Universidade do Minho
Escola de Economia e Gestão

Luís Carlos de Sousa Sá

Essays on Hospital Behaviour and Regulation

outubro de 2020

Essays on Hospital Behaviour and Regulation
Luís Sá
UMinho | 2020

REPÚBLICA PORTUGUESA | CIÊNCIA, TECNOLOGIA E ENSINO SUPERIOR
FCT Fundação para a Ciência e a Tecnologia
POCH
PORTUGAL 2020
UNIÃO EUROPEIA Fundo Social Europeu

Universidade do Minho
Escola de Economia e Gestão

Luís Carlos de Sousa Sá

Essays on Hospital Behaviour and Regulation

Tese de Doutoramento
Doutoramento em Economia

Trabalho realizado sob a orientação do
Professor Odd Rune Straume

outubro de 2020

Acknowledgements

I owe this to Odd Rune Straume.

I also gratefully acknowledge the financial support provided by the Portuguese Foundation for Science and Technology (FCT) through the PhD Studentship SFRH/BD/129073/2017, financed by National Funds of the FCT and the European Social Fund.

Statement of Integrity

I hereby declare having conducted this academic work with integrity. I confirm that I have not used plagiarism or any form of undue use of information or falsification of results along the process leading to its elaboration. I further declare that I have fully acknowledged the Code of Ethical Conduct of the University of Minho.

Ensaios Sobre Concorrência Hospitalar e Regulação

Resumo

Esta tese analisa o comportamento estratégico de hospitais em mercados regulados e concorrenciais. No capítulo 2, é apresentado um modelo dinâmico em que os tempos de espera aumentam quando a procura por cuidados hospitalares excede a oferta; os pacientes escolhem um hospital tendo em consideração os tempos de espera; e os hospitais são alvo de penalizações àqueles associadas. Tais penalizações reduzem os tempos de espera, mas políticas que fomentam a livre escolha dos pacientes têm o efeito contrário. Estes resultados são robustos a diferentes métodos de resolução, à estrutra das penalizações e à formalização da utilidade dos pacientes. Mais, ainda que as penalizações sejam mais eficazes na redução dos tempos de espera quando a sua estrutura é linear, o efeito negativo da escolha é mitigado por penalizações quadráticas. Estas conclusões são parcialmente derivadas da calibração do modelo com tempos de espera e elasticidades observados no Serviço Nacional de Saúde inglês. Os capítulos 3 e 4 dedicam-se à inércia na procura por cuidados hospitalares, que, de acordo com a recente literatura empírica, resulta do efeito conjunto de custos de mudança e da persistência das preferências dos doentes. No capítulo 3, desenvolve-se um modelo com dois hospitais semi-altruístas e detentores de procura *retida* no qual o efeito da redução de custos de mudança (exógenos) no bem-estar dos pacientes depende da tecnologia de produção dos hospitais e do seu grau de altruísmo. Se a substituibilidade (complementaridade) entre qualidade e volume de tratamentos for suficientemente fraca (forte) relativamente ao altruísmo, a qualidade média e a utilidade agregada dos doentes caem. Adicionalmente, se os hospitais forem capazes de os controlar, os custos de mudança serão máximos e o mercado perfeitamente segmentado. O capítulo 4 trata da relação entre escolhas de hospital presentes e futuras gerada pela inércia da procura e investiga o efeito das expectativas dos pacientes na qualidade dos cuidados de saúde. Pacientes com expectativas *míopes* escolhem um hospital observando apenas variáveis presentes; pacientes *ingénuos* prevêem incorretamente o futuro, assumindo que a qualidade se manterá inalterada; e pacientes racionais prevêem a evolução da qualidade. Conquanto seja mais alta na presença de pacientes ingénuos do que na de míopes, a qualidade oscila entre ser máxima ou mínima sob expectativas racionais. Este resultado aplica-se igualmente aos ganhos de saúde dos doentes, sugerindo que a racionalidade nem sempre os beneficia.

Palavras-chave: concorrência hospitalar; custos de mudança; expectativas racionais; regulação; tempos de espera.

Essays on Hospital Behaviour and Regulation

Abstract

This thesis analyses the behaviour of competing hospitals in regulated markets. Chapter 2 presents a dynamic model where waiting times increase if demand exceeds supply; patients choose a hospital based in part on waiting times; and hospitals incur waiting time penalties. Whereas policies based on penalties will lead to lower waiting times, policies that promote patient choice will instead lead to higher waiting times. These results are robust to different game-theoretic solution concepts, designs of the hospital penalty structure, and patient utility specifications. Furthermore, waiting time penalties are likely to be more effective in reducing waiting times if they are designed with a linear penalty structure, but the counterproductive effect of patient choice policies is smaller when penalties are convex. These conclusions are partly derived by calibration of the model based on waiting times and elasticities from the English National Health Service. Chapters 3 and 4 analyse demand inertia in hospital markets, which, recent empirical evidence indicates, results from switching costs and persistent patient preferences. Chapter 3 offers a model with two semi-altruistic hospitals with inherited demand, where the effect of lower (exogenous) switching costs on patient welfare depends on the hospitals' technology and degree of altruism: if cost substitutability (complementarity) between quality and output is sufficiently weak (strong) relative to altruism, average quality and aggregate patient utility decrease. Additionally, if the hospitals can set the switching costs incurred by their patients, the unique equilibrium is characterised by maximum switching costs and perfect history-based market segmentation. Chapter 4 deals with the link between current and future choices of hospital generated by demand inertia, and investigates the effect of *patient expectations* on quality provision. Myopic patients choose a hospital based on current variables alone, forward-looking but naïve patients take the future into account but assume that quality remains constant, and forward-looking and rational patients foresee the evolution of quality. While it is higher under naïve than myopic expectations, quality provision under rational expectations may be highest or lowest. This result also holds for patients' health gains, suggesting that rationality does not always benefit patients.

Keywords: hospital competition; rational expectations; regulation; switching costs; waiting times.

Contents

List of Abbreviations

EHR Electronic Health Records

NHS National Health Service

OECD Organisation for Economic Co-operation and Development

List of Tables

1. Introduction

Promoting competition has become a staple of the regulation of healthcare markets in the western world over the past three decades (Siciliani et al., 2017). Provided that benefits—not necessarily monetary—from treating additional patients exist and that these are free to choose their preferred provider, the primary aim of competition is to elicit lower prices and higher quality of care as providers strive to attract demand. Since the seminal contribution of Arrow (1963), however, it has been recognised that healthcare markets differ from most private good markets and that these differences imply that results from those other markets, like the benefits of competition, may not necessarily carry over to healthcare ones. Those differences lie not only in the demand and supply sides of the market simultaneously but also in the ubiquitousness of regulation. On the demand side, notable examples are the uncertainty about future health needs and the difficulty in assessing the quality of care. On the supply side, examples of those differences are the departure from pure profit-maximisation and competition on non-price attributes. Regulators or welfarist policymakers, in turn, have manifold objectives and enact a variety of policy interventions. From a public health perspective, those objectives often include fostering patient welfare and health gains, which requires, for example, expanding market coverage, improving the quality of care, and reducing patients' costs besides out-of-pocket expenditures, like travelling costs or the mismatch between a patient's diagnosis and the provider's specialty mix. This entails incentivising some behaviours and discouraging others, while maintaining market conditions sufficiently attractive for providers. To do this, the regulator's toolkit extends from imposing legal requirements, increasing payments, or fine-tuning payment schemes to less stringent—though not less powerful—approaches, as promoting patient choice to toughen competition.

Despite the growing importance of primary care, hospitals continue to be a key provider of healthcare and persist as the most prominent provider of specialised and acute care. This thesis, a collection of self-contained essays, explores the strategic interaction between hospitals when they operate in a competitive setting where particular phenomena of the demand and supply of healthcare are considered. Additionally, it explains how regulation affects the behaviour of competing hospitals and to what extent policy

interventions conflict and yield the intended outcomes.

The analyses presented in chapters 2–4 are based on models that share the same duopolistic and spatial framework, where prices are regulated and patients are insulated from out-of-pocket expenses. These overarching modelling *choices* reflect characteristics of competitive hospital markets that guide this thesis: the reduced number of hospitals in most patients' choice set (Gutacker et al., 2016); the importance of horizontal differentiation, which may have a clinical interpretation or capture the long-established salience of geographical differentiation (for example, Tay, 2003; Varkevisser et al., 2012; Gutacker et al., 2016); and a focus on non-price competition. Regarding this last-mentioned aspect, competition plays out on a vertical dimension in the three models. They also share the departure from pure profit-maximisation as the hospitals' objective and include similar functional forms to represent the technology of treatment production. How each chapter develops this common framework by formalising particular attributes relevant for the research question in consideration, however, differs a great deal.

Chapter 2 considers an objective, observable, and negative form of quality upon which hospitals rely to compete: waiting times. In hospital markets with no significant out-of-pocket expenses and where capacity constraints are binding, waiting times act as non-monetary prices and bring demand and supply into equilibrium by acting as a rationing device. This market stabilisation mechanism is welcome, but bringing a hospital market into equilibrium through waiting times is welfare-decreasing if those waits become excessively long. As waiting times delay the benefits of medical treatment and force patients to undergo a period of less-than-achievable health status, they reduce patient welfare and thus become a deserving target of policy intervention once they exceed values deemed acceptable by some clinical metric. While increased activity and hence shorter waiting times might generally be encouraged through higher (activity-based) payments, growing healthcare costs call for alternative policies. In chapter 2, we investigate the effectiveness of two commonly adopted policies in reducing waiting times; namely, waiting time penalties and the enhancement of patient choice of provider. This chapter's major contribution is to analyse hospital competition on waiting times in a framework where the waiting time-generating process—i.e., the dynamic evolution of supply and demand—is explicitly modelled. Importantly, it is this modelling approach that uncovers the link between patient choice policies, supply, and waiting times in the presence of penalties (or, more generally, provider disutility of waiting time). More specifically, by modelling explicitly the evolution of waiting times as a function of the demand and supply of treatments, we reveal the role of activity as an instrument to avoid tougher penalties and the implications for waiting times of its weakening by increased patient choice.

The two subsequent chapters adopt a broader definition of quality of care and deal with the existence of demand inertia (also referred to as choice persistence or loyalty) in the hospital industry, which has only recently been reported in the empirical literature. In a context where patients are free to choose their preferred provider, demand inertia, to the extent that it reflects patients' inability to adjust to changes in the environment, brings into question whether competition is in fact taking place; at least, in markets where patient-hospital relationships have already been built up. Chapter 3 considers such a mature market with asymmetrically split inherited demand, and explores one particular driver of inertia, switching costs. The chapter analyses the effect of facilitated switching on quality provision and patient welfare, whether it results intentionally from competition-enhancing policies, like the removal of obstacles to patient choice, or emerges, for example, as a side effect of the adoption of sharable Electronic Health Records (EHR). Similarly to that of chapter 2, the analysis assumes a short-run policy perspective, and this perspective is short-run in that the effect of facilitated switching is never of a magnitude such that market dominance is altered. We therefore focus on the reduction of market concentration. In addition to looking at the largely unexplored observation of patient inertia, this chapter's main methodological contribution is to advance a modelling of inertia that maps on the recent empirical evidence, which shows that it results from both switching costs and persistent horizontal patient preferences. This allows us to explain how the effect of lower switching costs interacts with the volatility of patient preferences; namely, how demand may flow to the lower-quality hospital when switching is facilitated and patients hence adjust their choice of provider according to their changing preferences. The chapter closes with a model extension that shows that, when hospitals are endowed with the ability to control their patients' switching costs, they are indeed capable of transforming the market structure, giving rise to local monopolies with full market coverage.

Chapter 4 offers a longer-run analysis as it is primarily concerned with the intertemporal implications of demand inertia. By linking present and future choices of hospital, inertia implies that whether patients anticipate the future and how sophisticated their foresight is play a role in hospital markets. In other words, how patients form their expectations about future health needs and the quality of care affects the competitiveness of hospital markets. We argue and formalise the notion that demand inertia and expectations are inextricable, revealing that quality provision is governed by the tension between the hospitals' incentive to build market share, which will be partly retained due to inertia, and the responsiveness of demand to quality, which dictates how effective it is in attracting patients in the first place and is crucially determined by patient expectations. This chapter considers and models the limiting case of rational expectations and then builds on the behavioural literature to model two additional types of patient expectations based on

departures from full rationality reported therein: present bias and the imperfect assessment of healthcare attributes. This is the first of the chapter's major contributions: to model expectations in the context of patient choice of hospital and to study the impact different types of expectations on the hospitals' incentives to provide quality. The other major contribution is to relate the results derived from a theoretical model of hospital competition to the novel literature on 'behavioural hazard' in healthcare by discussing those results in light of the effect of rationality (or lack thereof) on patients' health gains.

Finally, chapter 5 summarises the main results and their interpretations and discusses limitations and how to overcome them within the scope of future research.

Chapters 2–4 frequently adopt identical notation and variable names; all references made in each chapter relate to variables and expressions defined within it.

2. Dynamic Hospital Competition Under Rationing by Waiting Times[1]

2.1 Introduction

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

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

Two common approaches are to link maximum wait guarantees either to penalties or to competition (and patient choice) policies. The first approach was followed by Finland and England, which combined maximum waiting times with sanctions for failure to fulfil the guarantee. Targets with penalties were

[1] This chapter is co-authored with Luigi Siciliani and Odd Rune Straume and was published in the *Journal of Health Economics* as Sá, L., Siciliani, L., and Straume, O.R. (2019) Dynamic hospital competition under rationing by waiting times. *Journal of Health Economics*, 66, 260–282. https://doi.org/10.1016/j.jhealeco.2019.06.005 .

introduced in England in 2000–05 with political oversight from the Prime Ministerial Delivery Unit and the Health Care Commission. Senior health administrators risked losing their jobs if targets were not met. As a result, the proportion of patients waiting over six months was reduced by 6–9 percentage points (Propper, Sutton, et al., 2008). In 2010, maximum wait guarantees became a patient entitlement codified into the NHS Constitution, establishing a patient right to a maximum of 18 weeks from GP referral to treatment. In Finland, waiting time guarantees were combined with targets as part of the Health Care Guarantee in 2005, subsequently included in the 2010 Health Care Act. A National Supervisory Agency supervised the implementation of the guarantee through targets and penalised municipalities failing to comply. The number of patients waiting over six months was reduced from 12.6 per 1000 population in 2002 to 6.6 per 1000 in 2005 (Siciliani, Moran, and Borowitz, 2013).

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

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

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

removed from the waiting list than those added. A second feature of our model is that hospitals compete for patients, with hospitals with lower waiting times attracting more patients.

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

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

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

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

presence of waiting time penalties. Moreover, higher waiting penalties make patient choice policies even more counterproductive. We also show that a combined policy of more patient choice and higher waiting time penalties will lead to higher waiting times if the waiting time penalty is sufficiently high to begin with.

The above described results are derived analytically for the case of constant marginal provider disutility of waiting time; for example, because of linear waiting time penalties. For the case of convex waiting time penalties, a closed-form solution cannot be obtained, and our results are therefore numerically derived. To make the results more salient, we calibrate our model based on waiting times observed in the English National Health Service (NHS) for a common treatment (cataract surgery). The calibration is also informed by demand elasticities which have been estimated in the empirical literature (Martin and Smith, 1999; Sivey, 2012).

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

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

The rest of the chapter is organised as follows. In the next section, we present a brief overview of the literature and explain how we contribute to it. In section 2.3, we present the model, whereas the main analysis, based on the closed-loop solution, is given in section 2.4. Section 2.5 considers patient welfare.

Section 2.6 examines the robustness of our main result to non-linear patient utility in waiting time and distance. Finally, section 2.7 provides concluding remarks, including a discussion of how our main results relate to the empirical literature on patient choice and waiting times.

2.2 Related Literature

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

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

The second strand of the literature relates to hospital competition with fixed prices. Though most of this literature consists of studies using a static framework, there is a limited but growing literature that models hospital competition in a dynamic framework. It focuses, however, on incentives for quality provision rather than on waiting times.[2] Brekke et al. (2010, 2012) find that, if quality is modelled as a stock variable

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

which increases if quality investments are higher than its depreciation, or if demand is sluggish so that an increase in quality only partially translates into an increase in demand, then quality is higher under the open-loop solution if hospitals face increasing marginal treatment costs. Equilibrium quality instead coincide under the two solution concepts if marginal treatment costs are constant. Siciliani, Straume, and Cellini (2013) suggest that these results can be overturned in the presence of altruistic preferences, so that quality is higher under the closed-loop solution.

Our modelling of waiting times differs analytically from these previous contributions because the state variable (i.e., waiting time) of the rival enters the dynamic constraint of the maximisation problem of each provider. This is not the case when quality is modelled as a stock (as in Brekke et al., 2010) because neither the state nor control variable of the rival provider enters the quality stock function. It is also not the case when demand is modelled as sluggish (as in Brekke et al., 2012, or Siciliani, Straume, and Cellini, 2013) because demand depends on the control variable of the rival, not the state variable. Thus, because of these fundamental differences in the dynamic nature of the problems, the results from models of dynamic quality competition do not automatically carry over to the case of waiting times. In other words, if we want to study the effects of patient choice and competition on waiting times in a dynamic context, we cannot simply interpret waiting time as 'negative quality' and apply the results from the above mentioned studies of dynamic quality competition.

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

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

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

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

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

2.3 The Model

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

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

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

$$u(x, z_i, t) = v - w_i(t) - \tau|x - z_i|, \tag{2.1}$$

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

$$u(x, z_i, t) = v - k - w_i(t) - \tau|x - z_i|. \tag{2.2}$$

For patients with a positive outside option, we assume that k is sufficiently high such that some of these patients will strictly prefer the outside option to being treated by any of the two hospitals in the market. This implies that the relevant choice for each of these patients is between seeking treatment at the most preferred hospital or exercising the outside option. We will refer to this as the *monopolistic segment* of

the market. For all the patients without a valuable outside option, we assume that utility is maximised by seeking treatment at one of the hospitals. These patients therefore constitute the *competitive segment* of the market. By concentrating on cases where the competitive segment is fully covered, whereas the monopolistic segment is only partially covered, we ensure that total demand is elastic with respect to waiting times, implying that waiting times have a rationing effect on demand.

2.3.1 Demand for hospital treatment

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

$$v - w_i(t) - \tau x_C = v - w_j(t) - \tau(1 - x_C), \tag{2.3}$$

yielding

$$x_C(t) = \frac{1}{2} + \frac{w_j(t) - w_i(t)}{2\tau}. \tag{2.4}$$

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

$$v - w_i(t) - \tau x_M^i = k, \tag{2.5}$$

yielding

$$x_M^i(t) = \frac{v - k - w_i(t)}{\tau}. \tag{2.6}$$

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

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

$$D_i(w_i(t), w_j(t)) = N[\beta x_C(t) + (1 - \beta)x_M^i(t)] \tag{2.7}$$

and

$$D_j(w_i(t), w_j(t)) = N[\beta(1 - x_C(t)) + (1 - \beta)x_M^j(t)]. \tag{2.8}$$

2.3.2 Hospital objectives and treatment supply

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

T. The instantaneous objective function of Hospital i is assumed to be

$$\Pi_i(t) = T + pS_i(t) - C(S_i(t)) - \Phi(w_i(t)). \tag{2.9}$$

The cost of supplying hospital treatments is given by an increasing and strictly convex cost function $C(S_i(t)) = \frac{\gamma}{2}S_i(t)^2$, with $\gamma > 0$. The convexity of the cost function captures an important feature in the context of waiting times, namely that hospitals face capacity constraints.[3] The function $\Phi(w_i(t))$ captures the provider disutility of having positive waiting times. The disutility of waiting time is monetary if the hospital faces penalties levied by the regulator or reductions in funding. Alternatively, it is non-monetary if the hospital takes into the account the reputational damage of reporting long waiting times, or if the hospital is subject to a more stringent monitoring regime by the regulator. We assume that the disutility of waiting time takes the linear-quadratic form

$$\Phi(w_i(t)) = \alpha_1 w_i(t) + \frac{\alpha_2}{2} w_i(t)^2, \tag{2.10}$$

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

Hospital targets are set for broad areas of care, typically all elective (non-emergency) care. Only in recent years some more stringent maximum waiting times have been specified for prioritised areas of care, such as cancer patients or certain cardiac surgeries (Siciliani, Moran, and Borowitz, 2013). Although our model is specified for a specific treatment which is reimbursed with DRG price p, any increase in supply for a specific treatment will contribute to reduce waiting times and help to satisfy the targets across all elective care. In subsection 2.4.3, we calibrate the model for a specific treatment, cataract surgery. We choose this procedure because it has high volume and is correlated with waiting times for other high-volume procedures (such as hip and knee replacement; Siciliani et al., 2014). It has also similar demand elasticity to waiting across all elective care (Martin and Smith, 1999; Sivey, 2012).

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

Waiting times evolve dynamically over time according to

$$\frac{dw_i(t)}{dt} = \dot{w}_i(t) = \theta[D_i(w_i(t), w_j(t)) - S_i(t)] \tag{2.11}$$

and

$$\frac{dw_j(t)}{dt} = \dot{w}_j(t) = \theta[D_j(w_i(t), w_j(t)) - S_j(t)], \tag{2.12}$$

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

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

We assume that the hospitals maximise their payoffs over an infinite time horizon and have a common constant discount rate, ρ. Formally, the maximisation problem of Hospital i is given by

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

Although, in reality, hospitals do not plan their activity over an infinite time horizon, we argue that this is a reasonable approximation if hospitals are regarded as lasting institutions. Managerial and medical structures are periodically replaced, but the hospital's *mission*—to provide care given its production tech-

nology and the regulatory scheme it faces—is likely to remain the same over long periods of time. This is likely if hierarchies are substituted by others with similar objective functions.

2.3.3 Solution concepts

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

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

The appropriateness of each solution concept depends on the assumptions regarding the players' information set as well as commitment requirements. The open-loop solution implies that hospitals have no information concerning waiting times once the game starts or are committed to the supply plans computed at the beginning of the game, which might be considered an excessively stringent assumption. Due to regulatory requirements, hospitals periodically collect and report data on waiting times, upon which their activity may be conditioned.[4] Moreover, a setting in which hospitals adjust activity according to waiting times is more realistic and relevant for policy-making.[5] Thus, although the closed-loop solution is computationally much more demanding, it is based on a set of assumptions that are arguably more realistic and we will therefore conduct our main analysis under the assumption that hospital behaviour is characterised by closed-loop decision rules.

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

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

2.4 Treatment Supply and Waiting Times in the Closed-Loop Solution

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

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

2.4.1 Constant marginal provider disutility of waiting time

Suppose that the disutility of waiting time is given by (2.10) with $\alpha_1 > 0$ and $\alpha_2 = 0$. In this case, it can be shown (see Appendix 2.A.1) that the optimal supply rule for each hospital at time t is equal to the steady-state supply, S^{CL}, and given by

$$S_i(t) = S_j(t) = S^{CL} = \frac{p}{\gamma} + \frac{2\theta\tau\alpha_1}{\gamma\phi}, \tag{2.13}$$

where

$$\phi = \theta(2-\beta)N + 2\tau\rho - \frac{(\theta\beta N)^2}{\theta(2-\beta)N + 2\tau\rho} \in (0,1). \tag{2.14}$$

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

Proposition 2.1. *If the marginal provider disutility of waiting time is constant, the equilibrium is characterised by constant supply of treatments over time.*

This result is explained by the lack of strategic interaction between the hospitals when waiting time disutility is linear in waiting times. A unilateral increase in supply by Hospital i leads to an initial reduction in waiting times at this hospital. This will shift demand from the rival hospital and therefore will also reduce the waiting time at Hospital j. However, if $\alpha_2 = 0$, the reduction in waiting time at Hospital j does not affect the hospital's marginal disutility of waiting time, so that the hospital will not respond by changing its supply.[6]

The intuition behind each hospital's optimal supply rule is perhaps easier gained by re-writing (2.13) as

$$p + \frac{2\theta\tau\alpha_1}{\phi} = \gamma S_i. \tag{2.15}$$

On the one hand, a marginal increase in supply (i) generates more revenues and (ii) reduces the waiting time and its associated disutility. These two elements of the marginal benefit of supply are given by the two terms on the left-hand side of (2.15). On the other hand, increasing supply is costly, with the marginal cost of supply given by the right-hand side of (2.15). Each hospital offers a supply of treatments such that the marginal benefit is exactly offset by the marginal cost. This trade-off is key to understanding the main intuition behind most of our subsequently derived results.

It also follows directly from (2.15) that, in an interior-solution equilibrium, each hospital operates at a level where the price-cost margin is negative, implying that the marginal patient is unprofitable to treat.[7] This is a result of the disutility of waiting time, which gives each hospital an incentive to expand supply beyond the level where the price is equal to marginal treatment costs.

The corresponding steady-state waiting time is given by[8]

$$w^{CL} = \frac{\tau}{(1-\beta)N}\left\{N\left[\frac{\beta}{2} + (1-\beta)\left(\frac{v-k}{\tau}\right)\right] - \frac{p}{\gamma} - \frac{2\theta\tau\alpha_1}{\gamma\phi}\right\}. \tag{2.16}$$

We can see directly from (2.16) that the steady-state waiting time is decreasing in p and α_1, which is very intuitive. A higher price (p) makes the marginal patient more profitable (or less unprofitable) to treat,

[6] When $\alpha_2 = 0$, our differential game belongs to the class of the so-called linear-state games, which is characterised by the coincidence between the time path of controls and states under the open- and closed-loop solution concepts. The calibration in subsection 2.4.3 illustrates this general result.

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

[8] In Appendix 2.A.1, we show that a sufficiently large γ ensures that the steady-state is characterised by non-negative waiting times and a partially covered monopolistic segment.

whereas a higher waiting time penalty (α_1) increases the disutility of waiting time. In both cases, the hospitals have stronger incentives to increase supply and equilibrium waiting times will therefore go down.

Patient choice and waiting times

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

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

$$\frac{\partial w^{CL}}{\partial \tau} = -\frac{(1-\beta)x_M^{CL} + \frac{\tau}{N}\frac{\partial S^{CL}}{\partial \tau}}{1-\beta} < 0, \tag{2.17}$$

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

$$\frac{\partial S^{CL}}{\partial \tau} = N\theta^2\alpha_1 \frac{(1-\beta)[N\theta(2-\beta)+4\tau\rho]\theta N + (2-\beta)(\tau\rho)^2}{2\gamma(N\theta+\tau\rho)^2[N(1-\beta)\theta+\tau\rho]^2} > 0, \tag{2.18}$$

allowing us to establish the following result:

Proposition 2.2. *If the marginal provider disutility of waiting time is constant, a higher degree of patient choice leads to lower treatment supply and higher waiting times in the steady-state.*

The negative relationship between τ and w^{CL} is a consequence of two effects that work in the same direction. First, there is a direct demand effect. A reduction in τ increases total demand (and hence demand for each hospital) since a larger number of patients in the monopolistic segment chooses to opt for treatment (at the nearest hospital). A higher demand directly increases the waiting time at each hospital. This effect is given by the first term in the numerator of (2.17), and the size of this effect is smaller the larger the relative size of the competitive segment, β.

The second effect is related to how τ affects the demand responsiveness to waiting times in the competitive segment of demand, and is thus more directly related to the patient choice interpretation of the parameter τ. This is an indirect effect that works through changes in each hospital's incentive to affect waiting times through its treatment supply decision. Each hospital can lower its waiting time by increasing the supply of treatments, and the effect of a unilateral increase in treatment supply on the waiting time is given by a direct and an indirect (feedback) effect. For a given demand, an increase in treatment

supply will reduce the waiting time. However, a lower waiting time will increase demand and therefore dampen the initial reduction in the waiting time. Crucially, the strength of this feedback effect depends on how strongly demand responds to waiting time changes. A lower τ makes demand more responsive to changes in waiting times, which increases the feedback effect and therefore makes treatment supply a less effective instrument to reduce waiting times. Consequently, this *reduces* the marginal benefit of treatment supply and gives each hospital an incentive to reduce the supply of treatments. This effect is captured by the second term in the numerator of (2.17).

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

Combining patient choice policies with waiting time penalties

Suppose that policymakers aim at reducing hospital waiting times. Two commonly suggested policy options is to either directly target the perceived problem by introducing (or increasing) waiting time penalties, or to stimulate patient choice (e.g., by public reporting of waiting times) with the aim of achieving lower waiting times through increased intensity of competition between the hospitals. In our model, as Proposition 2.2 shows, only the former policy works, whereas the latter policy is counterproductive. Moreover, the former policy makes the latter policy more counterproductive. All else equal, the larger the waiting time penalties, the larger is the increase in steady-state waiting times as a result of more patient choice.

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

times is given by

$$\frac{\partial w^{CL}}{\partial \alpha_1} - \frac{\partial w^{CL}}{\partial \tau} = \frac{1}{N(1-\beta)} \left[(1-\beta) N x_M^{CL} + \tau \left(\frac{\partial S^{CL}}{\partial \tau} - \frac{2\theta\tau}{\gamma\phi} \right) \right]. \tag{2.19}$$

If we exclude the demand effect of lower travelling costs, thus focusing exclusively on the patient choice interpretation of τ, the overall effect of this dual policy on waiting times is given by the sign of the second term in the square brackets of on the right-hand side of (2.19). It can easily be shown that the sign of this effect is positive, implying higher waiting times, if

$$\alpha_1 > \frac{(N\theta + \tau\rho)\left((1-\beta) N\theta + \tau\rho\right)\left((2-\beta) N\theta + 2\tau\rho\right)\tau}{N\theta\left(N\theta(1-\beta)\left((2-\beta) N\theta + 4\tau\rho\right) + (2-\beta)\tau^2\rho^2\right)}. \tag{2.20}$$

Thus, a combined policy of increased patient choice and higher waiting time penalties is more likely to yield higher waiting times (and lower treatment supply) if the disutility of waiting time is sufficiently high to begin with. The reason is that the marginal effect of a higher waiting time penalty on waiting times is constant (as we can see from (2.16)), whereas the marginal effect of increased patient choice on waiting times is increasing in the disutility of waiting times. Consequently, the counterproductive effect of increased patient choice dominates for sufficiently high values of α_1. It can also be shown that, unless β is very close to 1, the right-hand side of (2.20) is decreasing in θ and approaches τ as $\theta \to \infty$. This implies that the scope for a waiting time increase as a result of the combined policy is larger the faster waiting times adjust to changes in supply.

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

Proposition 2.3. *Suppose that waiting time penalties are linear in waiting times. In this case, (i) the counterproductive effect of patient choice policies on treatment supply and waiting times is larger the higher the waiting time penalty. Furthermore, (ii) a combined policy of increased patient choice and higher waiting time penalties has an ambiguous effect on treatment supply and waiting times, but is more likely to be counterproductive the higher the initial waiting time penalty.*

2.4.2 Increasing marginal provider disutility of waiting time

Suppose that the disutility of waiting time is given by (2.10) with $\alpha_1 > 0$ and $\alpha_2 > 0$. In this scenario, a closed-form solution of supply and waiting times cannot be obtained. Our game belongs to the class of linear-quadratic differential games wherein the state variables enter the objective function quadratically,

while they enter the dynamic constraints linearly. Although the closed-loop solution of linear-quadratic games may generally be computed analytically, this is not always assured. This is the case of our model whose particular structure features both state variables entering the dynamic constraints and has algebraic properties that limit the tractability of its closed-loop solution.

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

Proposition 2.4. *If the marginal provider disutility of waiting time is increasing, the optimal closed-loop supply rule for Hospital i is given by:*

$$S_i(w_i, w_j, t) = \frac{p - \theta(\omega_1 + \omega_3 w_i(t) + \omega_5 w_j(t))}{\gamma}, \tag{2.21}$$

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

See Appendix 2.A.2 for the definition Ω and proof of Proposition 2.4.

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

The relationship between the treatment supply at Hospital i and the waiting time at Hospital j is determined be the sign of ω_5. Although it is not possible to unambiguously determine the sign of ω_5 analytically (see Appendix 2.A.2), our calibration results provided in the next subsection show that ω_5 is negative for all the parameter configurations considered. If ω_5 is negative, then hospitals' supply decisions are characterised by strategic *substitutability*, $\partial S_i(w_i, w_j)/\partial w_j > 0$, for which we provide the following intuition. Consider a unilateral increase in supply by Hospital i. This leads to lower waiting times at Hospital i, which in turn shifts demand from Hospital j to Hospital i, causing a reduction in waiting times also at Hospital j. A lower waiting time at Hospital j reduces its marginal disutility of waiting time, and thus its marginal benefit of supply. Hospital j will therefore optimally respond by reducing its supply of treatments. In other words, a supply increase by Hospital i triggers a supply decrease by Hospital j.

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

In Appendix 2.A.2, we also show that, if the initial waiting times are the same in both hospitals or if the average initial waiting time equals the steady-state waiting time, then waiting times, supply and demand in both segments of the market converge *monotonically* to the steady-state. In this case, if the condition $|\omega_3| > |\omega_5|$ holds, the equilibrium path to the steady-state is characterised by periods of increasing (decreasing) hospital activity and increasing (decreasing) waiting time, which is in line with Siciliani (2006) in a monopoly setting. Notice that $|\omega_3| > |\omega_5|$ implies that the own waiting time effect on hospital activity is larger than the effect of the waiting time of the competing hospital, which is both intuitive and confirmed by our calibration exercise below.[9]

However, *non-monotonic* convergence may also arise. In Appendix 2.A.2 we show that, if the average initial waiting time is above (below) the steady-state waiting time, the hospital with the shortest (longest) initial waiting time might experience a non-monotonic convergence along the equilibrium path, with the waiting time first increasing (decreasing) before decreasing (increasing) towards the steady-state. One policy implication is that short-run provider performance on waiting times may not be representative of its long-run one.

2.4.3 Calibration

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

Waiting time data for cataract surgery is obtained from the Hospital Episode Statistics published by NHS Digital. Table 2.1 reports the mean and median waiting times (in days) for a cataract procedure

[9] Additionally, it follows from equations (2.53) and (2.54) in Appendix 2.A.2 that $|\omega_3| > |\omega_5|$ is a sufficient (but not necessary) condition for convergence to be verified.

provided either by NHS hospitals or the independent sector (private hospitals treating publicly-funded patients).[10]

Table 2.1: Evolution of Waiting Times for Cataract Procedures in the English NHS

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

Waiting times have remained relatively stable in recent years. They coincide with a period in which NHS England (the main regulator) did not specify performance standards for non-emergency care (The King's Fund, 2017). We interpret this as a regime where no significant penalties have been imposed on providers with longer waits. Within our model this corresponds to the special case when there is no hospital disutility of waiting time ($\alpha_1 = \alpha_2 = 0$). We therefore use the data in Table 2.1 as a measure of waiting times in a steady-state with no penalties, which we denote by superscript s. To make the analysis consistent with the study of Propper et al. (2010), we employ the mean waiting time, measured in months, and focus on the financial year 2016-17, giving $w^s = 2.3$.

According to the National Schedule of Reference Costs from NHS Improvement, 234 NHS providers performed $286,596$ cataract procedures in the same year.[11] This gives a monthly average of approximately 100 procedures per provider, so that $S^s = D^s = 100$.

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

On the *demand* side, the key parameters are the potential demand, the size of the competitive segment, the demand responsiveness, the gross valuation of treatment, and the value of the outside option. These parameters are less easy to obtain but we infer them in the following way. According to OECD (2018), 10.5% of the UK population was covered by private health insurance in 2015. We assume that

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

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

patients with private insurance opt for private treatment and that publicly-funded cataract procedures account for about 90% of the market.[12] Given that the steady-state supply in each hospital is $S^s = 100$, potential demand across the two hospitals is then given by $N = 222$.

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

$$\frac{\partial D_i(w_i(t), w_j(t))}{\partial w_i(t)} \frac{w^s}{D^s} = -\frac{N(2-\beta)}{2\tau} \frac{w^s}{D^s} = -\frac{222(2-\beta)}{2\tau} \frac{2.3}{100} = -0.1. \tag{2.22}$$

We do not know how large is the competitive segment. In order to account for patient heterogeneity, we conduct the analysis for three different values, $\beta = \{0.2, 0.5, 0.8\}$. We start by assuming $\beta = 0.2$, so that the competitive segment accounts for 20% of potential demand and is therefore relatively small, and then check how the results differ when it is 50% and 80% (relatively large). If $\beta = 0.2$, then, from (2.22), the demand elasticity implies that $\tau = 45.954$. Moreover, from the demand equation evaluated at the steady-state,

$$D^s = N \left[\frac{\beta}{2} + (1-\beta) \left(\frac{v - k - w^s}{\tau} \right) \right], \tag{2.23}$$

we can recover the difference between the gross valuation of treatment and the value of the outside option: $v - k = 22.4308$. If $\beta = 0.5$, then, from (2.22), we obtain $\tau = 38.295$ and, from (2.23), we obtain $v - k = 17.653$. If $\beta = 0.8$, then, from (2.22), we obtain $\tau = 30.636$ and, from (2.23), we obtain $v - k = 10.028$. We have thus recovered the demand-side parameters for $\beta = \{0.2, 0.5, 0.8\}$.

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

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

We are interested in understanding provider behaviour in the presence of penalties. We therefore need to identify plausible values for α_1 and α_2 under a penalty regime. In order to do this, we make use of the open-loop solution, for which we can derive a closed-form solution for the steady-state waiting time when $\alpha_2 > 0$ (see Appendix 2.B). We denote variables in the open-loop steady-state by the superscript OL. Propper et al. (2010) find that the introduction of waiting time penalties in the English NHS in 2000-05

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

reduced the mean waiting time by 13 days (i.e., 0.43 months). Although this estimate refers to an earlier period, it provides us with a plausible order of magnitude if such penalties were re-introduced in 2016-17. We then use this figure to compute the difference between the steady-state waiting time in the model with no disutility of waiting time and the open-loop steady-state waiting time, which is given by

$$w^s - w^{OL} = 2.3 - \frac{\gamma\phi\tau}{(1-\beta)\gamma\phi N + 2\theta\tau^2\alpha_2}\left\{N\left[\frac{\beta}{2} + (1-\beta)\left(\frac{v-k}{\tau}\right)\right] - \frac{p}{\gamma} - \frac{2\theta\tau\alpha_1}{\gamma\phi}\right\} = 0.43. \tag{2.24}$$

Inserting the above described parameter values when $\beta = 0.2$, the solution to (2.24) has one degree of freedom and is given by

$$\alpha_2 = 30.5274 - 0.53486\alpha_1. \tag{2.25}$$

All α_1 and α_2 that satisfy (2.25) yield a reduction of 0.43 months in the open-loop steady-state waiting time compared to the case with no disutility of waiting time. We consider three disutility structures: (i) linear disutility ($\alpha_2 = 0$), yielding $\alpha_1 = 57.0826$; (ii) quadratic disutility ($\alpha_1 = 0$), yielding $\alpha_2 = 30.5274$; and (iii) an intermediate case in which $\alpha_1 = \frac{57.0826}{2}$ and $\alpha_2 = \frac{30.5274}{2}$.

We insert all parameter values and solve the system (2.38)–(2.40) in Appendix 2.A to yield ω_1, ω_3, and ω_5, which are plugged into (2.59) to obtain the closed-loop steady-state waiting time. With ω_1, ω_3, ω_5, and w^{CL}, we use (2.21) to retrieve the closed-loop steady-state supply. For the open-loop steady-state waiting time and supply, we insert the parameter values into equations (2.75) and (2.77) in Appendix 2.B.

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

Linear versus convex waiting time disutility

The results generated by the above described calibration procedure are summarised in Table 2.2.

Our calibration results confirm that, as explained in subsection 2.4.2, the dynamic interaction introduced by increasing marginal disutility of waiting time leads to longer steady-state waiting times. As the waiting time disutility becomes more convex (i.e., more weight is placed on the quadratic term), the longer is the waiting time and the lower is supply in the closed-loop steady-state. The reason is simply that a more convex disutility function increases the magnitude of each hospital's supply response to changes in the waiting time, which reinforces each hospital's incentive to reduce supply in order to provoke a supply increase by the rival hospital, which in turn benefits the former hospital in the form of a lower waiting time. This result has potentially interesting policy implications, as it suggests that *linear penalties* are more effective in reducing waiting times, all else equal. Notice also that the importance of the design of

Table 2.2: Calibration Results for a Waiting Time Elasticity of Demand of -0.1

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

the penalty structure is larger for higher values of the competitive segment, β. This is intuitive, since the strategic substitutability relies on the existence of a competitive segment, wherein changes in the waiting time at one hospital affect demand faced by the rival hospital. Thus, a larger relative size of the competitive segment will magnify the effects of strategic substitutability.

Besides confirming that they coincide when $\alpha_2 = 0$, another key insight from Table 2.2 is that the difference in waiting times under the open- and closed-loop solutions is very small (less than 1%) when $\alpha_2 > 0$. This suggests that, even with non-linear penalties, the less computationally demanding open-loop solution offers a close approximation of the closed-loop one.

Higher waiting time elasticity of demand

One may worry that the results from Table 2.2 are due to the low demand elasticity. We therefore extend the analysis under the assumption that the waiting time elasticity is higher. We consider two additional cases. First, we assume that the elasticity is -0.2, twice as large, which is the highest that has been reported in studies for England (see Iversen and Siciliani, 2011, for an overview). Second, we assume that the elasticity is -1. This is an upper bound. There is only one study from Australia which provides such a large estimate (Stavrunova and Yerokhin, 2011), and this is consistent with the features of the Australian health system where more than half of the population is treated privately. Tables 2.3 and 2.4 provide the results for waiting time elasticities of demand of -0.2 and -1, and they are derived following

the steps detailed above. We see that an increase in the waiting time elasticity of demand reinforces the relative effectiveness of linear (as opposed to convex) waiting time penalties. Still, the quantitative difference between steady-state waiting times in the open- and closed-loop solutions remains small.

Table 2.3: Calibration Results for a Waiting Time Elasticity of Demand of -0.2

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

Table 2.4: Calibration Results for a Waiting Time Elasticity of Demand of -1

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

Higher waiting times and hospital heterogeneity

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

Since long waiting times may be observed both at hospitals with high and low volumes, we recalibrate the model with the higher baseline waiting time ($w^s = 3.45$) and set steady-state supply respectively at $S^s = 300$ (high volume) and $S^s = 50$ (low volume) in Tables 2.5 and 2.6. This is in line with HES data that reveal significant dispersion in hospital volumes even at the upper tail of the waiting times distribution across all procedures.[13]

By repeating the steps outlined at the beginning of subsection 2.4.3, we obtain the results in Tables 2.5 and 2.6, which show that the effect of linear versus convex penalties is qualitatively similar to our previously derived results (in Tables 2.2–2.4). And again, the waiting times under the open-loop solution are very similar to those under closed-loop solution.

Table 2.5: Calibration Results for Larger Hospitals and a Higher Baseline Waiting Time

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

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

Table 2.6: Calibration Results for Smaller Hospitals and a Higher Baseline Waiting Time

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

Patient choice and waiting times

One of our main aims is to analyse the relationship between patient choice and waiting times. In line with the analysis in subsection 2.4.1, we therefore conduct comparative statics with respect to the patient choice parameter τ. The fourth and fifth columns of Table 2.7 show the effects (on steady-state waiting times and supply) of a 10% reduction in τ, with all other parameters kept unchanged from our main calibration analysis, which implies that the results displayed in Table 2.2 serve as a reference point of comparison. In the last two columns of Table 2.7, we report the equivalent effects of a combined policy package, where a 10% reduction in τ is accompanied by a 10% increase in waiting time penalties (equivalent to the last part of the analysis in subsection 2.4.1).

In qualitative terms, the effects of increased patient choice on steady-state waiting times and supply, as shown in the fourth and fifth columns of Table 2.7, confirm that the result stated in Proposition 2.2 generalises beyond the case of constant marginal disutility of waiting time. Regardless of the shape of the hospitals' waiting time disutility function, a reduction in τ leads to higher steady-state waiting times.[14]

However, even if more patient choice increases steady-state waiting times for all parameter configurations considered in Table 2.7, there is a clear pattern showing that this effect is quantitatively smaller if the waiting time disutility is more convex. The reason is that a reduction in τ has two counteracting

[14] In the open-loop solution, for which a closed-form solution can be derived also in the case of increasing marginal waiting time disutility (see Appendix 2.B), it is also easily shown that a reduction in τ leads to higher steady-state waiting times for all parameter values that are compatible with equilibrium existence.

Table 2.7: Steady-State Effects of Policy Reforms

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

[1] 10% reduction in τ

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

effects on steady-state supply when $\alpha_2 > 0$. On the one hand, a lower τ makes treatment supply a less effective instrument to reduce waiting times, as previously explained, which gives each hospital an incentive to reduce their supply. On the other hand, a lower τ also increases demand (from the monopolistic segment), which—all else equal—leads to higher waiting times. If the disutility of waiting time is strictly convex (i.e., if $\alpha_2 > 0$), such increase in waiting time increases the marginal disutility of waiting time and therefore increases the marginal benefit of supply. In other words, with a strictly convex waiting time disutility function, the waiting time increase due to increased patient choice is partly dampened by the hospitals' incentives to increase supply in response to higher waiting times. Indeed, the fifth column in Table 2.7 shows that steady-state supply increases for the parameter configurations with $\alpha_2 > 0$.

This illustrates another aspect of the inherent conflict between waiting time penalties and patient choice policies, as previously discussed in subsection 2.4.1. On the one hand, waiting time penalties are more effective in reducing waiting times when they are designed as linear penalties (as shown by Tables 2.2–2.6). On the other hand, the counterproductive effect of patient choice policies on waiting times is larger when penalties are linear (as shown by Table 2.7).

The last two columns of Table 2.7 show the effects of a policy package where the increased in patient

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

2.5 Patient Welfare

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

$$U = 2N\beta \int_0^{\frac{1}{2}} (v - w^{CL} - \tau x)dx + 2N(1-\beta) \int_0^{x_M^{CL}} (v - k - w^{CL} - \tau x)dx, \tag{2.26}$$

and the effect of *lower* travelling costs is

$$\frac{\partial U}{\partial \tau} = -2D^{CL}\frac{\partial w^{CL}}{\partial \tau} - N\left[\frac{\beta}{4} + (1-\beta)(x_M^{CL})^2\right]. \tag{2.27}$$

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

The above approach takes a utilitarian perspective. Suppose that a health authority or regulator (a Ministry of Health) is only interested in the *health* component of patient welfare (Gravelle and Siciliani, 2008c). This approach has been sometimes referred as the extra-welfarist approach since it ignores non-health components which affect patient utility. Aggregate health patient benefit, denoted B, at the symmetric steady-state, is

$$B = 2N\beta \int_0^{\frac{1}{2}} (v - w^{CL})dx + 2N(1-\beta) \int_0^{x_M^{CL}} (v - k - w^{CL})dx, \tag{2.28}$$

and the effect of lower travelling costs is

$$\frac{\partial B^W}{\partial \tau} = -2D^{CL}\frac{\partial w^{CL}}{\partial \tau} + 2(v - k - w^{CL})\frac{\partial S^{CL}}{\partial \tau}. \tag{2.29}$$

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

If providers' penalties are non-linear in waiting times, choice policies simultaneously increase waiting times and supply. Therefore, the effect on aggregate health benefit is in principle ambiguous. However, our calibration exercise shows that the supply effect is a second-order effect and that patient choice reduces aggregate health benefit also when $\alpha_2 > 0$. In more detail, Table 2.8 reports the percent change in B and U induced by a 10% reduction in τ, which is computed using the welfare values associated with Tables 2.2 and 2.7.

Table 2.8: Steady-State Effects of a 10% Reduction in Travelling Costs on Patient Welfare

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

2.6 Robustness

In order to facilitate analytical tractability, our model has a linear-quadratic structure. One implication is that patient (dis)utility is assumed to be linear in waiting times, and travelling costs are linear in distance. Here we will briefly evaluate whether our main result—that more patient choice leads to increased waiting times—is robust to a relaxation of these assumptions. Unfortunately, it is only possible to perform these robustness checks in the context of the open-loop solution. However, our previous analysis has shown that the open-loop solution is a very close approximation of the closed-loop solution in our setting. The two

solutions coincide if $\alpha_2 = 0$, and our calibration results show that the two solutions concepts produce quantitatively almost identical results if $\alpha_2 > 0$. More importantly, the positive relationship between patient choice and waiting times does not depend on the choice of the solution concept.

2.6.1 Non-linear patient disutility of waiting

Suppose that, in the patient utility functions (2.1) and (2.2), we replace w_i with a strictly increasing function $f(w_i)$. Total demand for Hospital i is then given by

$$D_i(w_i, w_j) = N\left\{\beta\left[\frac{1}{2} + \frac{f(w_j) - f(w_i)}{2\tau}\right] + (1-\beta)\left[\frac{v - k - f(w_i)}{\tau}\right]\right\}. \tag{2.30}$$

Let w^{OL} be the steady-state waiting time in the open-loop solution. In Appendix 2.B.1, we show that this solution exists if $f(\cdot)$ is either concave or convex with a sufficiently low degree of convexity. Furthermore, we also show that, under the conditions of equilibrium existence, $\partial w^{OL}/\partial\tau < 0$. Thus:

Proposition 2.5. *Regardless of whether patient utility is concave or convex in waiting time, the steady-state waiting time in the open-loop solution, if it exists, is increasing in the degree of patient choice.*

This result is not surprising, given the intuition behind the previously derived positive relationship between patient choice and steady-state waiting times, which is related to the responsiveness of demand to changes in waiting times. As long as increased patient choice makes demand more responsive to changes in waiting times, it becomes more difficult for each hospital to curb waiting times by unilaterally increasing supply, which in turn leads to longer steady-state waiting times at both hospitals. This mechanism only requires that patient utility decreases with longer waiting times; it does *not* depend on whether patient utility decreases at a faster or slower rate when waiting times increase. Thus, we conjecture that the result stated in Proposition 2.5 also holds in a closed-loop setting.

2.6.2 Non-linear patient disutility of travelling

Consider next the case where, in the patient utility functions (2.1) and (2.2), we replace $|x - z_i|$ with a strictly increasing function $g(|x - z_i|)$. This generalisation prevents a closed-form derivation of demand. However, by the Implicit Function Theorem, we can derive the demand responsiveness to waiting time as

$$\frac{\partial D_i(w_i(t), w_j(t))}{\partial w_i(t)} = -\frac{N}{\tau}\left(\frac{\beta}{\tau[g'(x_C(t)) + g'(1 - x_C(t))]} + \frac{(1-\beta)}{g'(x_M^i(t)))}\right) < 0 \tag{2.31}$$

and

$$\frac{\partial D_i(w_i(t), w_j(t))}{\partial w_j(t)} = \frac{N\beta}{\tau[g'(x_C(t)) + g'(1 - x_C(t))]} > 0. \tag{2.32}$$

Still using τ as an inverse measure of the degree of patient choice, we derive (see Appendix 2.B.2) the following result:

Proposition 2.6. *(i) The steady-state waiting time in the open-loop solution is increasing in the degree of patient choice if the patient disutility of travelling is either concave or not strongly convex in travelling distance. (ii) In the case of constant marginal provider disutility of waiting time, the open-loop steady-state waiting time is increasing in the degree of patient choice if it exists.*

Thus, unless patient utility is strongly convex in travelling distance, our main result holds also in the case of non-linear patient disutility of travelling. And it always holds in the case of linear waiting time penalties, given that the open-loop solution exists. The general condition stated in Proposition 2.6 covers, for example, the empirical specification of Sivey (2012), who assumes that the utility of English cataract patients is a function of the natural log of travel time.

2.7 Concluding Remarks

We have investigated whether increased competition through patient choice policies play a useful role in reducing waiting times and the extent to which such a role is altered in the presence of penalties for providers with long waits. Our main results suggest, perhaps surprisingly, that increased patient choice leads to *higher* waiting times and that patient choice policies are therefore *counterproductive* in this respect. Furthermore, in the presence of waiting time penalties, we have shown that larger penalties make patient choice policies even more counterproductive.

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

when patients' disutility of waiting is non-linear, and it also holds when patients' disutility of travelling is non-linear (though not too strongly convex).

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

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

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

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

Appendix 2.A Closed-Loop Solution

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

$$V^i(w_i, w_j) = \omega_0 + \omega_1 w_i + \omega_2 w_j + \frac{\omega_3}{2} w_i^2 + \frac{\omega_4}{2} w_j^2 + \omega_5 w_i w_j. \tag{2.33}$$

This value function must satisfy the Hamilton-Jacobi-Bellman (HJB) equation for Hospital i, which is given by[15]

$$\rho V^i(w_i, w_j) = \max \left\{ T + pS_i - \frac{\gamma}{2} S_i^2 - \alpha_1 w_i - \frac{\alpha_2}{2} w_i^2 + \theta \frac{\partial V^i}{\partial w_i}(D_i - S_i) + \theta \frac{\partial V^i}{\partial w_j}(D_j - S_j) \right\}. \tag{2.34}$$

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

$$S_i(w_i, w_j) = \frac{p - \theta(\omega_1 + \omega_3 w_i + \omega_5 w_j)}{\gamma}. \tag{2.35}$$

Substituting Hospital i's supply rule and the analogous supply rule for Hospital j into the HJB equation, together with (2.7)–(2.8), we obtain:

$$\begin{aligned} \rho V^i(w_i, w_j) = T + p \left[\frac{p - \theta(\omega_1 + \omega_3 w_i + \omega_5 w_j)}{\gamma} \right] - \frac{\gamma}{2} \left[\frac{p - \theta(\omega_1 + \omega_3 w_i + \omega_5 w_j)}{\gamma} \right]^2 - \alpha_1 w_i \\ - \frac{\alpha}{2} w_i^2 + \theta(\omega_1 + \omega_3 w_i + \omega_5 w_j) \left[\beta \left(\frac{1}{2} + \frac{w_j - w_i}{2\tau} \right) N + (1 - \beta) \left(\frac{v - k - w_i}{\tau} \right) N \right. \\ \left. - \frac{p - \theta(\omega_1 + \omega_3 w_i + \omega_5 w_j)}{\gamma} \right] \\ + \theta(\omega_2 + \omega_4 w_j + \omega_5 w_i) \left[\beta \left(\frac{1}{2} + \frac{w_i - w_j}{2\tau} \right) N + (1 - \beta) \left(\frac{v - k - w_j}{\tau} \right) N \right. \\ \left. - \frac{p - \theta(\omega_1 + \omega_3 w_j + \omega_5 w_i)}{\gamma} \right], \end{aligned} \tag{2.36}$$

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

which can be rewritten as

$$\begin{aligned}
&\left\{T+\frac{p^2}{2\gamma}+\sigma(\omega_1+\omega_2)+\frac{\theta^2}{2\gamma}\omega_1^2+\frac{\theta^2}{\gamma}\omega_1\omega_2-\rho\omega_0\right\}\\
&+w_i\left\{-\left[\rho+\frac{\theta(2-\beta)N}{2\tau}\right]\omega_1+\frac{\theta\beta N}{2\tau}\omega_2+\sigma(\omega_3+\omega_5)+\frac{\theta^2}{\gamma}\omega_1\omega_3+\frac{\theta^2}{\gamma}\omega_1\omega_5+\frac{\theta^2}{\gamma}\omega_2\omega_5-\alpha_1\right\}\\
&+w_j\left\{\frac{\theta\beta N}{2\tau}\omega_1-\left[\rho+\frac{\theta(2-\beta)N}{2\tau}\right]\omega_2+\sigma(\omega_4+\omega_5)+\frac{\theta^2}{\gamma}\omega_1\omega_4+\frac{\theta^2}{\gamma}\omega_1\omega_5+\frac{\theta^2}{\gamma}\omega_2\omega_3\right\}\\
&+w_i^2\left\{-\left[\frac{\rho}{2}+\frac{\theta(2-\beta)N}{2\tau}\right]\omega_3+\frac{\theta^2}{2\gamma}\omega_3^2+\frac{\theta\beta N}{2\tau}\omega_5+\frac{\theta^2}{\gamma}\omega_5^2-\frac{\alpha_2}{2}\right\}\\
&+w_j^2\left\{-\left[\frac{\rho}{2}+\frac{\theta(2-\beta)N}{2\tau}\right]\omega_4+\frac{\theta^2}{\gamma}\omega_3\omega_4+\frac{\theta\beta N}{2\tau}\omega_5+\frac{\theta^2}{2\gamma}\omega_5^2\right\}\\
&+w_iw_j\left\{\frac{\theta\beta N}{2\tau}(\omega_3+\omega_4)-\left[\rho+\frac{\theta(2-\beta)N}{\tau}\right]\omega_5+\frac{2\theta^2}{\gamma}\omega_3\omega_5+\frac{\theta^2}{\gamma}\omega_4\omega_5\right\}=0, \quad (2.37)
\end{aligned}$$

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

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

$$-\left[\frac{\rho}{2}+\frac{\theta(2-\beta)N}{2\tau}\right]\omega_3+\frac{\theta^2}{2\gamma}\omega_3^2+\frac{\theta\beta N}{2\tau}\omega_5+\frac{\theta^2}{\gamma}\omega_5^2-\frac{\alpha_2}{2}=0, \quad (2.38)$$

$$-\left[\frac{\rho}{2}+\frac{\theta(2-\beta)N}{2\tau}\right]\omega_4+\frac{\theta^2}{\gamma}\omega_3\omega_4+\frac{\theta\beta N}{2\tau}\omega_5+\frac{\theta^2}{2\gamma}\omega_5^2=0, \quad (2.39)$$

$$\frac{\theta\beta N}{2\tau}(\omega_3+\omega_4)-\left[\rho+\frac{\theta(2-\beta)N}{\tau}\right]\omega_5+\frac{2\theta^2}{\gamma}\omega_3\omega_5+\frac{\theta^2}{\gamma}\omega_4\omega_5=0. \quad (2.40)$$

2.A.1 Constant marginal provider disutility of waiting time

Consider first the closed-loop solution under constant marginal waiting time disutility. When $\alpha_2=0$, the system of equations (2.38)–(2.40) has a single candidate solution for which the value function is not convex with respect to w_i. The remaining five candidates have $\omega_3>0$ and cannot therefore constitute a solution the hospital's maximisation problem. The solution that yields a linear—hence, concave—value function with respect to w_i is $\omega_3=\omega_4=\omega_5=0$. This linearity of the value function with respect to waiting times is not surprising given the linear structure of the game when $\alpha_2=0$. Setting $\omega_3=\omega_5=0$ in (2.35), Hospital i's optimal supply rule becomes

$$S_i(w_i,w_j)=\frac{p-\theta\omega_1}{\gamma}, \quad (2.41)$$

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

With $\omega_3 = \omega_4 = \omega_5 = 0$, (2.37) simplifies to:

$$\left\{T + \frac{p^2}{2\gamma} + \sigma(\omega_1 + \omega_2) + \frac{\theta^2}{2\gamma}\omega_1^2 + \frac{\theta^2}{\gamma}\omega_1\omega_2 - \rho\omega_0\right\} + w_i\left\{-\left[\rho + \frac{\theta(2-\beta)N}{2\tau}\right]\omega_1 + \frac{\theta\beta N}{2\tau}\omega_2 - \alpha_1\right\} + w_j\left\{\frac{\theta\beta N}{2\tau}\omega_1 - \left[\rho + \frac{\theta(2-\beta)N}{2\tau}\right]\omega_2\right\} = 0. \quad (2.42)$$

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

$$\omega_1 = -\frac{\tau\alpha_1\left[2\rho\tau + \theta(2-\beta)N\right]}{2\left[\rho\tau + \theta(1-\beta)N\right]\left[\rho\tau + \theta N\right]} = -\frac{2\tau\alpha_1}{\phi}. \quad (2.43)$$

Inserting the expression for ω_1 into the optimal supply rule for hospitals i and j yields $S_i = S_j = S^{CL}$ as given by (2.13) in subsection 2.4.1. Using this result, the closed-loop steady-state waiting time is derived from the equations of motion (2.11)–(2.12), with $\dot{w}_i(t) = \dot{w}_j(t) = 0$. Simple algebra shows that $w_i = w_j = w^{CL}$ as given by (2.16) in subsection 2.4.1.

From (2.16), the steady-state waiting time is positive if and only if $p \leq \overline{p}$, given by

$$\overline{p} = \gamma N\left[\frac{\beta}{2} + (1-\beta)\left(\frac{v-k}{\tau}\right)\right] - \frac{2\theta\tau\alpha_1}{\phi}. \quad (2.44)$$

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

$$0 < \frac{v - k - w^{CL}}{\tau} < \frac{1}{2}. \quad (2.45)$$

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

$$\underline{p} = \frac{\beta\gamma N}{2} - \frac{2\theta\tau\alpha_1}{\phi}, \quad (2.46)$$

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

2.A.2 Increasing marginal provider disutility of waiting time

When $\alpha_2 > 0$, the solution to (2.38)–(2.40) depends on the root of a sixth degree polynomial, precluding the computation of an analytical solution. Assume, for now, that a solution exists and that it is such that (2.21) in Proposition 2.4 constitutes a Markov Perfect Nash Equilibrium.

From (2.38), two candidate solutions for ω_3 (as functions of ω_5) ensue:

$$\omega_3 = \frac{\gamma}{\theta^2}\left\{\left[\frac{\rho}{2}+\frac{\theta(2-\beta)N}{2\tau}\right] \pm \sqrt{\left[\frac{\rho}{2}+\frac{\theta(2-\beta)N}{2\tau}\right]^2 - \frac{2\theta^2}{\gamma}\left[\frac{\theta^2}{\gamma}\omega_5^2+\frac{\theta\beta N}{2\tau}\omega_5-\frac{\alpha_2}{2}\right]}\right\}. \tag{2.47}$$

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

$$\underline{\omega_5} = -\frac{\gamma}{2\theta^2}\left[\frac{\theta\beta N}{2\tau}+\sqrt{\left(\frac{\theta\beta N}{2\tau}\right)^2+\frac{2\theta^2\alpha_2}{\gamma}}\right] < 0, \tag{2.48}$$

$$\overline{\omega_5} = -\frac{\gamma}{2\theta^2}\left[\frac{\theta\beta N}{2\tau}-\sqrt{\left(\frac{\theta\beta N}{2\tau}\right)^2+\frac{2\theta^2\alpha_2}{\gamma}}\right] > 0. \tag{2.49}$$

Additionally, in order for (2.21) to be a Markov Perfect Nash Equilibrium, the value function must be bounded from above. A necessary and sufficient condition for this requirement to hold is that waiting times converge in equilibrium. Inserting (2.7), (2.8), (2.21), and the analogous supply rule for Hospital j into (2.11)–(2.12) yields the following system of differential equations:

$$\frac{\dot{w}_i}{\theta} = \left[-\frac{(2-\beta)N}{2\tau}+\frac{\theta}{\gamma}\omega_3\right]w_i + \left[\frac{\beta N}{2\tau}+\frac{\theta}{\gamma}\omega_5\right]w_j + N\left[\frac{\beta}{2}+(1-\beta)\left(\frac{v-k}{\tau}\right)\right] - \left(\frac{p-\theta\omega_1}{\gamma}\right), \tag{2.50}$$

$$\frac{\dot{w}_j}{\theta} = \left[\frac{\beta N}{2\tau}+\frac{\theta}{\gamma}\omega_5\right]w_i + \left[-\frac{(2-\beta)N}{2\tau}+\frac{\theta}{\gamma}\omega_3\right]w_j + N\left[\frac{\beta}{2}+(1-\beta)\left(\frac{v-k}{\tau}\right)\right] - \left(\frac{p-\theta\omega_1}{\gamma}\right). \tag{2.51}$$

The Jacobian of (2.50)–(2.51) is

$$J^{CL} = \theta\begin{bmatrix} -\frac{(2-\beta)N}{2\tau}+\frac{\theta}{\gamma}\omega_3 & \frac{\beta N}{2\tau}+\frac{\theta}{\gamma}\omega_5 \\ \frac{\beta N}{2\tau}+\frac{\theta}{\gamma}\omega_5 & -\frac{(2-\beta)N}{2\tau}+\frac{\theta}{\gamma}\omega_3 \end{bmatrix} \tag{2.52}$$

and its eigenvalues are

$$s_1 = \theta\left[-\frac{N}{\tau}+\frac{\theta}{\gamma}(\omega_3-\omega_5)\right] \tag{2.53}$$

and

$$s_2 = \theta\left[-\frac{(1-\beta)N}{\tau}+\frac{\theta}{\gamma}(\omega_3+\omega_5)\right]. \tag{2.54}$$

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

Using the expression for ω_3 as a function of ω_5, (2.47), the necessary condition $s_1 < 0 \wedge s_2 < 0 \wedge \omega_3 < 0$ is satisfied if $\omega_5 \in \Omega = \left(\max\{\underline{\omega_5}, \underline{\omega_5}'\}, \min\{\overline{\omega_5}, \overline{\omega_5}'\}\right)$, where

$$\underline{\omega_5}' = \frac{\gamma}{6\theta^2}\left[\rho - \frac{2\theta\beta N}{\tau} - \sqrt{\left(\rho - \frac{2\theta\beta N}{\tau}\right)^2 + \frac{12\theta^2}{\gamma}\left[\frac{\gamma N}{\theta\tau}\left(\rho + \frac{\theta(1-\beta)N}{\tau}\right) + \alpha_2\right]}\right] < 0, \tag{2.55}$$

$$\overline{\omega_5}' = \frac{\gamma}{6\theta^2}\left[-\left(\rho + \frac{2\theta\beta N}{\tau}\right) + \sqrt{\left(\rho + \frac{2\theta\beta N}{\tau}\right)^2 + \frac{12\theta^2}{\gamma}\left[\frac{\gamma(1-\beta)N}{\theta\tau}\left(\rho + \frac{\theta N}{\tau}\right) + \alpha_2\right]}\right] > 0. \tag{2.56}$$

Thus, provided that a solution to (2.38)–(2.40) exists, it constitutes a Markov Perfect Nash Equilibrium (or closed-loop equilibrium) if $\omega_5 \in \Omega$. Finally, an equilibrium with $\omega_5 = 0$ is ruled out by inspection of (2.38)–(2.40).

The eigenvalues given by (2.53)–(2.54) also provide confirmation that the supply rules derived in the previous subsection, under constant marginal disutility of waiting time, constitute a Markov Perfect Nash Equilibrium. It is straightforward to see from (2.53) and (2.54) that $s_1 < 0$ and $s_2 < 0$ when $\omega_3 = \omega_5 = 0$.

Transitional dynamics

In order to analyse the convergence to the steady-state in the closed-loop solution, we turn to its open-loop representation. That is, we derive time-profiles of the waiting time, supply, and demand from the optimal closed-loop supply rule. Let the superscript CL denote the closed-loop steady-state. The eigenvalues governing the system of differential equations (2.50)–(2.51), s_1 and s_2, are respectively associated with the eigenvectors $\nu_1 = c_1\begin{bmatrix}1 & -1\end{bmatrix}^T$ and $\nu_2 = c_2\begin{bmatrix}1 & 1\end{bmatrix}^T$, with $c_1, c_2 \in \mathbb{R}$. Setting $c_1 = c_2 = 1$, the solution of the system of differential equations (2.50)–(2.51) takes the form:

$$w_i(t) = C_1 e^{s_1 t} + C_2 e^{s_2 t} + w^{CL}, \tag{2.57}$$

$$w_j(t) = -C_1 e^{s_1 t} + C_2 e^{s_2 t} + w^{CL}, \tag{2.58}$$

where C_1 and C_2 are arbitrary constants. The closed-loop steady-state waiting time w^{CL} is retrieved by setting $\dot{w}_i = \dot{w}_j = 0$ in (2.50)–(2.51) and solving for w_i and w_j. This yields

$$w^{CL} = \frac{N\left[\frac{\beta}{2} + (1-\beta)\left(\frac{v-k}{\tau}\right)\right] - \left(\frac{p-\theta\omega_1}{\gamma}\right)}{\frac{(1-\beta)N}{\tau} - \frac{\theta}{\gamma}(\omega_3 + \omega_5)}. \tag{2.59}$$

Inserting the initial conditions $w_i(0) = w_{0i}$ and $w_j(0) = w_{0j}$ into (2.57)–(2.58) and solving for C_1 and C_2 gives $C_1 = \frac{w_{0i}-w_{0j}}{2}$ and $C_2 = \frac{w_{0i}+w_{0j}}{2} - w^{CL}$. Then, waiting times at Hospital i converge to the steady-state according to:

$$w_i(t) = \left(\frac{w_{0i} - w_{0j}}{2}\right) e^{s_1 t} + \left(\frac{w_{0i} + w_{0j}}{2} - w^{CL}\right) e^{s_2 t} + w^{CL}. \tag{2.60}$$

Consider, now, the dynamics of supply and demand. Inserting (2.60) and the analogous equation for $w_j(t)$ into (2.21) yields:

$$S_i(t) = \frac{\theta}{\gamma}\left[(\omega_5 - \omega_3)\left(\frac{w_{0i} - w_{0j}}{2}\right) e^{s_1 t} - (\omega_3 + \omega_5)\left(\frac{w_{0i} + w_{0j}}{2} - w^{CL}\right) e^{s_2 t}\right] + \frac{p - \theta[\omega_1 + (\omega_3 + \omega_5)w^{CL}]}{\gamma}. \tag{2.61}$$

Using (2.7), (2.60), and the analogous equation for $w_j(t)$, the dynamics of demand faced by Hospital i in the competitive and monopolistic segments are respectively given by

$$D_C^i(t) = \beta N\left[\frac{1}{2} + \left(\frac{w_{0j} - w_{0i}}{2\tau}\right) e^{s_1 t}\right] \tag{2.62}$$

and

$$D_M^i(t) = \frac{(1-\beta)N}{\tau}\left[v - k - w^{CL} + \left(\frac{w_{0j} - w_{0i}}{2}\right) e^{s_1 t} + \left(w^{CL} - \frac{w_{0i} + w_{0j}}{2}\right) e^{s_2 t}\right]. \tag{2.63}$$

If $w_{0i} = w_{0j}$, it follows from equations (2.60)–(2.63) that the dynamics of waiting times, supply, and demand are uniquely governed by s_2, and convergence is thus monotonic. By the same token, if the initial waiting times differ but their average equals the steady-state waiting time w^{CL}, dynamics are uniquely governed by s_1, and convergence is monotonic as well in this case. Note, additionally, that demand in the competitive segment always converges monotonically to $\beta N/2$.

For the transitional dynamics in the closed-loop solution under constant marginal disutility of waiting time, simply set $\omega_3 = \omega_5 = 0$ in equations (2.60)–(2.63). Constant hospital activity over time for $\alpha_2 = 0$ is then confirmed by (2.61).

Non-monotonic convergence

Equations (2.60)–(2.63) show that convergence to the steady-state depends on two, possible opposing, forces. It depends on whether a hospital's initial waiting time is longer than that of the rival, and whether the average initial waiting time in the market differs from the steady-state waiting time. When these

two conditions hold, the possibility of non-monotonic convergence arises. To see why non-monotonic convergence might occur, consider the equilibrium dynamics of waiting times described in (2.60). If the average initial waiting time is above (below) the steady-state, the first two terms have opposite signs for the hospital with the shorter (longer) waiting time. In both cases, whether or not non-monotonic convergence emerges depends on the relative size and speed of convergence (to zero) of each of those terms.

Differentiating (2.60) with respect to time and equating to zero yields a single critical point given by

$$t^* = \left(\frac{1}{s_1 - s_2}\right) ln \left[-\frac{s_2}{s_1}\left(\frac{w_{0i} + w_{0j} - 2w^{CL}}{w_{0i} - w_{0j}}\right)\right], \tag{2.64}$$

where s_1 and s_2 are given by (2.53) and (2.54). Convergence is non-monotonic for Hospital i if and only if $t^* \in \mathbb{R}^+$. With $s_1, s_2 < 0$, the first factor in (2.64) is negative if $|s_1| > |s_2|$. Thus, $t^* \in \mathbb{R}^+$ if and only if the second factor in (2.64) is defined and is negative, which requires that the expression in the square brackets lies between 0 and 1. It is possible to derive some easily interpretable conditions for this expression to be positive. Since $-\frac{s_2}{s_1} < 0$, we must have $\frac{w_{0i}+w_{0j}-2w^{CL}}{w_{0i}-w_{0j}} < 0$. Two cases then arise:

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

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

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

To conclude the proof, we consider the shape of (2.60). Evaluating its second-order derivative with respect to t at t^* yields the following results:

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

$$\left(\frac{s_1}{s_2}\right)^2 e^{(s_1-s_2)t^*}(w_{0i} - w_{0j}) < -(w_{0i} + w_{0j} - 2w^{CL}). \tag{2.65}$$

   Diving both sides by $(w_{0i} - w_{0j})$ reverses the inequality sign. Then, using (2.64), the inequality becomes $\frac{s_1}{s_2} > 1$, which is true.

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

$$\left(\frac{s_1}{s_2}\right)^2 e^{(s_1-s_2)t^*}(w_{0i} - w_{0j}) > -(w_{0i} + w_{0j} - 2w^{CL}). \tag{2.66}$$

Diving both sides by $(w_{0i} - w_{0j})$ does not reverse the inequality sign. Then, using (2.64), the inequality becomes $\frac{s_1}{s_2} > 1$, which is true.

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

Appendix 2.B Open-Loop Solution

Let $\mu_i(t)$ and $\lambda_i(t)$ denote, respectively, the costate variables associated with the dynamic equations of $w_i(t)$ and $w_j(t)$, given by (2.11) and (2.12), respectively, for Hospital i. That is, $\mu_i(t)$ is associated with Hospital i's waiting time and $\lambda_i(t)$ with that of the rival. The current-value Hamiltonian is

$$H_i = T + pS_i(t) - \frac{\gamma}{2}S_i(t)^2 - \alpha_1 w_i(t) - \frac{\alpha_2}{2}w_i(t)^2 + \mu_i(t)\theta[D_i(w_i(t), w_j(t)) - S_i(t)] + \lambda_i(t)\theta[D_j(w_i(t), w_j(t)) - S_j(t)]. \quad (2.67)$$

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

$$p - \gamma S_i(t) = \theta\mu_i(t), \quad (2.68)$$

$$\dot{\mu}_i(t) = \left[\rho + \frac{\theta(2-\beta)N}{2\tau}\right]\mu_i(t) - \frac{\theta\beta N}{2\tau}\lambda_i(t) + \alpha_1 + \alpha_2 w_i(t), \quad (2.69)$$

and

$$\dot{\lambda}_i(t) = \left[\rho + \frac{\theta(2-\beta)N}{2\tau}\right]\lambda_i(t) - \frac{\theta\beta N}{2\tau}\mu_i(t). \quad (2.70)$$

The solution must also satisfy the transversality conditions

$$\lim_{t\to\infty} e^{-\rho t}\mu_i(t)w_i(t) = 0 \quad \text{and} \quad \lim_{t\to\infty} e^{-\rho t}\lambda_i(t)w_j(t) = 0. \quad (2.71)$$

Optimality is established by concavity of the current-value Hamiltonian with respect to $S_i(t)$ and $w_i(t)$. Inserting the definition of demand (2.7) and the optimality condition for supply (2.68) into the dynamic constraint (2.11) yields

$$\dot{w}_i(t) = \theta N\left[\beta\left(\frac{1}{2} + \frac{w_j(t) - w_i(t)}{2\tau}\right) + (1-\beta)\left(\frac{v - k - w_i(t)}{\tau}\right)\right] - \theta\left(\frac{p - \theta\mu_i(t)}{\gamma}\right). \quad (2.72)$$

The Jacobian matrix of the symmetric system of equations (2.69), (2.70), and (2.72) is:

$$J^{OL} = \begin{bmatrix} -\frac{\theta(1-\beta)N}{\tau} & \frac{\theta^2}{\gamma} & 0 \\ \alpha_2 & \rho + \frac{\theta(2-\beta)N}{2\tau} & -\frac{\theta\beta N}{2\tau} \\ 0 & -\frac{\theta\beta N}{2\tau} & \rho + \frac{\theta(2-\beta)N}{2\tau} \end{bmatrix} \tag{2.73}$$

and has characteristic polynomial

$$P(s) = -s^3 + \left(2\rho + \frac{\theta N}{\tau}\right)s^2 + \left(\frac{\alpha_2\theta^2}{\gamma} + \left(\frac{\theta\beta N}{2\tau}\right)^2 - \left[\rho + \frac{\theta(2-\beta)N}{2\tau}\right]\left[\rho - \frac{\theta(2-3\beta)N}{2\tau}\right]\right)s$$
$$- \frac{\theta(1-\beta)N}{\tau}\left[\rho + \frac{\theta N}{\tau}\right]\left[\rho + \frac{\theta(1-\beta)N}{\tau}\right] - \frac{\alpha_2\theta^2}{\gamma}\left[\rho + \frac{\theta(2-\beta)N}{2\tau}\right]. \tag{2.74}$$

Since $P(s)$ is a third-degree polynomial whose factorisation is unfeasible, solving analytically for its roots yields little insight into the nature of the eigenvalues. According to the fundamental theorem of algebra, $P(s)$ has exactly three roots (real or complex). The coefficients of the cubic term and constant term are negative, while the coefficient of the quadratic term is positive. Although the sign of the coefficient of the linear term is ambiguous, it still follows that $P(-s)$ has a single change of sign—either between the second the the first powers or between the latter and the constant term. Thus, by Descartes' Rule of Signs, $P(s)$ has a single real negative root, which implies that the steady-state is a saddle point.

Let the superscript OL denote the symmetric open-loop steady-state in which $w_i(t) = w_j(t) = w^{OL}$, $\mu_i(t) = \mu_j(t) = \mu^{OL}$, and $S_i(t) = S_j(t) = S^{OL}$. Setting $\dot{w}(t) = \dot{\mu}(t) = \dot{\lambda}(t) = 0$ in equations (2.69), (2.70), and (2.72) and solving for the steady-state waiting time and costate variable gives

$$w^{OL} = \frac{\gamma\phi\tau}{(1-\beta)\gamma\phi N + 2\theta\tau^2\alpha_2}\left\{N\left[\frac{\beta}{2} + (1-\beta)\left(\frac{v-k}{\tau}\right)\right] - \frac{p}{\gamma} - \frac{2\theta\tau\alpha_1}{\gamma\phi}\right\} \tag{2.75}$$

and

$$\mu^{OL} = -\frac{2\tau}{\phi}(\alpha_1 + \alpha_2 w^{OL}), \tag{2.76}$$

where ϕ is defined by (2.14) in subsection 2.4.1. The corresponding steady-state supply is

$$S^{OL} = \frac{p}{\gamma} + \frac{2\theta\tau\left(\alpha_1 + \alpha_2 w^{OL}\right)}{\gamma\phi}. \tag{2.77}$$

The open-loop steady-state is characterised by a positive waiting time and a partially covered monopolistic segment if $p \in \mathcal{P} = (\max\{0, \underline{p}\}, \min\{\overline{p}_1, \overline{p}_2\})$, where

$$\underline{p} = \frac{\beta}{2}\gamma N - \frac{2\theta\tau}{\phi}\left[\alpha_1 + \alpha_2(v-k)\right], \tag{2.78}$$

$$\overline{p}_1 = \gamma N \left[\frac{\beta}{2} + (1-\beta)\left(\frac{v-k}{\tau}\right)\right] - \frac{2\theta\tau\alpha_1}{\phi}, \tag{2.79}$$

and

$$\overline{p}_2 = \frac{\gamma N}{2} - \frac{2\theta\tau}{\phi}\left[\alpha_1 + \alpha_2\left(v - k - \frac{\tau}{2}\right)\right]. \tag{2.80}$$

From (2.75), the waiting time is positive if and only if $p \leq \overline{p}_1$. Then, in order to have a partially covered monopolistic segment in the steady-state, the following condition must be satisfied:

$$0 < \frac{v - k - w^{OL}}{\tau} < \frac{1}{2}. \tag{2.81}$$

The lower bound is satisfied if $p > \underline{p}$, as defined by (2.78). Note that $\underline{p}$ may be negative, but $p \in \mathbb{R}^+$ must hold. Thus, $p > \max\{0, \underline{p}\}$. The upper bound, in turn, is satisfied if $p < \overline{p}_2$. Since $\overline{p}_1 > 0 \wedge \overline{p}_1 > \underline{p}$, $\mathcal{P}$ is non-empty when $\overline{p}_1 < \overline{p}_2$. Conversely, $\overline{p}_2$ only verifies $\overline{p}_2 > \underline{p}$, as it may be negative. If $\overline{p}_2 < 0$, parameters are such that $\underline{p} < \overline{p}_2 < 0 < \overline{p}_1$. Then, $\mathcal{P}$ is non-empty when $\overline{p}_2 < \overline{p}_1$ if and only if $\overline{p}_2 > 0$, which holds for a sufficiently large γ.

From (2.75), we derive

$$\frac{\partial w^{OL}}{\partial \tau} = -\frac{(1-\beta)x_M^{OL} + \frac{\tau}{N}\frac{\partial S^{OL}}{\partial \tau}}{1 - \beta + \frac{2\theta\tau^2\alpha_2}{\gamma\phi N}} < 0, \tag{2.82}$$

where

$$\frac{\partial S^{OL}}{\partial \tau} = N\theta^2(\alpha_1 + \alpha_2 w^{OL})\frac{(1-\beta)[N\theta(2-\beta) + 4\tau\rho]\theta N + (2-\beta)(\tau\rho)^2}{2\gamma(N\theta + \tau\rho)^2[N(1-\beta)\theta + \tau\rho]^2} > 0 \tag{2.83}$$

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

2.B.1 Non-linear patient disutility of waiting

Suppose that hospital demand is given by (2.30) in subsection 2.6.1. Defining the Hamiltonian as before, the optimality conditions in the symmetric steady-state are now given by

$$p - \gamma S^{OL} = \theta\mu^{OL}, \tag{2.84}$$

$$\left[\rho - \theta\frac{\partial D_i(w^{OL})}{\partial w_i}\right]\mu^{OL} - \theta\frac{\partial D_j(w^{OL})}{\partial w_i}\lambda^{OL} + \alpha_1 + \alpha_2 w^{OL} = 0, \tag{2.85}$$

and

$$\left[\rho - \theta\frac{\partial D_j(w^{OL})}{\partial w_j}\right]\lambda^{OL} - \theta\frac{\partial D_i(w^{OL})}{\partial w_j}\mu^{OL} = 0. \tag{2.86}$$

Using (2.30) and (2.85)–(2.86) to solve for μ^{OL} and λ^{OL}, we obtain

$$\mu^{OL} = -\frac{\tau}{2}\frac{2\rho\tau + \theta(2-\beta)N\frac{\partial f(w^{OL})}{\partial w}}{\left[\rho\tau + \theta(1-\beta)N\frac{\partial f(w^{OL})}{\partial w}\right]\left[\rho\tau + \theta N\frac{\partial f(w^{OL})}{\partial w}\right]}(\alpha_1 + \alpha_2 w^{OL}) < 0 \tag{2.87}$$

and

$$\lambda^{OL} = \left[\frac{\theta\beta N\frac{\partial f(w^{OL})}{\partial w}}{2\rho\tau + \theta(2-\beta)N\frac{\partial f(w^{OL})}{\partial w}}\right]\mu^{OL} < 0. \tag{2.88}$$

Using the dynamic constraint, (2.30), and (2.84), the steady-state waiting time is then implicitly defined by

$$N\left[\frac{\beta}{2} + (1-\beta)\left(\frac{v-k-f(w^{OL})}{\tau}\right)\right] - \frac{p-\theta\mu^{OL}}{\gamma} = 0. \tag{2.89}$$

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

$$\frac{\partial^2 H_i}{\partial w_i^2} = -\alpha_2 - \left[\frac{\theta(2-\beta)N}{2\tau}\mu_i - \frac{\theta\beta N}{2\tau}\lambda_i\right]\frac{\partial^2 f}{\partial w_i^2} \leq 0. \tag{2.90}$$

Evaluated at the steady-state, this expression becomes

$$-\alpha_2 + \frac{\left[\rho\tau(2-\beta) + 2\theta(1-\beta)N\frac{\partial f(w^{OL})}{\partial w}\right]\theta N(\alpha_1 + \alpha_2 w^{OL})}{2\left[\rho\tau + \theta(1-\beta)N\frac{\partial f(w^{OL})}{\partial w}\right]\left[\rho\tau + \theta N\frac{\partial f(w^{OL})}{\partial w}\right]}\frac{\partial^2 f(w^{OL})}{\partial w^2} \leq 0 \tag{2.91}$$

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

$$\frac{\partial^2 f(w^{OL})}{\partial w^2} \leq \frac{2\left[\rho\tau + \theta(1-\beta)N\frac{\partial f(w^{OL})}{\partial w}\right]\left[\rho\tau + \theta N\frac{\partial f(w^{OL})}{\partial w}\right]}{\rho\tau(2-\beta) + 2\theta(1-\beta)N\frac{\partial f(w^{OL})}{\partial w}}\frac{\alpha_2}{\theta N(\alpha_1 + \alpha_2 w^{OL})}. \tag{2.92}$$

Implicitly differentiating (2.89) with respect to w^{OL} and τ yields

$$\frac{\partial w^{OL}}{\partial \tau} = -\frac{(1-\beta)x_M^{OL} - \frac{\tau\theta}{N\gamma}\frac{\partial \mu^{OL}}{\partial \tau}}{(1-\beta)\frac{\partial f(w^{OL})}{\partial w} - \frac{\tau\theta}{N\gamma}\frac{\partial \mu^{OL}}{\partial w^{OL}}} < 0, \tag{2.93}$$

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

$$\frac{\partial \mu^{OL}}{\partial \tau} = -\frac{\partial f(w^{OL})}{\partial w} \frac{\theta N \Gamma(w^{OL})(\alpha_1 + \alpha_2 w^{OL})}{2\left[\rho\tau + \theta(1-\beta)N\frac{\partial f(w^{OL})}{\partial w}\right]^2 \left[\rho\tau + \theta N \frac{\partial f(w^{OL})}{\partial w}\right]^2} < 0, \tag{2.94}$$

$$\frac{\partial \mu^{OL}}{\partial w^{OL}} = -\frac{\tau}{2} \frac{\left[2\rho\tau + \theta(2-\beta)N\frac{\partial f(w^{OL})}{\partial w}\right]\alpha_2}{\left[\rho\tau + \theta(1-\beta)N\frac{\partial f(w^{OL})}{\partial w}\right]\left[\rho\tau + \theta N \frac{\partial f(w^{OL})}{\partial w}\right]} + \frac{\partial^2 f(w^{OL})}{\partial w^2} \frac{\tau\theta N \Gamma(w^{OL})(\alpha_1 + \alpha_2 w^{OL})}{2\left[\rho\tau + \theta(1-\beta)N\frac{\partial f(w^{OL})}{\partial w}\right]^2 \left[\rho\tau + \theta N \frac{\partial f(w^{OL})}{\partial w}\right]^2} \leq 0, \tag{2.95}$$

and

$$\Gamma(w^{OL}) = (\rho\tau)^2(2-\beta) + 4\rho\tau\theta(1-\beta)N\frac{\partial f(w^{OL})}{\partial w} + \theta^2(1-\beta)(2-\beta)N^2\left(\frac{\partial f(w^{OL})}{\partial w}\right)^2 > 0. \tag{2.96}$$

To show that (2.95) is always non-positive in the steady-state equilibrium, notice that the right-hand side of (2.95) is increasing in $\partial^2 f\left(w^{OL}\right)/\partial w^2$, while the second-order conditions dictate that $\partial^2 f\left(w^{OL}\right)/\partial w^2$ must be sufficiently low (cf. (2.92)). Replacing $\partial^2 f\left(w^{OL}\right)/\partial w^2$ in equation (2.95) with the right-hand side of (2.92), which is the maximum value of $\partial^2 f\left(w^{OL}\right)/\partial w^2$ that still ensures equilibrium existence, yields

$$\frac{\partial \mu^{OL}}{\partial w^{OL}} = -\frac{\rho\theta(\tau\beta)^2 N\frac{\partial f(w^{OL})}{\partial w}\alpha_2}{2\left[\rho\tau + \theta(1-\beta)N\frac{\partial f(w^{OL})}{\partial w}\right]\left[\rho\tau + \theta N\frac{\partial f(w^{OL})}{\partial w}\right]\left[\rho\tau(2-\beta) + 2\theta(1-\beta)N\frac{\partial f(w^{OL})}{\partial w}\right]} \leq 0. \tag{2.97}$$

This implies that $\partial\mu^{OL}/\partial w^{OL} \leq 0$, and thus $\partial w^{OL}/\partial\tau < 0$, for every specification of $f(w)$ that is compatible with equilibrium existence under open-loop rules.

2.B.2 Non-linear patient disutility of travelling

Suppose the patient utility function is redefined as indicated in subsection 2.6.2. The optimality conditions, evaluated at the symmetric steady-state, are the given by (2.84) and

$$\left[\rho + \frac{\theta\beta N}{2\tau g'\left(\frac{1}{2}\right)} + \frac{\theta(1-\beta)N}{\tau g'\left(x_M^{OL}\right)}\right]\mu^{OL} - \frac{\theta\beta N}{2\tau g'\left(\frac{1}{2}\right)}\lambda^{OL} + \alpha_1 + \alpha_2 w^{OL} = 0, \tag{2.98}$$

and

$$\left[\rho + \frac{\theta\beta N}{2\tau g'\left(\frac{1}{2}\right) + \frac{\theta(1-\beta)N}{\tau g'\left(x_M^{OL}\right)}}\right]\lambda^{OL} - \frac{\theta\beta N}{2\tau g'\left(\frac{1}{2}\right)}\mu^{OL}. \tag{2.99}$$

Using (2.98) and (2.99) to solve for μ^{OL} and λ^{OL}, we obtain:

$$\mu^{OL} = -\frac{\tau}{2}\frac{2g'\left(\frac{1}{2}\right)\rho\tau + \theta\beta N + \frac{2g'\left(\frac{1}{2}\right)\theta(1-\beta)N}{g'\left(x_M^{OL}\right)}}{\left[\rho\tau + \frac{\theta(1-\beta)N}{g'\left(x_M^{OL}\right)}\right]\left[g'\left(\frac{1}{2}\right)\rho\tau + \theta\beta N + \frac{g'\left(\frac{1}{2}\right)\theta(1-\beta)N}{g'\left(x_M^{OL}\right)}\right]}(\alpha_1 + \alpha_2 w^{OL}) < 0. \quad (2.100)$$

and

$$\lambda^{OL} = \left[\frac{\theta\beta N}{2g'\left(\frac{1}{2}\right)\rho\tau + \theta\beta N + \frac{2g'\left(\frac{1}{2}\right)\theta(1-\beta)N}{g'\left(x_M^{OL}\right)}}\right]\mu^{OL} < 0. \quad (2.101)$$

Using the dynamic constraint and (2.84), the steady-state waiting time is implicitly defined by

$$N\left[\frac{\beta}{2} + (1-\beta)x_M^{OL}\right] - \frac{p - \theta\mu^{OL}}{\gamma} = 0. \quad (2.102)$$

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

$$\frac{\partial^2 H_i}{\partial w_i^2} = -\alpha_2 + \left[\frac{\theta(1-\beta)N}{\tau}\frac{g''(x_M^i)}{[g''(x_M^i)]^2}\frac{\partial x_M^i}{\partial w_i}\right]\mu_i \leq 0. \quad (2.103)$$

Evaluated at the steady-state, this expression becomes

$$-\alpha_2 - \left[\frac{\theta(1-\beta)N}{\tau^2}\frac{g''\left(x_M^{OL}\right)}{[g'\left(x_M^{OL}\right)]^3}\right]\mu^{OL} \leq 0. \quad (2.104)$$

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

Implicitly differentiating (2.102) with respect to w^{OL} and τ yields

$$\frac{\partial w^{OL}}{\partial \tau} = -\frac{N(1-\beta)\frac{\partial x_M^{OL}}{\partial \tau} + \frac{\theta}{\gamma}\frac{\partial \mu^{OL}}{\partial \tau}}{N(1-\beta)\frac{\partial x_M^{OL}}{\partial w^{OL}} + \frac{\theta}{\gamma}\frac{\partial \mu^{OL}}{\partial w^{OL}}} \quad (2.105)$$

where

$$\frac{\partial x_M^{OL}}{\partial \tau} = -\frac{g(x_M^{OL})}{\tau g'\left(x_M^{OL}\right)} < 0, \quad (2.106)$$

$$\frac{\partial x_M^{OL}}{\partial w^{OL}} = -\frac{1}{\tau g'\left(x_M^{OL}\right)} < 0, \quad (2.107)$$

$$\frac{\partial \mu^{OL}}{\partial \tau} = -\frac{\left[\Delta_1 - \Delta_2\frac{g'\left(x_M^{OL}\right)g''\left(x_M^{OL}\right)}{[g'\left(x_M^{OL}\right)]^2}\right](\alpha_1 + \alpha_2 w^{OL})}{\left[\rho\tau + \frac{\theta(1-\beta)N}{g'\left(x_M^{OL}\right)}\right]^2\left[g'\left(\frac{1}{2}\right)\rho\tau + \theta\beta N + \frac{g'\left(\frac{1}{2}\right)\theta(1-\beta)N}{g'\left(x_M^{OL}\right)}\right]^2}, \quad (2.108)$$

$$\Delta_1 = \left[4g'\left(\frac{1}{2}\right)\rho\tau + \theta\beta N + \frac{2g'\left(\frac{1}{2}\right)\theta(1-\beta)N}{g'\left(x_M^{OL}\right)}\right]\left[g'\left(\frac{1}{2}\right)\left[\frac{\theta(1-\beta)N}{g'\left(x_M^{OL}\right)}\right]^2 + \frac{(\theta N)^2\beta(1-\beta)}{g'\left(x_M^{OL}\right)}\right] + \frac{2\left[g'\left(\frac{1}{2}\right)\rho\tau\right]^2\theta(1-\beta)N}{g'\left(x_M^{OL}\right)} + g'\left(\frac{1}{2}\right)(\rho\tau)^2\theta N\beta > 0, \quad (2.109)$$

$$\Delta_2 = g'\left(\frac{1}{2}\right)\left[4g'\left(\frac{1}{2}\right)\rho\tau + 2\theta\beta N + \frac{2g'\left(\frac{1}{2}\right)\theta(1-\beta)N}{g'\left(x_M^{OL}\right)}\right]\left[\frac{\theta(1-\beta)N}{g'\left(x_M^{OL}\right)}\right]^2 + \frac{2\left[g'\left(\frac{1}{2}\right)\rho\tau\right]^2\theta(1-\beta)N}{g'\left(x_M^{OL}\right)} + \left[2g'\left(\frac{1}{2}\right)\rho\tau + \theta\beta N\right]\frac{(\theta N)^2\beta(1-\beta)}{g'\left(x_M^{OL}\right)} > 0, \quad (2.110)$$

$$\frac{\partial\mu^{OL}}{\partial w^{OL}} = -\frac{\alpha_2\tau}{2}\frac{2g'\left(\frac{1}{2}\right)\rho\tau + \theta\beta N + \frac{2g'\left(\frac{1}{2}\right)\theta(1-\beta)N}{g'\left(x_M^{OL}\right)}}{\left[\rho\tau + \frac{\theta(1-\beta)N}{g'\left(x_M^{OL}\right)}\right]\left[g'\left(\frac{1}{2}\right)\rho\tau + \theta\beta N + \frac{g'\left(\frac{1}{2}\right)\theta(1-\beta)N}{g'\left(x_M^{OL}\right)}\right]} + \Delta_3\left[\frac{\theta(1-\beta)N(\alpha_1+\alpha_2 w^{OL})}{2[g'\left(x_M^{OL}\right)]^3}\right]g''\left(x_M^{OL}\right), \quad (2.111)$$

and

$$\Delta_3 = \frac{\left[\begin{array}{c}\left[2g'\left(\frac{1}{2}\right)\rho\tau + \theta\beta N + \frac{2g'\left(\frac{1}{2}\right)\theta(1-\beta)N}{g'\left(x_M^{OL}\right)}\right]^2 \\ -\left[2g'\left(\frac{1}{2}\right)\rho\tau + \frac{2g'\left(\frac{1}{2}\right)\theta(1-\beta)N}{g'\left(x_M^{OL}\right)}\right]\left[g'\left(\frac{1}{2}\right)\rho\tau + \theta\beta N + \frac{g'\left(\frac{1}{2}\right)\theta(1-\beta)N}{g'\left(x_M^{OL}\right)}\right]\end{array}\right]}{\left[\rho\tau + \frac{\theta(1-\beta)N}{g'\left(x_M^{OL}\right)}\right]^2\left[g'\left(\frac{1}{2}\right)\rho\tau + \theta\beta N + \frac{g'\left(\frac{1}{2}\right)\theta(1-\beta)N}{g'\left(x_M^{OL}\right)}\right]^2} > 0. \quad (2.112)$$

If $g''\left(x_M^{OL}\right) \leq 0$, the expressions on the right-hand side of (2.108) and (2.111) are unambiguously negative, which implies that $\partial w^{OL}/\partial\tau < 0$ for every concave function g. If instead $g''\left(x_M^{OL}\right) > 0$, a negative sign of $\partial\mu^{OL}/\partial\tau$ and $\partial\mu^{OL}/\partial w^{OL}$, which implies $\partial w^{OL}/\partial\tau < 0$, requires that $g''\left(x_M^{OL}\right)$ is sufficiently low.

3. Hospital Competition With Switching Costs: Quality, Patient Welfare, and Market Segmentation

3.1 Introduction

Free choice of hospital is becoming widespread. While, in the United States, it has been a structural feature of the health care 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 health care), competition in hospital markets is expected to play out through channels other than prices. Ranging from the 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 can 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 incen-

[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 a 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.

tives (or conditions) do hospitals have to carry out costly quality investments? To the extent that quality affects patient utility and that inertia reflects some degree of patients' inability to adjust to changes in the environment, another question ensues: how does reduced inertia affect patient welfare? The first part of this chapter addresses these questions from a policy perspective: it investigates whether policies that reduce exogenous switching costs, a driver of inertia, play a useful role in improving quality provision and patient welfare. An additional, related question arises when switching costs are endogenously determined: how are the incentives for investing in quality affected by the hospitals' ability to set switching costs? The second part of this chapter tackles this question by modelling the hospitals' strategic and joint decisions of quality and switching costs. Therefore, this chapter's main contribution is to explore the broader role of switching costs (and inertia more generally) in the context of hospital competition.

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 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 to transfer their medical records between providers. Second, some procedures are hospital-specific investments in that patients undergo medical procedures that are intertemporally linked and might be rendered useless if the patient switches providers and treatment

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

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]

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.

The adoption of shareable Electronic Health Records (EHR), electronic records of an individual patient's history of contact with the health care system (Oderkirk, 2017), has the potential to reduce switching costs. If these result mainly from 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 health care professionals whom they have not contacted before if these professionals can easily learn their medical history.

Throughout the chapter, a reduction in switching costs may be interpreted as the result of a policy based on the market-wide adoption of shareable EHR.[4] This is not, however, 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 chapter offers an alternative measure. By specifically reflecting a

[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 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. The potential of EHR systems to reduce switching costs, however, is often overlooked, and the emphasis is placed on their benefits to medical research and cost savings.

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, quality information made increasingly available in the public domain, a staple of choice policies, might reduce the uncertainty about quality at alternative hospitals.

To model the demand for health care faced by each provider, we 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 our 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.

We obtain several findings regarding the unintended effects of lower (exogenous) switching costs on patient welfare, which we define below as average quality and aggregate patient utility. 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 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 one, contributing to higher average quality and aggregate patient utility. Conversely, if cost substitutability is sufficiently weak (or cost complementarity sufficiently strong) relative altruism, 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, hurting the majority of patients in the market; consequently, average quality and total patient utility may also decrease. Importantly, the relationship between hospital-level quality, average quality, and total patient utility is not straightforward. Even if lower switching costs lead to higher quality at *both* hospitals *and* higher aggregate utility, average quality may nonetheless fall owing to the redistribution of patients between hospitals.

When hospitals can set the switching costs their inherited patients incur, local monopolies arise and the market is perfectly segmented according to the history of patient-hospital relationships. The intuition behind this result is simple. Because the marginal inherited patient is always beneficial to treat, setting maximum switching costs and hence retaining all inherited patients is a strictly dominant strategy for each hospital. This, in turn, implies that the incentives for quality provision change a great deal. Quality ceases to be an instrument to attract demand, and it is only offered above the minimum required threshold for altruistic reasons, provided that the hospitals are sufficiently altruistic. If cost substitutability is strong relative to altruism, then imposing switching costs and quality provision are substitutable strategies for each hospital. Setting maximum switching costs allows hospitals to retain all of their previous patients while offering minimum quality and hence avoiding quality provision costs completely.

The rest of the chapter is organised as follows. The next section offers an overview of the literature on patient inertia and explains how this chapter relates to it. In section 3.3, we describe how inertia shapes demand and discuss the assumptions underlying hospital preferences and production. In section 3.4, the model is solved for the Nash Equilibrium, and equilibrium quality and demand are characterised. Section 3.5 investigates the effect of switching costs on patient welfare. In section 3.6, we consider endogenous switching costs. Finally, section 3.7 offers concluding remarks and discusses policy implications.

3.2 Related Literature

This chapter 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 our knowledge, Jung et al. (2011) were the first to look at patient-level inertia. They model a hypothetical choice 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 health care plan choice. 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, 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 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 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. This 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 owing 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]

Importantly, Irace (2018) also looks at welfare. In a counterfactual scenario with no switching costs, he estimates an expected mortality 3% below the observed mortality rate. This reduction in mortality results only from 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 chapter considers these feedback effects and reveals that it is precisely the patients' increased ability to switch (and hence reduce mismatch costs) that, under some conditions, undermines the positive effect of lower switching costs on welfare through higher quality provision (cf. section 3.5.1).

The second strand of the literature is that on theoretical models of hospital competition under regulated

[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 we present below may indeed be interpreted as dealing with first-order state dependence.

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. 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, Straume, and Cellini (2013) investigate an information-related form of patient inertia. In both studies, demand adjusts sluggishly to changes in quality. Because health care quality is neither easily nor immediately observable, only a fraction of patients become aware of quality changes, and, consequently, only a fraction of any potential change in demand is realised. With pure profit-maximising providers, Brekke et al. (2012) show that a reduced degree of sluggishness increases quality. The intuition for this result is simple. Less sluggish beliefs make demand more responsive to quality changes, and, with a positive payment-cost margin, this gives providers incentives to increase quality. Siciliani, Straume, and Cellini (2013) show that this result may be overturned if providers are semi-altruistic. Like in Brekke et al. (2011), the effect of reduced sluggishness 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 the model with exogenous switching costs we present below, 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 model with exogenous switching costs. They find that quality is independent of switching costs, 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. Our 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.

3.3 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 speciality 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 owing to public provision of health care 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. We assume throughout that the gross valuation of treatment is $v > \tau$, so that the market is always fully covered.

The history of patient relationships with the two neighbouring hospitals is as follows: σ_H patients visited Hospital H in the preceding episode of care, while the remaining $\sigma_L = 1 - \sigma_H$ patients 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. If they decide to demand treatment from the hospital they have not used in the preceding episode of care, these patients incur an exogenous

switching cost s.[6] The remaining $1-\mu$ patients have unchanged preferences for treatment characteristics and choose the same hospital as before. 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 $[\sigma_H, 1]$. The model thus 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, we refer to σ_i as Hospital i's inherited demand. It should be emphasised that we consider a one-period model and that the issue of how the inherited hospital-patient relationships are formed is not formally addressed.[7] One possible interpretation for asymmetric inherited demand is the case of a former local monopolist whose incumbency has not been eroded. Indeed, we show in section 3.4 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.

3.3.1 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

$$v + q_H - \tau x = v + q_L - \tau(1 - x) - s \tag{3.1}$$

or, explicitly,

$$\hat{x}_{|H} = \frac{1}{2} + \frac{q_H - q_L + s}{2\tau}. \tag{3.2}$$

[6] More realistically, one may conjecture that patients have different switching costs. The main feature upon which 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 we adopt preserves that feature and additionally allows for a richer specification of horizontal patient preferences, while still keeping the analysis tractable.

[7] Following Klemperer (1995), we 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 chapter. 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

$$v + q_H - \tau x - s = v + q_L - \tau(1 - x) \tag{3.3}$$

or, explicitly,

$$\hat{x}_{|L} = \frac{1}{2} + \frac{q_H - q_L - s}{2\tau}. \tag{3.4}$$

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 (3.2) and (3.4). 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 to get half of the patients with changing preferences who are tied to Hospital L. Finally, fractions $(1 - \mu)\sigma_H$ and $(1 - \mu)\sigma_L$ of patients have unchanged preferences and again choose hospitals H and L. Combining demand from the two types of patients, it may be easily shown that total demand facing Hospital i is given by

$$D_i(q_i, q_j) = \frac{\mu}{2\tau}[\tau + q_i - q_j + (\sigma_i - \sigma_j)s] + (1 - \mu)\sigma_i, \quad i, j = H, L; \quad i \neq j; \tag{3.5}$$

provided that $|q_H - q_L| < \tau - s$.[8] Notice how inertia shapes demand: both hospitals have some captive patients owing to preference persistency, and the hospital with higher inherited demand has a demand bonus simply because switching hospitals is costly for patients.

3.3.2 Hospital objectives

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

$$\Omega_i(q_i, q_j) = T + \tilde{p}D_i(q_i, q_j) - C[q_i, D_i(q_i, q_j)] + \alpha B[q_i, D_i(q_i, q_j)], \quad i, j = H, L; \quad i \neq j; \tag{3.6}$$

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

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

$$C[q_i, D_i(q_i, q_j)] = (cq_i + k)D_i(q_i, q_j) + \frac{\gamma}{2}q_i^2, \quad i, j = H, L; \quad i \neq j; \tag{3.7}$$

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. These positive returns to hospital volume are generally attributed to learning-by-doing or quality-enhancing scale economies, which capture the idea that health care 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 the result of 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 (3.7) 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, the departure from pure profit-

maximisation may arise from 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.[9] The aggregate benefit to patients treated at hospitals H and L is respectively given by

$$B_H[q_H, D_H(q_H, q_L)] = \mu\sigma_H \int_0^{\hat{x}_{|H}} (v + q_H - \tau x)dx + \mu\sigma_L \int_0^{\hat{x}_{|L}} (v + q_H - \tau x - s)dx + (1-\mu) \int_0^{\sigma_H} (v + q_H - \tau x)dx \quad (3.8)$$

and

$$B_L[q_L, D_L(q_H, q_L)] = \mu\sigma_H \int_{\hat{x}_{|H}}^1 [v + q_L - \tau(1-x) - s]dx + \mu\sigma_L \int_{\hat{x}_{|L}}^1 [v + q_L - \tau(1-x)]dx + (1-\mu) \int_{\sigma_H}^1 [v + q_L - \tau(1-x)]dx. \quad (3.9)$$

It is instructive to see how semi-altruistic preferences affect incentives to provide quality. Differentiating (3.8) and (3.9) 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

$$\frac{\partial B_i[q_i, D_i(q_i, q_j)]}{\partial q_i} = \frac{\mu}{4\tau}(2v + q_i + q_j - \tau - s) + D_i > 0, \quad i, j = H, L; \quad i \neq j. \quad (3.10)$$

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 (3.10).

Finally, we make the following restrictions on parameter values:

$$c > c_{min} \equiv \max\left\{\left(\alpha - \frac{2\tau\gamma}{\mu}\right), \left(\frac{3\alpha}{4} - \frac{\tau\gamma}{\mu}\right), \left(\frac{2\alpha}{3} - \frac{2\tau\gamma}{3\mu}\right)\right\} \quad (3.11)$$

and

$$\sigma_H - \sigma_L < \min\left\{1, \frac{\tau - s}{|\alpha - c|\phi}\right\}, \quad (3.12)$$

where

$$\phi = \frac{2[\tau - (\tau - s)\mu]}{\mu\left[\frac{2\tau\gamma}{\mu} - (2\alpha - 3c)\right]} > 0. \quad (3.13)$$

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

Condition (3.11) 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 (3.12) ensures that the demand function (3.5) 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).

3.4 Inherited Demand, Quality, and Market Dominance

Using (3.7) and (3.10), maximisation of Ω_i with respect to q_i yields the first-order condition

$$\frac{\mu}{2\tau}\left[p - cq_i + \alpha\left(\frac{2v + q_i + q_j - \tau - s}{2}\right)\right] + (\alpha - c)D_i - \gamma q_i = 0, \quad i, j = H, L; \quad i \neq j; \tag{3.14}$$

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 (3.5) into the pair of equations given by (3.14) and solving for q_i yields the candidate equilibrium quality levels[10]

$$q_i^* = \max\left\{0, \frac{p + (\alpha - c)\left[\tau - s - (\alpha - 2c)\frac{\phi}{2}\right] + \alpha\left(v - \frac{\tau + s}{2}\right)}{\frac{2\tau\gamma}{\mu} - (\alpha - c)} + (\alpha - c)\phi\sigma_i\right\}, \quad i, j = H, L. \tag{3.15}$$

Suppose first that $c > 0$. It follows immediately from (3.15) that Hospital H offers lower quality than Hospital L if cost substitutability is stronger than the hospitals' altruism. Both the marginal cost and the marginal altruistic benefit from quality depend positively on current demand, which, in turn, depends positively on inherited demand. 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

[10] If the cost function is sufficiently convex in quality, Ω_i is concave in the region in which (3.5) holds and the second-order conditions are satisfied. Concavity of Ω_i requires that $\gamma - \frac{\mu}{2\tau}\left(\frac{3}{2}\alpha - 2c\right) > 0$, which is always true given (3.11). This, however, does not suffice to show that (3.15) defines a Nash Equilibrium. Hospitals may unilaterally deviate from those strategies by choosing a quality level outside the range in which (3.5) holds. 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 (3.15) defines a Nash equilibrium. Klemperer (1987), Beggs and Klemperer (1992), and To (1996) provide the analogous argument in the case of multi-period price competition. In the remainder of the analysis, we focus on strictly positive quality levels, which requires that $\tilde{p}$ is high enough.

higher quality increases the utility of more patients). Which of these effects dominates depends on the size of α and c. If $c < 0$, the result is clear-cut. Under cost complementarity, Hospital H has a higher marginal altruistic benefit and a lower marginal cost of quality. It will thus offer higher quality.

Having derived the optimum quality levels, one may now look at demand. Proposition 3.1 below describes how inertia shapes demand and how it counteracts or strengthens the effect of quality as a demand shifter.

Proposition 3.1. *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 $\underline{c} = \frac{2\tau[\alpha-(\tau-s)\gamma]}{2\tau+(\tau-s)\mu} < \alpha$.

Proof. Follows directly from (3.15) and the comparison between σ_H and D_H evaluated at the equilibrium quality levels. □

It is instructive to characterise the mechanism underlying the results of Proposition 3.1 in detail, as it sheds light on the interplay between quality provision, horizontal preferences, and switching costs, thus facilitating the interpretation of subsequently derived results. Recall that some patients (with changing preferences) face a strong mismatch between their preferences and the horizontal attributes of the hospital they are tied to, and they 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 caused by the mismatch between patient preferences and the hospital's attributes. Owing 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 higher demand and thus implies that the hospital with higher inherited demand will retain its position as market leader regardless of the values of c and α.

3.5 The Effect of Switching Costs on Patient Welfare

To investigate the effect of facilitated switching on patient welfare, defined below as average quality and aggregate patient utility, one needs first to characterise the effect of lower switching costs on hospital-level quality. The effect of a marginal change in switching costs on equilibrium quality is given by

$$\frac{\partial q_i^*}{\partial s} = -\frac{\alpha/2}{\frac{2\tau\gamma}{\mu} - (\alpha - c)} + \frac{(\alpha - c)(\sigma_i - \sigma_j)}{\frac{2\tau\gamma}{\mu} - (2\alpha - 3c)}, \quad i, j = H, L; \quad i \neq j. \tag{3.16}$$

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 is weaker since the utility of these patients is reduced by an amount s. From (3.10), the lower the switching cost, the stronger is the altruistic incentive the two hospitals have to increase quality. We shall henceforth refer to this as the *patient utility effect*.[11] There is also a *demand effect*, which can easily be seen from (3.5). 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 3.2. *Provided that the cost function is sufficiently convex in quality, there exist two threshold values of* c, $c_{HH} \in (c_{min}, \alpha)$ *and* $c_{HL} > \alpha$, *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* $(c_{HL} - c_{HH})$ *are increasing in* μ *and decreasing in* τ.

See Appendix 3.A for a proof and the definitions of c_{HH} and c_{HL}.

Consider 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

[11] Alternatively, this change in the marginal benefit from attracting patients who were previously treated at the rival hospital may be interpreted as a change in 'patient acquisition' costs.

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. In the presence of volume-outcome effects, a particular case of $c < \alpha$, this last-mentioned result naturally caries over. The underlying mechanism, however, differs slightly, as 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 $c < \alpha$. Conversely, 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.

Proposition 3.2 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. In addition, 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. When fewer patients have persistent preferences (higher μ) or travelling/mismatch costs are lower (lower τ), demand is more responsive to quality. Also, Hospital H's switching cost-induced demand advantage from patients with changing preferences is greater. Increased demand responsiveness implies that the altruistic incentive to attract patients (when s falls) 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. 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.

Finally, it is interesting to see how the results in Proposition 3.2 change with the hospitals' degree of profit-orientation. If hospitals are pure profit-maximisers (i.e., $\alpha = 0$), the patient utility effect vanishes, and the sign of the demand effect is uniquely determined by c and inherited demand. Thus, if hospitals are pure profit-maximisers, it is certain that some patients will enjoy lower quality after a switching costs reduction. Notice also 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.

We are now in a position to characterise the effect of lower switching costs on patient welfare.

3.5.1 Average quality

Hospital-level quality being 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 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

$$\frac{\partial \overline{q}}{\partial s} = (q_H^* - q_L^*)\frac{\partial D_H^*}{\partial s} + \frac{\partial q_H^*}{\partial s} D_H^* + \frac{\partial q_L^*}{\partial s} D_L^* = \\ (\alpha - c)\phi(\sigma_H - \sigma_L)\frac{\partial D_H^*}{\partial s} - \frac{\alpha/2}{\frac{2\tau\gamma}{\mu} - (\alpha - c)} + \frac{(\alpha - c)(\sigma_H - \sigma_L)(D_H^* - D_L^*)}{\frac{2\tau\gamma}{\mu} - (2\alpha - 3c)}, \quad (3.17)$$

where

$$\frac{\partial D_H^*}{\partial s} = \frac{\mu}{2\tau}\left[\frac{2(\alpha - c)}{\frac{2\tau\gamma}{\mu} - (2\alpha - 3c)} + 1\right](\sigma_H - \sigma_L) > 0. \quad (3.18)$$

Lower switching costs have a twofold effect on average quality. First, there is a patient redistribution effect as a reduction in switching costs always decreases market concentration by shifting demand from the high- to the low-volume hospital.[12] Importantly, 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

[12] $\frac{\partial D_H^*}{\partial s} > 0$ if $c > -\frac{2\tau\gamma}{\mu}$, which always holds given (3.11).

are weighted by each hospital's demand. The effect of lower switching costs on average quality may be stated as follows.

Proposition 3.3. *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*

$$\frac{\alpha/2}{\frac{2\tau\gamma}{\mu} - (\alpha - c_{\overline{q}})} = \frac{2[\tau - (\tau - s)\mu](\sigma_H - \sigma_L)^2(\alpha - c_{\overline{q}})\left(\frac{2\tau\gamma}{\mu} + c_{\overline{q}}\right)}{\tau\left[\frac{2\tau\gamma}{\mu} - (2\alpha - 3c_{\overline{q}})\right]^2}, \tag{3.19}$$

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

$$s > \frac{\tau}{\mu}\left[\frac{\frac{2\tau\gamma}{\mu} - (2\alpha - 3c_{HH})}{2(\sigma_H - \sigma_L)\left(\frac{2\tau\gamma}{\mu} + c_{HH}\right)} - (1 - \mu)\right], \tag{3.20}$$

with c_{HH} as given by the greater root in (3.31) in Appendix 3.A.

Proof. See Appendix 3.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 (3.16) 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. A lower s, however, 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.

Crucially, 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 3.3.

Corollary 3.1. *Provided that switching costs are initially sufficiently high, lower switching costs reduce average quality while increasing quality at both hospitals if* $c_{HH} < c < c_{\bar{q}}$.

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 some patients switching from Hospital H to Hospital L suffices to reduce average quality.

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 (higher) average quality in the presence of cost complementarity (substitutability).

3.5.2 Aggregate patient utility

Up until this point, the analysis of the effect of lower switching costs has focused mainly on quality provision; however, switching costs affect patient welfare through other channels. Consider now a more comprehensive measure of patient welfare, aggregate patient utility, defined as $W = B_H + B_L$, with B_H and B_L as given in equations (3.8)–(3.9). The effect of a marginal change in switching costs on total patient utility is given by

$$\frac{\partial W}{\partial s} = \frac{\partial q_H^*}{\partial s} D_H^* + \frac{\partial q_L^*}{\partial s} D_L^* - \mu \left[\sigma_H(1 - \hat{x}_{|H}) + \sigma_L \hat{x}_{|L}\right] + (q_H^* - q_L^*) \frac{\partial D_H^*}{\partial s} + \mu\tau \left[\sigma_H(1 - 2\hat{x}_{|H}) \frac{\partial \hat{x}_{|H}}{\partial s} + \sigma_L(1 - 2\hat{x}_{|L}) \frac{\partial \hat{x}_{|L}}{\partial s}\right] + \mu s \left(\sigma_H \frac{\partial \hat{x}_{|H}}{\partial s} - \sigma_L \frac{\partial \hat{x}_{|L}}{\partial s}\right). \quad (3.21)$$

Lower switching costs have a fivefold effect on total utility. 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, all else equal, total switching costs increase because more patients switch. These effects are respectively given

by the terms on the right-hand side of (3.21). It turns out that the three last-mentioned effects cancel out, and equation (3.21) can be rewritten as

$$\frac{\partial W}{\partial s} = -\frac{\alpha/2}{\frac{2\tau\gamma}{\mu} - (\alpha - c)} + \frac{(\alpha - c)(\sigma_H - \sigma_L)(D_H^* - D_L^*)}{\frac{2\tau\gamma}{\mu} - (2\alpha - 3c)} - \mu\left[\sigma_H(1 - \hat{x}_{|H}) + \sigma_L\hat{x}_{|L}\right]. \quad (3.22)$$

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 aggregate utility may be stated as follows.

Proposition 3.4. *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*

$$\frac{\alpha/2}{\frac{2\tau\gamma}{\mu} - (\alpha - c_W)} + \frac{\mu}{2}\left(1 - \frac{s}{\tau}\right) = \frac{2[\tau - (\tau - s)\mu](\sigma_H - \sigma_L)^2(\alpha - c_W)\left[\frac{2\tau\gamma}{\mu} - (\alpha - 2c_W)\right]}{\tau\left[\frac{2\tau\gamma}{\mu} - (2\alpha - 3c_W)\right]^2}, \quad (3.23)$$

such that lower switching costs reduce aggregate utility if $c < c_W$.

Proof. See Appendix 3.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 aggregate utility is clearly positive. If $c < \alpha$, 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 when a utilitarian approach is adopted.

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 total utility, 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 total utility 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 total utility than it must be to reduce average quality, if required at all. The following result may therefore be established.

Corollary 3.1. *Lower switching costs reduce average quality but increase aggregate patient utility if* $c_W < c < c_{\bar{q}}$.

Again, semi-altruistic hospital preferences have an uncertain impact on welfare. The negative effect of lower switching costs on aggregate utility may materialise even in the presence of a sufficiently weak degree of cost substitutability when hospitals are semi-altruistic. If hospitals are pure profit-maximisers, however, then lower switching costs only reduce aggregate utility in the presence of sufficiently strong cost complementarity. Moreover, recall that, for lower switching costs to reduce average when hospitals are pure profit-maximisers, any degree of cost complementarity suffices.

3.6 Endogenous Switching Costs and Market Segmentation

In this section, we relax the assumption of exogenous switching costs and allow hospitals to set the switching cost their inherited patients incur if they decide to switch. Thus, besides setting q_i, Hospital i now sets s_i as well. For clarity of exposition, we interpret Hospital i's ability to control s_i as its ability to restrict its patients' access to their medical records or use and exchange thereof. In other words, s_i reflects the extent to which Hospital i practices *data blocking*.[13]

Proceeding analogously to section 3.3.1, it may be shown that total demand facing Hospital H is again given by $D_H(q_H, q_L) = \mu\sigma_H\hat{x}_{|H} + \mu\sigma_L\hat{x}_{|L} + (1-\mu)\sigma_L$, with s replaced by s_H and s_L, respectively, in the expressions for $\hat{x}_{|H}$ and $\hat{x}_{|L}$ as given by equations (3.2) and (3.4). Since $D_L(q_H, q_L) = 1 - D_H(q_H, q_L)$, total demand facing Hospital i may be written more explicitly as

$$D_i(q_i, q_j) = \frac{\mu}{2\tau}[\tau + q_i - q_j + \sigma_i s_i - \sigma_j s_j] + (1-\mu)\sigma_i, \quad i, j = H, L; \quad i \neq j; \tag{3.24}$$

provided that $s_L - \tau < |q_H - q_L| < \tau - s_H$.[14]

From the hospitals' perspective, aggregate patient benefit is

$$\begin{aligned} B_H[q_H, D_H(q_H, q_L)] = \mu\sigma_H \int_0^{\hat{x}_{|H}} (v + q_H - \tau x)dx \\ + \mu\sigma_L \int_0^{\hat{x}_{|L}} (v + q_H - \tau x - s_L)dx + (1-\mu)\int_0^{\sigma_H} (v + q_H - \tau x)dx \end{aligned} \tag{3.25}$$

[13] For example, the U.S. Department of Health and Human Services defines information blocking in health care as a 'practice by a health care provider that is likely to interfere with access, exchange, or use of electronic health information'.

[14] As before, switching from both hospitals only occurs in equilibrium if the preferences for treatment characteristics of some patients outweigh the endogenously set switching costs.

and

$$B_L[q_L, D_L(q_H, q_L)] = \mu\sigma_H \int_{\hat{x}_{|H}}^{1} [v + q_L - \tau(1-x) - s_H]dx + \mu\sigma_L \int_{\hat{x}_{|L}}^{1} [v + q_L - \tau(1-x)]dx + (1-\mu) \int_{\sigma_H}^{1} [v + q_L - \tau(1-x)]dx. \quad (3.26)$$

It is easily seen from (3.24)–(3.26) that a unilateral increase in switching costs has a simple demand retention effect. All else equal, a higher s_i increases demand from inherited patients (with changing preferences), thus increasing revenues, aggregate patient benefit, and treatment production costs. Unlike quality provision, there is no direct cost from increasing switching costs; only the indirect cost of treating additional inherited patients. In turn, this implies that each hospital's optimum level of switching cost is uniquely determined by the sign of the marginal payoff of *inherited* demand; or, in other words, by whether treating the marginal inherited patient is beneficial. This assertion is formalised by the following first-order conditions for hospitals H and L. Using equations (3.24)–(3.26), maximisation of (3.6) with respect to s_i yields

$$\frac{\mu\sigma_H}{2\tau} \left[p + (\alpha - c)q_H + \alpha(v - \tau\hat{x}_{|H})\right] \gtrless 0 \quad (3.27)$$

and

$$\frac{\mu\sigma_L}{2\tau} \left[p + (\alpha - c)q_L + \alpha[v - \tau(1 - \hat{x}_{|L})]\right] \gtrless 0. \quad (3.28)$$

Importantly, the above pair of first-order conditions reveals that, while the sign of the marginal payoff of inherited demand is independent of s_i, it depends on q_i.[15] Thus, from each hospital's perspective, the optimum switching cost will either be minimum ($s_i = 0$) or maximum ($s_i \to \infty$) depending on whether the hospital offers a quality level such that the marginal inherited patient becomes financially unprofitable to treat to an extent that the hospital finds it optimal to avoid that patient.

The unique equilibrium with non-negative quality provision when switching costs are endogenous is presented in the following proposition.

Proposition 3.5. *If hospitals can set the switching costs their inherited patients incur, the unique equilibrium with non-negative quality provision and full market coverage is characterised by maximum switching costs ($s_i^* \to \infty$), perfect history-based market segmentation ($D_i = \sigma_i$), and* $q_i^* = \max\left\{0, \frac{(\alpha-c)\sigma_i}{\gamma}\right\}$.

Proof. See Appendix 3.D. □

[15] Note that s_i indeed enters the expressions for $\hat{x}_{|H}$ and $\hat{x}_{|L}$ on the left-hand side of (3.27)–(3.28). The full market coverage assumption, however, ensures that the third term on the left-hand side of each of conditions (3.27)–(3.28) is always positive.

To grasp the mechanism driving the results presented in Proposition 3.5, notice that retaining the marginal patient through a unilateral increase in switching costs is always beneficial when $\alpha > c$. In this case, the marginal altruistic incentive to increase s_i dominates the cost of treating the marginal inherited patient for all s_i, and hospitals, therefore, have incentives to retain all of these patients by setting maximum switching costs. Each hospital becomes a monopolist on its base of previous patients, and quality affects only the utility of captive patients and treatment production costs. Strictly positive quality provision arises for uniquely altruistic motives as $\alpha > c$ implies that the hospitals' valuation of the utility their captive patients derive from quality provision outweighs the cost of treating them.

When $\alpha < c$, the indirect marginal cost of s_i outweighs the marginal altruistic benefit, making it *a priori* ambiguous whether hospitals will continue to retain all of their inherited patients. Suppose initially that quality provision by Hospital i is sufficiently high to make the marginal inherited patient so costly to treat that doing so becomes detrimental from the hospital's perspective (i.e., $\partial\Omega_i/\partial s_i < 0$). However, the hospital would only offer such quality level if attracting a 'new' patient (i.e., a patient who chose Hospital j in the past) were sufficiently more beneficial than retaining an inherited one. This, in turn, would only be true if Hospital i set a switching cost so high that it retained a share of its inherited patients large to the extent that the hospital, owing to its semi-altruistic preferences, preferred to 'internalise' the switching cost of a new patient to 'internalising' the travelling/mismatch costs of its inherited patients. These two conditions are clearly incompatible.[16] Then, if Hospital i offers a quality level such that the marginal inherited patient is (always) beneficial to treat, it will again find it optimal to retain all of its inherited patients by setting maximum switching costs. Once total lock-in is implemented, quality provision by each hospital ceases to be an instrument to attract new demand. In this case, switching costs and quality provision become perfectly substitutable strategies for each hospital because either a higher s_i or a higher q_i yields only additional demand from inherited patients. The difference between these two strategies lies in the costs of enacting each of them. Besides the cost of treating an additional patient, an increase in switching costs is costless, while quality provision, conversely, also requires a fixed investment and implies an increase in the cost of treatment of all patients. Hospitals are thus better off by setting maximum switching costs and minimum quality (i.e., $q_i = 0$), avoiding quality provision costs entirely.

Therefore, setting maximum switching costs is a strictly dominant strategy, and each hospital always

[16] This relationship is more easily illustrated by the case of profit-oriented hospitals and cost substitutability between quality and output, which implies $c > \alpha = 0$. For such a hospital, the benefit from treating an additional patient, inherited or new, is the same, and no patient is profitable to treat if $p < cq_i$. Only in this case, does the hospital prefer to avoid all of its inherited patients and to impose no switching costs. However, if no patient is profitable to treat, the hospital would have never offered such a quality level in the first place.

finds it optimal to become a monopolist on its base of previous patients independently of the relative size of α and c and the degree of asymmetry in inherited demand. Interestingly, the ability to control their inherited patients' switching costs allows hospitals to replicate the limiting case of a market where all patients have fixed preferences for treatment characteristics (i.e., $\mu = 0$).

Finally, these results allow for a brief thought experiment. Suppose that the market is initially characterised by local monopolies as described in Proposition 3.5. Suppose also that a policy capable of preventing hospitals from imposing switching costs is implemented; for example, the prohibition of data blocking. In this case, it is reasonable to assume that patients would nonetheless incur some switching costs, not controlled by the hospitals, and a good approximation of this scenario would be the model of section 3.3. Let us then consider the welfare implications of breaking up those local monopolies. Provided that the prospective payment is sufficiently high, introducing *de facto* competition is generally beneficial to patient welfare. If $\alpha < c$, any prospective payment that elicits positive quality leads to higher average quality—which is null under local monopolies—, and even a moderately high prospective payment yields the same outcome if $\alpha > c$. In the case of aggregate patient utility, a welfare improvement is more easily achieved. By granting patients the ability to adjust their choice of hospital according to their changing preferences, such a policy reduces the asymmetry in the market and thus leads to lower aggregate travelling/mismatch costs. Importantly, this reduction in travelling/mismatch cost more than compensates for the newly generated switching costs, which implies that, even if only a marginal increase in quality results from breaking up the local monopolies, the impact on total patient utility is positive.

3.7 Discussion and Concluding Remarks

Employing a duopoly model that maps on the recent empirical evidence on choice persistence in the hospital industry, this chapter explored the role of switching costs in the context of hospital competition. First, it investigated the effect of lower exogenous switching costs on the quality of elective hospital treatments and patient welfare. Second, it analysed how quality provision and market structure are affected by hospitals' ability to impose switching costs on their patients.

While lower exogenous switching costs always reduce market concentration, the impact on quality provision depends crucially on the technology of production of hospital treatments and 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 health

care industry are taken into account—in this chapter, 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 grants the high-volume hospital the demand advantage that allows it to offer higher quality. In this case, facilitated switching triggers a patient outflow at the high-volume hospital, despite its higher quality, which leaves the majority of patients in the market worse off in terms of quality provision, contributing to lower average quality and aggregate utility. These results from the first part of the chapter have several policy implications which are discussed in the following.

First, the adverse effects of lower switching costs may require cost complementarity to materialize. This suggests that knowledge of hospital production attributes is key to anticipating the effect of 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 speciality or department level) but also that the institutional setting may play a role.

Second, perhaps surprisingly, average quality might fall even if lower switching costs lead to higher quality at the hospital level owing to a demand redistribution effect. In his empirical study, Irace (2018) identifies a demand redistribution effect, which consists in patients switching to higher-quality hospitals in a counterfactual scenario where switching costs are absent. Our model shows that such demand adjustment may also occur in the opposite direction. 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 initially large enough, this effect dominates and lower switching costs reduce average quality. This illustrates an important point regarding policies aimed at reducing switching cost-induced inertia. They affect both the hospital-side (via quality provision) and the patient-side of the market (via their ability to adjust), and these two effects may be conflicting. While lower switching costs may trigger

a quality increase, which all else equal leads to higher patient welfare, they also increase patients' ability to adjust to changes in their horizontal preferences (or diagnostic), which in turn may drive them toward lower-quality hospitals. 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.

Third, different measures of patient welfare yield distinct conclusions. We have shown that lower switching costs might increase total patient utility while reducing average quality. If policymakers are mostly concerned with clinical outcomes and indicators, the average quality will arguably be a more appropriate measure of welfare and lower switching more likely to be deemed welfare-decreasing. If, conversely, 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.

Consider, finally, the case of endogenous switching costs analysed in the second part of the chapter. Maximum switching costs and the emergence of local monopolies are strong theoretical results. Under the EHR and data blocking interpretation of s_i, however, these results are in line with the findings on 'information silos', data systems that exchange no data with similar systems, presented by Miller and Tucker (2014). They show that hospitals facing a greater commercial cost from allowing data outflow, those belonging to larger hospital systems, are less likely to exchange patient information externally with other hospitals, thereby creating information silos. The driver of this behaviour is the prospect that a more efficient information flow may cause patients to switch to rival hospitals. Thus, to the extent that data blocking can be interpreted as a form of endogenous switching costs, the local monopolies described in Proposition 3.5 may also be interpreted as information silos.

Appendix 3.A Proof of Proposition 3.2

From equation (3.16), let the effect of a marginal change in switching costs on equilibrium quality at Hospitals H and L be written, respectively, as

$$q_H^{*\prime}(c) = -\frac{\alpha/2}{\frac{2\tau\gamma}{\mu} - (\alpha - c)} + \frac{(\alpha - c)(\sigma_H - \sigma_L)}{\frac{2\tau\gamma}{\mu} - (2\alpha - 3c)} \tag{3.29}$$

and

$$q_L^{*\prime}(c) = -\frac{\alpha/2}{\frac{2\tau\gamma}{\mu} - (\alpha - c)} - \frac{(\alpha - c)(\sigma_H - \sigma_L)}{\frac{2\tau\gamma}{\mu} - (2\alpha - 3c)}, \tag{3.30}$$

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

$$c_{HH} = \alpha - \frac{4\tau\gamma(\sigma_H - \sigma_L) + 3\alpha\mu}{4\mu(\sigma_H - \sigma_L)} \pm \frac{\sqrt{8\tau\gamma(\sigma_H - \sigma_L)[2\tau\gamma(\sigma_H - \sigma_L) + \alpha\mu] + (\alpha\mu)^2[9 - 8(\sigma_H - \sigma_L)]}}{4\mu(\sigma_H - \sigma_L)} \quad (3.31)$$

and

$$c_{HL} = \alpha - \frac{4\tau\gamma(\sigma_H - \sigma_L) - 3\alpha\mu}{4\mu(\sigma_H - \sigma_L)} \pm \frac{\sqrt{8\tau\gamma(\sigma_H - \sigma_L)[2\tau\gamma(\sigma_H - \sigma_L) - \alpha\mu] + (\alpha\mu)^2[9 + 8(\sigma_H - \sigma_L)]}}{4\mu(\sigma_H - \sigma_L)}. \quad (3.32)$$

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. 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 (3.11). This is true if, for example, the following inequality holds:

$$\alpha - \frac{4\tau\gamma(\sigma_H - \sigma_L) + 3\alpha\mu}{4\mu(\sigma_H - \sigma_L)} - \frac{\sqrt{8\tau\gamma(\sigma_H - \sigma_L)[2\tau\gamma(\sigma_H - \sigma_L) + \alpha\mu] + (\alpha\mu)^2[9 - 8(\sigma_H - \sigma_L)]}}{4\mu(\sigma_H - \sigma_L)} < \frac{2\alpha}{3} - \frac{2\tau\gamma}{3\mu}. \quad (3.33)$$

The above inequality can be written as

$$\frac{4\tau\gamma(\sigma_H - \sigma_L)}{3} + \left[3 - \frac{4(\sigma_H - \sigma_L)}{3}\right]\alpha\mu + \sqrt{8\tau\gamma(\sigma_H - \sigma_L)[2\tau\gamma(\sigma_H - \sigma_L) + \alpha\mu] + (\alpha\mu)^2[9 - 8(\sigma_H - \sigma_L)]} > 0, \quad (3.34)$$

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 (3.11). The corresponding three inequalities can be, respectively, written as

$$\sqrt{8\tau\gamma(\sigma_H - \sigma_L)[2\tau\gamma(\sigma_H - \sigma_L) + \alpha\mu] + (\alpha\mu)^2[9 - 8(\sigma_H - \sigma_L)]} > |4\tau\gamma(\sigma_H - \sigma_L) - 3\alpha\mu| \iff 8(\sigma_H - \sigma_L)\alpha\mu(4\tau\gamma - \alpha\mu) > 0, \quad (3.35)$$

$$8(\tau\gamma)^2 + 2\tau\gamma\alpha\mu - (\alpha\mu)^2 > 0, \quad (3.36)$$

and

$$(\sigma_H - \sigma_L)(4\tau\gamma - \alpha\mu)[4(\sigma_H - \sigma_L)\tau\gamma + (2 \ + \sigma_H - \sigma_L)\alpha\mu] > 0. \quad (3.37)$$

The three inequalities hold simultaneously if $\gamma > \frac{\alpha\mu}{4\tau}$. Given (3.11), the denominators on the right-hand side of (3.29) 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 in (c_{min}, α), and it is given by the greater root in (3.31).

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.[17] Note again that, given (3.11), the denominators on the right-hand side of (3.30) 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

$$\sqrt{8\tau\gamma(\sigma_H - \sigma_L)[2\tau\gamma(\sigma_H - \sigma_L) - \alpha\mu] + (\alpha\mu)^2[9 + 8(\sigma_H - \sigma_L)]} > |4\tau\gamma(\sigma_H - \sigma_L) - 3\alpha\mu|, \tag{3.38}$$

which simplifies to

$$8(\sigma_H - \sigma_L)\alpha\mu(2\tau\gamma + \alpha\mu) > 0, \tag{3.39}$$

revealing that it is always satisfied. The solution to $q_L^{*\prime}(c) = 0$ is therefore in (α, ∞), and it is given by the greater root in (3.32).

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 $c > c_{HL}$.

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

$$\frac{\partial c_{HH}}{\partial \mu} = \frac{\tau\gamma}{\mu^2}\left(1 - \frac{4\tau\gamma(\sigma_H - \sigma_L) + \alpha\mu}{\sqrt{8\tau\gamma(\sigma_H - \sigma_L)[2\tau\gamma(\sigma_H - \sigma_L) + \alpha\mu] + (\alpha\mu)^2[9 - 8(\sigma_H - \sigma_L)]}}\right) > 0, \tag{3.40}$$

$$\frac{\partial c_{HL}}{\partial \mu} = \frac{\tau\gamma}{\mu^2}\left(1 - \frac{4\tau\gamma(\sigma_H - \sigma_L) - \alpha\mu}{\sqrt{8\tau\gamma(\sigma_H - \sigma_L)[2\tau\gamma(\sigma_H - \sigma_L) - \alpha\mu] + (\alpha\mu)^2[9 + 8(\sigma_H - \sigma_L)]}}\right) > 0, \tag{3.41}$$

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

$$\frac{\partial(c_{HL}-c_{HH})}{\partial\mu}=\frac{\tau\gamma}{\mu^2}\left(\frac{4\tau\gamma(\sigma_H-\sigma_L)+\alpha\mu}{\sqrt{8\tau\gamma(\sigma_H-\sigma_L)[2\tau\gamma(\sigma_H-\sigma_L)+\alpha\mu]+(\alpha\mu)^2[9-8(\sigma_H-\sigma_L)]}}\right.$$
$$\left.-\frac{4\tau\gamma(\sigma_H-\sigma_L)-\alpha\mu}{\sqrt{8\tau\gamma(\sigma_H-\sigma_L)[2\tau\gamma(\sigma_H-\sigma_L)-\alpha\mu]+(\alpha\mu)^2[9+8(\sigma_H-\sigma_L)]}}\right)>0, \quad (3.42)$$

$$\frac{\partial c_{HH}}{\partial\tau}=-\frac{\mu}{\tau}\frac{\partial c_{HH}}{\partial\mu}<0, \quad (3.43)$$

$$\frac{\partial c_{HL}}{\partial\tau}=-\frac{\mu}{\tau}\frac{\partial c_{HL}}{\partial\mu}<0, \quad (3.44)$$

and

$$\frac{\partial(c_{HL}-c_{HH})}{\partial\tau}=-\frac{\mu}{\tau}\frac{\partial(c_{HL}-c_{HL})}{\partial\mu}<0. \quad (3.45)$$

Note that the term on the right-hand side of (3.42) is positive since

$$[4\tau\gamma(\sigma_H-\sigma_L)+\alpha\mu]^2\{8\tau\gamma(\sigma_H-\sigma_L)[2\tau\gamma(\sigma_H-\sigma_L)-\alpha\mu]+(\alpha\mu)^2[9+8(\sigma_H-\sigma_L)]\}$$
$$-[4\tau\gamma(\sigma_H-\sigma_L)-\alpha\mu]^2\{8\tau\gamma(\sigma_H-\sigma_L)[2\tau\gamma(\sigma_H-\sigma_L)+\alpha\mu]+(\alpha\mu)^2[9-8(\sigma_H-\sigma_L)]\}$$
$$=(\sigma_H-\sigma_L)(4\alpha\mu)^2\{[4(\sigma_H-\sigma_L)\tau\gamma]^2+(\alpha\mu)^2+8\tau\gamma\alpha\mu\}>0. \quad (3.46)$$

Appendix 3.B Proof of Proposition 3.3

Using equations (3.5), (3.17), and (3.18), the effect of a marginal change in switching costs on average quality may be written as

$$\overline{q}'(c)=-\frac{\alpha/2}{\frac{2\tau\gamma}{\mu}-(\alpha-c)}+\frac{2[\tau-(\tau-s)\mu](\sigma_H-\sigma_L)^2(\alpha-c)\left(\frac{2\tau\gamma}{\mu}+c\right)}{\tau\left[\frac{2\tau\gamma}{\mu}-(2\alpha-3c)\right]^2}, \quad (3.47)$$

where prime denotes the derivative with respect to s.

Following the proof of Proposition 3.2 in Appendix 3.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 (3.47) 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, we proceed in three steps: (i) we show that the graph of $\overline{q}'(c)$ first approaches the horizontal axis from above as c increases in (c_{min}, α); (ii) we show that $\overline{q}'(c)$ has either one or three roots in (c_{min}, α); and (iii) we 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 (3.19). 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 (3.20) 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,

$$\overline{q}'(c_{HH}) > 0 \iff q'_{\overline{q}}(c_{HH}) > q^{*\prime}_H(c_{HH}). \tag{3.48}$$

This condition may be rewritten as

$$\frac{2[\tau - (\tau - s)\mu](\sigma_H - \sigma_L)^2(\alpha - c_{HH})\left(\frac{2\tau\gamma}{\mu} + c_{HH}\right)}{\tau\left[\frac{2\tau\gamma}{\mu} - (2\alpha - 3c_{HH})\right]^2} > \frac{(\alpha - c_{HH})(\sigma_H - \sigma_L)}{\frac{2\tau\gamma}{\mu} - (2\alpha - 3c_{HH})}. \tag{3.49}$$

Solving for s yields (3.20). This concludes the proof of Proposition 3.3.

Appendix 3.C Proof of Proposition 3.4

The proof of Proposition 3.4 is analogous to that of Proposition 3.3.

Using equations (3.2), (3.4), (3.5), and (3.22), the effect of a marginal change in switching costs on total patient welfare may be written as

$$W'(c) = -\left[\frac{\alpha/2}{\frac{2\tau\gamma}{\mu} - (\alpha - c)} + \frac{\mu}{2}\left(1 - \frac{s}{\tau}\right)\right] + \frac{2[\tau - (\tau - s)\mu](\sigma_H - \sigma_L)^2(\alpha - c)\left[\frac{2\tau\gamma}{\mu} - (\alpha - 2c)\right]}{\tau\left[\frac{2\tau\gamma}{\mu} - (2\alpha - 3c)\right]^2}, \quad (3.50)$$

where prime denotes the derivative with respect to s.

Following the proof of Proposition 3.2 in Appendix 3.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 (3.50) 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, we proceed in three steps: (i) we show that the graph of $W'(c)$ first approaches the horizontal axis from above as c increases in (c_{min}, α); (ii) we show that $W'(c)$ has either one or three roots in (c_{min}, α); and (iii) we 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 3.2 in Appendix 3.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,

$$\overline{q}'(c_W) > 0 \iff q'_{\overline{q}}(c_W) > W'(c_W). \tag{3.51}$$

This condition may be rewritten as

$$\frac{2[\tau - (\tau - s)\mu](\sigma_H - \sigma_L)^2(\alpha - c_W)^2}{\tau\left[\frac{2\tau\gamma}{\mu} - (2\alpha - 3c_W)\right]^2} > -\frac{\mu}{2}\left(1 - \frac{s}{\tau}\right), \tag{3.52}$$

which is always satisfied. From the proof of Proposition 3.3 in Appendix 3.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 (3.23).

Appendix 3.D Proof of Proposition 3.5

Maximisation of Hospital i's objective function (3.6) with respect to s_i and q_i, with D_i and B_i given by equations (3.24)–(3.26), yields, after manipulation, the following first-order conditions:

$$p + (\alpha - c)q_H + \alpha(v - \tau\hat{x}_{|H}) \gtrless 0, \tag{3.53}$$

$$p + (\alpha - c)q_L + \alpha[v - \tau(1 - \hat{x}_{|L})] \gtrless 0, \tag{3.54}$$

$$\frac{\mu}{2\tau}\left[p+(\alpha-c)q_H+\alpha(v-\tau\hat{x}_{|H})+\frac{\alpha\sigma_L(s_H-s_L)}{2}\right]+(\alpha-c)D_H=\gamma q_H, \tag{3.55}$$

$$\frac{\mu}{2\tau}\left[p+(\alpha-c)q_L+\alpha[v-\tau(1-\hat{x}_{|L})]-\frac{\alpha\sigma_H(s_H-s_L)}{2}\right]+(\alpha-c)D_L=\gamma q_L. \tag{3.56}$$

Full market coverage for $q_i \geq 0$ implies that the third term on the left-hand side of each of conditions (3.53) and (3.54) is always positive, which in turn implies that the sign of the expressions on the left-hand side of each of those conditions is independent of s_i. Thus, if an equilibrium exists, it will be characterized by minimum (i.e., $\partial\Omega_i/\partial s_i < 0$ and $s_i^* = 0$) or maximum (i.e., $\partial\Omega_i/\partial s_i > 0$ and $s_i^* \to \infty$) switching costs.

If $\alpha > c$, it follows immediately that $\partial\Omega_i/\partial s_i > 0$. With $s_i^* \to \infty$, no patient switches, and $D_i = \sigma_i$; thus, equations (3.55) and (3.56) fail to define the equilibrium quality levels. Maximization of (3.6) with respect to q_i when $D_i = \sigma_i$ yields $q_i^* = \frac{(\alpha-c)\sigma_i}{2}$.

To show that this is the unique equilibrium with non-negative quality provision also when $c > \alpha$, we first rule out other candidate equilibria and then compute the equilibrium by construction.

Start by noticing that equation (3.24) holds when one or both hospitals set switching costs equal to zero, so that (3.55)–(3.56) continue to define equilibrium quality levels. Suppose Hospital H sets $s_H = 0$. This is the case if the left-hand side of equation (3.53) is negative (i.e., $\partial\Omega_H/\partial s_H < 0$). Inserting (3.53) into (3.55) shows that the latter equation is only satisfied with $q_H \geq 0$ for $s_L < 0$, which cannot be true. Conducting the analogous steps for Hospital L and the pair of equations (3.54)–(3.56), as well as considering $s_H = s_L = 0$ simultaneously, yields the same result. Thus, an equilibrium characterized by non-negative quality provision, if it exists, must have $s_i > 0 \forall i = H, L$.

Suppose now that Hospital L sets a positive s_L such that none of its inherited patients switches (i.e., $\partial\Omega_H/\partial s_H > 0$). In this case, equations (3.55)–(3.56) no longer define equilibrium quality levels, because (3.24) fails to hold. Total demand facing Hospital H is now given by $D_H = \mu\sigma_H\hat{x}_{|H}+(1-\mu)\sigma_H$. The first-order condition defining the optimum s_H continues to be given by (3.53), but the first-order condition defining the optimum q_H is now given by

$$\frac{\mu\sigma_H}{2\tau}\left[p+(\alpha-c)q_H+\alpha(v-\tau\hat{x}_{|H})\right]+(\alpha-c)D_H=\gamma q_H. \tag{3.57}$$

Because $(\alpha-c)D_H < 0$, $q_H \geq 0$ requires that the expression on the left-hand side of (3.53) is strictly positive, implying $s_H \to \infty$. Again, fixing Hospital H's strategy deriving Hospital L's best response yields the analogous result.

Therefore, $s_i \to \infty$ is Hospital i's best response to $s_j \to \infty$, and $(s_H \to \infty, s_L \to \infty)$ is the unique Nash Equilibrium with non-negative quality provision and full market coverage.

4. Quality Provision in Hospital Markets With Demand Inertia: The Role of Patient Expectations[1]

4.1 Introduction

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

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

[1] This chapter is co-authored with Odd Rune Straume.

is costly, choosing a particular hospital in the present has an impact on the utility patients will derive from treatment at different hospitals in their choice set in the future, thereby affecting their current choice.[2]

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

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

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

[2] See section 3.1 for an in-depth discussion of switching costs in hospital markets.

such a task is unavailable, these patients are naïve in the sense that they resort to the simple rule-of-thumb of expecting that quality will remain constant.

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

To make the analysis of the evolution of quality more comprehensive, we assume that the hospitals are motivated and allow for both cost substitutability and complementarity between quality and output in hospital production. If the degree of cost substitutability is sufficiently strong, higher demand increases the marginal cost of quality provision. This, in turn, implies that higher quality in the present foretells lower quality in the future or, more specifically, that a current unilateral quality increase reduces the future quality difference. A current unilateral quality increase yields a demand advantage, which, owing to inertia, partially carries over into the future, increasing the marginal cost of quality and thus reducing the incentives for quality provision. Similarly, if there is cost complementarity (or if the degree of cost substitutability is sufficiently weak), higher demand reduces the marginal cost of quality and implies that a current unilateral quality increase widens both the current and the future quality differences.[3] This link between present and future quality, and the fact that only rational patients observe it, partly explain our results.

We show that patient expectations affect quality provision only through the responsiveness of demand to quality, with higher responsiveness leading to higher quality provision. While demand is always more responsive when patients are forward-looking but naïve than when patients are myopic, demand re-

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

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

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

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

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

The rest of the chapter is organised as follows. In the next section, we relate our study to several strands of literature. In section 4.3, we present the model and, in section 4.4, derive the equilibrium quality levels in the two-period game. Our primary analysis is given in sections 4.5, 4.6, and 4.7, where

we explore the role of patient expectations thoroughly, compare quality provision with the benchmark of a market without demand inertia, and investigate the effect of lower switching costs. Finally, as well as concluding remarks, section 4.8 provides a discussion of the implications of forward-looking and rational behaviour to patient utility.

4.2 Related Literature

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

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

share induced by switching costs, leading to higher quality provision.[4]

In the context of quality competition in primary care, Gravelle and Masiero (2000) present a two-period model where myopic patients incur switching costs. Contrary to our results, they show that quality is unaffected by switching costs. Within the hospital competition literature, two studies consider an information-related form of inertia. Arising from the complexity of assessing the quality of care, demand sluggishness implies that, at each point in time, only a fraction of patients become aware of quality changes and hence only a fraction of any potential change in demand materialises. Weaker sluggishness therefore makes demand more responsive to quality. With profit-maximising providers and a positive payment-cost margin, as in Brekke et al. (2012), increased demand responsiveness leads to higher quality. Siciliani, Straume, and Cellini (2013), however, show that semi-altruistic hospital preferences may overturn this result. Increased demand responsiveness leads to lower quality provision if the prospective payment is sufficiently below unit costs and the financial incentive to avoid patients dominates the altruistic incentive to attract them.[5] Although demand responsiveness to quality also plays a crucial role in our model, our analysis differs significantly from those of Brekke et al. (2012) and Siciliani, Straume, and Cellini (2013). First, they model inertia in a multiperiod framework where expectations are unexplored. Second, they focus on exogenous changes in parameters that affect demand responsiveness and on how this, in turn, impacts quality provision, given hospital preferences and technology. Here, we mainly investigate how patient expectations determine demand responsiveness endogenously and show that hospital preferences and technology may themselves affect demand responsiveness.

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

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

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

year of childbirth, fathers' monthly medical care utilisation rises by 11% upon becoming eligible for coinsurance, despite childbirth being an expected event that contributes a great deal to deductible fulfilment. Absent liquidity constraints and controlling for health shocks, these fluctuations in healthcare utilisation are consistent with some degree of myopic behaviour since a fully forward-looking consumer would respond to his expected end-of-year price rather than to the spot price, thereby smoothing consumption over the year. Myopic behaviour instead implies that consumers perceive changes in coverage as changes in prices and hence adjust consumption accordingly. Dalton et al. (2020) provide even stronger evidence of myopic behaviour. They find that consumers completely ignore the future prices of prescription drugs under Medicare Part D, whose nonlinear contract design includes an initial coverage region followed by a coverage gap (the 'doughnut hole'). Drug purchases are initially constant and drop sharply once the coverage gap is reached, implying an estimated discount rate that is consistent with full myopia (i.e., equal to zero). A similar pattern of drug consumption under Medicare Part D may be found in Sacks et al. (2017), Einav et al. (2015), and Abaluck et al. (2018). In the latter two studies, however, the estimated discount rates indicate some degree of forward-looking behaviour, which is considerably higher in Einav et al. (2015). Additional evidence of forward-looking behaviour comes from Aron-Dine et al. (2015). They find that initial medical care utilisation is lower for employees who join a health insurance plan with an annual deductible later in the year. Because their deductible is less likely to be reached, individuals who enrol later face a higher expected end-of-year price. Their lower initial utilisation under the plan therefore suggests that they do respond to future prices. Interestingly, Aron-Dine et al. (2015) find similar results for prescription drug consumption under Medicare Part D. Looking at the German public health insurance system, Farbmacher et al. (2017) also report evidence of forward-looking behaviour. After the introduction of a one-time co-payment, initial outpatient care demand falls for some consumers, while it is unresponsive for the relatively sick, who should expect future needs to exceed a single visit and thus be less sensitive to the co-payment.

4.3 The Model

Consider a healthcare market with two providers, henceforth referred to as hospitals. In each of two periods, $t = 1, 2$, the two hospitals, indexed $i = A, B$, are located at either endpoint of the unit line segment $[0, 1]$. Let Hospital A be located at 0 and Hospital B at 1. Locations on the line segment reflect the characteristics and preferences for elective hospital treatment supplied in this market. The line segment

may be thought of as a geographical space or a disease space. In the former case, a patient's location on the line is simply her residence or workplace, while the location of a hospital is simply the place where its facilities were built. In the latter case, a patient's location on the line is a medical condition or a diagnosis, and the location of a hospital is the speciality mix (i.e., the treatments and services) it offers.

Patients have a gross valuation of treatment $v > 0$, demand a single unit of treatment from one of the hospitals in each period, and are arrayed with unit density along the line segment. They incur a travelling or mismatch cost τ per unit of distance between their location and that of the chosen hospital, but bear no out-of-pocket expenses either due to public provision of healthcare or to (social or private) health insurance coverage.[6] Patients derive utility from the quality of treatment, q_t^i, to which hospitals resort to attract demand in each period. There is a lower bound on treatment quality that represents the minimum quality hospitals are allowed to offer, with quality below this threshold being interpreted as malpractice. For simplicity, we assume that the lower bound on quality is equal to zero. The gross valuation of treatment v is high enough so that the market is always fully covered.

Following the empirical analyses of Raval and Rosenbaum (2018) and Irace (2018), we model demand inertia in the style of Klemperer (1987). In the first period, all patients are new in the market, meaning that no patient is tied to any of the hospitals. Patients choose a hospital based on their horizontal preferences and the quality levels offered in the market. In the second period, however, the patient population consists of three different segments. (i) A fraction λ of the patients leave the market and are replaced by new patients with the same density and who are also uniformly distributed along the unit line segment. (ii) Another fraction μ of the existing patients have preferences for treatment characteristics that are independent of their first-period preferences (i.e., their location on the unit line is re-drawn at the start of the second period). 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. We will henceforth refer to them as patients with *changing preferences*. The parameter μ may, therefore, be interpreted as an inverse measure of the persistence of patient preferences over time. Patients with changing preferences who choose to demand treatment from the hospital they have not used in the first period incur an exogenous switching cost s. (iii) The remaining $(1 - \lambda - \mu)$ patients have unchanged preferences for treatment characteristics (i.e., their location on the line segment equals the first-period location) and choose the same hospital in both periods.[7] Thus, we measure demand inertia in two different

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

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

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

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

$$u_t(x_t, z^i) = v + q_t^i - \tau|x_t - z^i| - I_i s, \quad i, j = A, B; \tag{4.1}$$

where $I_i = 1$ in the second period if the patient has changing preferences, chose Hospital i in the first period, and chooses Hospital j in the second period; $I_i = 0$ otherwise.[8]

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

$$C\left(q_t^i, D_t^i\right) = (cq_t^i + k)D_t^i + \frac{\gamma}{2}(q_t^i)^2, \quad i, j = A, B; \quad i \neq j; \tag{4.2}$$

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

If $c > 0$, a certain level of quality is more costly to achieve when more patients are treated, implying that the marginal cost of quality is increasing in demand. In this case, hospital production exhibits *cost substitutability* between quality and output. This is a reasonable assumption if quality results from the investment in medical equipment and highly skilled staff. For example, offering an additional diagnostic test amounts to an increase in quality and requires a fixed investment in equipment and/or staff but also increases the cost of diagnosing each patient. On the other hand, if $c < 0$, the more patients a hospital treats, the less costly it is to provide each additional unit of quality, and the marginal cost of quality is decreasing in demand. In this case, quality and output are *cost complements*, reflecting the positive relationship between demand and quality observed when, all else equal, high-volume hospitals provide higher quality and generate better treatment outcomes than low-volume hospitals.[9]

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

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

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

$$\Omega_t^i = T + \tilde{p}D_t^i - C\left(q_t^i, D_t^i\right) + \alpha(v + q_t^i)D_t^i. \tag{4.3}$$

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

Finally, we impose the following restriction on parameter values:

$$c > c_{min} = \max\left\{\alpha - \frac{2\tau\gamma}{3(\lambda+\mu)}, \alpha - \tau\gamma\right\}. \tag{4.4}$$

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

4.4 Equilibrium Quality Provision

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

4.4.1 The second period

Consider the different groups of patients in turn. A fraction λ of patients were not in the market in the first period and are not therefore tied to any of the hospitals. The new patient who is indifferent between

seeking treatment at Hospital A and Hospital B is located at $\hat{x}$, given by

$$\hat{x} = \frac{1}{2} + \frac{q_2^A - q_2^B}{2\tau}. \tag{4.5}$$

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

A fraction μD_1^A of patients sought treatment from Hospital A in the first period and now have preferences for treatment characteristics that are uniformly distributed along the line segment $[0,1]$. The patient who was previously treated at Hospital A and is now indifferent between seeking treatment at Hospital A and Hospital B is located at $\hat{x}_{|A}$, given by

$$\hat{x}_{|A} = \frac{1}{2} + \frac{q_2^A - q_2^B + s}{2\tau}. \tag{4.6}$$

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

Similarly, a fraction μD_1^B of patients sought treatment from Hospital B in the first period and now have preferences for treatment characteristics that are uniformly distributed along the line segment $[0,1]$. The patient who was previously treated at Hospital B and is now indifferent between seeking treatment at Hospital A and Hospital B is located at $\hat{x}_{|B}$, given by

$$\hat{x}_{|B} = \frac{1}{2} + \frac{q_2^A - q_2^B - s}{2\tau}. \tag{4.7}$$

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

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

$$D_2^i(q_2^i, q_2^j) = \frac{\lambda+\mu}{2\tau}(\tau + q_2^i - q_2^j) + \frac{\mu}{2\tau}(D_1^i - D_1^j)s + (1-\lambda-\mu)D_1^i, \quad i,j = A,B; \quad i \neq j; \tag{4.8}$$

provided that $|q_2^A - q_2^B| < \tau - s$.[10]

Taking first-period demand as given, Hospital i maximises

$$\Omega_2^i(q_2^i, q_2^j) = T + [p + (\alpha - c)q_2^i]D_2^i(q_2^i, q_2^j) - \frac{\gamma}{2}(q_2^i)^2, \quad i,j = A,B; \quad i \neq j; \tag{4.9}$$

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

where $p = \tilde{p} - k + \alpha v$. Maximisation of (4.9) with respect to q_2^i yields the candidate equilibrium quality levels

$$q_2^{i*} = \frac{p + (\alpha - c)\left[\tau - \frac{\mu s}{\lambda + \mu}\right] - (\alpha - c)^2 \phi}{\frac{2\tau\gamma}{\lambda+\mu} - (\alpha - c)} + (\alpha - c)\phi D_1^i, \quad i = A, B, \tag{4.10}$$

where

$$\phi = \frac{2\tau(1 - \lambda - \mu + \frac{\mu s}{\tau})}{(\lambda + \mu)\left[\frac{2\tau\gamma}{\lambda+\mu} - 3(\alpha - c)\right]} > 0. \tag{4.11}$$

The parameter restriction given in (4.4) ensures that the second-order condition is always satisfied, provided that (4.8) holds. However, this is insufficient to prove that the pair of strategies (4.10) define an equilibrium in the second-period subgame. It must be ensured that hospitals do not deviate and serve only their captive patients with fixed preferences, thus choosing a quality level outside the range in which (4.8) holds. As Klemperer (1987) notes, the deviation is not beneficial if $\lambda + \mu$ is large enough and the difference between first-period demands is sufficiently small. In the next subsection, we show that a symmetric pure-strategy candidate subgame perfect equilibrium exists and assume $\lambda + \mu$ is such that it indeed is an equilibrium.

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

$$q_2^* = \frac{p + \frac{(\alpha - c)\tau}{\lambda + \mu}}{\frac{2\tau\gamma}{\lambda+\mu} - (\alpha - c)}. \tag{4.12}$$

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

$$\frac{\partial \Omega_2^i(q_2^*)}{\partial D_1^i} = \phi\left(\frac{\lambda + \mu}{\tau}\right)\left(\frac{\tau\gamma}{\lambda + \mu} - \alpha + c\right)[p + (\alpha - c)q_2^*] > 0, \quad i = A, B. \tag{4.13}$$

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

4.4.2 The first period

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

$$\sum_{t=1}^{2} \Omega_t^i(q_1^i, q_1^j) = T + [p + (\alpha - c)q_1^i] D_1^i(q_1^i, q_1^j) - \frac{\gamma}{2}(q_1^i)^2 + \Omega_2^i[D_1^i(q_1^i, q_1^j)], \quad i, j = A, B; \quad i \neq j. \tag{4.14}$$

The first- and second-order conditions of the hospital's maximisation problem are respectively given by[11]

$$[p + (\alpha - c)q_1^i] \frac{\partial D_1^i}{\partial q_1^i} + (\alpha - c) D_1^i - \gamma q_1^i + \frac{\partial \Omega_2^i}{\partial D_1^i} \frac{\partial D_1^i}{\partial q_1^i} = 0 \tag{4.15}$$

and

$$\gamma - 2(\alpha - c)\frac{\partial D_1^i}{\partial q_1^i} > \left(\frac{\lambda + \mu}{\tau}\right)\left(\frac{\tau\gamma}{\lambda + \mu} - \alpha + c\right)\left[(\alpha - c)\phi \frac{\partial D_1^i}{\partial q_1^i}\right]^2 + \left[p + (\alpha - c)q_1^i + \frac{\partial \Omega_2^i}{\partial D_1^i}\right] \frac{\partial^2 D_1^i}{\partial (q_1^i)^2}, \tag{4.16}$$

where $i, j = A, B$ and $i \neq j$. Applying symmetry and using (4.13), first-period equilibrium quality, q_1^*, is implicitly defined by

$$\left[p + (\alpha - c)q_1^* + \phi\left(\frac{\lambda + \mu}{\tau}\right)\left(\frac{\tau\gamma}{\lambda + \mu} - \alpha + c\right)[p + (\alpha - c)q_2^*]\right] \frac{\partial D_1^i}{\partial q_1^i} + \frac{\alpha - c}{2} = \gamma q_1^*. \tag{4.17}$$

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

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

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

patient anticipates that he will choose Hospital B and incur the switching cost if x exceeds that threshold. With probability $1-\lambda-\mu$, the patient has persistent preferences (i.e., he is located at y also in the second period) and will again choose Hospital A. Then, the expected total utility (first-period utility plus expected second-period utility) of the patient located at y which results from choosing Hospital A in the first period is

$$(v+q_1^A-\tau y)+\mu\left[\begin{array}{c}\int_0^{\frac{1}{2}+\frac{q_E^A-q_E^B+s}{2\tau}}(v+q_E^A-\tau x)dx\\+\int_{\frac{1}{2}+\frac{q_E^A-q_E^B+s}{2\tau}}^1[v+q_E^B-\tau(1-x)-s]dx\end{array}\right]+(1-\lambda-\mu)(v+q_E^A-\tau y). \tag{4.18}$$

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

$$[v+q_1^B-\tau(1-y)]+\mu\left[\begin{array}{c}\int_0^{\frac{1}{2}+\frac{q_E^A-q_E^B-s}{2\tau}}(v+q_E^A-\tau x-s)dx\\+\int_{\frac{1}{2}+\frac{q_E^A-q_E^B-s}{2\tau}}^1[v+q_E^B-\tau(1-x)]dx\end{array}\right]+(1-\lambda-\mu)[v+q_E^B-\tau(1-y)]. \tag{4.19}$$

Equating (4.18) and (4.19) implicitly defines the location of the patient who is indifferent between the two hospitals. Using the fact that this patient has $y=D_1^A(q_1^A,q_1^B)$, we solve for first-period demands

$$D_1^A(q_1^A,q_1^B)=\frac{1}{2}+\frac{q_1^A-q_1^B}{2\tau(2-\lambda-\mu)}+\left[\frac{1-\lambda-\mu+\frac{\mu s}{\tau}}{2\tau(2-\lambda-\mu)}\right](q_E^A-q_E^B) \tag{4.20}$$

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

$$\frac{\partial D_1^i(q_1^i,q_1^j)}{\partial q_1^i}=\frac{1}{2\tau(2-\lambda-\mu)}+\left[\frac{1-\lambda-\mu+\frac{\mu s}{\tau}}{2\tau(2-\lambda-\mu)}\right]\frac{\partial(q_E^i-q_E^j)}{\partial q_1^i},\quad i,j=A,B;\quad i\neq j. \tag{4.21}$$

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

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

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

on current demand and hence quality, naïve patients resort to the rule-of-thumb of expecting that quality is the same in both periods.

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

4.5 Patient Expectations and Quality Provision

In this section, we analyse how the different types of patient expectations affect each hospital's incentives for quality provision. We do so by deriving the demand responsiveness to quality, (4.21), under each of our three assumptions regarding patient expectations. We then proceed by performing a ranking of equilibrium quality levels based on these expectations. Notice that patient expectations have no effect on the second-period decisions, which allows us to focus only incentives for quality provision in the first period.

4.5.1 Myopic patients

If patients are myopic and ignore the future, demand responsiveness to quality is the same as it would be if all patients leave the market after the first period (i.e., $\lambda = 1$ and $\mu = 0$), which implies that (4.21) reduces to

$$\frac{\partial D_1^i(q_1^i, q_1^j)}{\partial q_1^i} = \frac{1}{2\tau}, \quad i, j = A, B; \quad i \neq j. \tag{4.22}$$

Thus, with myopic patients, demand responsiveness to quality is the same as in a static version of the model and demand inertia plays no role.[12]

4.5.2 Forward-looking but naïve patients

If patients expect first-period quality to prevail in the second period, this implies that $\partial(q_E^i - q_E^j)/\partial q_1^i = 1$, which in turn implies that (4.21) reduces to

$$\frac{\partial D_1^i(q_1^i, q_1^j)}{\partial q_1^i} = \frac{1}{2\tau}\left[1 + \frac{\mu s}{\tau(2 - \lambda - \mu)}\right], \quad i, j = A, B; \quad i \neq j. \tag{4.23}$$

[12] With myopic patients, although demand inertia plays no role in determining the demand responsiveness to quality, it still plays a role in determining the hospitals' incentives for quality provision, as can be seen from (4.17). The importance of demand inertia for equilibrium quality provision is analysed in section 4.6.

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

However, the presence of switching costs introduces a third effect that makes demand more responsive to quality if patients are forward-looking but naïve. More precisely, the presence of switching costs increases the relative importance of expected quality differences in the future. To illustrate this mechanism, consider the case of a marginal increase in first-period quality by Hospital A with $q_1^A > q_1^B$. While such a quality increase would increase demand for Hospital A, a patient located sufficiently close to Hospital B would still prefer to remain with that hospital, because the lower travelling costs outweigh the foregone quality improvement. However, if such a patient is forward-looking, he anticipates that, with probability μ, his location on the line will not remain the same in the future, but will be randomly drawn from a uniform distribution. Since the expected value of a uniform distribution on $[0, 1]$ is $1/2$, and since the patient expects that first-period quality differences will persist in the second period, he consequently expects that, with probability μ, his preferred choice of hospital in the future will be Hospital A and not Hospital B. However, since $s > 0$ makes it costly to switch from the low-quality to the high-quality hospital in the future, the patient might find it preferable to choose Hospital A already today, and this choice is more likely the higher the switching costs. In other words, when patients are naïve and expect quality differences to persist, the presence of switching costs *increases* demand responsiveness to quality because of patients' fear of being locked-in to the 'wrong' hospital in the future.

4.5.3 Forward-looking and rational patients

If patients have rational expectations, they know that hospitals will set quality according to (4.10) and therefore anticipate that the quality difference in the second period will be

$$q_E^i - q_E^j = (\alpha - c)\phi[2D_1^i(q_1^i, q_1^j) - 1], \quad i, j = A, B; \quad i \neq j, \tag{4.24}$$

which implies

$$\frac{\partial(q_E^i - q_E^j)}{\partial q_1^i} = 2(\alpha - c)\phi \frac{\partial D_1^i(q_1^i, q_1^j)}{\partial q_1^i}, \quad i, j = A, B; \quad i \neq j. \tag{4.25}$$

Inserting (4.25) into (4.21) and solving for $\partial D_1^i(q_1^i, q_1^j)/\partial q_1^i$ yields[13]

$$\frac{\partial D_1^i(q_1^i, q_1^j)}{\partial q_1^i} = \frac{1}{2\tau(2 - \lambda - \mu) - 2(1 - \lambda - \mu + \frac{\mu s}{\tau})(\alpha - c)\phi} \gtrless \frac{1}{2\tau}, \quad i, j = A, B; \quad i \neq j. \tag{4.26}$$

Forward-looking and rational patients not only anticipate that they will be (partially or totally) tied to their first-period hospital but also correctly anticipate how quality investments in the present affect future quality. This implies that the responsiveness of demand to quality in the first period depends on two additional factors, namely *provider motivation* and *technology*. These two factors determine the relationship between demand and the marginal cost of quality provision for each hospital. More specifically, higher demand increases (reduces) the marginal cost of quality provision if $c > (<)\, \alpha$. Under rational expectations, this has important implications for how a change in the current quality difference between hospitals informs patients' beliefs about future quality differences. From (4.25) we see that a unilateral quality increase by Hospital i will increase the expected quality difference between Hospital i and Hospital j in the future only if $\alpha > c$, and *reduce* the expected future quality difference otherwise. Furthermore, since $\partial \left(q_E^i - q_E^j\right)/\partial q_1^i$ is monotonically increasing in α and monotonically decreasing in c, it follows from (4.21) that the demand responsiveness to quality is also monotonically increasing in α and monotonically decreasing in c.

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

[13] Positive demand responsiveness requires that

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

Notice, however, that demand responsiveness with rational patients may be lower than with myopic patients, even in the case where higher demand reduces the marginal cost of quality provision (i.e., $c < \alpha$). In other words, patients may correctly anticipate that a marginal increase in the quality of Hospital i will increase the future quality difference and still be less attracted by that increase than they would if they were myopic and ignored the future. A necessary condition for this to happen is that patients expect that the quality advantage of Hospital i will decrease over time, i.e., that $\partial\left(q_E^i - q_E^j\right)/\partial q_1^i < 1$, which implies that quality becomes relatively less important than travelling/mismatch costs for forward-looking patients.[14]

4.5.4 The effect of patient expectations on equilibrium quality

We are now ready to summarise the effect of patient expectations on equilibrium quality provision. From (4.17), we know that equilibrium quality is increasing in demand responsiveness and that this is the only channel through which patient expectations influence quality provision. Therefore, to establish under which type of expectations quality is higher, it suffices to compare the magnitudes of the demand responsiveness. We have shown that demand is more responsive to quality when patients are forward-looking but naïve than when patients are myopic, implying that quality is higher in the former case.

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

The above analysis is summarised as follows.

Proposition 4.1. *(i) If patients are forward-looking but naïve, equilibrium quality is always higher than if patients are myopic. (ii) Provided that the cost function is sufficiently convex in quality, if patients are*

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

forward-looking and rational, equilibrium quality is

1. *higher than if patients are naïve if*

$$c < c' = \alpha - \frac{2\tau\gamma}{\frac{2(1-\lambda-\mu+\frac{\mu s}{\tau})(2-\lambda-\mu+\frac{\mu s}{\tau})}{(2-\lambda-\mu)} + 3(\lambda+\mu)}; \tag{4.27}$$

2. *lower than if patients are myopic if*

$$c > c'' = \alpha - \frac{2\tau\gamma}{\frac{2(1-\lambda-\mu+\frac{\mu s}{\tau})^2}{(1-\lambda-\mu)} + 3(\lambda+\mu)}; \tag{4.28}$$

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

Proof. Follows directly from a comparison of (4.22), (4.23) and (4.23). A sufficiently high γ ensures that the second-order condition in (4.16) is satisfied for values of c such that the set $\left(\max\left\{c_{\min}, c_R\right\}, c'\right)$ is non-empty. □

4.6 The Effect of Demand Inertia on Quality Provision

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

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

$$\frac{1}{2\tau}[p + (\alpha - c)q^N] + \frac{\alpha - c}{2} = \gamma q^N, \tag{4.29}$$

yielding

$$q^N = \frac{p + (\alpha - c)\tau}{2\tau\gamma - (\alpha - c)}. \tag{4.30}$$

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

Comparing the first-order conditions (4.17) and (4.30), we see that there are two additional effects influencing quality provision in a market with demand inertia. First, there is a *competition effect*, given by the third term in square brackets on the left-hand side of (4.17). Since first-period demand is always valuable in the second period, hospitals have incentives to invest in quality to build market share. All else equal, the competition effect always leads to higher quality. Second, there is a *patient foresight effect* affecting demand responsiveness, which, in turn, determines how effective a quality increase is in attracting demand. In general, the foresight effect may either reinforce or counteract the competition effect, depending on whether patients' expectations about the second period lead to higher or lower demand responsiveness relative to a market without inertia.

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

$$\left[\gamma - (\alpha - c)\frac{\partial D_1^i}{\partial q_1^i}\right](q_1^* - q^N) = \left(\frac{\partial D_1^i}{\partial q_1^i} - \frac{1}{2\tau}\right)[p + (\alpha - c)q^N] + \frac{\phi}{\tau}[\tau\gamma - (\lambda + \mu)(\alpha - c)][p + (\alpha - c)q_2^*]\frac{\partial D_1^i}{\partial q_1^i}. \quad (4.31)$$

Notice from (4.21) that demand responsiveness to quality in a market without inertia is equal to $1/2\tau$. The left-hand side of (4.31) is monotonic in q_1^* and q^N, and the first-period second-order condition ensures that the term in square brackets is positive.[15] Consequently, $q_1^* < q^N$ if the right-hand side of (4.31) is negative, which requires that

$$\frac{\frac{1}{2\tau} - \frac{\partial D_1^i}{\partial q_1^i}}{\frac{\partial D_1^i}{\partial q_1^i}} > \frac{(\phi/\tau)[\tau\gamma - (\lambda + \mu)(\alpha - c)][p + (\alpha - c)q_2^*]}{p + (\alpha - c)q^N}. \quad (4.32)$$

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

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

The above inequality shows that quality is lower than in a market without inertia if the foresight effect (given by the left-hand side) more than compensates for the competition effect (given by the right-hand side), which requires that demand responsiveness is sufficiently lower than in a market without inertia (i.e., sufficiently lower than $1/2\tau$). More specifically, equilibrium quality is lower than in the benchmark if the difference in demand responsiveness—which measures the difference in the effectiveness of a quality increase in attracting patients—exceeds the relative payoff of demand—which measures how beneficial that increase is in future terms.[16]

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

Proposition 4.2. *Under demand inertia, equilibrium quality is lower than in the benchmark case of a market without inertia if the following three conditions are all satisfied:*

(i) patients are forward-looking and rational,

(ii) c *is above a unique threshold in* (α, ∞)*, implicitly defined by*

$$\frac{\frac{1}{2\tau} - \frac{\partial D_1^i}{\partial q_1^i}}{\frac{\partial D_1^i}{\partial q_1^i}} = \frac{(\phi/\tau)[\tau\gamma - (\lambda+\mu)(\alpha-c)][p+(\alpha-c)q_2^*]}{p+(\alpha-c)q^N}, \tag{4.33}$$

where $\partial D_1^i/\partial q_1^i$ *is given by (4.26), and*

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

$$\tau(\lambda+\mu)(\tau(1-\lambda-\mu)-4s\mu)+2s\mu(\tau+s\mu)>0. \tag{4.34}$$

Proof. See Appendix 4.A. □

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

Since the demand responsiveness may fall below $1/2\tau$ only in case of rational expectations, this is a necessary but not sufficient condition for quality to be lower than in the benchmark. According to Proposition 4.2, two more conditions are needed. First, the degree of cost substitutability needs to be sufficiently

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

strong relative to the degree of provider motivation to ensure that the foresight effect is sufficiently strong (cf. Proposition 4.1). To grasp why, recall that only if a first-period unilateral quality increase yields a sufficiently large decrease in the second-period quality difference, will demand responsiveness be low enough. In addition, the demand inertia parameters need to satisfy the condition given by (4.34). It is easily seen that this condition is always satisfied if the switching costs are sufficiently low (i.e., if s is sufficiently close to zero). Notice that, for $c > \alpha$, lower switching costs contribute to reducing both the foresight effect and the competition effect. It reduces the foresight effect because it reduces the cost of being locked-in to the 'wrong' hospital *in the second period*, thus increasing the demand responsiveness to quality in the first period. But it also reduces the competition effect because it weakens the hospitals' ability to lock in patients by offering higher quality in the first period. However, it turns out that the reduction in the competition effect is larger than the reduction in the foresight effect, which explains why the third condition in Proposition 4.2 holds for sufficiently low values of s.

4.7 The Effect of Switching Costs on Quality Provision

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

Implicit differentiation of (4.17) yields

$$\frac{\partial q_1^*}{\partial s} = \frac{\begin{bmatrix} \left(\frac{\lambda+\mu}{\tau}\right)\left(\frac{\tau\gamma}{\lambda+\mu} - \alpha + c\right)\left[p + (\alpha - c)q_2^*\right]\frac{\partial D_1^i}{\partial q_1^i}\frac{\partial \phi}{\partial s} \\ + \left[p + (\alpha - c)q_1^* + \phi\left(\frac{\lambda+\mu}{\tau}\right)\left(\frac{\tau\gamma}{\lambda+\mu} - \alpha + c\right)\left[p + (\alpha - c)q_2^*\right]\right]\frac{\partial^2 D_1^i}{\partial q_1^i \partial s} \end{bmatrix}}{\gamma - (\alpha - c)\frac{\partial D_1^i}{\partial q_1^i}}, \tag{4.35}$$

where

$$\frac{\partial \phi}{\partial s} = \frac{2\mu}{(\lambda + \mu)\left[\frac{2\tau\gamma}{\lambda+\mu} - 3(\alpha - c)\right]} > 0. \tag{4.36}$$

Lower switching costs generally have a twofold effect on quality. First, because fewer patients will be locked-in when switching is less costly, lower switching costs reduce the benefit of a marginal increase in

first-period quality in terms of second-period payoffs. Thus, lower switching costs unambiguously dampen the competition effect, which, all else equal, leads to lower quality. Second, the effect of lower switching costs on demand responsiveness—and hence on the extent to which quality is effective in attracting demand—depends on the type of patient expectations. *A priori*, these two effects may either reinforce or counteract each other; however, it immediately follows that lower switching costs will lead to higher quality only if they make demand sufficiently more elastic.

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

While forward-looking but naïve patients anticipate the lock-in effect of switching costs, they expect quality to remain constant. Since these patients expect a unilateral quality increase to yield a long-lasting quality difference, the less locked-in they anticipate to be, the less attracted they are by such an increase. A lower s implies that 'correcting' the first-period choice of hospital in the second period is less costly, which implies that lower switching costs reduce the relative importance of (present and future) quality differences. In other words, from the perspective of naïve patients, lower switching costs reduce the benefit of being locked-in to the 'right' hospital (cf. subsection 4.5.2). This leads to lower demand responsiveness and reinforces the effect of the weaker incentives to invest in quality in terms of second-period payoffs. Thus, lower switching costs also lead to lower quality when patients are forward-looking but naïve.

When patients have rational expectations, provider motivation and technology again play a role. More specifically, the effect of switching costs on demand responsiveness depends on whether a unilateral quality increase today increases or reduces the quality difference in the future, which in turn depends on the sign of $(\alpha - c)$. Using (4.26), we derive

$$\frac{\partial^2 D_1^i}{\partial q_1^i \partial s} = 4\mu(\alpha - c)\frac{\phi}{\tau}\left(\frac{\partial D_1^i}{\partial q_1^i}\right)^2 \gtrless 0. \tag{4.37}$$

Inserting (4.37) into (4.35) yields

$$\frac{\partial q_1^*}{\partial s} = \frac{\phi}{\tau}\frac{\partial D_1^i}{\partial q_1^i}\left[\frac{\mu}{\gamma - (\alpha - c)\frac{\partial D_1^i}{\partial q_1^i}}\right]\left[\begin{array}{c}\frac{[\tau\gamma - (\lambda+\mu)(\alpha-c)][p+(\alpha-c)q_2^*]}{\tau\left(1-\lambda-\mu+\frac{\mu s}{\tau}\right)} \\ +4(\alpha - c)\left(\gamma q_1^* - \frac{\alpha - c}{2}\right)\end{array}\right] \gtrless 0. \tag{4.38}$$

If a first-period quality increase by Hospital i increases the expected quality difference between Hospital i and Hospital j in the second period (i.e., if $c < \alpha$), lower switching costs reduce demand responsiveness.

The intuition for this result is similar to that of the case of naïve patients. The less locked-in patients anticipate to be, the less attracted they are by a quality difference that carries over into the future, since adjusting their choices in the second period is less costly. Therefore, the two above-mentioned effects go in the same direction, and lower switching costs again lead to lower quality.

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

We summarise the above results in the following proposition.

Proposition 4.3. *Lower switching costs lead to lower quality if patients are myopic or forward-looking but naïve, but lead to higher quality if patients are rational and the degree of cost substitutability between quality and output is sufficiently high.*

Proof. See Appendix 4.B. □

4.8 Discussion and Concluding Remarks

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

Using this analysis, we show how patient expectations shape the responsiveness of demand for hospital care to quality and obtain three main results.

We find that, unless patients are rational and cost substitutability is sufficiently strong, quality provision is generally higher than in the benchmark of a market without inertia and, simultaneously, policies based on switching cost reductions are counterproductive. The co-existence of these two results is intuitive. If demand inertia leads to higher quality provision, weakening it by reducing switching costs is an ill-advised policy intervention. A closer inspection of our results, however, suggests that the link between demand inertia, patient expectations, and quality is not that simple. For some parameter values, demand inertia leads to lower quality *and* lower switching costs are nonetheless counterproductive.[17] In this case, for intermediate degrees of cost substitutability, rational patients' foresight of a reduction in the future quality difference (brought about by a current unilateral quality increase) makes demand responsiveness low enough to induce hospitals to offer lower quality than in a market without inertia. This same future reduction in the quality difference, conversely, does not suffice to persuade patients to take advantage of the present and future quality differences by reversing their choices if switching costs fall, thereby making demand sufficiently more responsive and triggering higher quality provision.

It is our first main result, based on a quality ranking, whose implications are more far-reaching. By ranking quality provision according to the type of patient expectations, we reveal that quality is always higher when patients are naïve than when they are myopic, while the relative position of quality when patients are rational ranges from highest to lowest, depending on the hospitals' technology and motivation. Perhaps surprisingly, these findings are connected to the concept of 'behavioural hazard', defined as the misuse of healthcare and the ensuing welfare losses caused by departures from forward-looking and perfectly rational patient behaviour (Baicker et al., 2015). Such departures are now well documented in the literature (cf. section 4.2), but the evidence on their impact on patient utility is less conclusive. The overall reduction in healthcare utilisation generated by myopic behaviour when compared with fully forward-looking behaviour reported by Guo and Zhang (2019) is concentrated in elective and preventive care, with emergency care showing no response. As for the results of Dalton et al. (2020), whereas there is little difference between fully myopic and fully rational behaviour in terms of quantity, there is a significant change in the composition of drugs consumed. In conjunction, these pieces of evidence suggest that the effect of deviations from perfect rationality on patient utility is generally ambiguous. While we do not study the misuse of healthcare, we do show that different types of patient expectations provide contrasting

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

incentives for hospitals to invest in the quality of care, which, in turn, affects patients' health gains. In the symmetric equilibrium of our model, patient expectations affect aggregate patient utility uniquely through first-period quality. This implies that Proposition 4.1 is also a *ranking of patient utility* according to the type of expectations and, consequently, that full rationality does not necessarily lead to better outcomes for patients.

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

Appendix 4.A Proof of Proposition 4.2

The proof that $q_1^* > q^N$ when patients are myopic or forward-looking but naïve follows directly from equations (4.22), (4.23), and (4.32).

To establish the conditions under which $q_1^* < q^N$ when patients are forward-looking and rational, we

use equations (4.26) and (4.30) to rewrite, after some manipulation, condition (4.32) as

$$\frac{[2\tau\gamma-(\alpha-c)]\left[\tau\gamma-(\lambda+\mu)(\alpha-c)\right]}{1-\lambda-\mu+\frac{\mu s}{\tau}}>\left[2\tau\gamma-(\lambda+\mu)(\alpha-c)\right]\left[\begin{array}{c}\frac{(1-\lambda-\mu)[2\tau\gamma-3(\lambda+\mu)(\alpha-c)]}{2\left(1-\lambda-\mu+\frac{\mu s}{\tau}\right)^{2}}\\-(\alpha-c)\end{array}\right]. \tag{4.39}$$

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

$$\frac{\partial LHS(c)}{\partial c}=\frac{[1+2(\lambda+\mu)]\tau\gamma-2(\lambda+\mu)(\alpha-c)}{1-\lambda-\mu+\frac{\mu s}{\tau}}>0 \tag{4.40}$$

and

$$\frac{\partial RHS(c)}{\partial c}=\frac{(1-\lambda-\mu)(\lambda+\mu)}{\left(1-\lambda-\mu+\frac{\mu s}{\tau}\right)^{2}}[4\tau\gamma-3(\lambda+\mu)(\alpha-c)]+2[\tau\gamma-(\lambda+\mu)(\alpha-c)]>0, \tag{4.41}$$

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

Recall, from condition (4.32), that $LHS(c)<RHS(c)$ may only hold if $\partial D_1^i/\partial q_1^i<1/2\tau$, which, in turn, requires that $c>c''$, with c'' given by equation (4.28) in Proposition 4.1. Then, because $LHS(c)$ and $RHS(c)$ are strictly increasing and convex in c,

$$LHS(c'')-RHS(c'')=\frac{[2\tau\gamma-(\alpha-c'')]\left[\tau\gamma-(\lambda+\mu)(\alpha-c'')\right]}{1-\lambda-\mu+\frac{\mu s}{\tau}}>0 \tag{4.42}$$

and

$$LHS(\alpha)-RHS(\alpha)=2\tau\mu s\left(\frac{\gamma}{1-\lambda-\mu+\frac{\mu s}{\tau}}\right)^{2}>0. \tag{4.43}$$

$LHS(c)<RHS(c)$ may only be true if c exceeds some unique threshold value in (α,∞) *and* $\partial^2 LHS(c)/\partial c^2<\partial^2 RHS(c)/\partial c^2$, which is true if the condition in (4.34) holds. The above mentioned threshold value is the unique solution to $LHS(c)=RHS(c)$ in (c_{min},∞).

Finally, note from (4.39) that this solution is independent of p, as well as that $q_t^*>0$ if p is sufficiently high. Thus, the set of values of c such that $q_1^*<q^N$ is non-empty and the symmetric pure strategy subgame perfect Nash equilibrium is characterised by an interior solution if p is sufficiently high.

Appendix 4.B Proof of Proposition 4.3

The proof that $\partial q_1^*/\partial s>0$ when patients are myopic or forward-looking but naïve follows directly from (4.35), given (4.22), (4.23), and (4.36).

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

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

$$\left[\gamma - (\alpha - c)\frac{\partial D_1^i}{\partial q_1^i}\right](q_1^* - q_2^*) = \left(\frac{\partial D_1^i}{\partial q_1^i} - \frac{\lambda+\mu}{2\tau}\right)[p + (\alpha - c)q_2^*] + \frac{\phi}{\tau}[\tau\gamma - (\lambda+\mu)(\alpha - c)][p + (\alpha - c)q_2^*]\frac{\partial D_1^i}{\partial q_1^i}. \tag{4.44}$$

The left-hand side of (4.44) is monotonic in q_1^* and q_2^*, and the second-period second-order condition ensures that the term in square brackets is positive. Thus, $q_1^* > q_2^*$ if

$$\frac{\lambda+\mu}{2\tau} - \frac{\partial D_1^i}{\partial q_1^i} < (\phi/\tau)[\tau\gamma - (\lambda+\mu)(\alpha - c)]\frac{\partial D_1^i}{\partial q_1^i}. \tag{4.45}$$

The above inequality is clearly satisfied under myopic and naïve patient expectations. Recall that $\partial D_1^i/\partial q_1^i > 1/2\tau$ under these two types of expectations and that the expression on the right-hand side of (4.45) is always positive.

Using (4.26), (4.45) is satisfied under rational expectations if

$$c > \alpha - \frac{2\tau\gamma}{3(\lambda+\mu)}\left[\frac{1 + \left(1 - \lambda - \mu + \frac{\mu s}{\tau}\right) - (\lambda+\mu)(2 - \lambda - \mu)}{1 + \frac{2}{3}\left(1 - \lambda - \mu + \frac{\mu s}{\tau}\right)\left(\lambda + \mu - \frac{\mu s}{\tau}\right) - (\lambda+\mu)(2 - \lambda - \mu)}\right]. \tag{4.46}$$

The term in square brackets is greater than 1, implying that the expression on the right-hand side of (4.46) is below c_{min}. Thus, regardless of the type of patient expectations, $q_1^* > q_2^* \quad \forall \quad c > c_{min} \implies (q_2^* > 0 \implies q_1^* > 0)$. This concludes the proof of (i).

Notice now that, given the first-period second-order condition and that $\partial D_1^i/\partial q_i^1 > 0$ for $c > c_R$, the sign of $\partial q_1^*/\partial s$ is uniquely determined by the sign of the last factor (in square brackets) in (4.38), which we now denote by σ. In addition, note that $\sigma < 0$ only holds for $c > \alpha$, given that, from the first-order condition defining first-period equilibrium quality, $\gamma q_1^* - (\alpha - c)/2 > 0$.

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

$$\lim_{c \to \tilde{c}^-} \sigma = \frac{\gamma p + \left[\frac{(\lambda+\mu)p}{\tau}\right]^2}{1 - \lambda - \mu + \frac{\mu s}{\tau}} - 2\left[\frac{(\lambda+\mu)p}{\tau}\right]^2 - \left[\frac{4(\lambda+\mu)\gamma p}{\tau}\right]q_1^*. \tag{4.47}$$

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

$$\frac{\gamma p+\left[\frac{(\lambda+\mu)p}{\tau}\right]^2}{1-\lambda-\mu+\frac{\mu s}{\tau}}-2\left[\frac{(\lambda+\mu)p}{\tau}\right]^2<0, \tag{4.48}$$

which is true provided that

$$p>\frac{\tau^2\gamma}{(\lambda+\mu)^2\left[2\left(1-\lambda-\mu+\frac{\mu s}{\tau}\right)-1\right]}. \tag{4.49}$$

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

5. Conclusion

The results of chapters 2–4 contribute to the now vast literature on the idiosyncrasies of competition between healthcare providers and the unintended outcomes of regulation in these markets.

We have first shown, in chapter 2, that patient choice policies lead to longer waiting times due to a demand effect and that it is reinforced by the reduced effectiveness of supply as an instrument to avoid tougher waiting time penalties. By making demand more responsive to waiting times, choice-enhancing policies diminish the benefit of increased activity in terms of a waiting time penalty reduction because short waits will attract more patients and thus offset the initial waiting time reduction. This second effect, importantly, stems directly from regulation—were waiting time penalties absent, it would vanish, and, while patient choice policies would still lead to longer waits, their effect would be weaker. As expected, but contrarily to choice policies, penalties work. What is perhaps surprising is that they are more effective in reducing waiting times when designed with a linear than with a convex structure since the latter reflect the imposition of harsher penalties for long waits. This is a result of the dynamic strategic substitutability in supply created by convex penalties and the incentive it gives each hospital to 'free-ride' on a rival's supply increase. Under convex penalties, hospitals will respond to lower treatment supply by the rival with increased supply, reducing their waiting times, diverting demand from the rival, and thus curbing the initial increase in waiting time caused by the supply reduction at the competing hospital. Consequently, convex penalties yield a unilateral incentive to reduce activity, which leads to longer waits.

Chapter 3 inquired into the possible unintended effects of lower switching costs. The majority of these effects relate to the benefit of having patients concentrated at a single hospital when there is strong (weak) cost complementarity (substitutability) between quality and output or provider altruism. In this case, higher volume begets higher quality. Seeking to use the increased ability to adjust granted by lower switching costs, some patients are willing to forgo the quality premium offered by high-volume hospitals to reduce the mismatch between their preferences and the horizontal characteristics of the chosen hospital. This decrease in market concentration, therefore, not only implies that some patients switch to lower-quality

hospitals but also that those who remain at high-volume, higher-quality ones will experience a reduction in the quality of care they receive. Furthermore, such a demand redistribution effect may bring about unintended consequences that are *not* related to cost complementarity or provider altruism and hence quality changes. Even if quality provision increases universally at the hospital level, average quality may nonetheless fall if a sufficiently large number of patients are driven to lower-quality hospitals. This reduction in mismatch costs is conceptually welfare-improving, but the extent to which it is desirable depends on the policymakers' objectives and, in fact, on its real-word interpretation. If, for example, it reflects a better match between a patient's diagnosis and the chosen hospital's specialty mix, then it is possible that the demand adjustment at least partially compensates for the forgone quality.

The least policy-oriented of the three, chapter 4 looked at expectations, whose role in the hospital industry follows from the existence of demand inertia. Accordingly, we have shown that quality provision is higher than in a market without inertia under both myopic and naïve patient expectations, whereas rational expectations are a necessary condition for the opposite result to arise. The chapter nevertheless offered a related policy analysis—only under rational expectations might switching costs reductions be beneficial. These results are driven by the link between strong cost substitutability and expectations, which determine the responsiveness of demand to quality. Only rational patients are aware that a unilateral quality increase in the present foretells a large decrease in the future quality difference when cost substitutability is sufficiently strong. In this case, demand responsiveness is low, and this is why quality may be lower than in a market without inertia. It is also in this case that lower switching costs, by providing rational patients with increased ability to exploit present and future quality differences, increase demand responsiveness and yield higher quality. The chapter's most innovative contribution, however, is the ranking of quality and patient health gains according to the type of expectations. Again, the results are driven by how expectations shape demand responsiveness. Naïve patients expect a quality increase to be long-lasting and are thus more sensitive to quality than their myopic counterparts, triggering higher provision. The responsiveness of demand from rational patients depends on the actual effect of a present quality increase on the future quality difference, which these patients observe and depends on the hospitals' motivation and technology. Unless the future quality difference increases enough, demand from rational patients is the least responsive, and quality the lowest. Additionally, in this case, both myopic and naïve patients are more sensitive to quality than they would be if they correctly anticipated the future quality difference, which prompts hospitals to offer higher quality, leads to larger health gains, and ultimately implies that deviations from perfect rationality may be beneficial. For these patients, ignorance is bliss. Chapter 4

also highlights how idiosyncrasies on one side of the market amplify or mitigate those on the other: the role of hospital technology and motivation is greatly diminished by the two behavioural deviations from rationality—present bias and the imperfect assessment of healthcare attributes—, we model as myopic and naïve patient expectations.

This thesis also offers insights into the pitfalls of regulation. Chapter 2 revealed an inherent conflict between two common policy interventions in markets where waiting times are a concern: waiting time penalties and patient choice policies. We demonstrated that the counterproductive effect of patient choice policies on waiting times is smaller when penalties are convex, whereas waiting time penalties are more effective in reducing waiting times when they are designed with a linear structure. Taken together, chapters 3 and 4 add to this discussion by revealing that conflict may lie in the scope of the *same* policy intervention, as market conditions and time frame matter. Consider a policy aimed at endowing patients with higher mobility within a health system through facilitated switching. In the static analysis of a mature and asymmetric market of chapter 3, lower switching costs increase average quality if cost substitutability is sufficiently strong. In the dynamic analysis of a symmetric market of chapter 4, for equivalently strong cost substitutability, lower switching costs always lead to lower quality provision if patients have myopic or forward-looking but naïve expectations; only under rational expectations would lower switching costs increase quality.

Let us now turn to limitations and the connections thereof to future research. The model presented in chapter 2 has arguably two main limitations—one conceptual and the other analytical—, which may be, as we show below, related. The former is the lack of a clear distinction between *perceived* waiting times, the figure patients use to make their choice of hospital, and the *actual* waiting time they experience, which is obviously only realised ex post and according to which penalties are posteriorly levied. In reality, waiting times indicators available to patients are often based on actual, historical waiting times data (e.g., average or median waiting times from referral to treatment in a previous period). Because these *same* indicators are commonly used by regulators to compute waiting time penalties, the inclusion of a single waiting time variable for each hospital is a reasonable assumption. However, if waiting time indicators are based on data from previous periods, how do current demand and supply affect them and to what extent is this impact reflected by the dynamic constraint? We address these issues below.

To derive the actual waiting times (used to compute the indicators), one must take into account the waiting list, and this is related to the model's analytical limitation: the 'reduced form' modelling of the dynamic equation of waiting times, which posits a positive and linear relationship between changes in

waiting lists and changes in waiting times. Siciliani (2008), conversely, demonstrates that the relationship between waiting lists and the *actual* waiting times is non-linear. While differential games with non-linear dynamic constraints do not have complete analytical solutions—to the point that even simulations of a closed-loop solution are precluded—, it is worth exploring the implications of a non-linear alternative specification both in terms of computation and interpretation. By doing so, we attempt to highlight how future research within this framework is limited. One such alternative is based on 'waiting times of patients treated' (Siciliani, Moran, and Borowitz, 2013), the actual wait experienced by patients computed once they are removed from the list. Suppose that in a given period t, the waiting time indicator of a hospital, which determines the penalty it incurs and is observed by its potential patients, is defined as the exact wait patients experienced from joining the list in period $t - w_i(t) - \hat{t}$ to receiving treatment in period $t-\hat{t}$, where $\hat{t}$ is a constant denoting the lag between waiting time realisation and its inclusion in published indicators. Following Siciliani (2008) and adopting the notation of chapter 2, such a waiting time may be defined as

$$y_i(t - w_i(t) - \hat{t}) = \int_{t-w_i(t)-\hat{t}}^{t-\hat{t}} S_i(h)dh, \tag{5.1}$$

where $y_i(t)$ is the waiting list. Differentiating with respect to time and using $\dot{y}_i(t) = D_i(t) - S(t)$ yields

$$\dot{w}_i(t) = \frac{D_i(t - w_i(t) - \hat{t}) - S_i(t - \hat{t})}{S_i(t - w_i(t) - \hat{t})}. \tag{5.2}$$

The above equation is highly non-linear and introducing it in a differential game would render the analysis intractable. However, it offers an expected yet important insight: increased activity reduces actual waiting times, whereas higher demand increases them. More importantly, it reveals that the primary mechanism in our framework with a linear dynamic constraint—namely, that supply is an instrument to avoid penalties but that the lower waiting times cause a demand increase which offsets the initial waiting time reduction—is similarly present in the exact, non-linear case. Although the linear specification required to make the model analytically tractable fails to capture the lag in the effect of demand and supply, it is nonetheless a good approximation of the dynamics of waiting times (observed and used to impose penalties) in the realistic cases where waiting times indicators are based on the actual waits of patients treated.

The main limitation of the two models presented in chapters 3 and 4 is the absence of a tension between a hospital's incentives to exploit locked-in patients through lower quality and compete aggressively for new patients, a hallmark of switching costs models. In the model of chapter 3; no patient is new; in the second period of the model of chapter 4, there is no game to play afterwards so that new patients may then be exploited; and in the first period of the model of the same chapter, all patients are new. One analytical

approach to introduce that tension is to model hospital competition in a fully dynamic framework, as in Beggs and Klemperer (1992) and To (1996). These two studies assume total lock-in—consumers choose a supplier once new and buy from it in all the subsequent periods they remain in the market—, which precludes the analysis of preference volatility and switching. Villas-Boas (2006) relaxes this assumption in a computationally demanding overlapping-generations model, where the effect of switching costs on prices may only be obtained analytically for the limiting case of marginally positive switching costs. Preliminary results from a fully fledged overlapping-generations version of chapter 4's model where patients live for two periods suggest similar challenges; however, a calibration in the spirit of that of chapter 2 could be used to make the results more salient.

Finally, it is instructive to discuss how the main results of chapters 3 and 4 would change in the presence of the tension between exploiting old patients and attracting new ones. To keep the discussion tractable, consider the above-mentioned overlapping-generations formulation where patients live for two periods. Given its asymmetry, the model of chapter 3 offers a type of off-steady-state analysis, which makes the discussion of the effect of lower switching costs in the same terms more complex. Still, one would expect that, in the presence of a cohort of new patients, a switching costs reduction would also yield the *competition* and *foresight* effects identified in chapter 4. If the latter dominated or if patients were myopic or naïve (as shown in chapter 4), there would exist an additional force driving quality downwards at both hospitals, besides the increase in marginal cost at the high-volume hospital. It would thus be possible that both quality paths would fall and that lower switching costs would more easily decrease (instantaneous) average quality. In the same framework, in the presence of a cohort of old, locked-in patients, steady-state quality under the three types of expectations considered in chapter 4 would be lower, but the mechanisms underlying each type and, crucially, the quality ranking should not change.

References

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

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

Arrow, K. J. (1963). Uncertainty and the welfare economics of medical care. *The American Economic Review*, *53*(5), 941–973.

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

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

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. (2010). Competition and quality in health care markets: A differential-game approach. *Journal of Health Economics*, *29*(4), 508–523. https://doi.org/10.1016/j.jhealeco.2010.05.004

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

https://doi.org/10.1007/978-88-470-5480-6_1

Brekke, K. R., Siciliani, L., and Straume, O. R. (2008). Competition and waiting times in hospital markets. *Journal of Public Economics*, *92*(7), 1607–1628. https://doi.org/10.1016/j.jpubeco.2008.02.003

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

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

Chen, Y., Meinecke, J., and Sivey, P. (2016). A theory of waiting time reporting and quality signaling. *Health Economics*, *25*(11), 1355–1371. https://doi.org/10.1002/hec.3222

Cooper, Z., Gibbons, S., Jones, S., and McGuire, A. (2011). Does hospital competition save lives? Evidence from the English NHS patient choice reforms. *The Economic Journal*, *121*(554), F228–F260. https://doi.org/10.1111/j.1468-0297.2011.02449.x

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

Dockner, E. J., Jorgensen, S., Long, N. V., and Sorger, G. (2000). *Differential games in economics and management science*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511805127

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

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

Gaynor, M., Moreno-Serra, R., and Propper, C. (2013). Death by market power: Reform, competition, and patient outcomes in the National Health Service. *American Economic Journal: Economic Policy*, *5*(4), 134–66. https://doi.org/10.1257/pol.5.4.134

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

Gravelle, H., and Siciliani, L. (2008a). Is waiting-time prioritisation welfare improving? *Health Economics*, *17*(2), 167–184. https://doi.org/10.1002/hec.1262

Gravelle, H., and Siciliani, L. (2008b). Optimal quality, waits and charges in health insurance. *Journal of Health Economics*, *27*(3), 663–674. https://doi.org/10.1016/j.jhealeco.2007.08.004

Gravelle, H., and Siciliani, L. (2008c). Ramsey waits: Allocating public health service resources when there is rationing by waiting. *Journal of Health Economics*, *27*(5), 1143–1154. https://doi.org/10.1016/j.jhealeco.2008.03.004

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

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

Hoel, M., and Sæther, E. M. (2003). Public health care with waiting time: The role of supplementary private health care. *Journal of Health Economics*, *22*(4), 599–616. https://doi.org/10.1016/S0167-6296(03)00007-9

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

Iversen, T. (1993). A theory of hospital waiting lists. *Journal of Health Economics*, *12*(1), 55–71. https://doi.org/10.1016/0167-6296(93)90040-L

Iversen, T. (1997). The effect of a private sector on the waiting time in a national health service. *Journal of Health Economics*, *16*(4), 381–396. https://doi.org/10.1016/S0167-6296(96)00518-8

Iversen, T., and Siciliani, L. (2011). Non-price rationing and waiting times. In S. Glied and P. C. Smith (Eds.), *The Oxford handbook of health economics.* Oxford: Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199238828.013.0027

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

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

March, M., and Schroyen, F. (2005). Can a mixed health care system be desirable on equity grounds? *The Scandinavian Journal of Economics*, *107*(1), 1–23. https://doi.org/10.1111/j.1467-9442.2005.00392.x

Martin, S., and Smith, P. C. (1999). Rationing by waiting lists: An empirical investigation. *Journal of Public Economics*, *71*(1), 141–164. https://doi.org/10.1016/S0047-2727(98)00067-X

Miller, A. R., and Tucker, C. (2014). Health information exchange, system size and information silos. *Journal of Health Economics*, *33*, 28–42. https://doi.org/10.1016/j.jhealeco.2013.10.004

Moscelli, G., Gravelle, H., and Siciliani, L. (2019a). *The effect of hospital choice and competition on waiting times and inequalities in waiting times* (Mimeo). University of York.

Moscelli, G., Gravelle, H., and Siciliani, L. (2019b). *Effects of market structure and patient choice on hospital quality for planned patients* (CHE Research Paper No. 162). Centre for Health Economics, University of York.

Moscelli, G., Gravelle, H., Siciliani, L., and Santos, R. (2018). Heterogeneous effects of patient choice

and hospital competition on mortality. *Social Science & Medicine*, *216*, 50–58. https://doi.org/10.1016/j.socscimed.2018.09.009

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

OECD. (2017). *Health at a glance 2017: OECD indicators*. Paris: OECD Publishing. https://doi.org/10.1787/health_glance-2017-en

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

Propper, C., Burgess, S., and Gossage, D. (2008). Competition and Quality: Evidence From the NHS Internal Market 1991–9. *The Economic Journal*, *118*(525), 138–170. https://doi.org/10.1111/j.1468-0297.2007.02107.x

Propper, C., Sutton, M., Whitnall, C., and Windmeijer, F. (2008). Did 'targets and terror' reduce waiting times in England for hospital care? *The B.E. Journal of Economic Analysis & Policy*, *8*(2). https://doi.org/10.2202/1935-1682.1863

Propper, C., Sutton, M., Whitnall, C., and Windmeijer, F. (2010). Incentives and targets in hospital care: Evidence from a natural experiment. *Journal of Public Economics*, *94*(3), 318–335. https://doi.org/10.1016/j.jpubeco.2010.01.002

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

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

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 & Manage-*

ment Strategy, *25*(3), 539–562. https://doi.org/10.1111/jems.12157

Siciliani, L. (2006). A dynamic model of supply of elective surgery in the presence of waiting times and waiting lists. *Journal of Health Economics*, *25*(5), 891–907. https://doi.org/10.1016/j.jhealeco.2005.12.002

Siciliani, L. (2008). A note on the dynamic interaction between waiting times and waiting lists. *Health Economics*, *17*(5), 639–647. https://doi.org/10.1002/hec.1286

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., Moran, V., and Borowitz, M. (Eds.). (2013). *Waiting time policies in the health sector: What works?* Paris: OECD Health Policy Studies, OECD Publishing. https://doi.org/10.1787/9789264179080-en

Siciliani, L., Moran, V., and Borowitz, M. (2014). Measuring and comparing health care waiting times in OECD countries. *Health Policy*, *118*(3), 292–303. https://doi.org/10.1016/j.healthpol.2014.08.011

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

Simões, J., Augusto, G., and Fronteira, I. (2017). Introduction of freedom of choice for hospital outpatient care in portugal: Implications and results of the 2016 reform. *Health Policy*, *121*(12), 1203–1207. https://doi.org/10.1016/j.healthpol.2017.09.010

Sivey, P. (2012). The effect of waiting time and distance on hospital choice for English cataract patients. *Health Economics*, *21*(4), 444–456. https://doi.org/10.1002/hec.1720

Stavrunova, O., and Yerokhin, O. (2011). An equilibrium model of waiting times for elective surgery in NSW public hospitals. *Economic Record*, *87*(278), 384–398. https://doi.org/10.1111/j.1475-4932.2011.00726.x

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

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

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

Villas-Boas, J. M. (2006). Dynamic competition with experience goods. *Journal of Economics & Management Strategy*, *15*(1), 37–66. https://doi.org/10.1111/j.1530-9134.2006.00091.x

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