NIPE Núcleo de Investigação em Políticas Económicas e Empresariais

#16
2019

WORKING PAPER

Luís Sá

“Hospital Competition Under Patient Inertia: Do Switching Costs Stimulate Quality Provision?”

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 «UID/ECO/03182/2019»

Universidade do Minho
Escola de Economia e Gestão

Hospital Competition Under Patient Inertia: Do Switching Costs Stimulate Quality Provision?[*]

Luís Sá[†]

December 2019

Abstract

Recent empirical evidence establishes previous use as a strong predictor of patient choice of hospital and indicates that switching costs explain a significant share of inertia in the hospital industry. In a model of competition between two semi-altruistic and horizontally differentiated hospitals with inherited demand, I investigate the effect of lower switching costs on quality provision and show that it depends on the hospitals' production technology and degree of altruism. If cost substitutability (complementarity) between quality and output is sufficiently weak (strong) relative to altruism, lower switching costs reduce quality at the high-volume hospital, average quality, and patient welfare. While milder patient preferences increase the scope for an increase in quality at both hospitals, it can only occur if hospitals are semi-altruistic. Finally, I show that the distribution of patients between hospitals matters. Even if hospital-level quality and patient welfare increase, lower switching costs may lead to lower average quality.

Keywords: Hospital competition; quality; switching costs; patient choice; volume-outcome effects; altruism.

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. I also thank FCT for the PhD Studentship SFRH/BD/129073/2017, financed by National Funds of the FCT and the European Social Fund (ESF). The comments and suggestions by Pedro Pita Barros, Lise Rochaix, and other participants in the 6th EuHEA PhD Student-Supervisor and Early Career Researcher Conference are greatly appreciated. Finally, a word of gratitude to Odd Rune Straume for the most valuable guidance.

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

1 Introduction

Free choice of hospital is becoming widespread. While, in the US, it has been a structural feature of the healthcare system in general and the hospital industry in particular, in Europe, where the sector is more tightly regulated, there is a general move towards the removal of constraints on the ability of patients to choose a hospital according to their preference (Siciliani et al., 2017). Although wider choice is increasingly common, there is one choice-related phenomenon whose implications have until recently received little or no attention from policymakers and from the literature on hospital choice and competition: the observation that patients tend to choose a hospital and repeatedly demand treatment from it regardless of whether the episodes of care are related. In other words, the idea that *patient inertia* (i.e., choice persistence or loyalty) exists in the hospital industry.

The premise underlying the benefits of free choice is that 'money will follow the patients', rewarding the more efficient providers. With patients often insulated from costs by third-party payers (i.e., private or social insurance and public provision of healthcare), competition in hospital markets is expected to play out through channels other than prices. Ranging from effectiveness of treatment to patient satisfaction, quality of care is arguably a key variable. Accordingly, the assumption underlying free choice—and supported by empirical evidence—is that patients are able to recognise quality and value it.[1]

Patient inertia then poses a question regarding quality provision. If patients are free to choose, able to recognise quality, value it, and are nonetheless strongly attached to a specific provider, what incentives do hospitals have to carry out costly quality investments? This paper addresses this question from a policy perspective: it investigates whether policies that reduce patient inertia play a useful role in improving quality provision. I present a model of hospital competition and show that a reduction in switching cost-induced inertia shifts demand from a high-volume to a low-volume hospital, thus affecting both the marginal cost and the marginal benefit from quality investments. Whether this leads higher or lower quality at either hospital depends crucially on hospital production attributes and the hospitals' degree of altruism (or, conversely, profit orientation).

[1] For example, Tay (2003) presents empirical evidence that quality is an important determinant of patient choice of hospital. Varkevisser et al. (2012) find that patients are sensitive to differences in quality as measured by public ratings and that hospitals with good reputation and low readmission rates attract more patients. Gutacker et al. (2016) report that patients choose hospitals that improve their self-reported health, although more conventional quality measures (readmission and mortality rates) are less important in determining patient choice of hospital. Relatedly, Gaynor et al. (2016) find that demand sensitivity to mortality rates increased substantially in England after the 2006 choice reform.

Evidence that patients are significantly more likely to demand treatment from a hospital they have previously visited is gradually emerging from studies on patient choice of hospital. Jung et al. (2011) estimate that the probability of a hospital being chosen for a future hospitalisation is 64 percentage points higher if the hospital was previously used. Shepard (2016) finds that patients are 5 times more likely to choose a hospital where they received outpatient care in the previous year. Raval and Rosenbaum (2018) report that the probability of a woman choosing a hospital for childbirth increases from 40% to 72% if she has previously given birth at that hospital. Irace (2018) finds that the hospital visited in the previous episode of care is 3.4 times more likely to be chosen for coronary artery bypass grafting (CABG) than an otherwise identical hospital. The two last-mentioned studies further show that patient inertia is explained by persistent unobserved patient heterogeneity (or persistent unobserved preferences) and state dependence. Unobserved heterogeneity denotes the case in which patients have strong and persistent preferences for a hospital that are generally unobservable to the empirical researcher. For patients with persistent horizontal preferences, repeated use of the same hospital is simply the utility maximising behaviour. State dependence, on the other hand, refers to the causal impact of past choices on current decisions.

When switching providers is costly, past use of a hospital affects the utility patients derive from treatment at different hospitals in the present and hence influence their current choice. There are several reasons why switching costs bring about state dependence in the context of patient choice of hospital.[2] First, patients incur monetary and time costs in order to transfer their medical records between providers. Second, some procedures are hospital-specific investments. Prior healthcare consumption may take the form of a previous investment in that patients undergo medical procedures which are intertemporally linked and might be rendered useless if the patient switches providers and treatment is restarted. This is the case, for instance, when a patient who started treatment at a different hospital is subjected to diagnostic tests at the new facility. Third, patients may find it optimal to repeatedly visit a hospital they have satisfactorily used in the past instead of risking an untested alternative. Assessing hospital quality is a demanding and complex task, and repeating past choices might be the optimal behaviour. Finally, patients may simply value an ongoing and close relationship with a provider. In this case, switching costs are the premium patients are willing to pay for familiarity with a given hospital, either in terms of a higher price or lower quality.[3]

[2] Raval and Rosenbaum (2018) in effect equate state dependence with switching costs.

[3] The interview-based study of Dutch patients' choice of hospital of Victoor et al. (2016) corroborates these hypotheses. Patients reported that knowledge of their medical history, trust in their physician, and familiarity as

Both switching costs and persistent preferences result in patient inertia, but the degree to which patient choice of hospital may be influenced by policy depends on the source of inertia being targeted. While policymakers have little or no influence over idiosyncratic preferences, there is arguably more scope for policy intervention if repeated use results from switching costs. To the extent that switching costs reflect constraints on patients' ability to adapt to the environment, policies that remove those constraints facilitate switching and reduce inertia.

The adoption of shareable Electronic Health Records (EHR), electronic records of an individual patient's history of contact with the healthcare system (Oderkirk, 2017), has the potential to reduce switching costs. If these are mainly due to the costs of transferring medical records between hospitals, EHR are likely to have a large impact. In a network of shared EHR, a patient's medical history can be readily retrieved even if the patient is visiting a hospital for the first time. Although to a lesser extent, shared EHR are likely to mitigate other forms of switching costs. The availability of test results from multiple sources increases the compatibility of treatment among providers by opening the possibility that patients are spared from duplicate procedures. Furthermore, patients may feel less uncertain about the effectiveness of treatment at a hospital they have not used before when their medical history is accessible, since the accuracy of diagnosis and the adequacy of treatment are generally increasing in the amount of information available. By the same token, patients may feel more familiar with healthcare professionals whom they have not contacted before if these professionals can easily learn their medical history.

Throughout the paper, a reduction in switching costs may be interpreted as the result of a policy based on the market-wide adoption of shareable EHR.[4] However, this is not the only possible interpretation. In the hospital competition literature that is based on models of spatial competition, travelling costs parameters are a standard measure of the degree of patient choice and competition intensity (see, for example, Brekke et al., 2011). This paper offers an alternative, perhaps more precise, measure. By specifically reflecting a situation wherein switching is facilitated, a reduction in switching costs may also be interpreted as the adoption of a broader scope of patient choice policies (Siciliani et al., 2017). For instance, information on quality made increasingly available in the public domain, a staple of choice policies, might reduce the uncertainty about quality at

some of the reasons why they sought treatment from the hospital they had previously used.

[4] A move towards the implementation of shareable EHR systems is already noticeable in several countries. According to Oderkirk (2017), 23 out of 28 surveyed OECD countries reported they were implementing or had implemented one country-wide EHR system in 2016. However, the potential of EHR systems to reduce switching costs is often overlooked, and emphasis is placed on their benefits to medical research and cost savings.

alternative hospitals.

To model the demand for healthcare faced by each provider, I use a Hotelling approach with two semi-altruistic hospitals located at each endpoint of the unit line segment. All patients have previously visited one of the two hospitals and are currently tied to that hospital. These patients form the *inherited* demand each hospital faces. Within my analytical framework, the modelling of patient inertia maps on the recent empirical evidence. A fraction of patients have persistent horizontal preferences (i.e., their current location on the unit line segment equals their inherited location), whereas the remaining patients have preferences that are newly drawn from a uniform distribution. Under regulated prices, patients choose a hospital based on their horizontal preferences, on the quality level offered by each hospital and, crucially, on the switching cost they incur if they demand treatment from the hospital they have not previously used.

I obtain several findings regarding the unintended effects of lower switching costs. The main mechanism through which quality provision is affected is related to the lock-in effect of switching costs when inherited demand is asymmetric. By reducing the number of locked-in patients, lower switching costs shift demand from the hospital with higher to the hospital with lower inherited demand, as patients switch in order to reduce the mismatch between their locations and that of the chosen hospital. The effect of this demand adjustment on quality provision at each of the two hospitals depends on the technology of production of hospital treatments—i.e., whether there is cost substitutability or complementarity between quality and output—and on the hospitals' degree of altruism. If there is cost substitutability, both the marginal cost and the marginal (altruistic) benefit from quality are increasing in current demand. When switching is facilitated, the marginal cost and the marginal benefit from quality decrease at the high-volume hospital (i.e., the hospital with higher inherited demand), whereas they increase at the low-volume one. If the degree of cost substitutability is sufficiently high relative to the degree of altruism, the change in the marginal cost dominates, and switching costs reductions are generally beneficial. Although quality may fall at the low-volume hospital, it unambiguously increases at the high-volume hospital, contributing to higher average quality and total patient welfare. If the degree of cost substitutability is sufficiently low relative to the degree of altruism instead, switching costs reductions may have more harmful effects. While quality is certain to increase at the low-volume hospital, lower switching costs lead to lower quality at the high-volume hospital if cost substitutability is sufficiently weak, hurting the majority of patients in the market. Consequently, average quality and patient welfare may also decrease. Importantly, these adverse effects may require cost complementarity between quality and output to arise. When hospital production exhibits this property, we are in the presence of *volume-outcome effects*, and the marginal cost of quality is decreasing in demand. In this case, besides reducing the marginal benefit from quality, lower switching costs increase the marginal cost at the high-volume hospital due to the demand loss, further contributing to lower quality at this hospital and hence to lower average quality and patient welfare.

Additionally, it is certain that lower switching costs will leave some patient worse off in terms of quality provision if hospitals are uniquely profit-oriented. However, quality may increase at both hospitals if they are semi-altruistic, and the scope for this simultaneous increase to occur is wider when fewer patients have persistent horizontal preferences or travelling costs are lower.

Finally, the relationship between hospital-level quality, average quality, and total patient welfare is not straightforward. Even if lower switching costs lead to higher quality at both hospitals, average quality may nonetheless fall due to a less efficient distribution of patients between hospitals. Likewise, lower switching may simultaneously increase patient welfare and reduce average quality.

The rest of the paper is organised as follows. The next section offers a brief overview of the literature and explains how this paper relates to it. In Section 3, I present the model, explain how patient inertia shapes demand, and discuss the assumptions underlying hospital preferences and production. In Section 4, the model is solved for the Nash Equilibrium, and equilibrium quality and demand are characterised. Section 5 investigates the effect of switching costs on hospital-level quality, average quality, and patient welfare. Finally, Section 6 offers concluding remarks and discusses policy implications.

2 Related Literature

This paper brings together two different strands of the literature. The first is the scarce but growing empirical literature on choice persistence in the hospital industry. To the best of my knowledge, Jung et al. (2011) were the first to look at patient-level inertia. They model a hypothetical choice of hospital for a surgical procedure of patients with a recent hospitalisation including a prior use indicator as a covariate. They find that previous use increases the probability of a hospital being chosen by 64 percentage points, which indicates the presence of strong choice persistence. Shepard (2016) studies adverse selection and moral hazard in healthcare plan choice when insurers compete on their networks of covered hospitals. In order to investigate whether patients with a propensity to choose high-quality, high-cost hospitals self-select into more generous plans, he first estimates a choice model which also includes a prior use indicator. Past use again emerges as a strong predictor of patient choice of hospital, increasing the probability of a hospital being chosen by five times to approximately 40%.

Past use coefficients capture both state dependence and persistent unobserved patient heterogeneity. Two recent studies with distinct approaches attempt to disentangle these two sources of patient inertia. Using data on choice of hospital for childbirth, Raval and Rosenbaum (2018) corroborate the earlier findings. When previous use is taken into account, the predicted share of women expected to return to a hospital increases from 40% to 72%. They then go on to investigate how much of this effect is due to switching costs and persistent unobserved heterogeneity. They estimate a choice model with hospital-patient fixed effects, which capture the effect of persistent preferences. This allows them to interpret the coefficient on the past use indicator as the switching cost. The inclusion of fixed effects roughly halves this coefficient, thus indicating the presence of both patient heterogeneity and switching costs. Using those estimates, they argue that switching costs account for approximately 40% of patient inertia. Differently, Irace (2018) makes use of quasi-exogenous shocks that induce patients to switch hospitals. He finds that patients who are admitted at a hospital they have never visited before during an emergency are more likely to choose that hospital in subsequent episodes of care than otherwise identical patients, which points to the presence of state dependence. Additionally, patients who return to the hospital they had been using before the emergency are more likely to choose that facility repeatedly, suggesting that unobserved heterogeneity also plays a role. Similarly, patients forced to try a new hospital during a temporary closure due to a natural disaster are less likely to return to the hospital they had been using than patients who did not seek hospital care during the closure. This too is indicative of state dependence.[5]

The second strand of the literature is that on theoretical models of hospital competition under regulated prices. Before turning to studies that specifically include some form of inertia in demand, consider the analysis of competition between semi-altruistic providers of Brekke et al. (2011). In a spatial model of hospital competition where patients choose a hospital based on the level of quality offered and their horizontal preferences, lower travelling costs increase demand responsiveness to quality changes. The effect on quality depends on whether the marginal patient is profitable to treat and on whether this effect reinforces or offsets the altruistic incentive to treat that patient.

[5] Specifically, this is indicative of first-order state dependence, meaning that the loyalty state of the patient is determined by the immediately preceding episode of care. If first-order state dependence is driven by switching costs, this implies that patients incur those costs if they switch to a hospital other than the one used in the preceding episode of care, even if they had visited that hospital before. The model I present below may indeed be interpreted as dealing with first-order state dependence.

If the degree of altruism is sufficiently strong, the marginal patient is so unprofitable to treat that the financial incentive to avoid her dominates, and quality falls in equilibrium.

Brekke et al. (2012) and Siciliani et al. (2013) investigate an information-related form of patient inertia. In both studies, demand adjusts sluggishly to changes in quality. Because healthcare quality is neither easily nor immediately observable, only a fraction of patients become aware of quality changes. Analytically, this implies that, at each point in time, only a fraction of any potential change in demand is realised. In a differential game with pure profit-maximising providers, Brekke et al. (2012) show that a reduced degree of sluggishness increases steady-state quality. The intuition for this result is simple. Less sluggish beliefs about quality make demand more responsive to quality changes. With a positive payment-cost margin, this gives providers incentives to increase quality. Siciliani et al. (2013) show that this result may be overturned if providers are semi-altruistic. Like in Brekke et al. (2011), the effect of less sluggish beliefs about quality depends on the financial and altruistic incentives to attract patients. If the per-treatment payment is sufficiently below unit costs, the former dominates, and less sluggishness leads to lower quality.

The effect of lower travelling costs, which may be interpreted as increased patient choice, and the effect of reduced demand sluggishness are qualitatively identical in these studies: they rely on the responsiveness of demand to quality. In my model, the mechanism through which facilitated switching affects quality is different. When switching costs fall, demand flows from the high- to the low-volume hospital. Because they depend on demand, both the marginal cost and the marginal benefit from quality change at each hospital in a way that is not related to demand responsiveness.

In a different institutional setting, Gravelle and Masiero (2000) analyse quality competition between horizontally differentiated, pure profit-maximising primary care providers in a two-period model. In the second period, patients may switch practices at a given common cost. They find that switching costs have no effect on quality, which is not surprising given the properties of their model. First, switching costs enter the demand functions additively and thus affect neither demand responsiveness to quality nor, consequently, the marginal revenue. Second, the marginal cost of quality is independent of demand. My model shares with theirs only the former feature. By adopting a more flexible cost function and considering semi-altruistic hospitals, both the marginal cost and the marginal benefit from quality depend on demand and hence on switching costs.[6]

[6] This does not always ensure that switching costs affect quality. Under symmetry, there are no demand differentials, and inherited demand advantages or disadvantages granted by switching costs cancel out. I argue that one particularly interesting effect of switching costs is that they affect market concentration, which requires an asymmetrically split market, and conduct my analysis accordingly.

3 The Model

Two hospitals, indexed $i = H, L$, are located at either endpoint of the unit line segment $[0, 1]$. Let Hospital H be located at 0 and Hospital L 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, for example, as the geographical space or the 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 specialty mix (i.e., the treatments and services) it offers.

Patients have a gross valuation of treatment $v > 0$ and demand a single unit of treatment from one of the hospitals. They are arrayed with unit density along the line segment and incur a travelling or mismatch cost τ per unit of distance between their location and that of the chosen hospital. Patients bear no out-of-pocket expenses either due to public provision of healthcare or to (social or private) health insurance coverage. Note that this last-mentioned feature is analytically equivalent to having hospitals charge the same regulated price. Patients derive utility from the quality of treatment, q_i, to which hospitals resort to attract demand. There is a lower bound $\underline{q}$ on treatment quality that represents the minimum quality hospitals are allowed to offer, with $q_i < \underline{q}$ being interpreted as malpractice. For simplicity, $\underline{q}$ is taken to be equal to zero. The gross valuation of treatment v is assumed to be high enough so that the market is always fully covered.[7]

I assume the following history of patient relationships with the two neighbouring hospitals: σ_H patients have visited Hospital H in the preceding episode of care, while the remaining $\sigma_L = 1 - \sigma_H$ patients have visited Hospital L. The two hospitals differ uniquely with respect to σ_i.

Patient inertia is modelled in the style of Klemperer (1987). A fraction μ of patients have preferences for treatment characteristics that are independent of the history of the game. 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 remaining $1 - \mu$ patients have unchanged preferences for treatment characteristics. That is, the location of these patients on the line segment equals their *past* location. Those who previously used Hospital H are uniformly distributed along $[0, \sigma_H]$, and those who previously used Hospital L are uniformly distributed along $[\sigma_H, 1]$. Patients who choose to demand treatment from the hospital they have not used in the preceding episode of care incur an exogenous switching cost s.[8] The model thus

[7] A sufficient condition for full market coverage is $v > \tau$.

[8] More realistically, one may conjecture that patients have different switching costs. The main feature upon which

maps on the empirical analyses of choice persistence of Raval and Rosenbaum (2018) and Irace (2018): both persistent preferences and switching costs (state dependence) induce inertia.

Since patients are tied to the hospitals, I refer to σ_i as Hospital i's inherited demand. It should be emphasised that I consider a one-period model and that the issue of how the inherited hospital-patient relationships are formed is not formally addressed.[9] One possible interpretation for asymmetric inherited demand is the case of a former local monopolist whose incumbency has not been eroded. Indeed, I show in Section 4 that patient inertia causes asymmetric market shares to persist even with otherwise identical hospitals. For clarity of exposition and without loss of generality, let $\sigma_H > \sigma_L$. Hence, Hospital H denotes the high-volume hospital (i.e., the hospital with higher inherited and current demand) and Hospital L denotes the low-volume hospital (i.e., the hospital with lower inherited and current demand).

3.1 Patient Utility and Demand

Consider the different groups of patients in turn. A fraction $\mu\sigma_H$ of patients sought treatment from Hospital H in the past 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 H and is now indifferent between seeking treatment at Hospital H and Hospital L is located at $\hat{x}_{|H}$, given by

$$v + q_H - \tau x = v + q_L - \tau(1 - x) - s \tag{1}$$

or, explicitly,

$$\hat{x}_{|H} = \frac{1}{2} + \frac{q_H - q_L + s}{2\tau}. \tag{2}$$

most of the subsequently derived results hinge—the fact that the hospital with larger inherited demand has a demand advantage—would indeed be present in a model with heterogeneous switching costs. The simpler formulation I adopt preserves that feature and additionally allows for a richer specification of horizontal patient preferences, while still keeping the analysis tractable.

[9] Following Klemperer (1995), I interpret this as a 'mature market', in which a patient's relationship with a hospital has already been built up. Multi-period switching cost models are common in the literature on price competition that analyses 'bargain-and-then-ripoffs' behaviour, whereby firms charge low prices early on to build a large market share and then exploit locked-in consumers by charging higher prices. More recently, single-period models have been used to study the implications for policymaking of firms having captive consumers in a variety of fields, rather than firms' incentives to engage in the above-mentioned behaviour, which resembles more closely the objective of this paper. For examples of such models, see Gehrig et al. (2011) and Shy and Stenbacka (2016).

Of these, hospitals H and L serve respectively $\mu\sigma_H\hat{x}_{|H}$ and $\mu\sigma_H(1-\hat{x}_{|H})$ patients. Additionally, Hospital H serves all of these patients if $q_H > q_L + \tau - s$ and none if $q_H < q_L - \tau - s$.

Similarly, a fraction $\mu\sigma_L$ of patients sought treatment from Hospital L in the past 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 L and is now indifferent between seeking treatment at Hospital H and Hospital L is located at $\hat{x}_{|L}$, given by

$$v + q_H - \tau x - s = v + q_L - \tau(1-x) \tag{3}$$

or, explicitly,

$$\hat{x}_{|L} = \frac{1}{2} + \frac{q_H - q_L - s}{2\tau}. \tag{4}$$

Of these, hospitals H and L serve respectively $\mu\sigma_L\hat{x}_{|L}$ and $\mu\sigma_L(1-\hat{x}_{|L})$ patients. Additionally, Hospital H serves all of these patients if $q_H > q_L + \tau + s$ and none if $q_H < q_L - \tau + s$.

The lock-in effect of switching costs is straightforward to see from (2) and (4). Hospital H may offer a quality level s units below that of Hospital L and still get half of its previous patients with changing preferences, whereas it has to offer a quality premium of s in order to get half of the patients with changing preferences who are tied to Hospital L.

Finally, a fraction $(1-\mu)\sigma_H$ of patients have unchanged preferences uniformly distributed on $[0,\sigma_H]$. These patients demanded treatment from Hospital H in the past and continue to do so provided that $v + q_H - \tau\sigma_H > v + q_L - \tau\sigma_L - s$. Similarly, the $(1-\mu)\sigma_L$ patients with persistent preferences who used Hospital L in the past continue to do so provided that $v + q_L - \tau\sigma_L > v + q_H - \tau\sigma_H - s$.

Combining demand from the two types of patients, it may be easily shown that total demand facing Hospital i is given by

$$D_i(q_i, q_j) = \frac{\mu}{2\tau}[\tau + q_i - q_j + (\sigma_i - \sigma_j)s] + (1-\mu)\sigma_i, \quad i,j = H,L; \quad i \neq j; \tag{5}$$

provided that $|q_H - q_L| < \tau - s$ and $|(q_H - \tau\sigma_H) - (q_L - \tau\sigma_L)| < s$.[10]

To make the distinction between the two sources of patient inertia more salient, I focus on equilibria wherein patients with persistent preferences always demand treatment from the hospital they have previously used. At the end of this section, I present the condition that ensures that switching occurs among patients with changing preferences but not among patients with fixed

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

preferences in an equilibrium where hospitals offer strictly positive quality. Notice how, in this case, patient inertia manifests: both hospitals have some captive patients due to persistent preferences, and the hospital with higher inherited demand has a demand bonus simply because switching hospitals is costly for patients.

3.2 Hospital Objectives

Hospitals simultaneously and independently choose quality levels to maximise a weighted sum of profits and aggregate patient benefit. Formally, Hospital i maximises:

$$\Omega_i(q_i, q_j) = T + \tilde{p} D_i(q_i, q_j) - C[q_i, D_i(q_i, q_j)] + \alpha B[q_i, D_i(q_i, q_j)], \quad i, j = H, L; \quad i \neq j; \tag{6}$$

where T denotes a lump-sum transfer that ensures that a no-liability constraint is satisfied, and $\tilde{p}$ denotes the per-treatment payment through which a third-party payer (e.g., a regulator or insurer) prospectively finances hospitals; $C[q_i, D_i(q_i, q_j)]$ is the cost of producing $D_i(q_i, q_j)$ units of treatment with quality q_i; $B[q_i, D_i(q_i, q_j)]$ is the total net benefit of patients treated at Hospital i; and $\alpha > 0$ captures the degree of altruism.

Treatment production costs are given by

$$C[q_i, D_i(q_i, q_j)] = (cq_i + k) D_i(q_i, q_j) + \frac{\gamma}{2} q_i^2, \quad i, j = H, L; \quad i \neq j; \tag{7}$$

where $c \lesseqgtr 0$ measures either the degree of cost substitutability or complementarity between quality and output, $k > \max\{0, -cq_i\}$ is the minimum unit cost of treatment, and $\gamma > 0$ gives the importance of the fixed investment cost. If $c > 0$, a certain level of quality is more costly to achieve when more patients are treated (i.e., the marginal cost of quality is increasing in demand). Hospital production hence 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. If $c < 0$, the more patients a hospital treats, the less costly it is to provide each additional unit of quality (i.e., the marginal cost of quality is decreasing in demand). Quality and output are cost complements, which suffices, in this analytical framework, to establish a positive relationship between demand and quality. Such link, observed in hospital production and well documented in the literature, is often referred to as the *volume-outcome* relationship. Volume-outcome effects are observed when, all else equal, high-volume hospitals provide better quality of care and thus generate better treatment outcomes than low-volume hospitals. 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. Hentschker and Mennicken (2018) and Avdic et al. (2019) present recent empirical evidence of volume-outcome effects. In particular, the latter show that this positive and causal relationship is due to learning-by-doing, with a significant share of the effect ascribed to current experience. Although their results suggest that cumulated experience also plays a role, they are in line with the earlier findings of Gaynor et al. (2005), who show that the effect of volume on outcome largely occurs contemporaneously. The cost function specification (7) therefore reflects this contemporaneous link.

Hospitals are assumed to have semi-altruistic preferences in the sense that they care, to some extent, about the utility their patients derive from treatment. In the hospital industry, departure from pure profit-maximisation may arise due to the structure of hospitals, wherein a managerial hierarchy and a medical one coexist. Physicians have long been recognised as acting, at least to some degree, in the interest of their patients, and hospital behaviour may then be thought as reflecting physician behaviour subject to a budget constraint imposed by managers.[11] The aggregate benefit to patients treated at hospitals H and L is respectively given by

$$\begin{aligned} B_H[q_H, D_H(q_H, q_L)] = \mu\sigma_H &\int_0^{\hat{x}_{|H}} (v + q_H - \tau x)dx \\ &+ \mu\sigma_L \int_0^{\hat{x}_{|L}} (v + q_H - \tau x - s)dx + (1-\mu)\int_0^{\sigma_H} (v + q_H - \tau x)dx \end{aligned} \quad (8)$$

and

$$\begin{aligned} B_L[q_L, D_L(q_H, q_L)] = \mu\sigma_H &\int_{\hat{x}_{|H}}^1 [v + q_L - \tau(1-x) - s]dx \\ &+ \mu\sigma_L \int_{\hat{x}_{|L}}^1 [v + q_L - \tau(1-x)]dx + (1-\mu)\int_{\sigma_H}^1 [v + q_L - \tau(1-x)]dx. \end{aligned} \quad (9)$$

It is instructive to see how semi-altruistic preferences affect incentives to provide quality. Differentiating (8) and (9) with respect to q_H and q_L respectively, one may show after some manipulation that the marginal altruistic benefit from quality is given by

$$\frac{\partial B_i[q_i, D_i(q_i, q_j)]}{\partial q_i} = \frac{\mu}{4\tau}(2v + q_i + q_j - \tau - s) + D_i > 0, \quad i, j = H, L; \quad i \neq j. \quad (10)$$

[11] See Brekke et al. (2011) and Siciliani et al. (2013) for a discussion of the assumption of semi-altruism in the general literature on healthcare supply and in the context of competition between healthcare providers in particular.

There is a twofold effect on aggregate patient benefit (at the hospital level). A marginal increase in quality simultaneously expands demand and increases the utility of each patient. These two effects are respectively captured by the two terms on the right-hand side of (10).

Finally, I make the following restrictions on parameter values:

$$c > c_{min} \equiv \max\left\{\left(\alpha - \frac{2\tau\gamma}{\mu}\right), \left(\frac{3\alpha}{4} - \frac{\tau\gamma}{\mu}\right), \left(\frac{2\alpha}{3} - \frac{2\tau\gamma}{3\mu}\right)\right\} \tag{11}$$

and

$$\sigma_H - \sigma_L < \min\left\{1, \frac{\tau - s}{|\alpha - c|\phi}, \frac{s}{|(\alpha - c)\phi - \tau|}\right\}, \tag{12}$$

where

$$\phi = \frac{2[\tau - (\tau - s)\mu]}{\mu\left[\frac{2\tau\gamma}{\mu} - (2\alpha - 3c)\right]} > 0. \tag{13}$$

Condition (11) imposes that the degree of cost substitutability is sufficiently strong or the degree of cost complementarity is sufficiently weak so that the second-order conditions of the hospitals' maximisation problems are satisfied and the solution is economically meaningful. Condition (12) ensures that the demand function (5) holds in equilibrium with strictly positive quality, which requires that the difference in inherited demand faced by the two hospitals is not too large. This, in turn, implies that equilibrium quality levels are such that neither hospital is chosen by all of its previous patients with changing preferences (i.e., switching occurs in equilibrium at both hospitals) and that all patients with fixed preferences continue to choose the same hospital.

4 Equilibrium Quality and Demand

Using (7) and (10), maximisation of Ω_i with respect to q_i yields the first-order condition

$$\frac{\mu}{2\tau}\left[p - cq_i + \alpha\left(\frac{2v + q_i + q_j - \tau - s}{2}\right)\right] + (\alpha - c)D_i - \gamma q_i = 0, \quad i, j = H, L; \quad i \neq j; \tag{14}$$

where $p = \tilde{p} - k$. The marginal benefit from quality is given by the increase in revenues ($\mu\tilde{p}/2\tau$) and in total patient surplus, and it includes an efficiency gain (cD_i) when $c < 0$. The marginal cost of quality includes the cost of treating additional patients ($\mu(cq_i + k)/2\tau$) and the marginal cost of quality investments (γq_i), as well as the increase in total treatment costs (cD_i) when $c > 0$.

Inserting D_i and D_j as defined in (5) into the pair of equations given by (14) and solving for q_i yields the candidate equilibrium quality levels

$$q_i^* = \max\left\{0, \frac{p + (\alpha - c)\left[\tau - s - (\alpha - 2c)\frac{\phi}{2}\right] + \alpha\left(v - \frac{\tau + s}{2}\right)}{\frac{2\tau\gamma}{\mu} - (\alpha - c)} + (\alpha - c)\phi\sigma_i\right\}, \quad i, j = H, L. \tag{15}$$

Equation (15) defines a Nash equilibrium if the first-order conditions specify hospitals' global best responses. If the cost function is sufficiently convex in quality, Ω_i is concave in the region in which (5) holds and the second-order conditions always hold locally. Concavity of Ω_i requires that $\gamma - \frac{\mu}{2\tau}\left(\frac{3}{2}\alpha - 2c\right) > 0$, which is always satisfied given (11). However, this is not sufficient to show that the first-order conditions define a Nash Equilibrium. Hospitals may unilaterally deviate from their strategies in the candidate equilibrium by choosing a quality level outside the range in which (5) holds. In particular, it must be ensured that no hospital would prefer to serve only its captive patients with fixed preferences. If μ is large enough and s is sufficiently small, deviation is not beneficial and (15) defines a Nash equilibrium.[12]

In the remainder of the analysis, I focus on strictly positive quality levels. As may be seen from (15), a sufficiently high prospective payment $\tilde{p}$ suffices to elicit strictly positive equilibrium quality.

Suppose first that $c > 0$. It follows immediately from (15) that Hospital H offers lower quality than Hospital L if cost substitutability is stronger than the hospitals' altruism. The intuition for this result is as follows. Both the marginal cost and the marginal altruistic benefit from quality depend on current demand, which, in turn, depends positively on inherited demand through the switching cost and the share of patients with persistent preferences. Hospital H has both a higher marginal cost of quality (because providing quality is more costly when more patients are treated) and a higher marginal altruistic benefit (because higher quality increases the utility of more patients). Which of these effects dominates depends on the size of α and c. However, if $c < 0$, the result is clear-cut. In the presence of volume-outcome effects, Hospital H has a higher marginal altruistic benefit and a lower marginal cost of quality. It will thus offer higher quality.

Having derived equilibrium quality, one may now look at demand. Proposition 1 below describes how patient inertia affects demand patterns and how it counteracts or strengthens the effect of quality as a demand shifter.

Proposition 1. *In equilibrium, quality and demand are characterised as follows:*

1. *if $c < \underline{c}$, then $q_H^* > q_L^*$ and $D_H(q_H^*, q_L^*) > \sigma_H$;*

2. *if $\underline{c} < c < \alpha$, then $q_H^* > q_L^*$ and $\frac{1}{2} < D_i(q_H^*, q_L^*) < \sigma_H$;*

3. *if $c > \alpha$, then $q_H^* < q_L^*$ and $\frac{1}{2} < D_H(q_H^*, q_H^*) < \sigma_H$;*

where $\underline{c} = \frac{2\tau[\alpha-(\tau-s)\gamma]}{2\tau+(\tau-s)\mu} < \alpha$.

[12] See Klemperer (1987), Beggs and Klemperer (1992), and To (1996) for the analogous argument in the case of multi-period price competition.

Proof. Follows directly from (15) and the comparison between σ_H and D_H evaluated at the equilibrium quality levels. □

To understand how demand evolves, recall that some patients face a strong mismatch between their preferences and the horizontal attributes of the hospital they are tied to and hence opt to switch. As explained above, if $c < \alpha$, Hospital H offers higher quality. For a value of c sufficiently below α, the quality difference is large enough to outweigh the demand loss due to the mismatch between patient preferences and the hospital's attributes. Due to its high quality, Hospital H attracts more patients who previously used Hospital L than those who switch from it, strengthening its position as market leader. Depending on the threshold value $\underline{c}$, this may occur with a sufficiently low degree of cost substitutability (when $\underline{c} > 0$) or may require sufficiently strong cost complementarity (when $\underline{c} < 0$). For intermediate degrees of cost substitutability or complementarity, Hospital H offers higher quality but not sufficiently high to attract enough patients to compensate for those who switch. Demand faced by this hospital declines, but it nonetheless amounts to more than half of the market. Finally, if cost substitutability is stronger than the hospitals' altruism, Hospital H offers lower quality, which reinforces the demand loss due to horizontal preferences. Patient inertia, however, ensures that it will retain its position as market leader.

Proposition 1 has therefore two immediate implications:

Corollary 1.1. *Due to patient inertia, the hospital with higher inherited demand will retain its position as market leader regardless of the values of* c *and* α*.*

Corollary 1.2. *In equilibrium, the majority of patients in the market is served by the hospital that offers lower (higher) quality if* $c > (<)\alpha$*.*

5 The Effect of Switching Costs

5.1 Hospital-level Quality

The effect of a marginal change in switching costs on equilibrium quality is given by

$$\frac{\partial q_i^*}{\partial s} = -\frac{\alpha/2}{\frac{2\tau\gamma}{\mu} - (\alpha - c)} + \frac{(\alpha - c)(\sigma_i - \sigma_j)}{\frac{2\tau\gamma}{\mu} - (2\alpha - 3c)}, \quad i, j = H, L; \quad i \neq j. \tag{16}$$

Lower switching costs affect quality directly through the change in patient utility. In the presence of switching costs, the altruistic incentive to attract patients who were previously treated at the neighbouring hospital are weaker, since the utility of these patients is reduced by an amount s.

From (10), the lower the switching cost, the stronger is the altruistic incentive the two hospitals have to increase quality. This is the *patient utility effect*. There is also a *demand effect*, which can easily be seen from (5). All else equal, lower switching costs shift demand from the high- to the low-volume hospital, and therefore change the marginal cost and the marginal benefit from quality. Unlike the patient utility effect, the demand effect affects hospitals differently, and its sign and magnitude depend on the strength of cost substitutability/complementarity relative to the degree of altruism. The effect of lower switching costs on hospital-level quality is formalised as follows.

Proposition 2. *Provided that the cost function is sufficiently convex in quality, there exist two threshold values of c, given by*

$$c_{HH} = \alpha - \frac{4\tau\gamma(\sigma_H - \sigma_L) + 3\alpha\mu}{4\mu(\sigma_H - \sigma_L)} + \frac{\sqrt{8\tau\gamma(\sigma_H - \sigma_L)[2\tau\gamma(\sigma_H - \sigma_L) + \alpha\mu] + (\alpha\mu)^2[9 - 8(\sigma_H - \sigma_L)]}}{4\mu(\sigma_H - \sigma_L)} \in (c_{min}, \alpha) \quad (17)$$

and

$$c_{HL} = \alpha - \frac{4\tau\gamma(\sigma_H - \sigma_L) - 3\alpha\mu}{4\mu(\sigma_H - \sigma_L)} + \frac{\sqrt{8\tau\gamma(\sigma_H - \sigma_L)[2\tau\gamma(\sigma_H - \sigma_L) - \alpha\mu] + (\alpha\mu)^2[9 + 8(\sigma_H - \sigma_L)]}}{4\mu(\sigma_H - \sigma_L)} > \alpha, \quad (18)$$

such that a reduction in switching costs leads to (i) lower quality at the high-volume hospital and higher quality at the low-volume hospital if $c < c_{HH}$; (ii) higher quality at both hospitals if $c_{HH} < c < c_{HL}$; and higher quality at the high-volume hospital and lower quality at the low-volume hospital if $c > c_{HL}$. Additionally, the threshold values c_{HH} and c_{HL} and the distance $c_{HH} - c_{HL}$ are increasing in μ and decreasing in τ.

Proof. See Appendix A. □

Start by considering Hospital H. Under cost substitutability, the marginal cost and the marginal benefit from quality change with demand in the same direction. The demand shift from the high- to the low-volume hospital decreases both the hospital's marginal cost and marginal altruistic benefit. If $c > \alpha$, the change in the marginal cost outweighs the change in the marginal altruistic benefit. This implies that the demand reduction contributes to higher quality. The demand effect thus reinforces the effect of increased patient utility (due to a lower s), and lower switching costs have a clear-cut positive effect on quality.

If $c < \alpha$ instead, the change in the marginal altruistic benefit dominates, and lower demand leads, all else equal, to lower quality. In this case, the patient utility and the demand effects go in opposite directions, and the net impact on quality is *a priori* ambiguous. In the presence of volume-outcome effects, a particular case of $c < \alpha$, this last-mentioned result naturally caries over. However, the underlying mechanism differs slightly. In this case, the demand shift leads to a lower marginal benefit and to a higher marginal cost of quality. For values of c sufficiently close to α, the demand effect is small, and the patient utility effect dominates. Hence, lower switching costs lead to higher quality even if $c < \alpha$. However, if there is sufficiently strong cost complementarity or, possibly, sufficiently weak cost substitutability, then the demand effect dominates, and lower switching costs lead to lower quality.

The analysis of Hospital L is analogous. Under cost substitutability, the demand shift increases its marginal cost and its marginal altruistic benefit. If $c > \alpha$, the demand inflow contributes to lower quality, implying that the demand effect counteracts the patient utility effect. Again, for values of c sufficiently close to α, the patient utility effect dominates, and lower switching costs lead to higher quality even if $c > \alpha$. If instead there is sufficiently strong cost substitutability, the demand effect dominates, which implies that lower switching costs reduce quality.

If $c < \alpha$, the change in the marginal benefit dominates, the two effects go in the same direction, and the impact on quality is clearly positive. This positive impact is reinforced if there are volume-outcome effects, since the inflow of patients reduces the marginal cost of quality.

Proposition 2 reveals that there is a set of values of c for which no patient is left worse off in terms of (changes in) quality provision after a reduction in switching costs. The link between this set and the intensity of patients' horizontal preferences sheds light on the impact of lower switching costs in different markets. When fewer patients have persistent preferences (higher μ) or travelling/mismatch costs are lower (lower τ), demand is more responsive to quality. Also, the share of Hospital H's demand advantage from patients with changing preferences is greater. Increased demand responsiveness implies that the altruistic incentive to attract patients (due to a lower s) is stronger because a marginal increase in quality will have a larger impact on demand. A greater demand advantage implies that, when switching costs fall, the resulting demand shift is stronger. Consequently, the above-mentioned patient utility and demand effects are simultaneously reinforced by a higher μ or a lower τ. The change in the demand effect dominates for Hospital H, whereas the change in the patient utility effect dominates for Hospital L. Thus, at Hospital H, a higher c is required for the utility effect to offset the demand effect, while, at Hospital L, a higher c is required for the demand effect to dominate. Because this outcome is more pronounced for the latter hospital, the set of values for which lower switching costs increase quality at both hospitals widens. The above analysis can be summarised as follows.

Corollary 2.1. *There is more scope for a quality increase at both hospitals in response to a reduction in switching costs when there are fewer patients with persistent horizontal preferences or travelling/mismatch costs are lower.*

It is also interesting to see how the results in Proposition 2 change with the hospitals' degree of profit-orientation. If hospitals are pure profit-maximisers, the patient utility effect vanishes, and the sign of the demand effect is uniquely determined by c and inherited demand. Under cost substitutability, lower switching costs lead to higher quality at the high-volume hospital but lower quality at low-volume one. Under cost complementarity, the reserve applies. The following result may therefore be established.

Corollary 2.2. *If hospitals are pure profit-maximisers, then it is certain that some patients will enjoy lower quality after a switching costs reduction.*

Notice that the implications of semi-altruistic hospital preferences are not straightforward. On the one hand, they create a set of values of c for which lower switching costs improve quality provision at both hospitals and open the possibility that quality increases at the high-volume hospital when quality and output are cost complements (i.e., when $c_{HH} < c < 0$). On the other hand, they allow for a quality decrease at the high-volume hospital when quality and output are cost substitutes (i.e., when $0 < c < c_{HH}$), implying that lower switching costs hurt the majority of patients in a situation where reduced market concentration would otherwise be beneficial.

5.2 Average Quality

The fact that hospital-level quality is affected differently implies that lower switching costs have heterogeneous effects on patients. To grasp the overall effect of a switching costs reduction on quality enjoyed by all patients in the market, define average equilibrium quality as the sum of qualities weighted by current demand, $\overline{q} = q_H^* D_H(q_H^*, q_L^*) + q_L^* D_L(q_H^*, q_L^*)$. The effect of a marginal change in switching costs on average quality is given by

$$\frac{\partial \overline{q}}{\partial s} = (q_H^* - q_L^*)\frac{\partial D_H^*}{\partial s} + \frac{\partial q_H^*}{\partial s} D_H^* + \frac{\partial q_L^*}{\partial s} D_L^* = \\ (\alpha - c)\phi(\sigma_H - \sigma_L)\frac{\partial D_H^*}{\partial s} - \frac{\alpha/2}{\frac{2\tau\gamma}{\mu} - (\alpha - c)} + \frac{(\alpha - c)(\sigma_H - \sigma_L)(D_H^* - D_L^*)}{\frac{2\tau\gamma}{\mu} - (2\alpha - 3c)}, \quad (19)$$

where

$$\frac{\partial D_H^*}{\partial s} = \frac{\mu}{2\tau}\left[\frac{2(\alpha - c)}{\frac{2\tau\gamma}{\mu} - (2\alpha - 3c)} + 1\right](\sigma_H - \sigma_L) > 0. \tag{20}$$

Lower switching costs have a twofold effect on average quality. First, there is a patient redistribution effect, since a reduction in switching costs always decreases market concentration by shifting demand from the high- to the low-volume hospital.[13] The sign of this redistribution effect depends on the initial quality difference. Second, as analysed above, quality changes at the hospital-level, and these changes are weighted by each hospital's demand. The effect of lower switching costs on average quality may be stated as follows.

Proposition 3. *Provided that the cost function is sufficiently convex in quality, there exists a threshold value of c, given by $c_{\overline{q}} \in (c_{min}, \alpha)$ and implicitly defined by*

$$\frac{\alpha/2}{\frac{2\tau\gamma}{\mu} - (\alpha - c_{\overline{q}})} = \frac{2[\tau - (\tau - s)\mu](\sigma_H - \sigma_L)^2(\alpha - c_{\overline{q}})\left(\frac{2\tau\gamma}{\mu} + c_{\overline{q}}\right)}{\tau\left[\frac{2\tau\gamma}{\mu} - (2\alpha - 3c_{\overline{q}})\right]^2}, \tag{21}$$

such that a reduction in switching costs leads to lower average quality if $c < c_{\overline{q}}$. Furthermore, $c_{\overline{q}} \in (c_{HH}, \alpha)$ if

$$s > \frac{\tau}{\mu}\left[\frac{\frac{2\tau\gamma}{\mu} - (2\alpha - 3c_{HH})}{2(\sigma_H - \sigma_L)\left(\frac{2\tau\gamma}{\mu} + c_{HH}\right)} - (1 - \mu)\right], \tag{22}$$

with c_{HH} as given in (17).

Proof. See Appendix B. □

If $c > \alpha$, Hospital H offers lower quality but a reduction in switching costs induces a quality increase. When switching costs fall, some patients switch from Hospital H to Hospital L, going from a lower to a higher quality hospital. All else equal, this leads to higher average quality. This effect is reinforced by the quality changes at the hospital level. Equation (16) implies that $|\partial q_H^*/\partial s| > |\partial q_L^*/\partial s|$ for $c > \alpha$. Since Hospital H treats more patients and its quality response is stronger, the weighted quality increase at Hospital H always dominates the weighted quality change at Hospital L. Thus, lower switching costs lead to higher average quality even if quality falls at Hospital L.

If $c < \alpha$ instead, it is Hospital L which offers lower quality. In this case, the demand adjustment contributes to lower average quality as patients switch from the higher to the lower quality hospital.

[13] $\frac{\partial D_H^*}{\partial s} > 0$ if $c > -\frac{2\tau\gamma}{\mu}$, which always holds given (11).

However, a lower value of s elicits a quality increase at Hospital L and an *a priori* indeterminate change in quality for the majority of patients in the market (those at Hospital H). Only for a value of c sufficiently below α, is the weighted increase in quality at Hospital L dominated by the patient redistribution effect, possibly in conjunction with a (weighted) reduction in quality at Hospital H.

Importantly, notice that a reduction in quality at the high-volume hospital is not a necessary condition for lower switching costs to have a negative impact on average quality. If the initial quality difference is large enough and patients switch from the higher to the lower quality hospital, there exists a set of values of c for which the redistribution of patients suffices to reduce average quality. The following result therefore ensues from Proposition 3.

Corollary 3.1. *If switching costs are sufficiently high to begin with, lower switching costs reduce average quality while increasing quality at both hospitals if* $c_{HH} < c < c_{\overline{q}}$.

To grasp why high initial switching costs are required to achieve an initial quality difference such that the redistribution effect outweighs the quality increases at the hospital level, recall that higher switching costs allow Hospital H to retain a greater demand advantage. It is this demand advantage that leads to a higher marginal benefit from quality—which dominates the higher marginal cost when $0 < c < \alpha$ or is indeed reinforced by the lower marginal cost when $c < 0$—, and hence to higher quality at Hospital H. The greater is the demand advantage, the higher is the quality premium offered by Hospital H when $c < \alpha$. If the initial switching costs are high enough, then the quality difference is so large that the fact that some patients switch from Hospital H to Hospital L suffices to drive average quality downward.

Finally, note that, under the assumption of semi-altruistic hospitals, the negative effect of lower switching costs on average quality arises for a sufficiently weak degree of cost substitutability or may instead require a sufficiently strong degree of cost complementarity. Conversely, if hospitals are pure profit-maximisers, lower switching costs always lead to lower average quality in the presence of cost complementarity.

5.3 Patient Welfare

Up until this point, the analysis of the effect lower switching costs has focused on the hospital-side of the market. However, switching costs affect patient welfare through other channels besides quality. Define total patient welfare as $W = B_H + B_L$, with B_H and B_L as given in equations

(8)-(9). The effect of a marginal change in switching costs on patient welfare is given by

$$\frac{\partial W}{\partial s} = \frac{\partial q_H^*}{\partial s} D_H^* + \frac{\partial q_L^*}{\partial s} D_L^* - \mu \left[\sigma_H(1-\hat{x}_{|H}) + \sigma_L \hat{x}_{|L}\right] + (q_H^* - q_L^*)\frac{\partial D_H^*}{\partial s}$$
$$+ \mu\tau \left[\sigma_H(1-2\hat{x}_{|H})\frac{\partial \hat{x}_{|H}}{\partial s} + \sigma_L(1-2\hat{x}_{|L})\frac{\partial \hat{x}_{|L}}{\partial s}\right] + \mu s \left(\sigma_H \frac{\partial \hat{x}_{|H}}{\partial s} - \sigma_L \frac{\partial \hat{x}_{|L}}{\partial s}\right). \quad (23)$$

Lower switching costs have a fivefold effect on total patient welfare. First, lower switching costs elicit changes in hospital-level quality. Second, there is a direct utility gain for the patients who switch because doing so becomes less costly. Third, a redistribution of demand occurs as patients switch from the high- to the low-volume hospital. Fourth and fifth, total travelling/mismatch costs change indeterminately and total switching costs increase. These effects are respectively given by the terms on the right-hand side of (23). It turns out that the three last-mentioned effects cancel out, and equation (23) can be rewritten as

$$\frac{\partial W}{\partial s} = -\frac{\alpha/2}{\frac{2\tau\gamma}{\mu} - (\alpha - c)} + \frac{(\alpha - c)(\sigma_H - \sigma_L)(D_H^* - D_L^*)}{\frac{2\tau\gamma}{\mu} - (2\alpha - 3c)} - \mu\left[\sigma_H(1-\hat{x}_{|H}) + \sigma_L \hat{x}_{|L}\right]. \quad (24)$$

Thus, the total effect of a switching costs reduction on patient welfare is uniquely determined by the increase in the utility of patients who switch and the weighted changes in quality at the hospital level. The effect of lower switching costs on total patient welfare may be stated as follows.

Proposition 4. *Provided that the cost function is sufficiently convex in quality, there exists a threshold value of c, given by $c_W \in (c_{min}, \min\{c_{HH}, c_{\overline{q}}\})$ and implicitly defined by*

$$\frac{\alpha/2}{\frac{2\tau\gamma}{\mu} - (\alpha - c_W)} + \frac{\mu}{2}\left(1 - \frac{s}{\tau}\right) = \frac{2[\tau - (\tau - s)\mu](\sigma_H - \sigma_L)^2(\alpha - c_W)\left[\frac{2\tau\gamma}{\mu} - (\alpha - 2c_W)\right]}{\tau\left[\frac{2\tau\gamma}{\mu} - (2\alpha - 3c_W)\right]^2}, \quad (25)$$

such that lower switching costs reduce total patient welfare if $c < c_W$.

Proof. See Appendix C. □

For given quality, lower switching costs always lead to higher patient welfare through the increase in the utility of patients who switch.

If $c > \alpha$, the weighted quality increase at Hospital H dominates the weighted change in quality at Hospital L, and the total impact of lower switching costs on patient welfare is clearly positive.

If $c < \alpha$, patient welfare can only decrease if the weighted reduction in quality at Hospital H is such that it outweighs the weighted increase in quality at Hospital L and the direct patient utility gain, which requires a value of c sufficiently below α. Thus, differently from the case of average quality, a reduction in quality at Hospital H is a necessary condition for patient welfare to fall.

Additionally, recall that the weighted reduction in quality at Hospital H must only dominate the weighted increase in quality at Hospital L, net of the negative demand adjustment effect, for switching costs to reduce average quality. Conversely, for switching costs to reduce patient welfare, the weighted reduction in quality at Hospital H must outweigh two counteracting effects. This implies that the value of c below which lower switching costs reduce patient welfare is less than the value of c below which lower switching costs reduce average quality. In other words, the decrease in quality at Hospital H must be stronger to reduce patient welfare than it must be to reduce average quality, if required at all. The following result may therefore be established.

Corollary 4.1. *Lower switching costs reduce average quality but increase total patient welfare if* $c_W < c < c_{\overline{q}}$.

Finally, notice that the negative effect of lower switching costs on total patient welfare may only require a sufficiently weak degree of cost substitutability when hospitals are semi-altruistic. Conversely, if hospitals are pure profit-maximisers, then lower switching costs only reduce patient welfare in the presence of sufficiently strong cost complementarity. Recall that, for lower switching costs to reduce average when quality hospitals are pure profit-maximisers, any degree of cost complementarity suffices.

6 Discussion and Concluding Remarks

By means of a duopoly model that maps on the recent empirical evidence on choice persistence (or loyalty) in the hospital industry, this paper has investigated the effect of lower switching costs on the quality of elective hospital treatments. While lower switching costs always reduce market concentration, the impact on quality provision depends crucially on the technology of production of hospital treatments and on the hospitals' degree of altruism. This result challenges the standard prediction that reduced market concentration is always welfare-improving. Once features that are characteristic, although not exclusive, to the healthcare industry are taken into account—in this paper, the departure from pure profit-maximisation and the existence of cost complementarity between quality and output, the so-called volume-outcome effects—, standard results may fail to arise. Whether lower switching costs are a by-product of the adoption of shared EHR or result intentionally from patient choice policies, there may be unintended consequences.

When the degree of cost substitutability between quality and output is low relative to the degree of altruism or when there is cost complementarity, switching costs act as 'minimum volume standards' for the high-volume hospital, ensuring that high quality is provided. In other words, the lock-in effect of switching costs grant the high-volume hospital the demand advantage that allows it to offer higher quality. This occurs because the marginal altruistic benefit from quality is increasing in demand and outweighs the marginal cost of quality (weak cost substitutability) or because the marginal cost is decreasing in demand (cost complementarity). If the degree of cost substitutability is sufficiently low relative to the degree of altruism or there is sufficiently strong cost complementarity, quality falls at the high-volume hospital when switching costs are reduced and patients switch. This leaves the majority of patients in the market worse off in terms of quality provision, contributing to lower average quality and total patient welfare. These results have several policy implications which are discussed in the following.

First, the adverse effects of lower switching costs may require cost complementarity to materialise. If they do and quality and output are cost substitutes, lower switching costs will be beneficial however small the degree of cost substitutability. This suggests that knowledge of hospital production attributes is key to anticipating the effect lower switching costs. Hentschker and Mennicken (2018) report a negative effect of volume on mortality rates in the case of German hip replacement patients. Avdic et al. (2019) also find a positive effect of volume on quality, with the results pointing towards a stronger effect for more complex types of advanced cancer surgery. Rachet-Jacquet et al. (2019), conversely, find no effect of volume on patient-reported health outcomes for hip replacement patients in the English NHS. Such mixed empirical evidence in turn suggests not only that lower switching costs may have heterogeneous effects at the sub-hospital-level (e.g., at the specialty or department level) but also that the institutional setting may play a role.

Second, my results indicate that policies aimed at reducing switching costs will be inaccurately assessed if the evaluation fails to consider the supply-side response. In a counterfactual scenario with no switching costs, Irace (2018) estimates an expected mortality 3% below the observed mortality rate. Importantly, this reduction in mortality is only due to a more efficient distribution of patients among hospitals, as patients switch to higher quality providers. Hospital quality is held fixed and feedback effects between demand and quality are ruled out. This paper has focused precisely on these feedback effects and revealed that lower switching costs may either increase or reduce quality, thus affecting patients through channels other than their increased ability to adjust to the environment.

Third, market conditions affect the impact of lower switching costs on quality at the hospital level. I have shown that there is set of values of the degree of cost substitutability/complementarity for which lower switching costs lead to higher quality at both hospitals and that this set is larger when patients' horizontal preferences are less stringent. This suggests that there is increased scope for lower switching costs to have no adverse effects in terms of quality changes at the hospital-level in markets where patients have greater geographical mobility, where there is stronger substitutability between hospitals or where patients' preferences are more volatile.

Fourth, perhaps surprisingly, average quality might fall even if lower switching costs lead to higher quality at the hospital level due to the demand redistribution effect identified by Irace (2018). My model shows that such demand adjustment may also occur in the opposite direction. In order to reduce the mismatch between their horizontal preferences and the attributes of the chosen hospital, patients may indeed switch to lower quality hospitals. If the quality differential is sufficiently large to begin with, this effect dominates and lower switching costs reduce average quality. A direct implication of this result is that, besides the evolution of quality at the hospital level, the redistribution of patients among hospitals should be considered within the scope of policy evaluation.

Finally, different measures of aggregate welfare yield distinct conclusions. I have shown that lower switching costs might increase total patient welfare but reduce average quality. If policymakers are mostly concerned with clinical outcomes and indicators, average quality will arguably be a more appropriate measure of welfare and lower switching more likely to be deemed welfare-decreasing. If, however, policymakers care about patient welfare more broadly—considering, for example, patient disutility of switching as well as clinical quality—, lower switching costs might be regarded as more beneficial.

References

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`

Beggs, A., and Klemperer, P. (1992). Multi-period competition with switching costs. *Econometrica*, *60*(3), 651–666. `https://doi.org/10.2307/2951587`

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`

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

Gaynor, M., Propper, C., and Seiler, S. (2016). Free to choose? Reform, choice, and consideration sets in the English National Health Service. *American Economic Review*, *106*(11), 3521–3557. http://dx.doi.org/10.1257/aer.20121532

Gaynor, M., Seider, H., and Vogt, W. B. (2005). The volume-outcome effect, scale economies, and learning-by-doing. *American Economic Review*, *95*(2), 243–247. https://doi.org/10.1257/000282805774670329

Gehrig, T., Shy, O., and Stenbacka, R. (2011). History-based price discrimination and entry in markets with switching costs: A welfare analysis. *European Economic Review*, *55*(5), 732–739. https://doi.org/10.1016/j.euroecorev.2010.09.001

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

Gutacker, N., Siciliani, L., Moscelli, G., and Gravelle, H. (2016). Choice of hospital: Which type of quality matters? *Journal of Health Economics*, *50*, 230–246. https://doi.org/10.1016/j.jhealeco.2016.08.001

Hentschker, C., and Mennicken, R. (2018). The volume–outcome relationship revisited: Practice indeed makes perfect. *Health Services Research*, *53*(1), 15–34. https://doi.org/10.1111/1475-6773.12696

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

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

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

Klemperer, P. (1995). Competition when consumers have switching costs: An overview with applications to industrial organization, macroeconomics, and international trade. *The Review of Economic Studies*, *62*(4), 515–539. https://doi.org/10.2307/2298075

Oderkirk, J. (2017). *Readiness of electronic health record systems to contribute to national health information and research* (Working Paper No. 99). OECD Health Working Papers, OECD Publishing, Paris. http://dx.doi.org/10.1787/9e296bf3-en

Rachet-Jacquet, L., Gutacker, N., and Siciliani, L. (2019). *The causal effect of hospital volume on health gains from hip replacement surgery* (CHE Research Paper No. 168). Centre for Health Economics, University of York.

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

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

Shy, O., and Stenbacka, R. (2016). Customer privacy and competition. *Journal of Economics & Management Strategy*, *25*(3), 539–562. https://doi.org/10.1111/jems.12157

Siciliani, L., Chalkley, M., and Gravelle, H. (2017). Policies towards hospital and GP competition in five European countries. *Health Policy*, *121*(2), 103–110. https://doi.org/10.1016/j.healthpol.2016.11.011

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

Tay, A. (2003). Assessing competition in hospital care markets: The importance of accounting for quality differentiation. *The RAND Journal of Economics*, *34*(4), 786–814. https://doi.org/10.2307/1593788

To, T. (1996). Multi-period competition with switching costs: An overlapping generations formulation. *The Journal of Industrial Economics*, *44*(1), 81–87. https://doi.org/10.2307/2950562

Varkevisser, M., van der Geest, S. A., and Schut, F. T. (2012). Do patients choose hospitals with high quality ratings? Empirical evidence from the market for angioplasty in the Netherlands. *Journal of Health Economics*, *31*(2), 371–378. https://doi.org/10.1016/j.jhealeco.2012.02.001

Victoor, A., Delnoij, D., Friele, R., and Rademakers, J. (2016). Why patients may not exercise their choice when referred for hospital care. An exploratory study based on interviews with patients. *Health Expectations*, *19*(3), 667–678. https://doi.org/10.1111/hex.12224

Appendix A Proof of Proposition 2

From equation (16), let the effect of a marginal change in switching costs on equilibrium quality at Hospitals H and L be written, respectively, as

$$q_H^{*\prime}(c) = -\frac{\alpha/2}{\frac{2\tau\gamma}{\mu} - (\alpha - c)} + \frac{(\alpha - c)(\sigma_H - \sigma_L)}{\frac{2\tau\gamma}{\mu} - (2\alpha - 3c)} \tag{A.1}$$

and

$$q_L^{*\prime}(c) = -\frac{\alpha/2}{\frac{2\tau\gamma}{\mu} - (\alpha - c)} - \frac{(\alpha - c)(\sigma_H - \sigma_L)}{\frac{2\tau\gamma}{\mu} - (2\alpha - 3c)}, \tag{A.2}$$

where primes denote derivatives with respect to s. Solving $q_H^{*\prime}(c) = 0$ and $q_L^{*\prime}(c) = 0$ yields, respectively, the two pairs of candidate solutions

$$\begin{aligned} c_{HH} = \alpha - {} & \frac{4\tau\gamma(\sigma_H - \sigma_L) + 3\alpha\mu}{4\mu(\sigma_H - \sigma_L)} \\ & \pm \frac{\sqrt{8\tau\gamma(\sigma_H - \sigma_L)[2\tau\gamma(\sigma_H - \sigma_L) + \alpha\mu] + (\alpha\mu)^2[9 - 8(\sigma_H - \sigma_L)]}}{4\mu(\sigma_H - \sigma_L)} \end{aligned} \tag{A.3}$$

and

$$\begin{aligned} c_{HL} = \alpha - {} & \frac{4\tau\gamma(\sigma_H - \sigma_L) - 3\alpha\mu}{4\mu(\sigma_H - \sigma_L)} \\ & \pm \frac{\sqrt{8\tau\gamma(\sigma_H - \sigma_L)[2\tau\gamma(\sigma_H - \sigma_L) - \alpha\mu] + (\alpha\mu)^2[9 + 8(\sigma_H - \sigma_L)]}}{4\mu(\sigma_H - \sigma_L)}. \end{aligned} \tag{A.4}$$

Consider, first, the two candidate solutions to $q_H^{*\prime}(c) = 0$. Start by noting that the discriminant is always positive, which implies that both roots are real. In order to show that the smaller root is not in the admissible set of values of c, (c_{min}, ∞), it suffices to show that it is less than any of the arguments on the right-hand side of (11). This is true if, for example, the following inequality holds:

$$\begin{aligned} \alpha - {} & \frac{4\tau\gamma(\sigma_H - \sigma_L) + 3\alpha\mu}{4\mu(\sigma_H - \sigma_L)} \\ & - \frac{\sqrt{8\tau\gamma(\sigma_H - \sigma_L)[2\tau\gamma(\sigma_H - \sigma_L) + \alpha\mu] + (\alpha\mu)^2[9 - 8(\sigma_H - \sigma_L)]}}{4\mu(\sigma_H - \sigma_L)} < \frac{2\alpha}{3} - \frac{2\tau\gamma}{3\mu}. \end{aligned} \tag{A.5}$$

The above inequality can be written as

$$\begin{aligned} & \frac{4\tau\gamma(\sigma_H - \sigma_L)}{3} + \left[3 - \frac{4(\sigma_H - \sigma_L)}{3}\right]\alpha\mu \\ & \qquad + \sqrt{8\tau\gamma(\sigma_H - \sigma_L)[2\tau\gamma(\sigma_H - \sigma_L) + \alpha\mu] + (\alpha\mu)^2[9 - 8(\sigma_H - \sigma_L)]} > 0, \end{aligned} \tag{A.6}$$

and it is always satisfied. Hence, the smaller root is ruled out. For the larger root to be in the admissible set of values of c, it must be greater than each of the three arguments on the right-hand side of (11). The corresponding three inequalities can be, respectively, written as

$$\sqrt{8\tau\gamma(\sigma_H-\sigma_L)[2\tau\gamma(\sigma_H-\sigma_L)+\alpha\mu]+(\alpha\mu)^2[9-8(\sigma_H-\sigma_L)]} > |4\tau\gamma(\sigma_H-\sigma_L)-3\alpha\mu|$$
$$\iff 8(\sigma_H-\sigma_L)\alpha\mu(4\tau\gamma-\alpha\mu)>0, \quad (A.7)$$

$$8(\tau\gamma)^2+2\tau\gamma\alpha\mu-(\alpha\mu)^2>0, \quad (A.8)$$

and

$$(\sigma_H-\sigma_L)(4\tau\gamma-\alpha\mu)[4(\sigma_H-\sigma_L)\tau\gamma+(2\ +\sigma_H-\sigma_L)\alpha\mu]>0. \quad (A.9)$$

The three inequalities hold simultaneously if $\gamma>\frac{\alpha\mu}{4\tau}$. Given (11), the denominators on the right-hand side of (A.1) are positive, which implies that the solution to $q_H^{*\prime}(c)=0$ in (c_{min},∞) must be less than α. This solution is therefore as given in (17).

Finally, note that the lower bound on c simplifies to $c_{min}=\frac{2\alpha}{3}-\frac{2\tau\gamma}{3\mu}$ if $\gamma>\frac{\alpha\mu}{4\tau}$. With $lim_{c\to c_{min}^+}q_H^{*\prime}(c)=\infty$ and $lim_{c\to\infty}q_H^{*\prime}(c)=-(\sigma_H-\sigma_L)/3<0$, existence and uniqueness of c_{HH} in (c_{min},∞) imply that $q_H^{*\prime}(c)>0$ for $c_{min}<c<c_{HH}$.

Consider, now, the two candidate solutions to $q_L^{*\prime}(c)=0$. Note that the discriminant is always positive, and both roots are therefore real.[14] Note again that, given (11), the denominators on the right-hand side of (A.2) are positive, and the solution to $q_L^{*\prime}(c)=0$ in (c_{min},∞) must therefore be greater than α. Thus, in order to show that the solution is uniquely given by the larger root, it suffices to show that this root is greater than α, while the smaller root is less than α. These two conditions hold simultaneously provided that

$$\sqrt{8\tau\gamma(\sigma_H-\sigma_L)[2\tau\gamma(\sigma_H-\sigma_L)-\alpha\mu]+(\alpha\mu)^2[9+8(\sigma_H-\sigma_L)]} > |4\tau\gamma(\sigma_H-\sigma_L)-3\alpha\mu|, \quad (A.10)$$

which simplifies to

$$8(\sigma_H-\sigma_L)\alpha\mu(2\tau\gamma+\alpha\mu)>0, \quad (A.11)$$

revealing that it is always satisfied. The solution to $q_L^{*\prime}(c)=0$ is therefore as given in (18).

With $lim_{c\to c_{min}^+}q_L^{*\prime}(c)=-\infty$ and $lim_{c\to\infty}q_L^{*\prime}(c)=(\sigma_H-\sigma_L)/3>0$, existence and uniqueness of c_{HL} in (c_{min},∞) imply that $q_L^{*\prime}(c)>0$ if $c>c_{HL}$.

[14] Note that the discriminant is $[4(\sigma_H-\sigma_L)\tau\gamma]^2-8(\sigma_H-\sigma_L)\tau\gamma\alpha\mu+(\alpha\mu)^2[9+8(\sigma_H-\sigma_L)]>0\forall\gamma>0$.

It remains to show that c_{HH}, c_{HL}, and the distance $c_{HL}-c_{HH}$ are increasing in μ and decreasing in τ. These results follow immediately from

$$\frac{\partial c_{HH}}{\partial \mu} = \frac{\tau\gamma}{\mu^2}\left(1 - \frac{4\tau\gamma(\sigma_H-\sigma_L)+\alpha\mu}{\sqrt{8\tau\gamma(\sigma_H-\sigma_L)[2\tau\gamma(\sigma_H-\sigma_L)+\alpha\mu]+(\alpha\mu)^2[9-8(\sigma_H-\sigma_L)]}}\right) > 0, \quad \text{(A.12)}$$

$$\frac{\partial c_{HL}}{\partial \mu} = \frac{\tau\gamma}{\mu^2}\left(1 - \frac{4\tau\gamma(\sigma_H-\sigma_L)-\alpha\mu}{\sqrt{8\tau\gamma(\sigma_H-\sigma_L)[2\tau\gamma(\sigma_H-\sigma_L)-\alpha\mu]+(\alpha\mu)^2[9+8(\sigma_H-\sigma_L)]}}\right) > 0, \quad \text{(A.13)}$$

$$\begin{aligned}\frac{\partial (c_{HL}-c_{HH})}{\partial \mu} = \frac{\tau\gamma}{\mu^2}\Bigg(&\frac{4\tau\gamma(\sigma_H-\sigma_L)+\alpha\mu}{\sqrt{8\tau\gamma(\sigma_H-\sigma_L)[2\tau\gamma(\sigma_H-\sigma_L)+\alpha\mu]+(\alpha\mu)^2[9-8(\sigma_H-\sigma_L)]}} \\ &- \frac{4\tau\gamma(\sigma_H-\sigma_L)-\alpha\mu}{\sqrt{8\tau\gamma(\sigma_H-\sigma_L)[2\tau\gamma(\sigma_H-\sigma_L)-\alpha\mu]+(\alpha\mu)^2[9+8(\sigma_H-\sigma_L)]}}\Bigg) > 0, \quad \text{(A.14)}\end{aligned}$$

$$\frac{\partial c_{HH}}{\partial \tau} = -\frac{\mu}{\tau}\frac{\partial c_{HH}}{\partial \mu} < 0, \quad \text{(A.15)}$$

$$\frac{\partial c_{HL}}{\partial \tau} = -\frac{\mu}{\tau}\frac{\partial c_{HL}}{\partial \mu} < 0, \quad \text{(A.16)}$$

and

$$\frac{\partial (c_{HL}-c_{HH})}{\partial \tau} = -\frac{\mu}{\tau}\frac{\partial (c_{HL}-c_{HL})}{\partial \mu} < 0. \quad \text{(A.17)}$$

Note that the term on the right-hand side of (A.14) is positive since

$$\begin{aligned}&[4\tau\gamma(\sigma_H-\sigma_L)+\alpha\mu]^2\{8\tau\gamma(\sigma_H-\sigma_L)[2\tau\gamma(\sigma_H-\sigma_L)-\alpha\mu]+(\alpha\mu)^2[9+8(\sigma_H-\sigma_L)]\} \\ &\quad - [4\tau\gamma(\sigma_H-\sigma_L)-\alpha\mu]^2\{8\tau\gamma(\sigma_H-\sigma_L)[2\tau\gamma(\sigma_H-\sigma_L)+\alpha\mu]+(\alpha\mu)^2[9-8(\sigma_H-\sigma_L)]\} \\ &\qquad = (\sigma_H-\sigma_L)(4\alpha\mu)^2\{[4(\sigma_H-\sigma_L)\tau\gamma]^2+(\alpha\mu)^2+8\tau\gamma\alpha\mu\} > 0. \quad \text{(A.18)}\end{aligned}$$

Appendix B Proof of Proposition 3

Using equations (5), (19), and (20), the effect of a marginal change in switching costs on average quality may be written as

$$\overline{q}'(c) = -\frac{\alpha/2}{\frac{2\tau\gamma}{\mu}-(\alpha-c)} + \frac{2[\tau-(\tau-s)\mu](\sigma_H-\sigma_L)^2(\alpha-c)\left(\frac{2\tau\gamma}{\mu}+c\right)}{\tau\left[\frac{2\tau\gamma}{\mu}-(2\alpha-3c)\right]^2}, \quad \text{(B.1)}$$

where prime denotes the derivative with respect to s.

Following the proof of Proposition 2 in Appendix A, let $\gamma > \frac{\alpha\mu}{4\tau}$. Under this condition, the lower bound on c simplifies to $c_{min} = \frac{2\alpha}{3} - \frac{2\tau\gamma}{3\mu}$.

Note that $c > c_{min}$ implies that the expressions $\frac{2\tau\gamma}{\mu} - (\alpha - c)$ and $\frac{2\tau\gamma}{\mu} + c$ on the right-hand side of (B.1) are positive. Thus, if a solution to $\overline{q}'(c) = 0$ exists in (c_{min}, ∞), it must be in (c_{min}, α). Let $c_{\overline{q}}$ denote such solution.

Existence of $c_{\overline{q}}$ follows from the Intermediate Value Theorem, given that $\overline{q}'(c)$ is continuous in (c_{min}, ∞) and that $lim_{c \to c_{min}^+} \overline{q}'(c) = \infty$ and $\overline{q}'(\alpha) = -\frac{\alpha\mu}{4\tau\gamma} < 0$.

To show that $c_{\overline{q}}$ is unique, I proceed in three steps: (i) I show that the graph of $\overline{q}'(c)$ first approaches the horizontal axis from above as c increases in (c_{min}, α); (ii) I show that $\overline{q}'(c)$ has either one or three roots in (c_{min}, α); and (iii) I show that there exists one solution to $\overline{q}'(c) = 0$ which is not in (c_{min}, α), implying that there can only be one solution in (c_{min}, α).

Because $lim_{c \to c_{min}^+} \overline{q}'(c) = \infty$, $\overline{q}'(c)$ is continuous in (c_{min}, α), and there is at least one $c_{\overline{q}}$, the smallest possible value of $c_{\overline{q}}$ is obtained when the graph of $\overline{q}'(c)$ first approaches the horizontal axis from above.

Note now that $\overline{q}'(\alpha) = -\frac{\alpha\mu}{4\tau\gamma} < 0$ implies two possible shapes of the graph of $\overline{q}'(c)$ for values of c greater than the smallest possible $c_{\overline{q}}$. If the graph of $\overline{q}'(c)$ does not cross the horizontal axis again, $c_{\overline{q}}$ is unique. This occurs if $\overline{q}'(c)$ is always decreasing or if it has a minimum for some value of c greater than the smallest possible $c_{\overline{q}}$. If the graph of $\overline{q}'(c)$ does cross the horizontal axis once more (implying that $\overline{q}'(c)$ becomes positive), then it must cross the horizontal axis at least a third time because $\overline{q}'(\alpha) < 0$. Note that the solutions to $\overline{q}'(c) = 0$ are the roots of a third degree polynomial. Hence, $\overline{q}'(c)$ has either one or three real roots in (c_{min}, α).

If there is only one root, $c_{\overline{q}}$ is unique. If there are three solutions and one is not in (c_{min}, α), then, from above, there can only be one solution in (c_{min}, α). That is, $c_{\overline{q}}$ is unique. Existence of a solution to $\overline{q}'(c) = 0$ which is not in (c_{min}, α) is established as follows. Given that $\overline{q}'(c)$ is continuous in $\left(\alpha - \frac{2\tau\gamma}{\mu}, c_{min}\right)$ and that $lim_{c \to \left(\alpha - \frac{2\tau\gamma}{\mu}\right)^+} \overline{q}'(c) = -\infty$ and $lim_{c \to c_{min}^-} \overline{q}'(c) = \infty$, by the Intermediate Value Theorem, there exists at least one solution to $\overline{q}'(c) = 0$ in $\left(\alpha - \frac{2\tau\gamma}{\mu}, c_{min}\right)$. Thus, $c_{\overline{q}}$ is unique.

This concludes the proof that $c_{\overline{q}} \in (c_{min}, \alpha)$ is implicitly defined by (21). Existence and uniqueness of $c_{\overline{q}}$ in (c_{min}, ∞), together with $lim_{c \to c_{min}^+} \overline{q}'(c) = \infty$ and $\overline{q}'(\alpha) < 0$, imply that that $\overline{q}'(c) > 0$ if $c < c_{\overline{q}}$.

It remains to prove that $c_{\overline{q}} \in (c_{HH}, \alpha)$ if (22) is verified. Given that $c_{\overline{q}}$ is unique and that

$\overline{q}'(\alpha) < 0$, by the Intermediate Value Theorem, $\overline{q}'(c_{HH}) > 0$ suffices for $c_{\overline{q}} > c_{HH}$. Formally,

$$\overline{q}'(c_{HH}) > 0 \iff q'_{\overline{q}}(c_{HH}) > q_H^{*\prime}(c_{HH}). \tag{B.2}$$

This condition may be rewritten as

$$\frac{2[\tau-(\tau-s)\mu](\sigma_H-\sigma_L)^2(\alpha-c_{HH})\left(\frac{2\tau\gamma}{\mu}+c_{HH}\right)}{\tau\left[\frac{2\tau\gamma}{\mu}-(2\alpha-3c_{HH})\right]^2} > \frac{(\alpha-c_{HH})(\sigma_H-\sigma_L)}{\frac{2\tau\gamma}{\mu}-(2\alpha-3c_{HH})}. \tag{B.3}$$

Solving for s yields (22). This concludes the proof of Proposition 3.

Appendix C Proof of Proposition 4

The proof of Proposition 4 is analogous to that of Proposition 3.

Using equations (2), (4), (5), and (24), the effect of a marginal change in switching costs on total patient welfare may be written as

$$W'(c) = -\left[\frac{\alpha/2}{\frac{2\tau\gamma}{\mu}-(\alpha-c)}+\frac{\mu}{2}\left(1-\frac{s}{\tau}\right)\right] + \frac{2[\tau-(\tau-s)\mu](\sigma_H-\sigma_L)^2(\alpha-c)\left[\frac{2\tau\gamma}{\mu}-(\alpha-2c)\right]}{\tau\left[\frac{2\tau\gamma}{\mu}-(2\alpha-3c)\right]^2}, \tag{C.1}$$

where prime denotes the derivative with respect to s.

Following the proof of Proposition 2 in Appendix A, let $\gamma > \frac{\alpha\mu}{4\tau}$. Under this condition, the lower bound on c simplifies to $c_{min} = \frac{2\alpha}{3} - \frac{2\tau\gamma}{3\mu}$.

Note that $c > c_{min}$ implies that the expressions $\frac{2\tau\gamma}{\mu} - (\alpha - c)$ and $\frac{2\tau\gamma}{\mu} - (\alpha - 2c)$ on the right-hand side of (C.1) are positive. Thus, if a solution to $W'(c) = 0$ exists in (c_{min}, ∞), it must be in (c_{min}, α). Let c_W denote such solution.

Existence of c_W follows from the Intermediate Value Theorem, given that $W'(c)$ is continuous in (c_{min}, ∞) and that $lim_{c\to c_{min}^+} W'(c) = \infty$ and $W'(\alpha) = -\left[\frac{\alpha\mu}{4\tau\gamma} + \frac{\mu}{2}\left(1-\frac{s}{\tau}\right)\right] < 0$.

To show that c_W is unique, I proceed in three steps: (i) I show that the graph of $W'(c)$ first approaches the horizontal axis from above as c increases in (c_{min}, α); (ii) I show that $W'(c)$ has either one or three roots in (c_{min}, α); and (iii) I show that there exists one solution to $W'(c) = 0$ which is not in (c_{min}, α), implying that there can only be one solution in (c_{min}, α).

Because $lim_{c\to c_{min}^+} W'(c) = \infty$, $W'(c)$ is continuous in (c_{min}, α), and there is at least one c_W, the smallest possible value of c_W is obtained when the graph of $W'(c)$ first approaches the horizontal axis from above.

Note now that $W'(\alpha) = -\left[\frac{\alpha\mu}{4\tau\gamma} + \frac{\mu}{2}\left(1 - \frac{s}{\tau}\right)\right] < 0$ implies two possible shapes of the graph of $W'(c)$ for values of c greater than the smallest possible c_W. If the graph of $W'(c)$ does not cross the horizontal axis again, c_W is unique. This occurs if $W'(c)$ is always decreasing or if it has a minimum for some value of c greater than the smallest possible c_W. If the graph of $W'(c)$ does cross the horizontal axis once more (implying that $W'(c)$ becomes positive), then it must cross the horizontal axis at least a third time because $W'(\alpha) < 0$. Note that the solutions to $W'(c) = 0$ are the roots of a third degree polynomial. Hence, $W'(c)$ has either one or three real roots in (c_{min}, α).

If there is only one root, c_W is unique. If there are three solutions and one is not in (c_{min}, α), then, from above, there can only be one solution in (c_{min}, α). That is, c_W is unique. Existence of a solution to $W'(c) = 0$ which is not in (c_{min}, α) is established as follows. Given that $W'(c)$ is continuous in $\left(\alpha - \frac{2\tau\gamma}{\mu}, c_{min}\right)$ and that $lim_{c \to \left(\alpha - \frac{2\tau\gamma}{\mu}\right)^+} W'(c) = -\infty$ and $lim_{c \to c_{min}^-} W'(c) = \infty$, by the Intermediate Value Theorem, there exists at least one solution to $W'(c) = 0$ in $\left(\alpha - \frac{2\tau\gamma}{\mu}, c_{min}\right)$. Thus, c_W is unique.

Existence and uniqueness of c_W in (c_{min}, ∞), together with $lim_{c \to c_{min}^+} W'(c) = \infty$ and $W'(\alpha) < 0$, imply that $W'(c) > 0$ if $c < c_W$.

It remains to prove that $c_W \in (c_{min}, \min\{c_{HH}, c_{\overline{q}}\})$. First, $W'(c) > 0$ requires that $q_H^{*\prime}(c) > 0$. From the proof of Proposition 2 in Appendix A, $q_H^{*\prime}(c) > 0$ for $c < c_{HH}$. Then it must be that $c_W < c_{HH}$. Second, given that $c_{\overline{q}}$ is unique and that $\overline{q}'(\alpha) < 0$, by the Intermediate Value Theorem, $\overline{q}'(c_W) > 0$ implies that $c_{\overline{q}} > c_W$. Formally,

$$\overline{q}'(c_W) > 0 \iff q'_{\overline{q}}(c_W) > W'(c_W). \tag{C.2}$$

This condition may be rewritten as

$$\frac{2[\tau - (\tau - s)\mu](\sigma_H - \sigma_L)^2(\alpha - c_W)^2}{\tau\left[\frac{2\tau\gamma}{\mu} - (2\alpha - 3c_W)\right]^2} > -\frac{\mu}{2}\left(1 - \frac{s}{\tau}\right), \tag{C.3}$$

which is always satisfied. From the proof of Proposition 3 in Appendix B, $c_{\overline{q}} \lesseqgtr c_{HH}$. Hence, this concludes the proof that $c_W \in (c_{min}, \min\{c_{HH}, c_{\overline{q}}\})$ is implicitly defined by (25).

Most Recent Working Paper

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
NIPE WP 15/2018	**Vítor Castro** and Rodrigo Martins, “Economic and political drivers of the duration of credit booms”, 2018
NIPE WP 14/2018	**Arash Rezazadeh** and **Ana Carvalho,** “Towards a survival capabilities framework: Lessons from the Portuguese Textile and Clothing industry”, 2018
NIPE WP 13/2018	**Areal, Nelson** and **Ana Carvalho**, “Shoot-at-will: the effect of mass-shootings on US small gun manufacturers”, 2018
NIPE WP 12/2018	**Rezazadeh, Arash** and **Ana Carvalho**, “A value-based approach to business model innovation: Defining the elements of the concept”, 2018