(\frac{1}{2}\right)\rho\tau + 2\theta\beta N + \frac{2g'\left(\frac{1}{2}\right)\theta(1-\beta)N}{g'\left(x_M^{OL}\right)}\right]\left[\frac{\theta(1-\beta)N}{g'\left(x_M^{OL}\right)}\right]^2 + \frac{2\left[g'\left(\frac{1}{2}\right)\rho\tau\right]^2\theta(1-\beta)N}{g'\left(x_M^{OL}\right)} + \left[2g'\left(\frac{1}{2}\right)\rho\tau + \theta\beta N\right]\frac{(\theta N)^2\beta(1-\beta)}{g'\left(x_M^{OL}\right)} > 0, \tag{B.37}$$

$$\frac{\partial \mu^{OL}}{\partial w^{OL}} = -\frac{\alpha_2\tau}{2}\frac{2g'\left(\frac{1}{2}\right)\rho\tau + \theta\beta N + \frac{2g'\left(\frac{1}{2}\right)\theta(1-\beta)N}{g'\left(x_M^{OL}\right)}}{\left[\rho\tau + \frac{\theta(1-\beta)N}{g'\left(x_M^{OL}\right)}\right]\left[g'\left(\frac{1}{2}\right)\rho\tau + \theta\beta N + \frac{g'\left(\frac{1}{2}\right)\theta(1-\beta)N}{g'\left(x_M^{OL}\right)}\right]} + \Delta_3\left[\frac{\theta(1-\beta)N(\alpha_1 + \alpha_2 w^{OL})}{2[g'\left(x_M^{OL}\right)]^3}\right] g''\left(x_M^{OL}\right), \tag{B.38}$$

and

$$\Delta_3 = \frac{\left[\begin{array}{c} \left[2g'\left(\frac{1}{2}\right)\rho\tau + \theta\beta N + \frac{2g'\left(\frac{1}{2}\right)\theta(1-\beta)N}{g'\left(x_M^{OL}\right)}\right]^2 \\ -\left[2g'\left(\frac{1}{2}\right)\rho\tau + \frac{2g'\left(\frac{1}{2}\right)\theta(1-\beta)N}{g'\left(x_M^{OL}\right)}\right]\left[g'\left(\frac{1}{2}\right)\rho\tau + \theta\beta N + \frac{g'\left(\frac{1}{2}\right)\theta(1-\beta)N}{g'\left(x_M^{OL}\right)}\right] \end{array}\right]}{\left[\rho\tau + \frac{\theta(1-\beta)N}{g'\left(x_M^{OL}\right)}\right]^2\left[g'\left(\frac{1}{2}\right)\rho\tau + \theta\beta N + \frac{g'\left(\frac{1}{2}\right)\theta(1-\beta)N}{g'\left(x_M^{OL}\right)}\right]^2} > 0 \qquad \text{(B.39)}$$

If $g''\left(x_M^{OL}\right) \leq 0$, the expressions on the right-hand side of (B.35) and (B.38) are unambiguously negative, which implies that $\partial w^{OL}/\partial\tau < 0$ for every concave function g. If instead $g''\left(x_M^{OL}\right) > 0$, a negative sign of $\partial\mu^{OL}/\partial\tau$ and $\partial\mu^{OL}/\partial w^{OL}$, which implies $\partial w^{OL}/\partial\tau < 0$, requires that $g''\left(x_M^{OL}\right)$ is sufficiently low.