NIPE Núcleo de Investigação em Políticas Económicas e Empresariais

#03
2020

WORKING PAPER

Luís Sá
Odd Rune Straume

# "Quality provision in hospital markets with demand inertia: The role of patient expectations"

https://www.eeg.uminho.pt/pt/investigar/nipe

FCT Fundação para a Ciência e a Tecnologia
MINISTÉRIO DA EDUCAÇÃO E CIÊNCIA

«This paper is financed by National Funds of the FCT – Portuguese Foundation for Science and Technology within the project «UIDB/03182/2020»

Universidade do Minho
Escola de Economia e Gestão

# Quality provision in hospital markets with demand inertia: The role of patient expectations*

Luís Sá[†]and Odd Rune Straume[‡]

July 2020

## Abstract

The presence of switching costs and persistent patient preferences generates demand inertia and links current and future choices of hospital. Using a model of hospital competition with demand inertia, we investigate the effect of *patient expectations* (whether and how patients anticipate the future) on quality provision. We consider three types of expectations. Myopic patients choose a hospital based on current variables alone, forward-looking but naïve patients take the future into account but assume that quality remains constant, and forward-looking and rational patients foresee the evolution of quality. We rank equilibrium quality provision and show that it is higher under naïve than myopic expectations, while equilibrium quality under rational expectations may be highest or lowest. This result also holds for patient welfare, suggesting that rationality does not always benefit patients. We also show that only under rational expectations may quality be lower than in a market without inertia and switching cost reductions beneficial.

*Keywords*: Hospital competition; myopic behaviour; forward-looking behaviour; rational expectations; switching costs.

*JEL Classification*: I11, I18, L13, L51.

[*] This paper is financed by National Funds of the FCT (Portuguese Foundation for Science and Technology) within the project UID/ECO/03182/2019. Sá acknowledges the PhD Studentship SFRH/BD/129073/2017 awarded by the FCT and financed by National Funds of the FCT and the European Social Fund. We thank Pedro Pita Barros and Lise Rochaix for the discussion which gave rise to this paper.

[†] Corresponding author. Department of Economics/NIPE, University of Minho, Campus de Gualtar, 4710-057 Braga, Portugal. E-mail: luis.sa@eeg.uminho.pt.

[‡] Department of Economics/NIPE, University of Minho, Campus de Gualtar, 4710-057 Braga, Portugal; and Department of Economics, University of Bergen. E-mail: o.r.straume@eeg.uminho.pt

# 1 Introduction

Motivated by the observation that patients tend to choose a hospital and repeatedly demand treatment from it, even during unrelated episodes of care, recent empirical literature provides evidence of demand inertia in hospital markets (Jung 2011; Shepard, 2016; Raval and Rosenbaum, 2018; Irace, 2018). Like travelling distance and quality of care, prior utilisation emerges as a key determinant of hospital choice, and its effect has been shown to result both from persistent patient preferences *and* from switching costs (Raval and Rosenbaum, 2018; Irace, 2018). Persistent preferences denote the time-invariant horizontal preferences some patients have for hospital characteristics. Absent significant changes in the market, and upon realising that their tastes or health needs have remained constant, repeated utilisation of the same hospital may be the optimal behaviour for these patients.

Preference persistency, however, does not fully explain the magnitude of demand inertia. Even when their preferences change, patients may still find it optimal to choose the same hospital repeatedly if switching is costly, and there is a variety of reasons why switching costs arise in hospital markets. First, there may be monetary and opportunity costs incurred by patients in order to have their medical records transferred across providers. Second, because evaluating hospital quality is a time-consuming and complex task, switching costs may reflect the risk of trying an untested, alternative provider. Third, switching costs might arise from the need to undergo duplicate procedures, such as diagnostic tests, when patients restart treatment after switching providers. Fourth, switching costs may also be the premium patients are willing to pay, either in terms of higher prices or lower quality, for familiarity with their chosen hospital. Switching costs, therefore, induce state dependence; i.e., a causal impact of current on future choices. If switching is costly, choosing a particular hospital in the present has an impact on the utility patients will derive from treatment at different hospitals in their choice set in the future, thereby affecting their current choice.

Both sources of demand inertia create a link between the choices patients make at different points in time. If the choices patients make are intertemporally linked, these choices will be affected by whether or not patients anticipate the future, as well as the degree of sophistication of their foresight — what we refer to as *patient expectations.* If patient preferences were completely independent across time and switching costs inexistent, meaning that there would be no intertemporal link, current choices would be unaffected by whether and how patients anticipate future ones. In other words, the role of patient expectations and demand inertia are inextricable.

In the present paper we analyse a hospital market where switching costs and persistent horizontal patient preferences generate demand inertia and investigate how different types of patient expectations affect quality provision by two competing hospitals. In the context of patient choice of hospital, rational expectations imply that patients take the future into account and are able to correctly assess the evolution of the determinants of their choices. In our framework more specifically, where demand inertia is present, forward-looking and rational patients know that they will demand hospital care in the future with some positive probability, anticipate that their preferences may change over time, are aware of the lock-in effect of switching costs, and foresee future quality. Regarding the latter aspect, these patients not only know that higher quality attracts higher demand in the present and that part of this demand will be locked-in, but also predict how this locked-in demand affects future quality. In turn, understanding the link between current and future quality, via demand, requires some knowledge of hospital objectives and technology.

Departures from fully rational behaviour may occur because patients are present-biased or because they have incorrect beliefs about the link between current and future quality (Baicker et al., 2015). We look at present-bias by considering myopic patients, who ignore the future and base their choice of hospital on current observable variables only. We also look at incorrect beliefs about future quality by allowing for the possibility that patients are forward-looking but naïve. In this case, the difference from full rationality lies not on whether patients anticipate the future but on how they do it. Similarly to forward-looking and rational patients, forward-looking but naïve ones anticipate the possibility of having persistent preferences and the existence of switching costs. They fail, however, in foreseeing future quality. Because predicting the evolution of hospital quality is cognitively complex or because the information required to carry out such a task is unavailable, these patients are naïve in the sense that they resort to the simple rule-of-thumb of expecting that quality will remain constant.

To study the demand for hospital care when there is inertia, we present a two-period model where patients choose a hospital based on the level of quality offered, their horizontal preferences, and, possibly, a switching cost. In the second period, patients who remain in the market either have new or the same preferences as in the first period and incur a switching cost if they decide to demand treatment from the hospital they did not choose previously. In the first period, all patients are new in the market, implying that there are no switching costs and that horizontal preferences affect first-period utility only to the extent that they represent contemporaneous tastes. If patients are forward-looking, however, their choices are also conditioned on what might happen in the second period; namely, the possibility that their preferences may change and that they might want to switch (i.e., patients may see themselves tied to the 'wrong' hospital) and the evolution of the quality difference between the two hospitals. It is, therefore, in the first period that patient expectations play a role in determining the demand for hospital care and hence in affecting the incentives for quality provision.

To make the analysis of the evolution of quality more comprehensive, we assume that the hospitals are motivated and allow for both cost substitutability and complementarity between quality and output in hospital production. If the degree of cost substitutability is sufficiently strong, higher demand increases the marginal cost of quality provision. This, in turn, implies that higher quality in the present foretells lower quality in the future or, more specifically, that a current unilateral quality increase reduces the future quality difference. A current unilateral quality increase yields a demand advantage, which, owing to inertia, partially carries over into the future, increasing the marginal cost of quality and thus reducing the incentives for quality provision. Similarly, if there is cost complementarity (or if the degree of cost substitutability is sufficiently weak), higher demand reduces the marginal cost of quality and implies that a current unilateral quality increase widens both the current and the future quality differences.[1] This link between present and future quality, and the fact that only rational patients observe it, partly explain our results.

We show that patient expectations affect quality provision only through the responsiveness of demand to quality, with higher responsiveness leading to higher quality provision. While demand is always more responsive when patients are forward-looking but naïve than when patients are myopic, demand responsiveness under rational expectations depends on the actual relationship between present and future quality. The more rational patients anticipate a current quality increase to be offset (or more than offset) in terms of the future quality difference, the less attracted by it these patients are. This is why demand responsiveness to quality is decreasing (increasing) in the degree of cost substitutability (complementarity) when patients are rational. Consequently, demand responsiveness and quality under rational expectations are ranked highest, lowest, or in between the cases of forward-looking but naïve and myopic expectations, depending on the degree of cost substitutability/complementarity.

[1] In this case, naturally, the lower the degree of cost complementarity is or the higher the degree of cost substitutability is, the smaller is the magnitude of the increase in the future quality difference caused by a current quality increase.

This first main result has important implications for patient welfare. In a symmetric equilibrium, the type of which we focus our analysis on, expectations affect aggregate patient utility uniquely through quality. Thus, when we rank quality according to the type of expectations, we are also raking patient welfare. This implies that full rationality does not necessarily make patients better off.

Our second main result relates to the effect of demand inertia on quality provision and its connection with patient expectations. We show that, compared with the benchmark of a market without demand inertia, quality provision is determined by two additional effects. First, there is a pro-quality effect of competition for market share, because current demand is valuable in the future and will be partially locked-in. Second, there is a patient foresight effect, capturing the size of demand responsiveness under the different types of patient expectations relative to the benchmark. The foresight effect vanishes when patients are myopic and reinforces the competition effect when they are forward-looking but naïve. It may instead outweigh the competition effect if patients foresee that a unilateral quality increase will yield a sufficiently large reduction in the future quality difference. Rational expectations and strong cost substitutability are, therefore, necessary (but not sufficient) conditions for demand responsiveness to be low enough to dominate the competition effect and quality to be lower than in a market without inertia.

The intuition behind our third and final result mirrors that which we have just described. We look at the outcome of a policy aimed at reducing inertia and show that lower switching costs are generally counterproductive. Lower switching costs reduce the competition effect and thus can only lead to higher quality if they increase demand responsiveness to the extent that it more than compensates for that reduction. This turns out to be the case only when patients are rational and a unilateral quality increase today causes a sufficiently large reduction in the future quality difference.

The rest of the paper is organised as follows. In the next section, we relate our study to several strands of literature. In Section 3, we present the model and, in Section 4, derive the equilibrium quality levels in the two-period game. Our primary analysis is given in sections 5, 6, and 7, where we explore the role of patient expectations thoroughly, compare quality provision with the benchmark of a market without demand inertia, and investigate the effect of lower switching costs. Finally, as well as concluding remarks, Section 8 provides a discussion of the implications of forward-looking and rational behaviour to patient welfare.

## 2 Related literature

The recent empirical literature that documents choice persistence in the hospital industry motivates our study. Jung et al. (2011) estimate that the probability of a hospital being chosen for a hypothetical hospitalisation is 64 percentage points higher if the hospital was previously used, and Shepard (2016) finds that patients are five times more likely to choose a hospital where they received outpatient care in the previous year. Two subsequent studies corroborate these results and show that demand inertia results from both switching costs (or state dependence) and persistent patient preferences (or unobserved patient heterogeneity). Raval and Rosenbaum (2018) report that previous use increases the predicted share of women expected to return to a hospital for childbirth from 40% to 72%. Additionally, they show that the effect of previous utilisation, the switching cost, falls in magnitude but is statistically robust to the inclusion of hospital-patient fixed effects, which capture the effect of persistent preferences. More specifically, they estimate that the effect of switching costs accounts for roughly 40% of demand inertia. Irace (2018) resorts to quasi-exogenous shocks that induce patients to switch hospitals. He finds that patients admitted at a hospital they have never visited before during an emergency are more likely to return to that hospital in subsequent episodes of care. This is indicative of switching costs and is also true for patients forced to try a new hospital during a temporary closure because of a natural disaster. Conversely, patients who do return to the hospital they had been using before the emergency are more likely to choose it repeatedly, which points to preference persistency.

Much earlier, Klemperer (1987) established a framework to analyse price competition in markets with switching costs where some patients have persistent horizontal preferences. One of the key insights it provides, and that is well-established in the switching costs literature (Villas-Boas, 2015), is that rational consumers' realisation that a higher price in the future follows a lower price in the present makes demand less elastic, contributing to higher prices. While the analogous result may be present in our model, it also allows for the possibility that higher quality in the future follows higher quality in the present. When anticipated by patients, this makes demand more elastic and reinforces the effect of competition for market share induced by switching costs, leading to higher quality provision.[2]

[2] For example, Klemperer (1987) shows that prices are always above the no-inertia case if consumers are rational and all of those who bought in the first period have unchanged preferences. In our model, however, under the same conditions, quality provision may be *higher* than in a market without demand inertia owing to the relationship between hospital technology and motivation.

In the context of quality competition in primary care, Gravelle and Masiero (2000) present a two-period model where myopic patients incur switching costs. Contrary to our results, they show that quality is unaffected by switching costs. Within the hospital competition literature, two studies consider an information-related form of inertia. Arising from the complexity of assessing the quality of care, demand sluggishness implies that, at each point in time, only a fraction of patients become aware of quality changes and hence only a fraction of any potential change in demand materialises. Weaker sluggishness, therefore, makes demand more responsive to quality. With profit-maximising providers and a positive payment-cost margin, as in Brekke et al. (2012), increased demand responsiveness leads to higher quality. Siciliani et al. (2013), however, show that semi-altruistic hospital preferences may overturn this result. Increased demand responsiveness leads to lower quality provision if the prospective payment is sufficiently below unit costs and the financial incentive to avoid patients dominates the altruistic incentive to attract them.[3] Although demand responsiveness to quality also plays a crucial role in our model, our analysis differs significantly from those of Brekke et al. (2012) and Siciliani et al. (2013). First, they model inertia in a multiperiod framework where expectations are unexplored. Second, they focus on exogenous changes in parameters that affect demand responsiveness and on how this, in turn, impacts quality provision, given hospital preferences and technology. Here, we mainly investigate how patient expectations determine demand responsiveness endogenously and show that hospital preferences and technology may themselves affect demand responsiveness.

To the best of our knowledge, no study has explored the link between patient expectations and choice of provider. There is, however, a growing empirical literature on healthcare utilisation under nonlinear health insurance contracts, which sheds light on whether consumers take the future into account in the broader healthcare context. Brot-Goldberg et al. (2017) study healthcare utilisation by employees who were required to switch from free full-coverage to a nonlinear, high-deductible insurance plan. They report that annual utilisation decreases by 17.9% in response to the plan change, and, importantly, it does so almost entirely while consumers are still under the deductible (i.e., before coinsurance eligibility). This result holds even for the sickest of consumers, who should anticipate reaching the coinsurance arm of the plan with near certainty and thus face lower end-of-year prices. Guo and Zhang (2019) show that, during the year of childbirth, fathers' monthly medical care utilisation rises by 11% upon becoming eligible for coinsurance, despite childbirth being

[3] Brekke et al. (2011) investigate this mechanism thoroughly. For an overview of the literature on quality competition in healthcare markets, see Brekke et al. (2014).

an expected event that contributes a great deal to deductible fulfilment. Absent liquidity constraints and controlling for health shocks, these fluctuations in healthcare utilisation are consistent with some degree of myopic behaviour since a fully forward-looking consumer would respond to his expected end-of-year price rather than to the spot price, thereby smoothing consumption over the year. Myopic behaviour instead implies that consumers perceive changes in coverage as changes in prices and hence adjust consumption accordingly. Dalton et al. (2020) provide even stronger evidence of myopic behaviour. They find that consumers completely ignore the future prices of prescription drugs under Medicare Part D, whose nonlinear contract design includes an initial coverage region followed by a coverage gap (the 'doughnut hole'). Drug purchases are initially constant and drop sharply once the coverage gap is reached, implying an estimated discount rate that is consistent with full myopia (i.e., equal to zero). A similar pattern of drug consumption under Medicare Part D may be found in Sacks et al. (2017), Einav et al. (2015), and Abaluck et al. (2018). In the latter two studies, however, the estimated discount rates indicate some degree of forward-looking behaviour, which is considerably higher in Einav et al. (2015). Additional evidence of forward-looking behaviour comes from Aron-Dine et al. (2015). They find that initial medical care utilisation is lower for employees who join a health insurance plan with an annual deductible later in the year. Because their deductible is less likely to be reached, individuals who enrol later face a higher expected end-of-year price. Their lower initial utilisation under the plan, therefore, suggests that they do respond to future prices. Interestingly, Aron-Dine et al. (2015) find similar results for prescription drug consumption under Medicare Part D. Looking at the German public health insurance system, Farbmacher et al. (2017) also report evidence of forward-looking behaviour. After the introduction of a one-time co-payment, initial outpatient care demand falls for some consumers, while it is unresponsive for the relatively sick, who should expect future needs to exceed a single visit and thus be less sensitive to the co-payment.

# 3 The model

Consider a health care market with two providers, henceforth referred to as hospitals. In each of two periods, $t = 1, 2$, the two hospitals, indexed $i = A, B$, are located at either endpoint of the unit line segment $[0, 1]$. Let Hospital A be located at 0 and Hospital B at 1. Locations on the line segment reflect the characteristics and preferences for elective hospital treatment supplied in this market. The line segment may be thought of as a geographical space or a disease space. In the former case, a patient's location on the line is simply her residence or workplace, while the location of a hospital is simply the place where its facilities were built. In the latter case, a patient's location on the line is a medical condition or a diagnosis, and the location of a hospital is the speciality mix (i.e., the treatments and services) it offers.

Patients have a gross valuation of treatment $v > 0$, demand a single unit of treatment from one of the hospitals in each period, and are arrayed with unit density along the line segment. They incur a travelling or mismatch cost $\tau$ per unit of distance between their location and that of the chosen hospital, but bear no out-of-pocket expenses either due to public provision of healthcare or to (social or private) health insurance coverage.[4] Patients derive utility from the quality of treatment, $q_t^i$, to which hospitals resort to attract demand in each period. There is a lower bound on treatment quality that represents the minimum quality hospitals are allowed to offer, with quality below this threshold being interpreted as malpractice. For simplicity, we assume that the lower bound on quality is equal to zero. The gross valuation of treatment $v$ is high enough so that the market is always fully covered.

Following the empirical analyses of Raval and Rosenbaum (2018) and Irace (2018), we model demand inertia in the style of Klemperer (1987). In the first period, all patients are new in the market, meaning that no patient is tied to any of the hospitals. Patients choose a hospital based on their horizontal preferences and the quality levels offered in the market. In the second period, a fraction $\lambda$ of the patients leave the market and are replaced by new patients with the same density and who are also uniformly distributed along the unit line segment. Another fraction $\mu$ of the existing patients have preferences for treatment characteristics that are independent of their first-period preferences. These patients are uniformly distributed along $[0, 1]$ and may be interpreted as patients who now reside or work in a different place or patients who have developed another, unrelated, disease. The parameter $\mu$ may, therefore, be interpreted as an inverse measure of the persistence of patient preferences over time. Patients with changing preferences who choose to demand treatment from the hospital they have not used in the first period incur an exogenous switching cost $s$. The remaining $(1 - \lambda - \mu)$ patients have unchanged preferences for treatment characteristics (i.e., their location on the line segment equals the first-period location) and choose the same hospital in both periods.[5] Thus, we measure demand inertia in two different ways: the

[4] The latter feature is analytically equivalent to having hospitals charge the same regulated price.

[5] As Villas-Boas (2015) suggests, this could be explicitly modelled by adding an infinitely high switching cost for these patients.

cost of switching providers ($s$) and the persistence of patient preferences $(1 - \lambda - \mu)$.

In the first period, patients know that they will leave the market with probability $\lambda$, have different preferences in the second period with probability $\mu$, and have persistent preferences with the remaining probability $1-\lambda-\mu$. These probabilities are independent of the first-period choice of hospital. Under these assumptions, the utility, in period $t$, of a patient located at $x_t$ who demands treatment from Hospital $i$, located at $z^i$, is given by

$$u_t(x_t, z^i) = v + q_t^i - \tau|x_t - z^i| - I_i s, \quad i, j = A, B; \tag{1}$$

where $I_i = 1$ in the second period if the patient has changing preferences, chose Hospital $i$ in the first period, and chooses Hospital $j$ in the second period; $I_i = 0$ otherwise.[6]

Hospitals are prospectively financed by a third-party payer (e.g., a regulator or insurer) that offers a price $\tilde{p}$ for each unit of treatment supplied and a lump-sum transfer, $T$, which ensures that a no-liability constraint is satisfied. Total treatment production costs are given by

$$C\left(q_t^i, D_t^i\right) = (cq_t^i + k)D_t^i + \frac{\gamma}{2}(q_t^i)^2, \quad i, j = A, B; \quad i \neq j; \tag{2}$$

where $c \lessgtr 0$ measures either the degree of cost substitutability (if $c > 0$) or complementarity (if $c < 0$) between quality and output, $k > \max\{0, -cq_t^i\}$ is the minimum unit cost of treatment, $\gamma > 0$ is a quality investment cost parameter, and $D_t^i$ is the demand for Hospital $i$ in period $t$ (or the number of treatments produced).

If $c > 0$, a certain level of quality is more costly to achieve when more patients are treated, implying that the marginal cost of quality is increasing in demand. In this case, hospital production exhibits *cost substitutability* between quality and output. This is a reasonable assumption if quality results from the investment in medical equipment and highly skilled staff. For example, offering an additional diagnostic test amounts to an increase in quality and requires a fixed investment in equipment and/or staff but also increases the cost of diagnosing each patient. On the other hand, if $c < 0$, the more patients a hospital treats, the less costly it is to provide each additional unit of quality, and the marginal cost of quality is decreasing in demand. In this case, quality and output are *cost complements*, reflecting the positive relationship between demand and quality observed when, all else equal, high-volume hospitals provide higher quality and generate better treatment

[6] For patients with persistent preferences, $x_1 = x_2$.

outcomes than low-volume hospitals.[7]

Additionally, we assume that hospitals are *motivated* in the sense that they care, to some extent, about the gross utility their patients derive from treatment. Specifically, we assume that Hospital $i$ ignores the travelling/mismatch and switching costs of its patients but attaches a weight $\alpha > 0$, denoting the degree of provider motivation, to the remaining part of their aggregate utility $(v + q_t^i)D_t^i$. Per-period payoff of Hospital $i$ is thus given by

$$\Omega_t^i = T + \tilde{p}D_t^i - C\left(q_t^i, D_t^i\right) + \alpha(v + q_t^i)D_t^i. \tag{3}$$

For simplicity and without loss of generality, there is no discounting. Furthermore, whereas hospitals have rational expectations, we allow for different types of patient expectations, which will be detailed later.

Finally, we impose the following restriction on parameter values:

$$c > c_{min} := \max\left\{\alpha - \frac{2\tau\gamma}{3(\lambda + \mu)}, \alpha - \tau\gamma\right\} \tag{4}$$

This restriction ensures that the second-order condition of the hospitals' maximisation problems in the second period and in a market without demand inertia are satisfied, as well as that the games we consider have economically meaningful, interior solutions. It simply implies that the degree of cost substitutability must be sufficiently strong or the degree of cost complementarity sufficiently weak. Throughout the analysis, we also assume the existence of interior-solution equilibria, i.e., $q_t^i > 0$, which requires that $\tilde{p}$ is sufficiently high.

# 4 Equilibrium quality provision

In each period, hospitals simultaneously and independently choose quality levels to maximise the total (present and future) value of a weighted sum of profits and aggregate gross patient utility. First-period quality levels result in first-period demands, with $D_1^A + D_1^B = 1$. Second-period quality levels and payoffs depend on these demands, which fully capture the outcome of the first period. To take into account this dependence, we solve the game backwards for a pure-strategy

[7] These positive returns to hospital volume are generally attributed to learning-by-doing or quality-enhancing scale economies, which capture the idea that healthcare providers become increasingly efficient as the number of times they perform a certain procedure rises. For recent empirical evidence of volume-outcome effects, see Avdic et al. (2019).

subgame-perfect Nash equilibrium.

## 4.1 The second period

Consider the different groups of patients in turn. A fraction $\lambda$ of patients were not in the market in the first period and are not therefore tied to any of the hospitals. The new patient who is indifferent between seeking treatment at Hospital A and Hospital B is located at $\hat{x}$, given by

$$\hat{x} = \frac{1}{2} + \frac{q_2^A - q_2^B}{2\tau}. \tag{5}$$

Hospitals A and B serve respectively $\lambda\hat{x}$ and $\lambda(1-\hat{x})$ of these patients. Additionally, Hospital A serves all of these patients if $q_2^A > q_2^B + \tau$ and none if $q_2^A < q_2^B - \tau$.

A fraction $\mu D_1^A$ of patients sought treatment from Hospital A in the first period and now have preferences for treatment characteristics that are uniformly distributed along the line segment $[0,1]$. The patient who was previously treated at Hospital A and is now indifferent between seeking treatment at Hospital A and Hospital B is located at $\hat{x}_{|A}$, given by

$$\hat{x}_{|A} = \frac{1}{2} + \frac{q_2^A - q_2^B + s}{2\tau}. \tag{6}$$

Hospitals A and B serve respectively $\mu D_1^A \hat{x}_{|A}$ and $\mu D_1^A(1-\hat{x}_{|A})$ of these patients. Additionally, Hospital A serves all of these patients if $q_2^A > q_2^B + \tau - s$ and none if $q_2^A < q_2^B - \tau - s$.

Similarly, a fraction $\mu D_1^B$ of patients sought treatment from Hospital B in the first period and now have preferences for treatment characteristics that are uniformly distributed along the line segment $[0,1]$. The patient who was previously treated at Hospital B and is now indifferent between seeking treatment at Hospital A and Hospital B is located at $\hat{x}_{|B}$, given by

$$\hat{x}_{|B} = \frac{1}{2} + \frac{q_2^A - q_2^B - s}{2\tau}. \tag{7}$$

Hospitals A and B serve respectively $\mu D_1^B \hat{x}_{|B}$ and $\mu D_1^B(1-\hat{x}_{|B})$ of theses patients. Additionally, Hospital A serves all of these patients if $q_2^A > q_2^B + \tau + s$ and none if $q_2^A < q_2^B - \tau + s$.

Finally, the remaining fractions $(1-\lambda-\mu)D_1^A$ and $(1-\lambda-\mu)D_1^B$ of the patients choose, respectively, Hospital A and Hospital B in both periods. Combining demand from the three types of patients, it may be easily shown that total demand facing Hospital $i$ in the second period is given by

$$D_2^i(q_2^i, q_2^j) = \frac{\lambda+\mu}{2\tau}(\tau + q_2^i - q_2^j) + \frac{\mu}{2\tau}(D_1^i - D_1^j)s + (1-\lambda-\mu)D_1^i, \quad i,j = A,B; \quad i \neq j; \tag{8}$$

provided that $|q_2^A - q_2^B| < \tau - s$.[8]

Taking first-period demand as given, Hospital $i$ maximises

$$\Omega_2^i(q_2^i, q_2^j) = T + [p + (\alpha - c)q_2^i]D_2^i(q_2^i, q_2^j) - \frac{\gamma}{2}(q_2^i)^2, \quad i,j = A,B; \quad i \neq j; \tag{9}$$

where $p := \tilde{p} - k + \alpha v$. Maximisation of (9) with respect to $q_2^i$ yields the candidate equilibrium quality levels

$$q_2^{i*} = \frac{p + (\alpha-c)\left[\tau - \frac{\mu s}{\lambda+\mu}\right] - (\alpha-c)^2\phi}{\frac{2\tau\gamma}{\lambda+\mu} - (\alpha-c)} + (\alpha-c)\phi D_1^i, \quad i = A,B, \tag{10}$$

where

$$\phi := \frac{2\tau(1-\lambda-\mu+\frac{\mu s}{\tau})}{(\lambda+\mu)\left[\frac{2\tau\gamma}{\lambda+\mu} - 3(\alpha-c)\right]} > 0. \tag{11}$$

The parameter restriction given in (4) ensures that the second-order condition is always satisfied, provided that (8) holds. However, this is insufficient to prove that the pair of strategies (10) define an equilibrium in the second-period subgame. It must be ensured that hospitals do not deviate and serve only their captive patients with fixed preferences, thus choosing a quality level outside the range in which (8) holds. As Klemperer (1987) notes, the deviation is not beneficial if $\lambda+\mu$ is large enough and the difference between first-period demands is sufficiently small. In the next section, we show that a symmetric pure-strategy candidate subgame perfect equilibrium exists and assume $\lambda+\mu$ is such that it indeed is an equilibrium.

Applying symmetry ($D_1^A = D_1^B = 1/2$), equilibrium quality in the second period becomes

$$q_2^* = \frac{p + \frac{(\alpha-c)\tau}{\lambda+\mu}}{\frac{2\tau\gamma}{\lambda+\mu} - (\alpha-c)}. \tag{12}$$

Before turning to the first-period subgame, one must take into account the inter-period dependence by analysing the effect of first-period demand on second-period payoffs. In a symmetric

[8] Switching only occurs in equilibrium if $s < \tau$, so that the preferences for treatment characteristics of some patients outweigh the switching cost.

equilibrium, it is given by

$$\frac{\partial \Omega_2^i(q_2^*)}{\partial D_1^i} = \phi \left( \frac{\lambda + \mu}{\tau} \right) \left( \frac{\tau \gamma}{\lambda + \mu} - \alpha + c \right) [p + (\alpha - c) q_2^*] > 0, \quad i = A, B. \tag{13}$$

Because the marginal patient is always beneficial to treat in the second period ($p + (\alpha - c)q_2^* > 0$), first-period demand has an unambiguously positive effect on second-period payoffs. This gives hospitals an additional incentive to invest in quality in the first period and attract demand, since it will be partially locked-in.

## 4.2 The first period

Anticipating the effect of first-period quality choices in the second period, hospitals maximise the present value of total payoffs. Formally, Hospital $i$ maximises

$$\sum_{t=1}^{2} \Omega_t^i(q_1^i, q_1^j) = T + [p + (\alpha - c) q_1^i] D_1^i(q_1^i, q_1^j) - \frac{\gamma}{2}(q_1^i)^2 + \Omega_2^i[D_1^i(q_1^i, q_1^j)], \quad i, j = A, B; \quad i \neq j. \tag{14}$$

The first- and second-order conditions of the hospital's maximisation problem are respectively given by[9]

$$[p + (\alpha - c) q_1^i] \frac{\partial D_1^i}{\partial q_1^i} + (\alpha - c) D_1^i - \gamma q_1^i + \frac{\partial \Omega_2^i}{\partial D_1^i} \frac{\partial D_1^i}{\partial q_1^i} = 0 \tag{15}$$

and

$$\gamma - 2(\alpha - c) \frac{\partial D_1^i}{\partial q_1^i} > \left( \frac{\lambda + \mu}{\tau} \right) \left( \frac{\tau \gamma}{\lambda + \mu} - \alpha + c \right) \left[ (\alpha - c) \phi \frac{\partial D_1^i}{\partial q_1^i} \right]^2 + \left[ p + (\alpha - c) q_1^i + \frac{\partial \Omega_2^i}{\partial D_1^i} \right] \frac{\partial^2 D_1^i}{\partial (q_1^i)^2}, \tag{16}$$

where $i, j = A, B$ and $i \neq j$. Applying symmetry and using (13), first-period equilibrium quality, $q_1^*$, is implicitly defined by

$$\left[ p + (\alpha - c) q_1^* + \phi \left( \frac{\lambda + \mu}{\tau} \right) \left( \frac{\tau \gamma}{\lambda + \mu} - \alpha + c \right) [p + (\alpha - c) q_2^*] \right] \frac{\partial D_1^i}{\partial q_1^i} + \frac{\alpha - c}{2} = \gamma q_1^*. \tag{17}$$

The term in square brackets is the total payoff (present plus future) of treating an additional patient in the first period, and it is always positive in equilibrium. Because treating an additional patient is always beneficial, the incentive to invest in quality depends on how strongly first-period demand responds to quality changes. This response, as we show below, is determined by patient

[9] To save notation, we omit function arguments whenever there is no ambiguity.

expectations.

Let expected quality in the second period, $q_E^i(q_1^i, q_1^j)$, be functions of first-period quality levels, which are observable to patients, and consider the first-period choice of hospital of a patient who is located at $y$. In the first period, the patient's utility from choosing Hospital A is $(v + q_1^A - \tau y)$. In the second period, with probability $\lambda$, the patient is not in the market and has zero utility. With probability $\mu$, the patient remains in the market and has preferences for treatment characteristics uniformly distributed on $[0,1]$. Conditional on having volatile preferences and choosing Hospital A in the first period, the patient anticipates that, for a given second-period location $x$, he will choose Hospital A in the second period if $v + q_E^A - \tau x > v + q_E^B - \tau(1-x) - s$; or, equivalently, if $x < 1/2 + (q_E^A - q_E^B + s)/2\tau$. Conversely, the patient anticipates that he will choose Hospital B and incur the switching cost if $x$ exceeds that threshold. With probability $1 - \lambda - \mu$, the patient has persistent preferences (i.e., he is located at $y$ also in the second period) and will again choose Hospital A. Then, the expected total utility (first-period utility plus expected second-period utility) of the patient located at $y$ which results from choosing Hospital A in the first period is

$$(v + q_1^A - \tau y) + \mu \begin{bmatrix} \int_0^{\frac{1}{2}+\frac{q_E^A - q_E^B + s}{2\tau}} (v + q_E^A - \tau x)dx \\ + \int_{\frac{1}{2}+\frac{q_E^A - q_E^B + s}{2\tau}}^1 [v + q_E^B - \tau(1-x) - s]dx \end{bmatrix} + (1-\lambda-\mu)(v + q_E^A - \tau y). \quad (18)$$

Analogously, the expected total utility from choosing Hospital B in the first period is

$$[v + q_1^B - \tau(1-y)] + \mu \begin{bmatrix} \int_0^{\frac{1}{2}+\frac{q_E^A - q_E^B - s}{2\tau}} (v + q_E^A - \tau x - s)dx \\ + \int_{\frac{1}{2}+\frac{q_E^A - q_E^B - s}{2\tau}}^1 [v + q_E^B - \tau(1-x)]dx \end{bmatrix} + (1-\lambda-\mu)[v + q_E^B - \tau(1-y)]. \quad (19)$$

Equating (18) and (19) implicitly defines the location of the patient who is indifferent between the two hospitals. Using the fact that this patient has $y = D_1^A(q_1^A, q_1^B)$, we solve for first-period demands

$$D_1^A(q_1^A, q_1^B) = \frac{1}{2} + \frac{q_1^A - q_1^B}{2\tau(2-\lambda-\mu)} + \left[\frac{1-\lambda-\mu+\frac{\mu s}{\tau}}{2\tau(2-\lambda-\mu)}\right](q_E^A - q_E^B) \quad (20)$$

and $D_1^B = 1 - D_1^A$, yielding

$$\frac{\partial D_1^i(q_1^i, q_1^j)}{\partial q_1^i} = \frac{1}{2\tau(2-\lambda-\mu)} + \left[\frac{1-\lambda-\mu+\frac{\mu s}{\tau}}{2\tau(2-\lambda-\mu)}\right]\frac{\partial(q_E^i - q_E^j)}{\partial q_1^i}, \quad i,j = A,B; \quad i \neq j. \quad (21)$$

Thus, demand responsiveness to quality in the first period depends in part on patients' expectations of how a unilateral quality increase affects the quality difference between the hospitals in the next period. In the following we will consider three different assumptions regarding patient expectations:

***(i) Myopic patients.*** If patients are myopic, they fully ignore the second period when making their first-period choice of hospital. Their decisions are therefore only based on observable first-period variables (qualities and travelling distance).

***(ii) Forward-looking but naïve patients.*** In this case, patients take the second period into account when making their first-period choice of hospital, anticipating the lock-in effect of switching costs and that their preferences may change, but fail to properly assess the evolution of quality. Specifically, given the complexity of evaluating hospital quality and, in particular, how future quality depends on current demand and hence quality, naïve patients resort to the rule-of-thumb of expecting that quality is the same in both periods.

***(iii) Forward-looking and rational patients.*** In this case, patients have rational expectations and correctly anticipate how quality investments today affect each hospital's incentives for quality investments in the future.

## 5 Patient expectations and quality provision

In this section, we analyse how the different types of patient expectations affect each hospital's incentives for quality provision. We do so by deriving the demand responsiveness to quality, (21), under each of our three assumptions regarding patient expectations. We then proceed by performing a ranking of equilibrium quality levels based on these expectations. Notice that patient expectations have no effect on the second-period decisions, which allows us to focus only incentives for quality provision in the first period.

### 5.1 Myopic patients

If patients are myopic and ignore the future, demand responsiveness to quality is the same as it would be if all patients leave the market after the first period (i.e., $\lambda = 1$ and $\mu = 0$), which implies that (21) reduces to

$$\frac{\partial D_1^i(q_1^i, q_1^j)}{\partial q_1^i} = \frac{1}{2\tau}, \quad i, j = A, B; \quad i \neq j. \tag{22}$$

Thus, with myopic patients, demand responsiveness to quality is the same as in a static version of the model and demand inertia plays no role.[10]

## 5.2 Forward-looking but naïve patients

If patients expect first-period quality to prevail in the second period, this implies that $\partial(q_E^i - q_E^j)/\partial q_1^i = 1$, which in turn implies that (21) reduces to

$$\frac{\partial D_1^i(q_1^i, q_1^j)}{\partial q_1^i} = \frac{1}{2\tau}\left[1 + \frac{\mu s}{\tau(2-\lambda-\mu)}\right], \quad i,j = A,B; \quad i \neq j. \tag{23}$$

Compared with the case of myopic patients, the presence of patients with naïve expectations introduces three additional effects on the demand responsiveness to quality. First, patients anticipate that they will also need treatment in the second period, thus having to 'travel' twice. This makes quality relatively less important than travelling/mismatch costs and leads, all else equal, to lower demand responsiveness to quality. This effect, however, is counteracted by the effect of patients' naïvety, since they expect a marginal change in quality to persist in the future; i.e., the benefit of higher quality is also 'counted twice'. In the absence of switching costs, these two effects cancel each other. In other words, $\partial D_1^i(q_1^i, q_1^j)/\partial q_1^i = 1/2\tau$ if $s = 0$, regardless of whether patients are myopic or forward-looking but naïve.

However, the presence of switching costs introduces a third effect that makes demand more responsive to quality if patients are forward-looking but naïve. More precisely, the presence of switching costs increases the relative importance of expected quality differences in the future. To illustrate this mechanism, consider the case of a marginal increase in first-period quality by Hospital A with $q_1^A > q_1^B$. While such a quality increase would increase demand for Hospital A, a patient located sufficiently close to Hospital B would still prefer to remain with that hospital, because the lower travelling costs outweigh the foregone quality improvement. However, if such a patient is forward-looking, he anticipates that, with probability $\mu$, his location on the line will not remain the same in the future, but will be randomly drawn from a uniform distribution. Since the expected value of a uniform distribution on $[0, 1]$ is $1/2$, and since the patient expects that first-period quality differences will persist in the second period, he consequently expects that, with probability $\mu$, his preferred choice of hospital in the future will be Hospital A and not Hospital B. However, since

[10] With myopic patients, although demand inertia plays no role in determining the demand responsiveness to quality, it still plays a role in determining the hospitals' incentives for quality provision, as can be seen from (17). The importance of demand inertia for equilibrium quality provision is analysed in Section 6.

$s > 0$ makes it costly to switch from the low-quality to the high-quality hospital in the future, the patient might find it preferable to choose Hospital A already today, and this choice is more likely the higher the switching costs. In other words, when patients are naïve and expect quality differences to persist, the presence of switching costs *increases* demand responsiveness to quality because of patients' fear of being locked-in to the 'wrong' hospital in the future.

### 5.3 Forward-looking and rational patients

If patients have rational expectations, they know that hospitals will set quality according to (10) and therefore anticipate that the quality difference in the second period will be

$$q_E^i - q_E^j = (\alpha - c)\phi[2D_1^i(q_1^i, q_1^j) - 1], \quad i,j = A,B; \quad i \neq j, \tag{24}$$

which implies

$$\frac{\partial(q_E^i - q_E^j)}{\partial q_1^i} = 2(\alpha - c)\phi \frac{\partial D_1^i(q_1^i, q_1^j)}{\partial q_1^i}, \quad i,j = A,B; \quad i \neq j. \tag{25}$$

Inserting (25) into (21) and solving for $\partial D_1^i(q_1^i, q_1^j)/\partial q_1^i$ yields[11]

$$\frac{\partial D_1^i(q_1^i, q_1^j)}{\partial q_1^i} = \frac{1}{2\tau(2 - \lambda - \mu) - 2(1 - \lambda - \mu + \frac{\mu s}{\tau})(\alpha - c)\phi} \gtrless \frac{1}{2\tau}, \quad i,j = A,B; \quad i \neq j. \tag{26}$$

Forward-looking and rational patients not only anticipate that they will be (partially or totally) tied to their first-period hospital but also correctly anticipate how quality investments in the present affect future quality. This implies that the responsiveness of demand to quality in the first period depends on two additional factors, namely *provider motivation* and *technology.* These two factors determine the relationship between demand and the marginal cost of quality provision for each hospital. More specifically, higher demand increases (reduces) the marginal cost of quality provision if $c > (<)\,\alpha$. Under rational expectations, this has important implications for how a change in the current quality difference between hospitals informs patients' beliefs about future quality differences. From (25) we see that a unilateral quality increase by Hospital $i$ will increase the expected quality difference between Hospital $i$ and Hospital $j$ in the future only if $\alpha > c$, and

[11] Positive demand responsiveness requires that

$$c > c_R := \alpha - \frac{2\tau\gamma}{3(\lambda + \mu) + \frac{2\left(1 - \lambda - \mu + \frac{\mu s}{\tau}\right)^2}{2 - \lambda - \mu}} \gtrless c_{\min}.$$

*reduce* the expected future quality difference otherwise. Furthermore, since $\partial \left( q_E^i - q_E^j \right) / \partial q_1^i$ is monotonically increasing in $\alpha$ and monotonically decreasing in $c$, it follows from (21) that the demand responsiveness to quality is also monotonically increasing in $\alpha$ and monotonically decreasing in $c$.

In order to illustrate the above stated mechanism, consider for example the case of profit-oriented hospitals and cost substitutability between quality and output, which implies $c > \alpha = 0$. In this case, if patients observe a unilateral quality increase by, say, Hospital A, they will rationally expect that the resulting shift in demand from Hospital B to Hospital A is going to increase the marginal cost of quality provision for Hospital A and reduce it for Hospital B, thus resulting in a weakening of Hospital A's incentives for quality provision in the future, and a corresponding strengthening of Hospital B's future incentives for quality provision, all else equal. Such expectations will make patients more reluctant to switch from Hospital B to Hospital A following a quality increase by the latter hospital, thus *reducing* the demand responsiveness to quality. The opposite logic obviously applies if $c < \alpha$.

Notice, however, that demand responsiveness with rational patients may be lower than with myopic patients, even in the case where higher demand reduces the marginal cost of quality provision (i.e., $c < \alpha$). In other words, patients may correctly anticipate that a marginal increase in the quality of Hospital $i$ will increase the future quality difference and still be less attracted by that increase than they would if they were myopic and ignored the future. A necessary condition for this to happen is that patients expect that the quality advantage of Hospital $i$ will decrease over time, i.e., that $\partial \left( q_E^i - q_E^j \right) / \partial q_1^i < 1$, which implies that quality becomes relatively less important than travelling/mismatch costs for forward-looking patients.[12]

### 5.4 The effect of patient expectations on equilibrium quality

We are now ready to summarise the effect of patient expectations on equilibrium quality provision. From (17), we know that equilibrium quality is increasing in demand responsiveness and that this is the only channel through which patient expectations influence quality provision. Therefore, to establish under which type of expectations quality is higher, it suffices to compare the magnitudes of the demand responsiveness. We have shown that demand is more responsive to quality when

[12] Recall that forward-looking patients anticipate that theyy may have to 'travel' twice, which makes quality relatively less important than travelling/mismatch costs and contributes to lower demand responsiveness. Only if the future quality difference is sufficiently large, will demand responsiveness be higher than when patients are myopic.

patients are forward-looking but naïve than when patients are myopic, implying that quality is higher in the former case.

Depending on how much a first-period quality increase is offset in the second period, demand responsiveness (and hence quality) when patients are rational may be lower than when patients are myopic, higher than when patients are naïve, or lie in between. Recall that, with rational patients, demand responsiveness is monotonically decreasing in $c$. If a first-period quality increase has no effect on the expected second-period quality difference (i.e., if $c = \alpha$), demand responsiveness is lower with forward-looking and rational patients than if patients are either myopic or naïve. Demand will be more responsive to quality under rational expectations only if a current unilateral quality increase produces a sufficiently large increase in the future expected quality difference between the hospitals. This requires sufficiently weak cost substitutability (or sufficiently strong cost complementarity).

The above analysis is summarised as follows.

**Proposition 1** *(i) If patients are forward-looking but naïve, equilibrium quality is always higher than if patients are myopic. (ii) Provided that the cost function is sufficiently convex in quality, if patients are forward-looking and rational, equilibrium quality is*

1. *higher than if patients are naïve if*

$$c < c' := \alpha - \frac{2\tau\gamma}{\frac{2(1-\lambda-\mu+\frac{\mu s}{\tau})(2-\lambda-\mu+\frac{\mu s}{\tau})}{(2-\lambda-\mu)} + 3(\lambda+\mu)}; \tag{27}$$

2. *lower than if patients are myopic if*

$$c > c'' := \alpha - \frac{2\tau\gamma}{\frac{2(1-\lambda-\mu+\frac{\mu s}{\tau})^2}{(1-\lambda-\mu)} + 3(\lambda+\mu)}; \tag{28}$$

*where* $\max\{c_{min}, c_R\} < c' < c'' < \alpha$.

**Proof.** Follows directly from a comparison of (22), (23) and (23). A sufficiently high $\gamma$ ensures that the second-order condition in (16) is satisfied for values of $c$ such that the set $(\max\{c_{\min}, c_R\}, c')$ is non-empty. ■

# 6 The effect of demand inertia on quality provision

In this section, we investigate how demand inertia affects incentives for quality provision. Our benchmark case of no demand inertia may be derived by setting (i) $\lambda = 0$, $\mu = 1$ and $s = 0$; or (ii) $\lambda = 1$ and $\mu = 0$. Although analytically equivalent, (i) and (ii) have different interpretations. In the former case, no patient leaves the market and there are no switching costs, but the preferences of all patients are reshuffled after the first period. In the latter case, all patients are replaced between periods, and hence there is no switching. In either case, there is no interaction between periods, patients' choices of hospital are independent, and demand is unaffected by expectations. This also illustrates that the role of patient expectations is unavoidably linked to the presence of demand inertia. Our choice of benchmark, thus, allows the analysis in this section to be interpreted as an analysis of the effect of patient expectations relative to a market wherein they play no role.

The first-order condition defining the symmetric equilibrium quality level in a market without demand inertia is given by

$$\frac{1}{2\tau}[p + (\alpha - c)q^N] + \frac{\alpha - c}{2} = \gamma q^N, \tag{29}$$

yielding

$$q^N = \frac{p + (\alpha - c)\tau}{2\tau\gamma - (\alpha - c)}. \tag{30}$$

Since the absence of demand inertia implies that equilibrium quality provision is equal in both periods, it is not immediately clear how a comparison with a model where equilibrium quality provision might differ over time should be interpreted. However, notice that equilibrium quality without demand inertia is higher than second-period quality provision in the presence of demand inertia; i.e., $q^N > q_2^*$. Our analytical strategy will therefore be to characterise under which conditions this inequality also holds with respect to first-period quality provision (i.e., $q^N > q_1^*$). If $q_1^* < q^N$, we can conclude that the presence of demand inertia unambiguously leads to a lower quality provision.

Comparing the first-order conditions (17) and (30), we see that there are two additional effects influencing quality provision in a market with demand inertia. First, there is a *competition effect*, given by the third term in square brackets on the left-hand side of (17). Since first-period demand is always valuable in the second period, hospitals have incentives to invest in quality to build market share. All else equal, the competition effect always leads to higher quality. Second, there is a *patient foresight effect* affecting demand responsiveness, which, in turn, determines how effective a quality increase is in attracting demand. In general, the foresight effect may either reinforce or counteract the competition effect, depending on whether patients' expectations about the second period lead to higher or lower demand responsiveness relative to a market without inertia.

Combining the two equilibrium conditions, we obtain, after some manipulations,

$$\left[\gamma-(\alpha-c)\frac{\partial D_1^i}{\partial q_1^i}\right](q_1^*-q^N)=\left(\frac{\partial D_1^i}{\partial q_1^i}-\frac{1}{2\tau}\right)[p+(\alpha-c)q^N]+\frac{\phi}{\tau}[\tau\gamma-(\lambda+\mu)(\alpha-c)][p+(\alpha-c)q_2^*]\frac{\partial D_1^i}{\partial q_1^i}. \tag{31}$$

Notice from (21) that demand responsiveness to quality in a market without inertia is equal to $1/2\tau$. The left-hand side of (31) is monotonic in $q_1^*$ and $q^N$, and the second-period second-order condition ensures that the term in square brackets is positive.[13] Consequently, $q_1^* < q^N$ if the right-hand side of (31) is negative, which requires that

$$\frac{\frac{1}{2\tau}-\frac{\partial D_1^i}{\partial q_1^i}}{\frac{\partial D_1^i}{\partial q_1^i}} > \frac{(\phi/\tau)[\tau\gamma-(\lambda+\mu)(\alpha-c)][p+(\alpha-c)q_2^*]}{p+(\alpha-c)q^N}. \tag{32}$$

The above inequality shows that quality is lower than in a market without inertia if the foresight effect (given by the left-hand side) more than compensates for the competition effect (given by the right-hand side), which requires that demand responsiveness is sufficiently lower than in a market without inertia (i.e., sufficiently lower than $1/2\tau$). More specifically, equilibrium quality is lower than in the benchmark if the difference in demand responsiveness—which measures the difference in the effectiveness of a quality increase in attracting patients—exceeds the relative payoff of demand—which measures how beneficial that increase is in future terms.[14]

We state the comparison of quality provision between markets with and without demand inertia in the following proposition.

**Proposition 2** *Under demand inertia, equilibrium quality is lower than in the benchmark case of a market without inertia if the following three conditions are all satisfied:*

*(i) patients are forward-looking and rational,*

[13] Under all of the three types of patient expectations considered, the second-order condition in the first period simplifies to

$$\gamma > 2(\alpha-c)\frac{\partial D_1^i}{\partial q_1^i}+\left(\frac{\lambda+\mu}{\tau}\right)\left(\frac{\tau\gamma}{\lambda+\mu}-\alpha+c\right)\left[(\alpha-c)\phi\frac{\partial D_1^i}{\partial q_1^i}\right]^2.$$

[14] Notice that by 'relative payoff of demand' we refer to the increase in second-period payoffs from treating an additional patient in the first period expressed in terms of the increase in payoffs from treating an additional patient in a market without inertia.

*(ii) c is above a unique threshold in $(\alpha, \infty)$, implicitly defined by*

$$\frac{\frac{1}{2\tau} - \frac{\partial D_1^i}{\partial q_1^i}}{\frac{\partial D_1^i}{\partial q_1^i}} = \frac{(\phi/\tau)[\tau\gamma - (\lambda+\mu)(\alpha - c)][p + (\alpha - c)q_2^*]}{p + (\alpha - c)q^N}, \tag{33}$$

*where $\partial D_1^i / \partial q_1^i$ is given by (26), and*

*(iii) the parameters determining the degree of demand inertia satisfy the following condition:*

$$\tau (\lambda + \mu) (\tau (1 - \lambda - \mu) - 4s\mu) + 2s\mu (\tau + s\mu) > 0. \tag{34}$$

**Proof.** See Appendix A. ■

Notice first that the presence of demand inertia can only lead to lower quality provision if patients have rational expectations. Since myopic patients fully ignore the second period, first-period demand responsiveness when patients are myopic is the same as in a market without inertia, which implies that the foresight effect vanishes and quality provision is higher than in the benchmark due to the competition effect. With forward-looking but naïve patients, demand is more responsive than in a market without inertia, which implies that the foresight effect is positive and hence *reinforces* the competition effect.

Since the demand responsiveness may fall below $1/2\tau$ only in case of rational expectations, this is a necessary but not sufficient condition for quality to be lower than in the benchmark. According to Proposition 2, two more conditions are needed. First, the degree of cost substitutability needs to be sufficiently strong relative to the degree of provider motivation to ensure that the foresight effect is sufficiently strong (cf. Proposition 1). To grasp why, recall that only if a first-period unilateral quality increase yields a sufficiently large decrease in the second-period quality difference, will demand responsiveness be low enough. In addition, the demand inertia parameters need to satisfy the condition given by (34). It is easily seen that this condition is always satisfied if the switching costs are sufficiently low (i.e., if $s$ is sufficiently close to zero). Notice that, for $c > \alpha$, lower switching costs contribute to reducing both the foresight effect and the competition effect. It reduces the foresight effect because it reduces the cost of being locked-in to the 'wrong' hospital *in the second period*, thus increasing the demand responsiveness to quality in the first period. But it also reduces the competition effect because it weakens the hospitals' ability to lock in patients by offering higher quality in the first period. However, it turns out that the reduction in the competition effect is larger than the reduction in the foresight effect, which explains why the third condition in Proposition 2 holds for sufficiently low values of $s$.

## 7 The effect of switching costs on quality

In this section, we take a more policy-oriented perspective and investigate how expectations affect the impact on quality of a policy intervention aimed at facilitating switching, which we measure by a reduction in $s$. Switching may be facilitated, for example, by the adoption of a market-wide network of shareable Electronic Health Records, allowing patients to transfer their medical records between providers easily, or by the publication of quality indicators in the public domain by regulators, which reduces patients' uncertainty associated with trying an alternative provider. Since neither patient expectations nor switching costs affect second-period quality levels in a symmetric equilibrium, we again focus on the first period.

Implicit differentiation of (17) yields

$$\frac{\partial q_1^*}{\partial s} = \frac{\left[\begin{array}{c} \left(\frac{\lambda+\mu}{\tau}\right)\left(\frac{\tau\gamma}{\lambda+\mu} - \alpha + c\right)[p + (\alpha - c)q_2^*]\frac{\partial D_1^i}{\partial q_1^i}\frac{\partial \phi}{\partial s} \\ + \left[p + (\alpha - c)q_1^* + \phi\left(\frac{\lambda+\mu}{\tau}\right)\left(\frac{\tau\gamma}{\lambda+\mu} - \alpha + c\right)[p + (\alpha - c)q_2^*]\right]\frac{\partial^2 D_1^i}{\partial q_1^i \partial s} \end{array}\right]}{\gamma - (\alpha - c)\frac{\partial D_1^i}{\partial q_1^i}}, \tag{35}$$

where

$$\frac{\partial \phi}{\partial s} = \frac{2\mu}{(\lambda + \mu)\left[\frac{2\tau\gamma}{\lambda+\mu} - 3(\alpha - c)\right]} > 0. \tag{36}$$

Lower switching costs generally have a twofold effect on quality. First, because fewer patients will be locked-in when switching is less costly, lower switching costs reduce the benefit of a marginal increase in first-period quality in terms of second-period payoffs. Thus, lower switching costs unambiguously dampen the competition effect, which, all else equal, leads to lower quality. Second, the effect of lower switching costs on demand responsiveness—and hence on the extent to which quality is effective in attracting demand—depends on the type of patient expectations. *A priori*, these two effects may either reinforce or counteract each other; however, it immediately follows that lower switching costs will lead to higher quality only if they make demand sufficiently more elastic.

Myopic patients ignore that they will be (at least partially) locked-in to their first-period provider and only take into account observable variables that affect their first-period utility when choosing a hospital. This implies that demand responsiveness is unaffected by switching costs and, in turn, that the change in quality is uniquely determined by the weakened competition effect. Therefore, lower switching costs unambiguously lead to lower quality when patients are myopic.

While forward-looking but naïve patients anticipate the lock-in effect of switching costs, they expect quality to remain constant. Since these patients expect a unilateral quality increase to yield a long-lasting quality difference, the less locked-in they anticipate to be, the less attracted they are by such an increase. A lower $s$ implies that 'correcting' the first-period choice of hospital in the second period is less costly, which implies that lower switching costs reduce the relative importance of (present and future) quality differences. In other words, from the perspective of naïve patients, lower switching costs reduce the benefit of being locked-in to the 'right' hospital (cf. Section 5.2). This leads to lower demand responsiveness and reinforces the effect of the weaker incentives to invest in quality in terms of second-period payoffs. Thus, lower switching costs also lead to lower quality when patients are forward-looking but naïve.

When patients have rational expectations, provider motivation and technology again play a role. More specifically, the effect of switching costs on demand responsiveness depends on whether a unilateral quality increase today increases or reduces the quality difference in the future, which in turn depends on the sign of $(\alpha - c)$. Using (26), we derive

$$\frac{\partial^2 D_1^i}{\partial q_1^i \partial s} = 4\mu(\alpha - c)\frac{\phi}{\tau}\left(\frac{\partial D_1^i}{\partial q_1^i}\right)^2 \gtrless 0. \tag{37}$$

Inserting (37) into (35) yields

$$\frac{\partial q_1^*}{\partial s} = \frac{\phi}{\tau}\frac{\partial D_1^i}{\partial q_1^i}\left[\frac{\mu}{\gamma - (\alpha - c)\frac{\partial D_1^i}{\partial q_1^i}}\right]\left[\begin{array}{c} \frac{[\tau\gamma-(\lambda+\mu)(\alpha-c)][p+(\alpha-c)q_2^*]}{\tau\left(1-\lambda-\mu+\frac{\mu s}{\tau}\right)} \\ +4(\alpha - c)\left(\gamma q_1^* - \frac{\alpha-c}{2}\right) \end{array}\right] \gtrless 0. \tag{38}$$

If a first-period quality increase by Hospital $i$ increases the expected quality difference between Hospital $i$ and Hospital $j$ in the second period (i.e., if $c < \alpha$), lower switching costs reduce demand responsiveness. The intuition for this result is similar to that of the case of naïve patients. The less locked-in patients anticipate to be, the less attracted they are by a quality difference that carries over into the future, since adjusting their choices in the second period is less costly. Therefore, the two above-mentioned effects go in the same direction, and lower switching costs again lead to lower quality.

If patients instead expect that a marginal increase in first-period quality by Hospital $i$ will be overturned in the second period, thus leading to a future *reduction* in the quality difference between Hospital $i$ and Hospital $j$, weaker lock-in makes patients *more* sensitive to quality in the first period. This happens when $c > \alpha$. In this case, rational patients know that a first-period quality increase by one hospital will increase the marginal cost of quality at that hospital, which implies that the quality difference between the two hospitals will decrease over time. All else equal, when switching is less costly, patients may take advantage of such differences by choosing the hospital that offers higher quality in the first period and reversing their choice in the second period at a lower cost. This is why lower switching costs increase demand responsiveness in the first period, offsetting the weakened competition effect. If $c$ is initially such that a first-period unilateral quality increase produces a sufficiently large reduction in the future quality difference, then a reduction in switching costs increases the patients' scope for exploiting quality differences to an extent where the increase in demand elasticity dominates the reduction in the competition effect, leading to an increase in equilibrium quality provision.

We summarise the above results in the following proposition.

**Proposition 3** *Lower switching costs lead to lower quality if patients are myopic or forward-looking but naïve, but lead to higher quality if patients are rational and the degree of cost substitutability between quality and output is sufficiently high.*

**Proof.** See Appendix B. ■

## 8 Discussion and concluding remarks

In this paper, we argue that demand inertia and patient expectations are inextricable in hospital markets and investigate their combined effect on quality provision. We start by exploring the behaviour of three types of patients differing with respect to whether and how they anticipate the future. Myopic patients ignore the future entirely, forward-looking but naïve patients assume that hospital quality remains constant over time, whereas forward-looking and rational patients correctly foresee hospitals' strategic quality investments. Using this analysis, we show how patient expectations shape the responsiveness of demand for hospital care to quality and obtain three main results.

We find that, unless patients are rational and cost substitutability is sufficiently strong, quality provision is generally higher than in the benchmark of a market without inertia and, simultaneously, policies based on switching cost reductions are counterproductive. The co-existence of these two results is intuitive. If demand inertia leads to higher quality provision, weakening it by reducing switching costs is an ill-advised policy intervention. A closer inspection of our results, however, suggests that the link between demand inertia, patient expectations and quality is not that simple. For some parameter values, demand inertia leads to lower quality *and* lower switching costs are nonetheless counterproductive.[15] In this case, for intermediate degrees of cost substitutability, rational patients' foresight of a reduction in the future quality difference (brought about by a current unilateral quality increase) makes demand responsiveness low enough to induce hospitals to offer lower quality than in a market without inertia. This same future reduction in the quality difference, conversely, does not suffice to persuade patients to take advantage of the present and future quality differences by reversing their choices if switching costs fall, thereby making demand sufficiently more responsive and triggering higher quality provision.

It is our first main result, based on a quality ranking, whose implications are more far-reaching. By ranking quality provision according to the type of patient expectations, we reveal that quality is always higher when patients are naïve than when they are myopic, while the relative position of quality when patients are rational ranges from highest to lowest, depending on the hospitals' technology and motivation. Perhaps surprisingly, these findings are connected to the concept of 'behaviour hazard', defined as the misuse of healthcare and the ensuing welfare losses caused by departures from forward-looking and perfectly rational patient behaviour (Baicker et al., 2015). Such departures are now well documented in the literature (cf. Section 2), but the evidence on their impact on patient welfare is less conclusive. The overall reduction in healthcare utilisation generated by myopic behaviour when compared with fully forward-looking behaviour reported by Guo and Zhang (2019) is concentrated in elective and preventive care, with emergency care showing no response. As for the results of Dalton et al. (2020), whereas there is little difference between fully myopic and fully rational behaviour in terms of quantity, there is a significant change in the composition of drugs consumed. In conjunction, these pieces of evidence suggest that the effect of deviations from perfect rationality on patient welfare is generally ambiguous. While we do not study the misuse of healthcare, we do show that different types of patient expectations provide

[15] For example, $\lambda = 0.1$, $\mu = 0.4$, $\tau = 0.7$, $s = 0.5$, $p = 10$, $\gamma = 5$, and $\alpha = 1$.

contrasting incentives for hospitals to invest in the quality of care, which, in turn, affects patient welfare. In the symmetric equilibrium of our model, patient expectations affect aggregate patient utility uniquely through first-period quality. This implies that Proposition 1 is also a *ranking of patient welfare* according to the type of expectations and, consequently, that full rationality does not necessarily lead to better outcomes for patients.

Discussions of the role of rationality commonly focus on the idea that deviations from fully rational behaviour make consumers act not in their best interest and that firms may find it beneficial to exploit those deviations. Our results indicate that the reverse might as well hold in hospital markets. To illustrate this point, suppose first that the degree of cost substitutability/complementarity is such that a unilateral increase in current quality yields a relatively larger increase in the future quality difference; i.e., $\partial \left( q_E^i - q_E^j \right) / \partial q_1^i > 1$. In this case, both myopic and naïve patients are less sensitive to current quality than they would be if they were aware that the larger quality difference in the present foretells an even larger quality difference in the future. In other words, both myopic and naïve patients fail to comprehend the true impact of the current unilateral quality increase on their total utility, which makes demand from these types of patients less responsive to quality. Hospitals thus exploit the lower demand responsiveness to offer lower quality, and, as expected, these departures from rationality are detrimental to patient welfare. Conversely, if the degree of cost substitutability is such that a unilateral increase in current quality yields a reduction in the future quality difference, myopic and naïve patients are *more* sensitive to quality than their rational counterparts. Because rational patients foresee the reduction in the future quality difference and its effect on their total expected utility, they are less sensitive to quality than they would be if they ignored the future. Myopic and naïve patients, differently, are oblivious to the future quality reduction and hence overestimate the impact of the current quality increase on their total utility, which leads to higher demand responsiveness and induces hospitals to invest in quality. In this case, therefore, the departures from rationality insulate patients from inferior quality provision by hindering the hospitals' ability to exploit the otherwise lower demand responsiveness.

## References

[1] Abaluck, J., Gruber, J., and Swanson, A. (2018). Prescription drug use under Medicare Part D: A linear model of nonlinear budget sets. *Journal of Public Economics*, 164, 106-138. https://doi.org/10.1016/j.jpubeco.2018.05.005

[2] Aron-Dine, A., Einav, L., Finkelstein, A., and Cullen, M. (2015). Moral hazard in health insurance: Do dynamic incentives matter? *The Review of Economics and Statistics*, 97(4), 725-741. https://doi.org/10.1162/REST_a_00518

[3] Avdic, D., Lundborg, P., and Vikström, J. (2019). Estimating returns to hospital volume: Evidence from advanced cancer surgery. *Journal of Health Economics*, 63, 81-99. https://doi.org/10.1016/j.jhealeco.2018.10.005

[4] Baicker, K., Mullainathan, S., and Schwartzstein, J. (2015). Behavioral hazard in health insurance. *The Quarterly Journal of Economics*, 130(4), 1623-1667. https://doi.org/10.1093/qje/qjv029

[5] 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. https://doi.org/10.1111/j.1530-9134.2011.00319.x

[6] 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). Milano: Springer. https://doi.org/10.1007/978-88-470-5480-6_1

[7] Brekke, K. R., Siciliani, L., and Straume, O. R. (2011). Hospital competition and quality with regulated prices. *The Scandinavian Journal of Economics*, 113 (2), 444-469. https://doi.org/10.1111/j.1467-9442.2011.01647.x

[8] Brot-Goldberg, Z. C., Chandra, A., Handel, B. R., and Kolstad, J. T. (2017). What does a deductible do? The impact of cost-sharing on health care prices, quantities, and spending dynamics. *The Quarterly Journal of Economics*, 132(3), 1261-1318. https://doi.org/10.1093/qje/qjx013

[9] Dalton, C. M., Gowrisankaran, G., and Town, R. J. (2020). Salience, myopia, and complex dynamic incentives: Evidence from Medicare Part D. *The Review of Economic Studies*, 87 (2), 822-869. https://doi.org/10.1093/restud/rdz023

[10] Einav, L., Finkelstein, A., and Schrimpf, P. (2015). The response of drug expenditure to nonlinear contract design: Evidence from Medicare Part D. *The Quarterly Journal of Economics*, 130(2), 841-899. https://doi.org/10.1093/qje/qjv005

[11] Farbmacher, H., Ihle, P., Schubert, I., Winter, J., and Wuppermann, A. (2017). Heterogeneous effects of a nonlinear price schedule for outpatient care. *Health Economics*, 26(10), 1234-1248. https://doi.org/10.1002/hec.3395

[12] Gravelle, H., and Masiero, G. (2000). Quality incentives in a regulated market with imperfect information and switching costs: Capitation in general practice. *Journal of Health Economics*, 19(6), 1067-1088. https://doi.org/10.1016/S0167-6296(00)00060-6

[13] Guo, A., and Zhang, J. (2019). What to expect when you are expecting: Are health care consumers forward-looking? *Journal of Health Economics*, 67, 102216. https://doi.org/10.1016/j.jhealeco.2019.06.003

[14] Irace, M. (2018). Patient loyalty in hospital choice: Evidence from New York (Working Paper No. 2018-52). University of Chicago, Becker Friedman Institute for Economics. http://dx.doi.org/10.2139/ssrn.3223702

[15] Jung, K., Feldman, R., and Scanlon, D. (2011). Where would you go for your next hospitalization? *Journal of Health Economics*, 30(4), 832-841. https://doi.org/10.1016/j.jhealeco.2011.05.006

[16] Klemperer, P. (1987). The competitiveness of markets with switching costs. *The RAND Journal of Economics*, 18(1), 138-150. https://doi.org/10.2307/2555540

[17] Raval, D., and Rosenbaum, T. (2018). Why do previous choices matter for hospital demand? Decomposing switching costs from unobserved preferences. *The Review of Economics and Statistics*, 100(5), 906-915. https://doi.org/10.1162/rest_a_00741

[18] Sacks, N. C., Burgess Jr., J. F., Cabral, H. J., and Pizer, S. D. (2017). Myopic and forward looking behavior in branded oral anti-diabetic medication consumption: An example from Medicare Part D. *Health Economics*, 26(6), 753-764. https://doi.org/10.1002/hec.3355

[19] Shepard, M. (2016). Hospital network competition and adverse selection: Evidence from the Massachusetts Health Insurance Exchange (Working Paper No. 22600). National Bureau of Economic Research. http://dx.doi.org/10.3386/w22600

[20] 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. https://doi.org/10.1016/j.jedc.2013.05.002

[21] Villas-Boas, J. M. (2015). A short survey on switching costs and dynamic competition. *International Journal of Research in Marketing*, 32(2), 219-222. https://doi.org/10.1016/j.ijresmar.2015.03.001

## Appendix A: Proof of Proposition 2

The proof that $q_1^* > q^N$ when patients are myopic or forward-looking but naïve follows directly from equations (22), (23), and (32).

To establish the conditions under which $q_1^* < q^N$ when patients are forward-looking and rational, we use equations (26) and (30) to rewrite, after some manipulation, condition (32) as

$$\frac{[2\tau\gamma - (\alpha - c)]\,[\tau\gamma - (\lambda + \mu)\,(\alpha - c)]}{1 - \lambda - \mu + \frac{\mu s}{\tau}} > [2\tau\gamma - (\lambda + \mu)\,(\alpha - c)] \left[ \begin{array}{c} \frac{(1-\lambda-\mu)[2\tau\gamma - 3(\lambda+\mu)(\alpha-c)]}{2\left(1-\lambda-\mu+\frac{\mu s}{\tau}\right)^2} \\ -\,(\alpha - c) \end{array} \right]. \tag{A1}$$

Let $LHS\,(c)$ and $RHS\,(c)$ denote the left-hand and right-hand sides of the above inequality. It is straightforward to see that $LHS(c)$ and $RHS(c)$ are quadratic functions of $c$. From

$$\frac{\partial LHS\,(c)}{\partial c} = \frac{[1 + 2\,(\lambda + \mu)]\,\tau\gamma - 2\,(\lambda + \mu)\,(\alpha - c)}{1 - \lambda - \mu + \frac{\mu s}{\tau}} > 0 \tag{A2}$$

and

$$\frac{\partial RHS\,(c)}{\partial c} = \frac{(1 - \lambda - \mu)\,(\lambda + \mu)}{\left(1 - \lambda - \mu + \frac{\mu s}{\tau}\right)^2}\,[4\tau\gamma - 3\,(\lambda + \mu)\,(\alpha - c)] + 2\,[\tau\gamma - (\lambda + \mu)\,(\alpha - c)] > 0, \tag{A3}$$

we see that $LHS(c)$ and $RHS(c)$ are strictly increasing in $(c_{\min}, \infty)$.

Recall, from condition (32), that $LHS(c) < RHS(c)$ may only hold if $\partial D_1^i/\partial q_1^i < 1/2\tau$, which, in turn, requires that $c > c''$, with $c''$ given by equation (28) in Proposition 1. Then, because

$LHS(c)$ and $RHS(c)$ are strictly increasing and convex in $c$,

$$LHS(c'') - RHS(c'') = \frac{[2\tau\gamma - (\alpha - c'')]\,[\tau\gamma - (\lambda + \mu)\,(\alpha - c'')]}{1 - \lambda - \mu + \frac{\mu s}{\tau}} > 0 \tag{A4}$$

and

$$LHS(\alpha) - RHS(\alpha) = 2\tau\mu s \left( \frac{\gamma}{1 - \lambda - \mu + \frac{\mu s}{\tau}} \right)^2 > 0. \tag{A5}$$

$LHS(c) < RHS(c)$ may only be true if $c$ exceeds some unique threshold value in $(\alpha, \infty)$ *and* $\partial^2 LHS\,(c)\,/\partial c^2 < \partial^2 RHS\,(c)\,/\partial c^2$, which is true if the condition in (34) holds. The above mentioned threshold value is the unique solution to $LHS(c) = RHS(c)$ in $(c_{\min}, \infty)$.

Finally, note from (A1) that this solution is independent of $p$, as well as that $q_t^* > 0$ if $p$ is sufficiently high. Thus, the set of values of $c$ such that $q_1^* < q^N$ is non-empty and the symmetric pure strategy subgame perfect Nash equilibrium is characterised by an interior solution if $p$ is sufficiently high.

## Appendix B: Proof of Proposition 3

The proof that $\partial q_1^*/\partial s > 0$ when patients are myopic or forward-looking but naïve follows directly from (35), given (22), (23), and (36).

To prove that $\partial q_1^*/\partial s < 0$ if $c$ is sufficiently high and patients are forward-looking and rational in an interior solution, we proceed in two steps: $(i)$ we prove that positive equilibrium quality in the second-period subgame ensures that first-period equilibrium quality is also positive; $(ii)$ we prove that there is a set of values of $c$ such that $\partial q_1^*/\partial s < 0$ and equilibrium quality is positive in both periods provided that $p$ is sufficiently high.

Combining the first-order conditions defining first- and second-period equilibrium qualities and rearranging yields

$$\begin{aligned}\left[\gamma - (\alpha - c)\frac{\partial D_1^i}{\partial q_1^i}\right](q_1^* - q_2^*) = \left(\frac{\partial D_1^i}{\partial q_1^i} - \frac{\lambda + \mu}{2\tau}\right)[p + (\alpha - c)q_2^*]& \\ +\frac{\phi}{\tau}[\tau\gamma - (\lambda + \mu)(\alpha - c)][p + (\alpha - c)q_2^*]\frac{\partial D_1^i}{\partial q_1^i}.&\end{aligned} \tag{B1}$$

The left-hand side of (B1) is monotonic in $q_1^*$ and $q_2^*$, and the second-period second-order condition ensures that the term in square brackets is positive. Thus, $q_1^* > q_2^*$ if

$$\frac{\lambda+\mu}{2\tau} - \frac{\partial D_1^i}{\partial q_1^i} < (\phi/\tau)[\tau\gamma - (\lambda+\mu)(\alpha-c)]\frac{\partial D_1^i}{\partial q_1^i}. \tag{B2}$$

The above inequality is clearly satisfied under myopic and naïve patient expectations. Recall that $\partial D_1^i/\partial q_1^i > 1/2\tau$ under these two types of expectations and that the expression on the right-hand side of (B2) is always positive.

Using (26), (B2) is satisfied under rational expectations if

$$c > \alpha - \frac{2\tau\gamma}{3(\lambda+\mu)}\left[\frac{1+\left(1-\lambda-\mu+\frac{\mu s}{\tau}\right)-(\lambda+\mu)(2-\lambda-\mu)}{1+\frac{2}{3}\left(1-\lambda-\mu+\frac{\mu s}{\tau}\right)\left(\lambda+\mu-\frac{\mu s}{\tau}\right)-(\lambda+\mu)(2-\lambda-\mu)}\right] \tag{B3}$$

The term in square brackets is greater than 1, implying that the expression on the right-hand side of (B3) is below $c_{min}$. Thus, regardless of the type of patient expectations, $q_1^* > q_2^* \quad \forall \quad c > c_{min} \implies (q_2^* > 0 \implies q_1^* > 0)$. This concludes the proof of $(i)$.

Notice now that, given the second-period second-order condition and that $\partial D_1^i/\partial q_i^1 > 0$ for $c > c_R$, the sign of $\partial q_1^*/\partial s$ is uniquely determined by the sign of the last factor (in square brackets) in (38), which we now denote by $\sigma$. In addition, note that $\sigma < 0$ only holds for $c > \alpha$, given that, from the first-order condition defining first-period equilibrium quality, $\gamma q_1^* - (\alpha - c)/2 > 0$.

Let $\tilde{c} := \alpha + p(\lambda+\mu)/\tau$ denote the unique value of $c$ such that $q_2^* = 0$. Then,

$$\lim_{c\to\tilde{c}^-} \sigma = \frac{\gamma p + \left[\frac{(\lambda+\mu)p}{\tau}\right]^2}{1-\lambda-\mu+\frac{\mu s}{\tau}} - 2\left[\frac{(\lambda+\mu)p}{\tau}\right]^2 - \left[\frac{4(\lambda+\mu)\gamma p}{\tau}\right] q_1^*. \tag{B4}$$

A sufficient condition for $\lim_{c\to\tilde{c}^-} \sigma < 0$ is simply

$$\frac{\gamma p + \left[\frac{(\lambda+\mu)p}{\tau}\right]^2}{1-\lambda-\mu+\frac{\mu s}{\tau}} - 2\left[\frac{(\lambda+\mu)p}{\tau}\right]^2 < 0, \tag{B5}$$

which is true provided that

$$p > \frac{\tau^2\gamma}{(\lambda+\mu)^2\left[2\left(1-\lambda-\mu+\frac{\mu s}{\tau}\right)-1\right]}. \tag{B6}$$

Since $q_1^*$ is strictly increasing in $p$, it follows that $\lim_{c\to\tilde{c}^-} \sigma < 0$ and hence $\lim_{c\to\tilde{c}^-}(\partial q_1^*/\partial s) < 0$ if $p$ is sufficiently high. Then, by continuity of $\partial q_1^*/\partial s$ in $c$, there exists a non-empty set of values of $c$ contained in $(\alpha, \tilde{c})$ such that $\partial q_1^* / \partial s < 0$ and the symmetric pure strategy subgame perfect Nash equilibrium is characterised by an interior solution if $p$ is sufficiently high.

# *Most Recent Working Paper*

| | |
|---|---|
| NIPE WP 03/2020 | Luís Sá and Odd Rune Straume, Quality provision in hospital markets with demand inertia: The role of patient expectations, 2020 |
| NIPE WP 02/2020 | Rosa-Branca Esteves, Liu Qihong and Shuai, J., Behavior-Based Price Discrimination with Non-Uniform Distribution of Consumer Preferences, 2020 |
| NIPE WP 01/2020 | Diogo Teixeira and **J. Cadima Ribeiro,** “Residents’ perceptions of the tourism impacts on a mature destination: the case of Madeira Island”, 2020 |
| NIPE WP 17/2019 | Liao, R. C., **Loureiro, G.**, and Taboada, A. G., “Women on Bank Boards: Evidence from Gender Quotas around the World”, 2019 |
| NIPE WP 16/2019 | **Luís Sá,** “Hospital Competition Under Patient Inertia: Do Switching Costs Stimulate Quality Provision?”, 2019 |
| NIPE WP 15/2019 | **João Martins** and **Linda G. Veiga**, “Undergraduate students’ economic literacy, knowledge of the country’s economic performance and opinions regarding appropriate economic policies”, 2019 |
| NIPE WP 14/2019 | **Natália P. Monteiro, Odd Rune Straume** and **Marieta Valente,** “Does remote work improve or impair firm labour productivity? Longitudinal evidence from Portugal”, 2019 |
| NIPE WP 13/2019 | **Luís Aguiar-Conraria,** Manuel M. F. Martins and **Maria Joana Soares**, “ Okun’s Law Across Time and Frequencies”, 2019 |
| NIPE WP 12/2019 | Bohn, F., and **Veiga, F. J.,** “Political Budget Forecast Cycles”, 2019 |
| NIPE WP 11/2019 | **Ojo, M. O., Aguiar-Conraria, L**. and **Soares, M. J., “**A Time-Frequency Analysis of Sovereign Debt Contagion in Europe”, 2019 |
| NIPE WP 10/2019 | Lommerud, K. E., Meland, F. and **Straume, O. R.**, “International outsourcing and trade union (de-) centralization”, 2019 |
| NIPE WP 09/2019 | **Carvalho, Margarita** and **João Cerejeira**, “Level Leverage decisions and manager characteristics”,2019 |
| NIPE WP 08/2019 | **Carvalho, Margarita** and **João Cerejeira**, “Financialization, Corporate Governance and Employee Pay: A Firm Level Analysis”, 2019 |
| NIPE WP 07/2019 | **Carvalho, Margarita** and **João Cerejeira,** “Mergers and Acquisitions and wage effects in the Portuguese banking sector”, 2019 |
| NIPE WP 06/2019 | Bisceglia, Michele, Roberto Cellini, Luigi Siciliani and **Odd Rune Straume**, “Optimal dynamic volume-based price regulation”, 2019 |
| NIPE WP 05/2019 | Hélia Costa and **Linda Veiga**, “Local labor impact of wind energy investment: an analysis of Portuguese municipalities”, 2019 |
| NIPE WP 04/2019 | **Luís Aguiar-Conraria,** Manuel M. F. Martins and **Maria Joana Soares**, “ The Phillips Curve at 60: time for time and frequency”, 2019 |
| NIPE WP 03/2019 | **Luís Aguiar-Conraria,** Pedro C. Magalhães and Christoph A. Vanberg, "What are the best quorum rules? A Laboratory Investigation", 2019 |
| NIPE WP 02/2019 | **Ghandour, Ziad R**., "Public-Private Competition in Regulated Markets", 2019 |
| NIPE WP 01/2019 | **Alexandre, Fernando**, Pedro Bação and **Miguel Portela**, "A flatter life-cycle consumption profile", 2019 |
| NIPE WP 21/2018 | **Veiga, Linda**, Georgios Efthyvoulou and Atsuyoshi Morozumi, "Political Budget Cycles: Conditioning Factors and New Evidence", 2018 |
| NIPE WP 20/2018 | **Sá, Luís**, Luigi Siciliani e **Odd Rune Straume**, "Dynamic Hospital Competition Under Rationing by Waiting Times", 2018 |
| NIPE WP 19/2018 | Brekke, Kurt R., Chiara Canta, Luigi Siciliani and **Odd Rune Straume**, "Hospital Competition in the National Health Service: Evidence from a Patient Choice Reform", 2018 |
| NIPE WP 18/2018 | Paulo Soares Esteves, **Miguel Portela** and António Rua, "Does domestic demand matter for firms' exports?", 2018 |
| NIPE WP 17/2018 | **Alexandre, Fernando,** Hélder Costa, **Miguel Portela** and Miguel Rodrigues, "Asymmetric regional dynamics: from bust to recovery", 2018 |
| NIPE WP 16/2018 | **Sochirca, Elena** and Pedro Cunha Neves, "Optimal policies, middle class development and human capital accumulation under elite rivalry", 2018 |