Welfare effects of outcome-based prices in the English National Health Service

Enrico Maria Camarda*

Job Market Paper

December 2021

Abstract

Optimal regulation of health care systems aims to incentivize clinical quality while containing cost. Some recent studies have highlighted the potential benefits of reforming the structure of hospital reimbursement tariffs ("prices") to achieve this objective. In this paper, I explore the potential of outcome-based prices, i.e. a new form of price regulation where prices are based on the level of quality of outcomes ("quality-based" prices). Using panel data on hip replacements from England, this paper develops and estimates a structural model of hospital demand and supply of quality which allows for heterogeneity in productivity and non-profit motives. I use this model to evaluate the social welfare effects of quality-based prices vis-à-vis *uniform* prices. I show that a system with quality-based prices that rewards for higher quality can lead to a social welfare improvement of up to 37% of the welfare gap with the social optimum. The improvement is driven by an intensification of competition and a re-allocation of patients and resources to the most efficient providers. However, I show that in presence of increasing marginal costs of quality and multiple hospital characteristics, the role of re-allocation in improving welfare can be limited, while indirect effects due to competition and measurement error in quality have an important impact on welfare.

JEL-code: I1, L5, L2, L3, I3
Key words: Regulation, Health Economics, Quality of outcomes, Supply-side incentives

*KU Leuven, Belgium and FWO, Belgium, enricomaria.camarda@kuleuven.be
Acknowledgements: I would like to thank Sebastian Fleitas, Frank Verboven, Jan De Loecker and Jo Van Biesebroeck for their helpful comments, I am also grateful to Martin Hackmann, Chad Syverson, Jon Kolstad, Ashvin Gandhi for the useful conversations. Most recent version HERE

1 Introduction

Provision of affordable high quality health care is a challenge for many health care systems. Regulated hospital prices can play a part in achieving this objective, but have typically not been designed to reflect quality or its related cost, thus reducing their ability to signal important information. Only recently, the effects of price regulation on quality of care and its mechanisms have started to be explored, for example in Hackmann (2019), Eliason et al. (2018), Eliason et al. (2020), Handel et al. (2021) and Einav et al. (2018). This literature mostly studied the effect *uniform* prices can have on individual *input* decisions, for example on staffing or time in the hospital (number of bed-days). Less attention has been given to the effect on *outcomes* and the importance of heterogeneity across providers. In my paper, I cover this gap by investigating the mechanisms through which regulated prices directly linked to clinical *outcomes* affect total welfare when we consider differences across providers and their competitive interactions.

The focus of my paper is on quality-of-outcomes based prices (or quality-based prices).[1] In such a system, providers of a treatment are paid differently depending on their quality of clinical outcome, i.e. the specific level of clinical improvement coming of their patients. Hospitals can choose their quality and if quality is higher/lower than a certain threshold the price for the surgery will be higher/lower. This type of prices are under-studied: while some attention has been devoted to the effect of *uniform* prices, little consideration has been given to prices linked to outcomes. The difference is that while the former incentivizes all providers, the latter can better target only the best (in outcomes) hospitals. Outcome-based prices have two main channels through which they can improve welfare: i) more incentives can be given to better and typically more productive hospitals (as in Chandra et al. (2016)); ii) other firms would change their quality in response, without receiving higher prices, because quality is a strategic complement. In my paper, I use empirical estimates of cost and preferences to investigate the relative size of these welfare effects for both rewards and punishments linked to quality. This analysis highlights their potential appeal to policy makers who are discussing about introducing quality-based prices in the UK and other countries.

The setting of my study is hip replacement surgeries performed by National Health Service ("NHS") hospitals in England. They consist in substituting a damaged hip joint with an artificial one. The choice of this setting is motivated by some appealing characteristics. Hip replacements are common, non-urgent procedures for which detailed data is available on cost and clinical outcomes. The English NHS is a prime example of a managed competition system where hospitals compete with each other for patients. Prices are set by the government, patients choose where to receive treatment and hospitals, which are not-for-profit organizations, compete in quality to attract patients. Crucially, it is the government that pays for the surgeries and the price is *uniform* and not linked to patients

[1] From now on in the rest of the paper I will use "quality-based prices" to indicate this type of prices that vary with the level of clinical outcomes after surgery. Unless otherwise specified I will consider prices that vary non-linearly above or below a quality threshold.

clinical outcomes.[2] The welfare effects of a change to this system allowing prices to vary by clinical outcomes is the focus of my study.

To quantify these effects, I develop and estimate a model of partial equilibrium with demand and supply decisions in a health care market. For the demand side, following Berry (1994) I estimate a model of patients' preferences where patients trade off waiting time for quality.[3] For the supply side, I develop a model where hospitals choose to produce different levels of quantity and quality given their own level of productivity and the quality chosen by the other hospitals. Using surgery-level cost data, I estimate the cost function corresponding to these production decisions. Following Arellano and Bond (1991) I estimate marginal costs, marginal costs of quality and productivity.

Combining these estimates I retrieve the hospitals non-profit motives. I back them out from their first order conditions in quality, assuming Bertrand-Nash competition in quality (as in Fan, 2013).[4] They are the wedge between marginal costs, calculated using estimates of the cost function, and marginal revenue of quality, calculated using estimates of patients' preferences. With all the retrieved parameters, I can then fully characterize hospitals and social planner objective functions (which include the cost of raising public funds and externalities[5]).

Then, I use the model to simulate changes in price regulation involving quality-based prices and uniform prices. I perform the following counterfactuals: i) assessing the welfare loss from the current uniform prices, ii) comparing the welfare impact of uniform prices and quality-based prices and iii) assessing the relative importance of re-allocation, competition and measurement error in quality on the effect of quality-based prices on welfare. I validate some of the results of my counterfactuals with an analysis of a pilot reform introducing quality-based punishments in hip replacement in the NHS.

My first finding, analyzing this health care market, is that marginal costs are increasing in quantity and in quality. The fact that quality is costly is an important feature to justify a financial reward for quality. On average, I find a marginal cost of around 5,000GBP (the cost of one additional patient) and an average marginal cost of quality of around 300GBP per patient (the average cost of one additional unit of quality divided by the average number of patients across the hospitals).[6] These are averages across hospitals, the marginal costs just described increase in the number of surgeries and quality offered, respectively. I also

[2] A pilot reform to link prices to clinical outcomes has been implemented for hip replacements and I discuss this in Appendix I.

[3] Waiting time is not endogeneized in the model. While this is a simplifying modeling choice it is supported by two considerations. The waiting time is for all of the orthopedics department and I assume that changes in hip replacements volumes would not affect waiting times in orthopedics, especially given the relatively small changes in prices I consider in my counterfactuals. Additionally, waiting times are unlikely to be affected by small changes in demand as they do not work as traditional queues: they match the needs of hospitals and patients, additional patients can be accommodated in between patients.

[4] The non-profit motives enter as a linear function of quality in the utility function of the hospitals.

[5] The estimates for these two components are taken from other studies. Externalities consist in the loss imposed on family members and society for a poor quality procedure that limits patients in their self-reliance.

[6] These estimates are in line with expectations emerged in talks with industry experts. Note that the regulated price is around £5,500 for one hip replacement.

find a significant heterogeneity in productivity across hospitals. Some hospitals are almost three times more productive than their least productive counterparts. Such a dispersion highlights the importance of allocating production to the most productive providers.

My second finding is that there are important non-profit motives influencing hospital quality decisions. This limits the impact of quality-based prices, contrary to a simpler setting where hospitals are profit maximizers. In fact, in the analysis of the welfare loss from current uniform prices, I find that non-profit motives, not only lead to a higher average level of quality, but also partially compensate for market power. The non-profit motives are approximately 200GBP per patient.

My third and main finding is that quality-based prices can outperform higher uniform prices in their welfare effects. Higher uniform prices increase the marginal revenues of the hospitals leading to higher quality, but this comes at the cost of increasing government expenditure and prices for all hospitals. This is why they lead a smaller positive effect on welfare. Quality-based rewards, instead, target a smaller number of hospitals and generate a larger welfare improvement (with *optimal* quality-based prices 37% of the gap between the welfare achieved with current prices and optimal welfare). They achieve this through re-allocation and indirectly by intensifying competition and exploiting the presence of measurement error in quality.

Three main channels are at play. First, only some performers are rewarded by the government and incentivized to increase quality. The funds directed at them lead to relatively cheap higher quality, because these hospitals are typically the most productive. By reason of the increase in quality they attract more patients and in this sense quality-based prices lead to a re-allocation. Second, this increase in quality leads to more intense competition: in my model quality is a strategic complement and lower quality/lower productivity competitors react by also increasing quality even without being rewarded. Third, given that the regulator measures quality with an error, some hospitals are incentivized to increase quality as the error creates uncertainty and gives them a positive probability of receiving a reward, even if they may not receive a reward in the end. I call these last two channels "indirect channels" given that they lead to an increase in quality also from the “un-rewarded” hospitals.

Finally, in my counterfactuals I assess the relative importance of these three channels comparing results across different local markets and with and without allowing for uncertainty. My analysis is limited by the fact that the hospitals in different markets have different characteristics and cost structures. Nevertheless, I can find indications that the three channels have similar importance in determining the welfare effects of quality-based prices. First, to understand the importance of the measurement error I perform my counterfactuals with and without measurement error in quality. Second, to understand the importance of competition I compare the results of my counterfactuals in local monopolies *vis-à-vis* competitive markets. Third, I can learn from studying the impact of quality-based punishments, where prices are below a baseline price for lower quality levels. In this

case the indirect channels have an opposite effect to the re-allocation one. While the last one has still a positive effect on welfare, the indirect channels lead to a decrease in quality without generating government savings. The net effect on welfare in null or negative which confirms again the importance of the indirect channels *vis-à-vis* the re-allocation channel.

Exploring the importance of the re-allocation channel is particularly interesting in a context with increasing marginal costs and horizontal differentiation (due to the fact that hospitals have other characteristics beyond quality). These factors limit the benefits from re-allocating to the most productive hospitals. When quality and quantity increase, marginal costs also increase, which means that the most productive hospitals are not necessarily the ones with the lowest marginal costs. There is an un-coupling between marginal costs and productivity. Productivity determines the slope and intercepts of marginal cost curves, but having the highest productivity does not mean having the lowest marginal costs. For this reason, rewards may lead to large increases in cost as quality increases, making large quality increases and induced patients re-allocation not-optimal even when the best are the most productive. Additionally, when the best hospitals do not capture the entire market, they must perform worse in other dimensions, therefore the net benefit of a switch to these hospitals would be limited by losses in these other dimensions.

Relation to the literature: The main contribution of this paper is to assess the welfare impact of "quality of outcomes"-based regulated prices. I show that prices linked to quality-of-outcomes can improve welfare. My work is related to few papers that studied the effect of price regulation on decisions related to quality: Hackmann (2019), Eliason et al. (2020), Eliason et al. (2018), Einav et al. (2018) and Handel et al. (2021).[7] In particular, while Hackmann (2019) focused on the effects on welfare of a uniform price increase, I compare the effect of a uniform price increase to the effect of quality-based prices. Additionally, I extend his framework because I use a measure of quality of outcomes rather than a measure of quality of process (number of nurses per patient). Interestingly, he shows that a uniform increase in regulated prices can incentivize quality more efficiently than increasing the number of competitors. In my study, I show that quality-based prices can be even more efficient than a uniform price increase and can have different re-allocative effects across hospitals. I find that quality-based rewards can improve welfare through the three channels of re-allocation, competition and measurement error, even in presence of market power.

A second contribution is in the re-allocation literature. My study is related to the seminal papers of Olley and Pakes (1996) as well as Foster et al. (2008). More recently, Chandra et al. (2016) have studied the re-allocation to the best providers in the health care sector in the U.S.. They found that the best hospitals grow over time and happen to also be typically the most productive. In this paper, I study how quality-based prices can be a tool in fostering this type of re-allocation. While in my analysis I also find that the

[7] The paper is also related to Drenove (2003) that studies the effect of report cards in hospital care. Additionally, there is a vibrant literature on the effect of competition and choice in the English NHS, for example: Gaynor et al. (2016), Santos et al. (2017), Gutacker et al. (2016).

best hospitals tend to be the most productive, I directly analyze the welfare impact from re-allocation. In line with the literature, I find that re-allocation to the most productive has a positive effect on welfare. However, contrary to a traditional setting, I also find that this effect can be severely limited in presence of increasing marginal costs and horizontal differentiation.

Methodologically, contrary to previous studies, I am able to separately identify higher costs due to inefficiency and higher cost due to quality. This is because, in my paper, quality is defined as an improvement in clinical outcomes. While other measures, like number of nurses/doctors per patient, are typically used in health care analysis as measures of quality, most of these metrics could be confused with measures of productivity. For example, number of nurses per patients could be seen as number of workers per level of output. I address this problem by estimating a cost function where the output consists of quantity and quality of outcomes. My supply analysis is related to Grieco and McDevitt (2016) who use the methodology of Ackerberg et al. (2016) ("ACF"), from the productivity literature. They are the only paper to estimate a production function of a bundle of quantity and quality of outcomes in a health care setting. Compared to them, I applied methods from the productivity literature addressing additional issues of serially correlated measurement error in quality and input price differentials. To the best of my knowledge, it is also the first study to apply this methodology to a cost function.

My final contribution is in the identification of hospital non-profit motives and the productivity levels of the different hospitals. In particular, using demand estimates, cost estimates and assuming competition in quality I can determine the level of non-profit motives separately from cost and productivity. This extends previous studies, in particular Gaynor and Vogt (2003) and Hackmann (2019). My study is the only one with Hackmann (2019) to estimate them separately. I do so by combining demand and production estimation approaches (similarly to De Loecker and Scott, 2016) in a context of publicly funded hospitals where prices are not chosen by hospitals. My findings indicate that, in my analysis, non-profit motives are important in determining the level of quality provided and they are larger for hospitals with higher level of market power. In this light, I find that not accounting for them would lead to overstating the role of price changes in improving welfare.

The rest of the paper is structured as follows. Section II briefly describes the institutional background, introducing the National Health Service and the market for hip replacement in England. Section III provides a modeling framework to understand hospitals decision and the social optimum. Section IV includes the partial equilibrium model of demand and supply and its estimation. In particular, subsection IV.1 describes the data used in this study. Subsections IV.2 and IV.3 specify the cost and demand model to be estimated and discusses the empirical results. Section V describes the counterfactual exercises. Section VI discusses some caveats for implementation beyond the context of this paper. Finally, I conclude in Section VII.

2 Institutional background

The context of my study is hip replacement in England. I cover below few features of the English health system to understand the rules and the incentives governing the decisions of hospitals. Most health care in England is financed through taxes and is free at the point of use. Briefly, 90%[8] of care services are paid by the public National Health Service and hospitals receive payments for their services in the form of regulated prices for each service they provide. They are incentivized to provide better quality as they have to compete with other hospitals to attract patients.

Competitive environment. In England the vast majority of treatments and surgeries are paid by the National Health Service ("NHS"). The English NHS covers 55 million people and is a leading example of managed competition. The NHS Trusts (that I will call "hospitals" in the rest of the paper) are public non-profit organizations that provide hospital care and are directly linked to the NHS.[9] In this system patients are free to choose to receive treatment where they want and care is completely free at the point of delivery -patients pay no co-payments to receive care. Doctors are salaried employees who generally work only for one hospital. It is local health authorities (Clinical Commissioning Groups or "CCGs") that purchase care services from hospitals on behalf of patients using as funding transfers that CCGs receive from the central government. In particular, CCGs pay a fixed regulated price for each treatment.

These prices are set nationally and are the same for all hospitals up a to a small local adjustment. They are based on the national average of average costs for each treatment, with the intention of incentivizing efficiency through a system of yardstick competition (Shleifer (1985)). In fact, the administration of each hospital is expected to balance the book and the success of hospitals CEO is closely linked to financial performance (Bloom *et al.* (2015)). In order to achieve this objective, hospitals need to deliver care with an average cost per treatment at least in line with the national average. They can control costs by minimizing waste or reduce quality.

In the light of these incentives, competition across hospitals plays an important role in safeguarding quality. While hospitals may want to "cut corners" to increase their margins, they have to provide quality services to attract patients and revenues. The effect of competition on quality in the English NHS has been documented by Gaynor *et al.* (2016) where the authors showed that patients care about quality and quality was higher in less concentrated markets. However, there is no price differentiation due to quality. This fea-

[8] Around 80% in the case of hip replacements.

[9] The role of private health care is very limited. Some private hospitals provide treatments and are paid by the NHS and a smaller number of hospitals offer care exclusively to patients with private insurance. However, they will not be object of study in this paper, because as highlight by Kelly & Stoye (2016) and by a 2014 UK Competition Authority report the privately-funded market for hip replacement is separate from the publicly-funded one. Private health care account for about 9% (2017 data) of the total health care expenditure covering an even smaller number of patients. In the case of hip replacement privately funded surgery are slighly more important representing up to 20% of the surgeries. Private hospitals that provide NHS treatment will be included in my analysis as outside option for patients in line with Kelly & Stoye (2016).

ture of price regulation in health care markets make them different from other markets and may cause a welfare loss. Higher value for patients and related higher costs are not reflected in prices.

Hip replacement. The specific context of my study is hip replacements in England. These procedures consist in surgically replacing damaged or diseased hip joints with artificial ones. This is a common surgery that many patients over 60 have to undergo. They are typically not emergency surgery for which patients can wait and shop around for where they prefer to receive care. In the last decade the NHS has started a special data collection program to measure and record the quality of clinical outcomes of patients specifically receiving hip and knee replacement.

This data has been used by the NHS to actually implement for the first time a form of quality-of-outcomes based prices, even if in a limited way. Starting from 2014 hospitals having their measure of quality three standard deviations below the mean quality would receive 10% less than the normal regulated price. The focus for my demand and cost function estimation is primary hip replacements, which is the first hip replacement a patient receives, contrary to revisions which consist of the second or third surgery for the same joint. For reference, the NHS spends around 250-300 million pounds a year for primary hip replacements. In Appendix C I discuss further details about the procedure and information collected from interview with doctors and documents from the NHS.

3 Quality choice and its determinants

I specify a model of quality competition where hospitals choose the quality of hip replacements. The quantity of surgeries is instead indirectly determined by the number of patients attracted by the hospitals quality, given that hospitals cannot turn away patients.

3.1 Hospital choice

Following the modelling framework in (Gaynor, 2006) or (Gaynor *et al.*, 2014) I consider a simple framework where hospitals maximize utility and compete in quality z_j with each other. Hospital utility is modeled as the sum of expected profits π_j and non-profit motives $v_j(z_j)$. Note that the time subscript is omitted in the following sub-sections for simplicity of notation.

Hospitals decide[10] their quality target z_j (simply "quality" hereafter), given fixed regulated prices $\bar{p}$, quality of competitors z_{-j} and potential market of size M. They deliver quantity q_j which is not chosen directly but depends on own quality z_j (targets) and the quality of the competitors z_{-j} (hospitals compete Nash-in-quality as in Fan (2013)). With regard to $v_j(z_j)$, I make three main assumptions. (i) More quality increases the non-profit

[10] In this paper I abstract away from complementaries across surgeries in hospital decisions.

motives of the hospital, $\frac{\partial v_j}{\partial z_j} > 0$. (ii) Hospitals internalize only the benefits given to their own patients $\frac{\partial v_j}{\partial z_{-j}} = 0$. (iii) Finally, hospital utility can be measured in British Pounds.

$$\max_{z_j} \quad U_j = \pi_j(z_j) + v_j(z_j), \tag{1}$$

$$\text{where} \quad \pi_j(z_j) = \bar{p} q_j(z_j, z_{-j}) - c(z_j, q_j) \quad \text{and} \quad q_j(z_j, z_{-j}) = s_j(z_j, z_{-j}) M.$$

The first order condition with respect to z_j for each hospital j is:[11]

$$\underbrace{\bar{p}\left\{\frac{\partial s_j}{\partial z_j} M\right\} + \frac{\partial v_j}{\partial z_j}}_{MVQ \quad or \quad \text{“}MR_z\text{”}} = \underbrace{\frac{\partial C_j}{\partial z_j} + \frac{\partial C_j}{\partial q_j}\left\{\frac{\partial s_j}{\partial z_j} M\right\}}_{MC_z}. \tag{2}$$

In Appendix D I show graphically the marginal revenue of quality ("MR_z") or marginal value of quality ("MVQ") and the marginal cost of quality MC_z. Assuming the cost of quality is convex in z_j then, its marginal cost is increasing in z_j, the intuition is that as a hospital increases quality it becomes more difficult to increase it. The MVQ instead is decreasing, because even if hospitals receive the same regulated price for each surgery, the ability to attract patients naturally declines as quality increases, because there would be simply no more patients to attract.[12] The quality chosen by the hospital j is then given by the intersection of the two curves.

3.2 Social Planner choice

A social planner maximizes total welfare setting directly the quality of each hospital.[13] Total welfare is composed by five components: consumer surplus (CS), producer surplus ($\pi_j(z_j) + v_j(z_j)$), minus government expenditure (GovExp) plus positive externalities ($\Psi_{Ext}(z_j)$) arising from treatment. These externalities materialize because a person who has full mobility benefits society and family members. She can work more (and pay more taxes) and would not require the additional attention from family members that would be required with limited mobility. I assume that the social planner knows patients preferences and marginal costs and can directly set qualities of the different hospitals. I also assume

[11] I am limiting my analysis to interior solutions assuming that services closures and null or negative quality are not possible.

[12] The same point can be captured considering distance: some patients may live too far to be attracted by any level of quality.

[13] I am abstracting here from optimal entry exit decisions because political constraints would prevent the NHS to freely opening and shutting down hospitals. The analysis is made here based on the existing network of hospitals.

that the social planner cannot force patients to go to a specific hospital (this is similar to the assumption in Decarolis, Polyakova and Ryan (2020)).

$$max_{z_1,\ldots,z_j} W = \underbrace{CS}_{\substack{\text{Consumer}\\ \text{Surplus}}} + \underbrace{\sum_{j=1}^{J} \Psi_{Ext}(z_j)}_{\text{Externalities}} + \underbrace{\sum_{j=1}^{J} \bar{p} q_j}_{\text{Revenues}} + \underbrace{\sum_{j=1}^{J} v_j(z_j)}_{\text{Non-profit motives}} - \underbrace{\sum_{j=1}^{J} C(z_j, q_j(z_j, z_{-j}))}_{\text{Costs of all firms in market}} - \underbrace{(1+\lambda)\sum_{j=1}^{J} \bar{p} q_j}_{\substack{\text{Government Expenditure}\\ \lambda = \text{distortionary cost of taxes}}}$$

The corresponding FOC's w.r.t z_j is:

$$\underbrace{\frac{\partial CS}{\partial z_j}}_{\Delta CS} + \underbrace{\frac{\partial \Psi_{Ext}}{\partial z_j}}_{\Delta Externalities} + \underbrace{\frac{\partial v_j}{\partial z_j}}_{\Delta \text{ non-profit motives}} =$$

$$= \underbrace{\frac{\partial C_j}{\partial z_j} + \frac{\partial C_j}{\partial q_j}\left\{\frac{\partial q_j(z_j, z_{-j})}{\partial z_j}\right\}}_{MC_z\uparrow \text{ production hosp. j}} + \underbrace{\sum_{-j} \frac{\partial C_{-j}}{\partial q_{-j}}\left\{\frac{\partial q_{-j}(z_j, z_{-j})}{\partial z_j}\right\}}_{\downarrow \text{ production hosp. -j}} \underbrace{+\lambda\left(\bar{p}\frac{\partial q_j(z_j, z_{-j})}{\partial z_j} + \sum_{-j} \bar{p}\frac{\partial q_{-j}(z_j, z_{-j})}{\partial z_j}\right)}_{\text{Distortionary tax effect (for net additional expenditure)}}.$$

The optimal social choice characterized by the FOC's describes a trade-off. Increasing quality of a hospital has a positive effect on the utility of patients, but it also leads to an increase in government expenditure and in the cost of that hospital. Costs increase both for the additional quality provided and the additional patients who would not otherwise have received surgery in the NHS hospitals.[14] Finally, the social planner should consider positive externalities arising from treatment, for example a well treated patient will be less likely to require more treatment, special support and will be more likely to engage in some job. I discuss in detail the quantification of these externalities in Section 6.

By comparing the first order conditions of the hospitals and the social planner it is possible to individuate why welfare losses can arise. On the side of the social benefits, beyond the presence of the externalities, hospitals marginal revenues are different from marginal consumer surplus and are likely to be smaller in concentrated markets. Secondly, the hospitals do not internalize the fact that the costs of the industry could be lower with different production configurations across providers.

3.3 The role of heterogeneity

Hospitals are different, they have different marginal revenues, marginal costs and non-profit motives.[15] The social planner takes into consideration this heterogeneity and prefers that patients receive treatment at the hospital (in the local market) with the lower costs without lowering the utility of the patients. The social planner trades off the additional

[14] These patients are considered in the logit model as choosing the outside good

[15] Additionally, they can have also different characteristics not included in this simplified version of the model, for example waiting time -which are included in the full model and considered as exogenously set.

utility for patients and hospitals (and externalities) from being treated at a hospital, with the additional costs attached to being treated at a certain hospital. In particular, the social planner increases or decreases the quality of hospitals to steer patients to different providers up to the point of optimality.

Given that the government uses a single fixed regulated price it cannot adjust for differences across providers. It cannot give different incentives to each hospitals even if hospitals are different from each other both in terms of costs, non-profit motives and quality. These differences, however, may play a role in determining the effect of fixed uniform prices. For example, hospitals with higher non-profits motives would provide higher quality than hospitals with lower ones *ceteris paribus*.

Quality-based prices allow to differentiate across providers in a way that uniform prices cannot. Exploiting their potential, however, is not straight forward. Rewarding higher quality hospitals can be appealing if they have lower marginal costs and lead to high levels of utility for patients and hospitals. However, higher quality providers may not be necessarily be the most productive ones because other drivers may be responsible for the higher quality, e.g. non-profit motives, size of the market or level of competition.[16] In fact, in the Appendix A I show that quality-based prices can be welfare enhancing depending on the value of a series of structural parameters determining preferences, costs an non-profit motives.

4 Modelling and estimation

Motivated by the results of the analysis of the effect of the pilot reform I now proceed with the modelling and estimation of demand and cost which allows me to perform counterfactual experiments.

From theory to estimation. Different elements in the first order condition with respect to quality z_j (2) that I described in Section 3 are primitives of my model and are object of my estimation. In particular:

Demand: $\frac{\partial s_j}{\partial z_j}$. This term is estimated using a Berry (1994) approach and using a geographical market definition that individuates several local markets for hip replacement.

Cost: $\frac{\partial C_j}{\partial q_j}, \frac{\partial C_j}{\partial z_j}$. These terms are estimated separately using a cost function estimation technique and detailed cost data.

Non-profit motives: $\frac{\partial v_j}{\partial z_j}$. I parametrize $v_j(z_j) = \mu_j \alpha_z^d z_j$ (where α_z^d is patient marginal utility from quality from demand estimation) and retrieve $\frac{\partial v_j}{\partial z_j} = \mu_j \alpha_z^d$ from the FOCs. Given that I have estimated all other elements present in the FOC I can simply determine this term as the difference between MR_z and MC_z. Additionally, μ_j could be seen as a sort of conduct parameter that represents the departure from simple profit maximizing

[16] Also if other hospital characteristics increase sufficiently the utility of patients to make up for the higher (inefficient) cost.

behavior.

4.1 Data

In this section I describe the data that I used in my study. In my demand analysis I use number of surgeries per hospital and hospital characteristics, like quality of outcomes and waiting time. In my cost function analysis I use data on costs quantity, quality by hospital as well as hospital and area characteristics, for example the degree of architectural barriers in different local communities. Quality of outcomes is of particular importance and it's captured by an indicator based on patients surveys.

Quality of outcomes data. The NHS has collected measures of quality of outcomes for hip and knee replacements (in England). The unique feature of this data is that it is one of the first and more complete sources about patients outcomes for a surgery beyond simple mortality rates. While most of the studies in health care markets use mortality as an inverse measure of quality, most surgeries are not life threatening and it is important to study quality in these cases.

The quality measure used in this study are based on patients' surveys, so called Patients Reported Outcome Measures ("PROMs") for hip replacements. This indicator captures outcomes in the form of clinical improvement after hip replacements for each hospital and for every year. This data is based on clinical improvements after the operations self-assessed by patients using questionnaires. To address the differences in patient characteristics that may affect the comparability of the results across hospitals, the NHS adjusts the results based on patients characteristics. The disadvantage of this adjustment procedure is that, to implement the procedure, the NHS does not consider hospitals with less than 30 participants to the surveys. There is more than one quality measure based on different surveys, in particular the ones used in this paper are Oxford hip score and EQ VAS.

Crucially, the measure obtained from this process is affected by measurement error. Hospitals can make mistakes in collecting the data and asking patients questions. Furthermore, if a hospital uses a low number of surveys will have a more uncertain measure of quality, contrary to a hospital that uses a large number of surveys, because the errors would compensate across many patients. Hospitals tend to have similar procedures in collecting data from the surveys and similar response rates over the years, creating a serial correlation in the measurement error. Interestingly, both the response rate of patients and collection procedures are different for the two measures considered in this study (Oxford hip score and the EQ VAS). Taking into account of this fact is important in my analysis and may help shed lights on how to deal with this type of data in future studies.

Costs data series. Cost and quantity data by procedure are captured by the Reference Cost Data collection. The data is available at a very disaggregated level for all hospitals in the country. In fact, it is possible to observe quantity and cost data by surgery, depart-

ment and level of complexity at an yearly frequency.[17][18] This level of completeness and disaggregation makes this dataset quite unique as in most jurisdictions this type of data may be available only for a subset of hospitals in a country. I used this data for the period going from 2012-13 to 2018-19.[19]

The cost data includes all fixed (in an accounting sense) and variable costs that were used to produce all the procedures classified under different types of surgeries.[20] In the case of hip replacement, it means the cost for the number of full-time equivalents doctors and nurses involved in the surgeries, as well as the prosthesis, the anesthetics, the use of the operating theater and the cost of the hospital bed days. For the inputs shared across multiple treatments or used over multiple years, accounting assumptions are made to assign portions of the costs for those inputs.

This is the first study that uses this data at surgery level and analyzes costs over multiple years. In particular, this was made possible by my work at the former English regulator Monitor during a secondment as consultant. Comparability of costs over time was not obviously given changes in coding procedures especially for the period before 2012-13.

The data shows that even for the same procedure at the same level of complexity, in the same department there is a large variation in average cost per procedure, which may also indicate differences in efficiency. This serves as motivation to take differences in efficiency into account, even if this is only a partial indication, because, for example, quality is not taken into account.[21]

Additional data sources. Apart from the data sources already described, I used several others source of data on hospital characteristics. I exploited data from the National Joint Register, a dataset that contains an account of all hip replacements paid by the NHS, including those operations performed at private hospitals. This data is used in demand estimation. I also used patient surveys on the experiences of patients in different hospitals as well as surveys of staff in all hospitals with regard to working conditions and whether they would recommend their hospital as a good place for work or for care. In my cost estimation, I also used demographic data of the areas around the hospitals, including age and gender of the population, level of deprivation of the elderly and presence of architectural barriers.

[17] I focus on elective procedures in orthopedic departments only, even if orthopedic surgeons may still perform the surgery in general surgery and then patients are kept in general surgery beds within a hospital. The operations performed outside the orthopedic department represent only a small fraction of the total number of surgeries and I exclude them from my analysis.

[18] Using the UK financial year April to March.

[19] The period of analysis for my study is limited to 2012/13 to 2018/19 because a different surgery categorization was used in the previous period and the quality of the data collection was also lower in the previous years.

[20] The inclusion of all costs is based on the fact that regulated prices are based on the national average of the unit costs per treatment/surgery. Roughly, prices are equal to the national average plus an inflation factor, minus an efficiency fact. Note price is around £6,000 for one hip replacement.

[21] In fact, the same cost data is used by the NHS to estimate the efficient frontier of the NHS hospitals, even if "productivity" improvements are narrowly defined with respect to physical output.

Table 1: Summary statistics

	2012/13	**2013/14**	**2014/15**	**2015/16**	**2016/17**	**2017/18**
Quality (Oxford measure)						
Average	20.86	20.94	21.11	21.25	21.44	21.94
Median	20.92	21.16	21.19	21.36	21.57	21.98
Standard deviation	1.11	1.09	1.08	1.18	1.19	1.19
No. of hospitals	136	135	131	132	131	130
Total cost (British Pounds)						
Average	863,978	895,001	770,097	794,217	754,580	691,594
Median	760,174	732,613	635,944	689,317	641,888	567,229
Standard deviation	792,191	872,800	745,816	715,937	700,336	696,244
No. of procedures						
Average	214	225	244	245	242	214
Median	170	182	213	217	204	182
Standard deviation	216	217	221	212	215	206
No. of hospitals	149	147	144	141	142	139
Average unit cost (British Pounds)						
Average	6,226.54	6,087.22	6,193.17	6,330.46	6,407.71	6,591.41
Median	6,073.27	5,749.88	6,044.27	6,096.70	6,071.90	6,402.15
Standard deviation	1,798.80	2,401.61	2,127.21	1,952.44	3,108.37	1,975.98

Note: Cost classifications changed slightly from 2012/13 to 2013/14

4.2 Local demand for hip replacements

Patients (indicated with the index p below) in the markets under study are considered to be maximizing the following utility function when they choose the hospital where to undergo primary hip replacement (where u_{jtp} is the utility of the patient, $\widetilde{z_{jt}}$ is the observed quality of outcomes, $wait_{jt}$ is waiting time for an orthopedic surgery, x'_{jt} and ν_{jt} are respectively observed and unobserved characteristics of the hospital and ϵ^d_{jtp} is i.i.d. EV1):

$$u_{jtp} = \delta_{jt} + \epsilon_{jtp} = \alpha^d_z \widetilde{z_{jt}} + x'_{jt}\beta - \alpha^d_w wait_{jt} + \nu_{jt} + \epsilon^d_{jtp}.$$

I restricted the set of hospitals patients can choose based on a geographic definition of the markets (13 local markets). Additionally, in my study I consider publicly funded private care to be the outside option for patients. These are private providers that accept payment from the NHS to perform surgeries.[22]

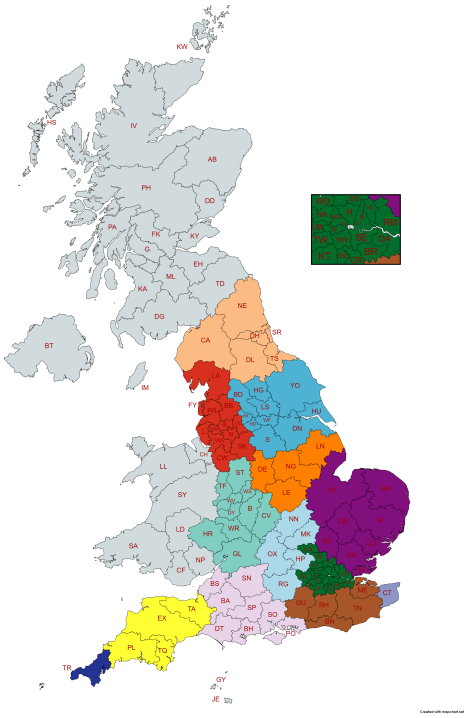

Figure 1: Local markets considered

[22]This arrangement between private hospitals and the NHS was established few years before the start of my sample because waiting times were too long for public hospitals as highlighted in Kelly and Stoye (2016). The introduction of this agreement between private hospitals and the NHS led to a market expansion, as described in Kelly & Stoye (2016). For this reason I think using these hospitals as outside option is appropriate. Notably, private hospitals for privately funded patients are not considered to belong to a different market as highlighted by a recent investigation of the competition authority in the UK as well as in Kelly & Stoye (2016) and Gutacker *et al.* (2016). Gutacker *et al.* (2016) the authors seem to suggest that, between the two options, patients have a preference for public hospitals.

Estimation. I estimate the following equation, where s_j, s_0 are the market shares of hospital j at time t for complexity patient c and the market share of the outside option :

$$log(\frac{s_{jct}}{s_{0ct}}) = \delta_{jct} = \alpha^d_{zc}\widetilde{z_{jt}} + x'_{jt}\beta_c - \alpha^d_{wc}wait_{jt} + \nu_{jct}. \tag{3}$$

I estimate equation (3) following Berry(1994) logit model. I assume that quality is observed with a measurement error η_{jt} that is not observed separately by the econometricians or the patients. The assumption is that patients can foresee observed quality. Additionally, one can think that these mistakes may be correlated over time because of how the hospitals manage the measurement process. Then, observed quality is $\widetilde{z_{jt}} = z_{jt}\eta_{jt}$.[23] Instead, the variable z_{jt} is the expected quality of hospital j chosen by the management of hospital j, i.e. a quality target for that hospital.

With regard to endogeneity, in this context I observe quality, contrary to a typical model with unobserved quality as Berry (1994), so endogeneity may represent a less important concern.[24] However, the unobserved component ν_{jt} may include potential sources of endogeneity in the following way. Firstly, general patient experience at the hospital and general satisfaction with the hospital can influence demand and be correlated with the clinical improvements in hip replacement. For this reason I include patient satisfaction survey results in my demand specification.[25] Secondly, there may be a feed-back-loop between quantity and quality with hospitals receiving more patients that become better at performing the operation because they do more procedures. However, this feedback loop is not a significant concern in the set of hospitals considered in my analysis, because for the hospitals in my sample these learning effects should already be exploited. In fact, the measures of quality used in this study are available only for hospitals that have at least thirty recorded surveys per year. This is the level at which Rhee et al. (2018) find that that potential learning effects are exhausted.[26]

In order to address the possible endogeneity concerns for my quality variable, I also estimate demand using instruments d_{jt} and the following moment condition:

$$\mathbb{E}[\zeta_{jt} \otimes d_{jt}] = 0.$$

I assume that $\nu_{jct} = \rho_d c\nu_{jct-1} + \zeta_{jct}$ following Sweeting (2011) and a suggestion

[23] Assuming $E[log(\eta_{jt})] = 0$ and $\mathbb{E}[log(z_{jt})log(\eta_{jt})] = 0$

[24] Similarly, Santos *et al.* (2017) and Gaynor *et al.* (2016) do not address time-varying quality endogeneity in their main specifications.

[25] I make the assumption that they are outside the control of the hospital in the short term and I abstract away about dynamic decisions over reputation building.

[26] Mehta et al. (2018) find a flatter learning curve for surgeons. However, these results are over the career of a surgeon, senior surgeons would be able to reach the required levels of surgeries. The presence of learning effects depends, then, on the hiring choices of hospitals that can hire "higher or lower quality/experience" doctors.

contained in Berry Levinsohn and Pakes (1995). The instruments d_{jt} include cost-side shifters[27] and lags of own hospital characteristics (waiting time and patient experience).[28] In Appendix F, I included the demand estimation results from my analysis following Ellickson et al. (2020) and Holmes (2011) which allows me estimate a parameter for distance.[29]

Table 2: Estimates of demand

Dependent variable: from logit model				
	(1)	(2)	(3)	(4)
Quality ($\widetilde{z}$)	0.11	0.09	0.09	0.07
	(0.02)	(0.04)	(0.02)	(0.03)
waiting time	-0.11	-0.08	-0.01	-0.02
	(0.09)	(0.09)	(0.09)	(0.09)
General patient experience	0.08	0.01	0.06	0.03
	(0.01)	(0.01)	(0.01)	(0.01)
Year fixed effects	Yes	Yes	Yes	Yes
Market fixed effects	Yes	Yes	Yes	Yes
No. obs	832	668	832	668
Hansen J stat.		0.31		0.24

Note: robust standard errors in parentheses. In (2)-(4) results using the instrumental variable approach described here. Easy patients in (1)-(2). Difficult patients in (3)-(4). Smaller data in (2)-(4) because of using lags and instruments not available for all years.

[27] These are productivity from the cost-side estimation, an index capturing the cost of the local labor market and satisfaction of staff. The index is relevant for the hiring of external staff like nurses. I will deflate my costs using this index in the cost function estimation.

[28] Following Hackmann (2019) I re-estimated the model only for all hospitals with bed occupancy below 90% to ensure my estimates are robust to the possible presence of capacity constraints. The estimate in the restricted sample are very close to the ones in the complete sample. A nested logit model with nests for inside good and outside good was also estimated, but the estimates for the nesting parameter were not robustly consistent with the nesting model assumptions.

[29] This model does not allow to use an instrumental variable approach.

4.3 Cost of hip replacements surgeries

The objective of this section is to present the cost function model (for primary hip replacement) to be estimated as well as the related estimation strategy and estimation results.

In my model hospitals choose quality and the corresponding inputs necessary to produce the chosen quality with the following timing:

- At the beginning of the period hospitals set targets for quality z_{jt} (and the implied quantity q_{jt}, as they cannot turn away patients). At the same time they choose the corresponding levels of inputs to produce those targets, after observing productivity from the previous period. Hospitals can observe the targets of the competitors.
- During the period consumers choose where to receive treatment after observing the quality targets and shocks to individual patient preferences (ϵ^q_{jtp} in section 4.2). Quantity and quality are produced and measurement error η_{jt} in quality is realized. A shock to productivity Ξ_{jt} is also realized.

I assume that hospitals are cost minimizing, when choosing their inputs. Each hospital chooses its inputs K_{jt}, L_{jt}, M_{jt} to minimize costs given the chosen level of quantity and quality q_{jt}, z_{jt} (where j is for the hospital and t is for the year). This is a reasonable assumption because in the NHS system[30] there is a very active market for CEOs and financial performance determines the career of CEOs across different hospitals in England.[31]

Consider the following maximization problem where the production function is $F(K_{jt}, L_{jt}, M_{jt}, \Omega_{jt}) = K_{jt}^{\beta_K} L_{jt}^{\beta_L} M_{jt}^{\beta_M} \Omega_{jt}$.[32] The production possibility frontier depends on the number of hospital beds, denoted by K_{jt},[33] the number (adjusted by seniority) of staff workers (doctors, nurses), denoted by L_{jt},[34] materials, denoted by M_{jt},[35] and productivity, denoted by Ω_{jt}. C_{jt} is cost and r_{jt}, w_{jt}, p_{jt} are input prices.

[30] Highlighted also in Bloom *et al.* (2015)

[31] However, my estimation strategy can allow for the presence of optimization errors in hiring and investment. I also include non-profits motives in the objective function of the hospitals, but this does not affect cost minimization.

[32] In future versions I will test the robustness of this model assuming that some inputs are used in fixed proportions.

[33] I do not model dynamic decisions in the stock of capital, because variation in the stock of capital is extremely small. There are mergers in the sample and I treat merged hospitals and merging hospitals as separate organizations.

[34] It is possible to extend the model to include different levels of seniority and distinguish between doctor and nurses. This would not impact the estimation strategy or the results of the cost function. More input prices would appear in the cost function specification. Time fixed effects would capture them as well, given that all salaries are set nationally.

[35] Input prices for all these categories are the same nationally and change over time, I capture this aspect using time fixed effect in my estimation.

$$\max_{K_{jt},L_{jt},M_{jt}} K_{jt}^{\beta_K} L_{jt}^{\beta_L} M_{jt}^{\beta_M} \Omega_{jt}$$

$$s.t.$$

$$C_{jt} = r_{jt}K_{jt} + w_{jt}L_{jt} + p_{jt}M_{jt}.$$

Each hospital transforms inputs K_{jt}, L_{jt}, M_{jt} in two outputs q_{jt}, z_{jt} according to the transformation function:

$$T(q_{jt}, z_{jt}) = F(K_{jt}, L_{jt}, M_{jt}, \Omega_{jt}). \tag{4}$$

$T(q_{jt}, z_{jt})$ determines how hospitals can use the different inputs to produce the two outputs, quantity and quality. In particular, the function specifies how additional inputs can be used for either higher quantity or higher quality. I make some assumptions about the parametrization of the transformation function. In particular, I assume that the exponent of q_{jt} is one and q_{jt} multiplies all other elements in the function.[36]

In this paper I use a trans-log cost function specification that corresponds to equation (4) (in Appendix B I provide a more detailed derivation). The cost function is in equation (5), where g is a linear function of input prices. In this specification, I assume that there is no interaction effect between quality and quantity because of the size of the hospitals included in my analysis, as discussed in the demand section.

$$log(C_{jt}) = \alpha_0 + \alpha_{q1} log(q_{jt}) + \alpha_{z1} log(z_{jt}) + \alpha_{q2}(log(q_{jt}))^2 + \alpha_{z2}(log(z_{jt}))^2 + g(r_{jt}, w_{jt}, p_{jt}) + \omega_{jt}. \tag{5}$$

The residual of this cost function depends on productivity: $\omega_{jt} = log\left(\frac{1}{\Omega_{jt}}\right)^{\frac{1}{\sum_b \beta_b}}$ where $\sum_b \beta_b = \beta_K + \beta_L + \beta_M$. In my model productivity Ω_{jt} captures all the aspects that may change the production possibilities given the same inputs, for example, the patient pathway, the optimization of the staff rooster or the procurement ability. As in most of the production function literature, I assume that differences in productivity can be summarized in this single index where all the elements affecting this index affect production possibilities in the same way.[37] In my model I also assume that productivity follows a Markov process with a shock Ξ_{jt}. Hospitals do not have control on productivity Ω_{jt} or on shocks to productivity Ξ_{jt}. Notably, productivity, as often assumed in the productivity literature, is here considered as an exogenous process. Even as an exogenous process, productivity is important in determining the persistency of quality, because hospitals are limited in their

[36] See Appendix B.

[37] This assumption is not necessary for estimation, but it is useful for explanatory purposes. I adopt an Arellano Bond technique to estimate the cost function which allows me to relax the assumption of scalar unobservable (no heterogeneity in adjustments costs, input prices and no optimization error in hiring or investment).

choices by the level of productivity they have.

As shown in equation (5) the cost function naturally includes not just outputs, but also the input prices. As anticipated, the main inputs are staff, bed usage and prostheses. Staff salaries are set nationally through a bargaining process and are the same for the entire nation, but may change over time. The cost of bed days and operating theater use is assumed to be the same across England, up to a local adjustment factor due to differences in rents.[38]

While salaries of staff are assumed the same across hospitals, the prices of prostheses are assumed to be heterogenous. Difference in prices may be due to two factors: the manifacturer and the quality of the prosthesis. Prosthesis prices vary with quality, better prostheses of the same manufacturer cost more than lower performance prostheses of the same company. Additionally, hospitals may only use only one specific provider. This may be due to long lasting relations between doctors or procurement officers and one specific manufacturer. The market of orthopedic prostheses is dominated by three multinational companies, and I assume each of the approximately 150 hospitals performing hip replacements does not have enough bargaining power to affect the price.

Estimation. Following the framework suggested by Grieco & McDevitt (2016) where health care providers produce bundles of quality and quantity[39] I specify the following trans-log model[40] for my cost function estimation (χ_{jt} is an idiosyncratic shock, η_{jt} is the measurement errors of quality[41]):

$$ln\, C_{jt} = \alpha_0 + \alpha_{q1} log(q_{jt}) + \alpha_{z1} log(\widetilde{z_{jt}}) + \alpha_{q2}(log(q_{jt}))^2 + \alpha_{z2}(log(\widetilde{z_{jt}}))^2 + g(r_{jt}, w_{jt}, p_{jt})$$

$$+\beta_{cc} Complexity\ mix("CC") - \alpha_{z1} log(\eta_{jt}) - \alpha_{z2} log(\widetilde{z_{jt}}) log(\eta_{jt}) + \alpha_{z2}(log(\eta_{jt}))^2 + \omega_{jt} + \chi_{jt}. \quad (6)$$

The trans-log specification comes from the fact that I expect quadratic costs in quality. Increasing quality is expected to become more costly at higher levels of quality. I include as control an index that captures what percentage of the surgeries is more complex (when patient have co-morbidities, for example obesity).

To estimate the cost function I need to address two main challenges: the presence of unobservable productivity and measurement error. Firstly, I need to account for differences in productivity. Productivity[42] cannot be observed by the econometricians and it may bias cost function estimates. It affects both quality and cost because it shifts the production possibility frontier. Secondly, as anticipated, quality of outcomes is observed with a measurement error. As a reminder, z is a quality target, not the quality observed

[38] See next section.

[39] This is the first example of the application of the methodology by Ackerberg *et al.* (2015) in the health care sector.

[40] As in Christensen & Greene (1976).

[41] Remember I assume $\mathbb{E}[log(\widetilde{z_{jt}}) log(\eta_{jt})] = 0$ and $E[log(\eta_{jt})^2] = 0$.

[42] Importantly, productivity Ω affects not only the number of patients treated, but also the level of quality of treatment.

by the econometricians. Hospitals are unable control the final observed quality, because different unexpected factors may influence the outcome measure. This is reflected by the presence of shocks/measurement errors. As already discussed, I assume that the final observed outcome ($\tilde{z}$) is a combination of the quality target and a measurement error.

To address the presence of unobservable productivity differences I adopt a ρ-differencing approach based on Arellano Bond (1991). I make the assumption that ω_{jt} follows an AR(1) process: $\omega_{jt} = \rho\omega_{jt-1} + \xi_{jt}$. Taking ρ differences of the variables I can eliminate ω_{jt} and use the following moments:

$$\mathbb{E}[\xi_{jt} + (\chi_{jt} - \rho\chi_{jt-1}) + (f(\eta_{jt}) - \rho f(\eta_{jt-1})) \otimes (\mathbf{w_{jt}})] = 0. \tag{7}$$

$\mathbf{w_{jt}}$ is the set of instruments I use and it includes lags of quantity, Waldfogel instruments, BLP instruments and non-profit shifters, $f(\eta_{jt}) = \alpha_{z1}log(\eta_{jt}) - \alpha_{z2}log(\widetilde{z_{jt}})log(\eta_{jt}) + \alpha_{z2}(log(\eta_{jt}))^2$. Non-profit shifters are foundation status[43] and teaching status[44] which determine different levels of non-profit motives. The choice of these instruments addresses the presence of serially correlated measurement errors.

Table 4 shows results using different instruments for quality that highlight the presence of serially correlated measurement error. If the measurement error in quality is persistent then $\mathbf{w_{jt}}$ in equation (6) cannot include lags of quality. A proper set of instruments could include, instead, the lags of another measure of quality "EQ VAS": a more imprecise measure of z_{jt} $\hat{z_{jt}}$, such that $log(\hat{z_{jt}}) = log(z_{jt}\mu_{jt})$. I assume that the measurement error μ_{jt} is independent of the measurement error η_{jt}, because the participation rates and accuracy is documented to be different between the two measures (Feng et al. (2014)). For identification I also use quality shifters like Waldfogel instruments capturing for example the percentage of the elderly population, the level of economic deprivation of the elderly or the presence of architectural barriers in the area. If, instead, I use as an alternative instrument that has a measurement error that is not credibly independent of η_{jt} the estimates are downward biased. In particular, using lagged EQ 5D, another measure of quality coming from patients surveys, I under-estimate the costs and obtain that a large percentage of marginal costs are negative. The reason is that such measure is highly correlated with quality (z_{jt}) and its measurement error cannot be considered to be independent of η_{jt}.

Finally, to account for the presence of input prices for bed space, temporary external staff and prostheses, I do the following. I divide my costs by the hospital specific Market Forces Factor ("MFF") to account for local differences in prices of land, rent or external staff. The MFF has been designed by the NHS specifically to capture these differences. I

[43] hospitals can be foundations and re-invest their profits in the hospital, alternatively they have to give the profits back to the NHS

[44] For robustness check I also included inspections by the quality commission CQC, that happen more often if the hospital does not meet certain quality criteria. Additionally they would put pressure on the hospital to improve their quality performance.

include time fixed effects to account for the permanent staff salaries that change over time, but are the same across the country. However, I do not observe prostheses prices and I do not include them in my estimation. This choice, however, does not depend on data availability alone.

The prostheses price differentials are due to two elements: quality differences and hospital procurement choices (for example a long standing relationship between one hospital and one manufacturer). Given that these inputs are not homogeneous in quality, including them would mean that the coefficient of the quality terms would capture the cost of better clinical outcomes net of input quality. To avoid this, I would choose not to include the prices even if I could observe them. The remaining differences due to procurement relations are instead assumed to be captured by the productivity term Ω_{jt}.[45]

Table 3: Estimates of the cost function

Dependent variable: Total Cost			
	(1)	(2)	(3)
q	1.21	1.24	1.25
	(0.03)	(0.03)	(0.05)
z	2.56	-16.18	-14.52
	(3.34)	(3.83)	(7.26)
q^2	-0.03	-0.03	-0.03
	(0.01)	(0.00)	(0.01)
z^2	-0.36	2.69	2.56
	(0.55)	(0.63)	(1.20)
Complexity "cc"	0.51	0.31	-0.11
	(0.05)	(0.07)	(0.11)
Year fixed effects	Yes	Yes	Yes
No. obs	695	695	695

Note: Clustered standard errors in parentheses.
(1) using EQ5D as instrument for quality and
(2) using EQ VAS, (3) only quality shifters

[45]Given that I use the Arellano-Bond strategy I have extra flexibility in the assumptions. I could assume that there is another unobservable capturing this aspect. I would only have to assume that it follows a separate AR(1) process with the same parameter ρ, as long as the innovation/shock in this process is orthogonal to my instruments.

Table 4: Marginal costs and non-profit motives in British Pounds

		Mean	90 perc.	75 perc.	50 perc.	25 perc.	10 perc.
Panel A: Costs	$\frac{\partial C}{\partial q}$	5,119	6,430	5,670	4,994	4,336	3,812
	$\frac{\partial C}{\partial z}$	99,055	179,632	125,008	82,008	52,822	34,790
	$\frac{\partial C}{\partial z}\frac{1}{q}$	295	402	347	291	243	193
Panel B: Non-profit	$\frac{\partial v}{\partial z}\frac{1}{q}$	225	420	306	207	124	25

Note: non-profit motives backed out from FOC of the hospitals, average price £5,500

4.4 Non-profit motives

I assume that hospitals have non-profit motives, they include purely altruistic motives as well as unforeseen health benefits for the patient, who may not fully understand the short and long term consequences of better care. I estimate them as the difference (in each period) between the monetary component of the marginal revenue of quality and marginal costs of quality (MC_z) in equation (2). I can observe that the marginal cost of quality is larger than the monetary component of the marginal revenue of quality. The difference shows that hospitals are providing quality beyond what would be optimal for profit maximizing firms. The size of these non-profit motives varies across hospitals and to better understand their size I reported in table 5 the non-profit motives per patient. Interestingly, I can observe that the non-profit motives are positively correlated with market shares: the hospitals with greater market shares have larger non-profit motives.

The benefit of having both demand and cost data is that it is possible to more carefully estimate marginal cost and quality elasticity without backing them out from the FOC's as in BLP(1995). This is especially true in the case in which hospitals are not profit maximizing. Not considering non-for-profit motives could lead to a biased estimates of marginal costs or elasticity of demand. If one was only estimating demand and was to back out marginal cost from the FOC's, without considering non-profit motives, it would underestimate marginal costs - potentially having marginal costs close to zero. This happens because marginal revenues alone would correspond to low costs, but, actually, higher costs are incurred because of the non-profit motives. If, instead, one was trying to back out $\frac{dq_j}{dz_j}$ from the FOC's using marginal costs, it would over-estimate patients' preferences for quality. This happens because marginal costs are higher than marginal revenues in presence of non-profit motives.

5 Counterfactual analysis

In this section i) I calculate the welfare loss arising from current uniform prices, ii) compare the welfare effects arising from higher uniform prices *vis-à-vis* the introduction of quality-based prices and finally iii) explore the importance of differences in efficiency and market

power in determining my results.

The choice of this paper to use simulations to study quality-based prices is threefold. Firstly, there is no large scale application of quality-based prices in hospital settings that would allow to study the issue empirically. Secondly, theoretical results about quality provision are not clear. Already Spence (1975) and White (1972) showed that regulation and market power may lead firms to provide different sub-optimal levels of quality (White especially showed this for health systems with administratively set prices). For this reason, when analyzing quality provisions studies as Crawford et al. (2018) have estimated the effects on quality in specific market configurations. Thirdly, simulations present advantages that empirical studies would not be able to offer. With simulations it is possible to get a measure of the potential welfare effect of different range of prices and also explore the channels driving the results.

Externalities and welfare loss from uniform prices. To evaluate the impact of different types of prices I need to fully specify a total welfare function. To complete the specification I discussed in previous sections I need to parametrize the positive externalities. Positive externalities arise from many health care surgeries/treatments, but in this case they are particularly evident. Improving mobility of patients is good not just for the patient, but for the health system that does not have to provide additional services for patients with limited mobility and for the potential complications arising from it. A better treatment would lead to less government expenditure later on as well as less need for intra-household support. Additionally, a person with improved mobility can be a more active member of society be it at work (paying more taxes) or in household production. I parametrize the externalities as $\Psi_{Ext}(z_j) = \psi\, z_j\, q_j(z_j, z_{-j})$.

To quantify the parameter ψ I use as reference the monetary value of Quality-Adjusted Life Years ("QALY") estimates related to hip replacements from other studies: Appleby *et al.* (2013) and Fordham *et al.* (2012).[46] QALYs indicate the utility arising from medical treatments due to longer life spans (more years) or increased quality of life, and 1 QALY corresponds to one full year of full quality of life. The UK government attaches to 1 QALY a value between £20,000 and £30,000.[47] This value according to the National Institute for Health and Clinical Excellence is meant to include both a monetary valuation of the private utility arising from treatment, but also potential social externalities.[48] Different values of ψ imply different sizes of the externalities. In particular, for higher values of ψ the externalities are larger.[49]

A value of $\psi = 150$ would be a conservative estimate and would be underpinned by the following calibration. The sum of consumer surplus, non-profit motives and externalities together would correspond to 0.8 QALY (on average) per patient at £20,000 per QALY

[46] Both studies indicate that hip replacements in England led, on average, to around 0.8 QALY in terms of increased quality of living over 5 years and 2.77 QALYs over 15 years.

[47] Irrespective of whether the QALY arises from longer life span or improved quality of life.

[48] Included in the NICE (the National Institute for Health and Clinical Excellence) guidelines.

[49] See Appendix D for an illustration.

per patient.[50] The value of hip replacement including private and social value is estimated in the mentioned studies at 0.8 and 2.77 QALYs. At the chosen value of £150, externalities represent 21% of the value generated by the hip replacements. At $\psi = 150$ the welfare loss is around 14%. Without externalities the welfare loss would be around 5%.[51]

Quality-based prices v. uniform prices. By introducing quality-based prices the government would change hospitals objective function. Instead of having fixed uniform prices, hospitals would face prices that depend on the observed clinical outcomes. Quality is observed with a measurement error and this leads to a term $P(\bar{z})$:

$$(1 + \underbrace{\tau}_{\substack{\text{reward } or \\ \text{punishment}}} \underbrace{P(\bar{z})}_{\substack{\text{Probability of} \\ \text{observing } \tilde{z} > or \tilde{z} < \bar{z}}}) \overbrace{\bar{p}}^{\text{fixed price}} .$$

Quality-based prices and hospital choice

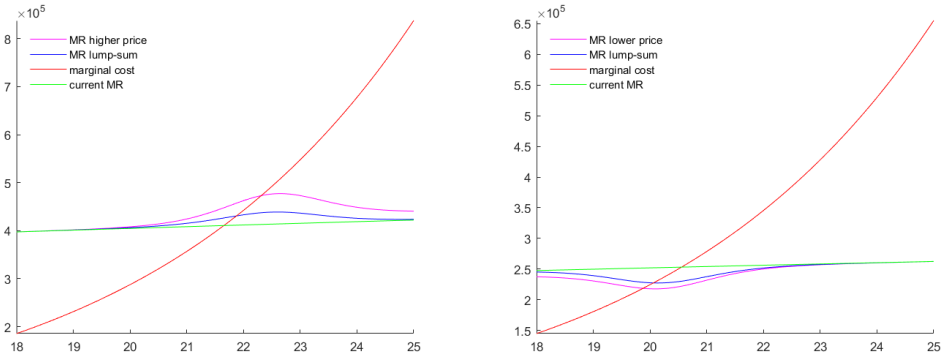

Figure 2: Reward above $\bar{z} = 22.5$ Punishment below $\bar{z} = 20.2$

The change in prices affects the marginal revenue of quality, in case of a reward above a threshold $\bar{z}$ or a punishment below another threshold $\bar{z}$ as in Figure 1. The bump is due to the fact that the derivative of the incentive is larger around that threshold.

Assuming that hospitals compete Bertrand-Nash in expected quality, I simulate the welfare effects of having either higher uniform prices or quality-based prices with different levels of rewards or punishments in correspondence to different quality threshold $\bar{z}$.

In this section, firstly I briefly present the results and, then, I discuss the mechanisms at play behind these results. The measure that I display in Table 5 is: $\frac{Welfare_{reform} - Welfare_{current}}{Welfare_{optimal} - Welfare_{current}}$, where $Welfare_{reform}$ is the level of social welfare achieved with the different reforms, either higher uniform prices or quality-based prices. The measure is meant to capture the per-

[50] It is a lower bound because this corresponds to a patient who had a low outcome (compared to the best outcome in the sample) needing 60 minutes a month (for five year) of extra care from family members or social workers, for £22 per hour. This amount per hour is indicated by the Department of Health when evaluating externalities, in terms of hours of work lost by employers because of illnesses of employees. 1 year could be rationalize either with a strong discounting or a combination of expected time of death and natural deterioration of health outcomes.

[51] To convert the utility of patient into monetary value I use the value of waiting time in NHS lists following the estimates in Propper (1990).

centage of the gap in welfare between the baseline current level of welfare and the optimal welfare that is reduced by the introduction of the different reforms. I present in Appendix E a decomposition of the welfare effects by their different components (consumer surplus, etc.), in Appendix G the break down of the impact on consumer surplus by category of patient.

Table 5: Welfare changes under different reforms-2013

% of the difference b/w optimum social welfare and baseline welfare	No m. error	W/ m. error
Panel A: Uniform price increases		
$-10\%\bar{p}$ uniform price increase	-10.7%	
$+10\%\bar{p}$ uniform price increase	8.5%	
$+15\%\bar{p}$ uniform price increase	11.9%	
$+20\%\bar{p}$ uniform price increase	14.8%	
Panel B: Rewards for high quality		
$+10\%\bar{p}$ above median z	9.0%	16.3%
$+20\%\bar{p}$ above median z	19.5%	23.5%
$+30\%\bar{p}$ above median z	25.6%	27.6%
$+40\%\bar{p}$ above median z	29.4%	30.4%
Panel C: Punishments for low quality		
$-10\%\bar{p}$ below mean z	-7.3%	-25.2%
$-10\%\bar{p}$ below 3 st. dev. from median z	1.0%	-0.0%

Note: Cost of public funds are assumed 30% of raised funds.

Uniform prices. The welfare effects are not very large, given the curvature and slope of the marginal revenues and marginal costs. Marginal revenues are relatively flat given that quality elasticity is low while marginal costs are steep, so a small change in marginal revenue does not lead to a large change in quality and welfare. Higher uniform prices, similarly to Hackmann (2019) have a relatively small, but positive effect on welfare: they lead to an increase in quality, but the effect is compensated by additional costs as well as higher levels of distortionary taxes needed to cover the related additional government expenditure.

Rewards for higher quality. To make comparisons easier, in Table 5 I included price reforms that would be approximately government expenditure neutral: for example a 10% uniform price increase *v.* a 20% reward for quality above average quality.[52]

Quality-based prices in the form of rewards have larger positive effects on welfare than higher uniform prices. This is driven by the fact that a smaller amount of public money

[52] Average and median are close in my sample. The reason why there may be a discrepancy in government expenditure in the two pricing alternatives is due to the fact that more patients may be attracted from the outside option in the case of the rewards.

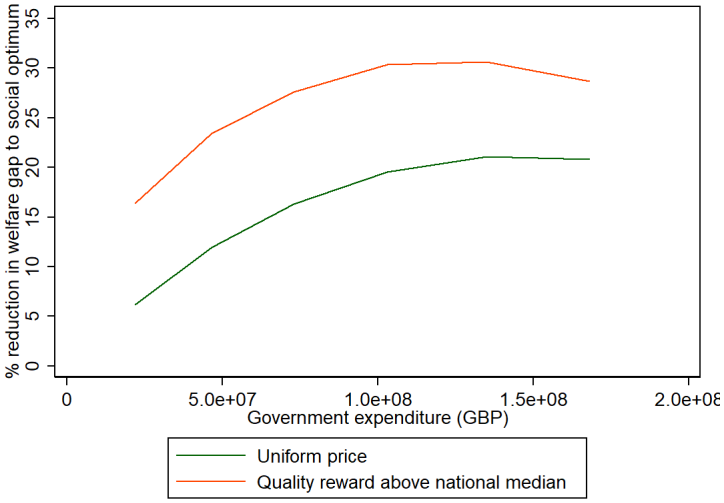

Figure 3: Uniform prices *v.* rewards (in 2013)

is used in a more effective way. There are two main channels that determine the benefits of rewards. On one side, there is the re-allocation to the best hospitals that happen to be typically also the most productive. On the other side, a more indirect effect is at play due to the presence of measurement error in quality and competition. I call this effect "indirect" because they lead to a change in quality from not directly affected hospitals.

Given that quality is a strategic complement, not just the rewarded hospitals increase quality, but also the other hospitals when they are competing with rewarded hospitals. From the government perspective this has the advantage of delivering an increase in quality without additional government expenditure. However, the net effect on welfare of this increase in competition depends on the costs of these competing hospitals. If their increase in costs is too high, the net effect may not be positive.

Another element at play in quality-based prices is the presence of measurement error in quality. This leads hospitals around the threshold to react even if they would not do it in absence of measurement error. In particular, hospital that would provide quality below the threshold but not too far from the threshold are also affected by the quality-based prices. They increase quality because in expectation they have higher expected marginal revenues, even if in the end some hospitals will not receive the rewards after not reaching the threshold. In this way quality-based prices lead to another "un-rewarded" increase in quality, which, in turn, also puts more competitive pressure on other hospitals. The effect of measurement error on welfare varies in the different scenarios but is around 20% of the effects of the reforms. For a more theoretical discussion about the impact of quality-based rewards see Appendix A.

Punishments for lower quality. Finally, punishments for lower quality have a negative or null effect on welfare. They lead to a decrease in quality: the hospitals providing the lowest quality would not have incentives to keep quality at the level afforded by uniform

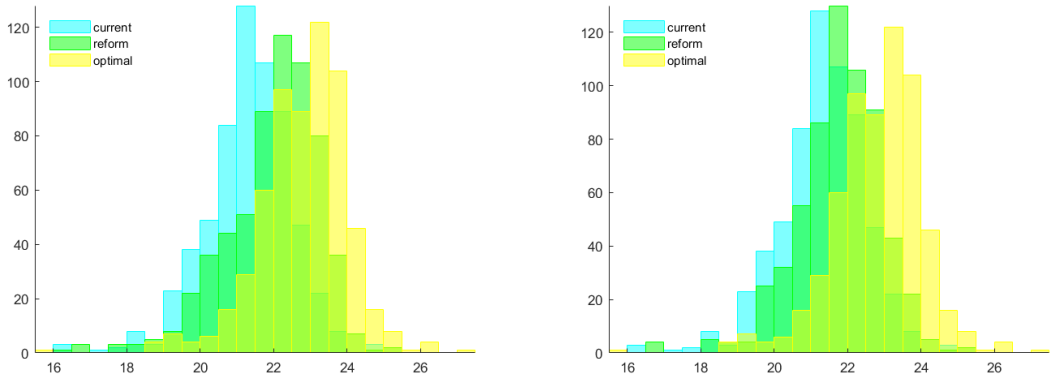

Figure 6: Note: Quality distribution w/ an increase in price of 10% above one standard deviation above median quality (left) or a 20% uniform price increase (right).

prices and they would lower their level.[53] Given that the worst hospitals are typically less productive this may be welfare enhancing. However, the effect is compensated by a decrease in competition and by the presence of measurement error that work in this case in the opposite directions.

With regard to welfare, in this case quality-based prices lead to reductions in costs and government expenditure, however, these savings are smaller than the value lost because of patients receiving lower quality in the markets affected by quality-based prices. The result is in line with an analysis of a pilot reform implemented by the NHS in 2014 reform which introduced punishments of $10\%\bar{p}$ for quality below 3 standard deviation from the mean. In a difference-in-differences analysis I found that the reform led to a small decrease in quality.I included the analysis in Appendix I. My analysis suggests that punishments may have the unintended consequence of reducing quality and welfare rather than increasing it.[54]

Interestingly, price discrimination based on lower quality here is not welfare enhancing. The reason is that the competition and measurement error effects dominate the impact of the punishments and leading even more hospitals to decrease quality even without a reduction in government expenditure. This can give us an indication of the relative size of the different effects.

The role of provider heterogeneity and re-allocation. In this subsection I discuss the role of provider heterogeneity on the performance of quality-based prices *vis-à-vis* uniform prices. In particular, I look at the role of productivity differences (contained in

[53] Importantly, the relation between payment and welfare effects is due to the fact that in this paper quality is costly and additional resources are needed to reach higher levels of quality.

[54] This point of view could be contrasted by another view: that hospitals are slacking or need to increase their productivity. In this paper I took the view that such control is harder to materialize. This is not an uncommon assumption in the productivity literature and and it is motivated by two empirical facts. On the one side, the UK has adopted a yardstick competition that should incentivize a convergence to the same level of productivity and unit cost, but such convergence has not materialize. On the other side, in recent years hospitals have challenged the NHS on their ability of increasing their productivity and have been able to prevent large decreases in uniform prices.

the residual of the cost function) and differences in other hospital characteristics.

Productivity differences are an important driver in differences in marginal costs, but their impact on welfare is limited by the fact that marginal costs are increasing. Productivity determines the intercept and slope of the marginal costs curves. Higher productivity hospitals have lower intercepts and less steep slope. However, given that marginal costs are increasing, the realized value of the marginal cost depends on the level of quality provided. For this reason, lower productivity hospitals can have lower realized marginal costs than more productive hospitals if they provide lower quality. The level of realized marginal cost is an equilibrium outcome and depend also on the level of the marginal revenues (determined by market size as well as other hospital characteristics). This implies a de-coupling between realized marginal cost and productivity. Contrary to what happens when marginal costs are constant, the most productive hospital may not be the one with the lowest marginal cost.

This has implication for the role that outcome-based prices can have in the re-allocation of resources to the best hospitals and how this re-allocation affects social welfare. Chandra *et al.* (2016) has highlighted how the health care sector is not different from other sectors in that we can observe a re-allocation to the best and typically most productive hospitals. In my paper I show how this re-allocation can be helped by outcome-based prices and enhance welfare, especially given that I do find a positive correlation between productivity and quality. However, I also want to highlight that this may not always be welfare enhancing and that re-allocation to the best and most productive hospitals may have a smaller role in improving social welfare in presence of increasing marginal costs. As marginal costs increase, it becomes not socially optimal to re-allocate further to the best and most productive.[55]

Additionally, hospitals have additional characteristics that are valued by consumers and this may also limit the benefits from re-allocation. When the best hospitals do not attract all the patients from competitors, they perform less positively in other dimensions, for example waiting time. For this reason, if a patient chooses to go a better hospital after the introduction of outcomes-based prices, she will experience an net improvement in consumer surplus. However, she will also lose in terms of waiting time, so the net increase is smaller when considering all hospital characteristics.[56]

Increasing marginal costs and other hospital characteristics also make less obvious what is the efficient scale of the hospital. While a social planner may be tempted to choose levels of production where hospitals are at the minimum of the marginal costs curves, this would not be optimal. On the one side, marginal costs are functions of both quantity and

[55] To the extreme, if patients have decreasing utility from quality, re-allocation to the best providers may not be optimal at all, even if the best hospitals are the most productive.

[56] This keeping waiting time fixed. If waiting time was to increase at the best hospitals and patients may even decide not to choose that hospital. In my paper, I do not model changes to waiting time, as they change in number of patients in small compared to the size of the orthopedics departments. Additionally, waiting time for surgery does not work as a typical queue: slots are designed considering the needs of the entire department as well as the needs of the patients.

quality: the quality offered at the minimum of these curves may be too low and lead to too important losses in consumer surplus. On the other side, heterogeneity in hospital characteristics imply that some hospitals should produce more simply because they have higher levels of hospital characteristics appealing to patients.

The role of competition. Quality-based prices affect the competitive dynamics between hospitals. As anticipated, rewards can increase quality competition, while punishments may diminish it. Quality is a strategic complement and even hospitals that are not directly rewarded or punished react to the change in quality by the hospitals that are directly affected. To show the importance of this channel I decomposed the aggregate effect of the introduction of quality-based prices into the effect in different markets with different numbers of competitors. This analysis is limited by the fact the hospitals in the different markets have different characteristics and cost structures. Nevertheless, the analysis can give some interesting indications.

As shown in the table, local monopolies are characterized by a smaller welfare effect when not considering the role of measurement error. This can be explained by the fact that there is no effect of competition on not directly affected providers. In the case of rewards, competition seem to have a positive effect on welfare because in the markets with more than one provider the welfare effect is larger.

Table 6: Welfare changes in different local markets

% of the difference b/w optimum social welfare and baseline welfare		
	No m. error	W/ m. error
Effect in competitive mkt #competitors=19		
+10%$\bar{p}$ uniform price increase	9.8%	-
+20%$\bar{p}$ above mean z	10.3%	14.9%
Effect in competitive mkt #competitors=13		
+10%$\bar{p}$ uniform price increase	9.6%	-
+20%$\bar{p}$ above mean z	12.6%	13.2%
Effect in local monopolies #competitors=5		
+10%$\bar{p}$ uniform price increase	7.9%	-
+20%$\bar{p}$ above mean z	5.2%	12.7%
Effect in local monopolies #competitors=1		
+10%$\bar{p}$ uniform price increase	-1.6%	-
+20%$\bar{p}$ above mean z	3.4%	6.8%

Note: cost of public funds $\lambda = 0.3$, threshold: $\bar{z} = 21.04$

The role of uncertainty. As previously discussed, the measurement error introduces a level of uncertainty in quality-based prices. Regulators are rewarding the achievements of hospitals based on observed quality, but hospitals do not have full control of observed

quality, because of the presence of measurement error. This uncertainty creates an additional potential channel for quality-based prices to increase welfare. In Figure 7 I show that for higher levels of rewards the role of uncertainty diminishes. This happens because as the size of the reward increases the marginal revenue of quality shifts upwards directly influencing more and more hospitals even without uncertainty.[57]

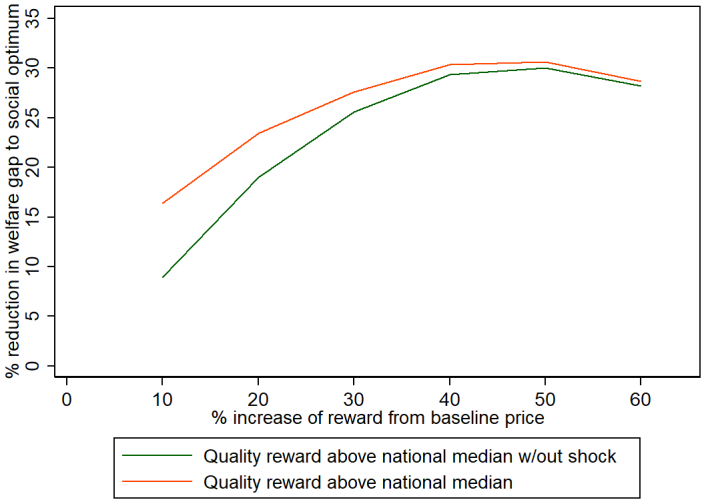

Figure 7: Role of shock in welfare effects (in 2013)

Optimal quality-based prices. The combination of this last effect on expectations, increased competition and the fact that higher quality hospitals may be more productive can be exploited to reach higher levels of welfare. In particular, using the threshold level for the rewards one can exploit these effects and obtain optimal quality-based prices. Threshold levels and reward sizes can be combined to minimize government expenditure and boost quality. A higher threshold reduces government expenditure, while a higher reward for few hospitals can enhance the strategic responses of other hospitals. A too high threshold would involve a too small number of hospitals, so it is not optimal to increase a threshold indefinitely. A higher reward would instead become more and more expensive for the government and lead to a smaller importance for uncertainty. I solve for the optimal combination of reward size and threshold level that reaches the maximum achievable by quality-based rewards. It would lead to cover 37% of the gap between the welfare with baseline prices and the social optimal welfare.

6 Policy implications: Beyond hip replacements

So far I have discussed quality-based prices in the context of hip replacements surgeries in the English NHS. Nevertheless, my framework could be applied to other surgeries in

[57] Settings with high rewards can lead to multiplicity of optimal private decisions and equilibria. I assume that the hospitals would always choose the highest quality for which $MR_z = MC_z$.

other health care systems that follow DRG-like[58] reimbursement schemes. In this section I highlight three considerations that should guide policy makers that wanted to implement quality based prices in other contexts.

Correlation between cost and quality. For quality-based prices to deliver positive effects on welfare, the re-allocation to the best providers should lead to a non-negative effect. In my paper this is the case because there is a positive correlation between quality and productivity: the best hospitals tend to be the most productive. Conditional on the level of quality, they have lower marginal costs than less productive hospitals. If this was not the case, the least productive hospitals would provide higher quality. It would likely lead to a smaller or even negative effect from re-allocation as their costs increase sharply with quality. One way to rule this out is if there is a positive correlation between quality and other hospital characteristics. This would reduce drastically the probability that lower productivity hospitals would provide higher quality. In fact, higher quality is higher due to high marginal revenues curves (due to other hospital characteristics, market size or non-profit motives) or lower marginal costs curves (which directly depends on productivity).

Input flexibility. In my paper, I assume that hospitals can freely change their level of inputs and adjust quality. This is reasonable in the context of hip replacements and for the small increases in prices I considered.[59] If quality-based prices were rolled out in more surgeries and in other contexts flexibility may become an issue. In particular, the availability and the importance of senior medical staff may vary impacting the ability of hospitals to increase quality in response to price changes. This would mean that some hospitals would be more able to increase quality than others, reducing the effect of quality-based prices and leading to a stronger re-allocation to the hospitals with higher flexibility. While this may still be socially optimal, it may lead to concerns in terms of inequality of access for the patients living far away from "flexible" hospitals.

In the specific context of this paper, this higher flexibility in hiring can depend on the ability to attract better medical staff and it is included in the productivity term. Given that doctors are paid the same at the same seniority level, differences in ability of doctors of the same grade would influence the level of productivity. Additionally, given that pay is not a relevant driver of choice, other characteristics of the hospitals may attract doctors differently: hospital reputation, working conditions and other characteristics (for example geographical location). Then, the ability of attracting better doctors would influence the meaning of re-allocation to the most productive hospitals. I regressed productivity on several characteristics, including type of hospital (teaching, specialist or not), reputation and staff satisfaction. They explain very little variation in productivity, but other characteristics may still influence doctors decisions and therefore productivity.

I want to highlight that there could be a different impact in urban areas *vis-à-vis*

[58] The Diagnosis Related Group system is widely adopted in many countries and it originated in the U.S. Medicare. In this system hospitals are paid a regulated fixed amount of money, a "price" for each surgery/ treatment they perform.

[59] This was confirmed with industry experts in relation to the ability of recruiting medical staff in the different English hospitals.

rural areas. Urban areas may be able to attract more doctors and hospitals with better reputation may be able to hire better senior doctors. This is not *per-se* an issue. As long as patients live in the same urban area, this will not impact significantly consumer welfare if other characteristics do not change. In rural areas, if the hospitals lose senior doctors to cities after the introduction of outcome-base prices this may impact consumer surplus because patients would have to travel too far. Doctors who are in rural areas, however, they may have strong geographical preferences and generally not be willing to move. To understand the issue I regressed changes in staff on local area dummies and I saw that rural areas had considerably less inflow and outflow of doctors per capita. As a consequence outcome-based rewards may not affect negatively rural areas, but consumers in rural areas may be less able to benefit from the rewards.[60]

Selection. In this paper patient selection is not considered to be a concern. I assume that hospitals do not decide to serve easier patients to boost their performance. This rests on four considerations. Firstly, the measure of quality considered is constructed to control for many patient characteristics, this should limit the selection ability of hospitals. Secondly, hospitals are not allowed to reject patients. Thirdly, even if they could use other channels like waiting time to select patients, I do not find evidence that patients observables significantly influence waiting time. Fourthly, the more difficult patients as shown in Appendix G are very elastic and are willing to wait longer for quality, limiting the ability of hospitals to use this channel.

This may not be the case in other contexts and hospitals may have other channels that they can exploit to select patients. For this reason, policy makers should restrict quality-based prices to high volume surgeries for which they can control for patient characteristics and monitor closely the introduction of such a price regime. The influence of hospitals may also happen in the data collection and affect the measurement error in quality. To avoid this, it is reasonable to think that policy makers should assign centrally the data collection to third parties to limit hospitals' influence.

7 Final remarks

Governments aimed for many years to create health care systems that combine high quality health care services with cost containment. Competition and price regulation have been used widely to achieve this goal. More recently, quality-based prices, that link payment to clinical performance, have been proposed and experimented by some regulators. In this paper I develop a framework to evaluate the impact of quality-based prices, taking into account the potential re-allocation effects across hospitals with different productivity levels. I show that quality-based prices lead to an improvement in welfare and I develop a model of supply and demand to quantify this positive effect in different simulations.

I apply my framework to hip replacements in the English NHS, a type of surgery

[60] Future versions of this paper will address this topic more organically.

for which the NHS has started using quality-based prices. In my simulations, the total welfare gains with quality-based prices are at least twice effective compared to uniform prices. The price regime effect relies on rewarding higher prices only to few hospitals and operates through three channels: re-allocation to the most productive, competition and measurement error.

I observe in my simulations that the importance of the three channels is similar and that the re-allocation channel may be less important than what it would be usually expected. This is due to the fact that hospitals have increasing marginal costs and multiple characteristics. This limit the benefit from re-allocation as costs increase also for the most productive hospitals and these hospitals may not be the best across all characteristics.

The importance of the channels of competition and measurement error *vis-à-vis* reallocation also explains that I find a negative effect on welfare of quality-based punishments for lower levels of quality. Punishments lead to lower government expenditure and help diverting patients from the worst and least productive providers. However, they also lead to "extra" lower quality because of lower competition and the presence of measurement error.

From a methodological perspective I apply methodologies from the productivity literature to a cost function in quality and quantity. They allow me to distinguish between inefficiency and cost of quality. Not considering this difference would lead to unreliable estimates for the cost of quality and misleading conclusions on the welfare effects of quality-based prices. In fact, to draw reliable results it is crucial to individuate the hospitals that provide quality more efficiently, not simply those that have higher or lower unit cost.

Finally, I point out some caveats that policy makers should consider when applying the promising tool of quality based prices. The level of productivity of the best hospitals, the flexibility of inputs and patient selection can change the welfare effect of re-allocation and the equality of access to care for all patients.

I hope that this analysis can help bring focus to the potential of outcome-based prices and on policies directed to improve welfare for patients in regulated health care markets.

References

Ackerberg, Daniel A, Caves, Kevin, & Frazer, Garth. 2015. Identification properties of recent production function estimators. *Econometrica*, **83**(6), 2411–2451.

Appleby, John, Poteliakhoff, Emmi, Shah, Koonal, & Devlin, Nancy. 2013. Using patient-reported outcome measures to estimate cost-effectiveness of hip replacements in English hospitals. *Journal of the Royal Society of Medicine*, **106**(8), 323–331.

Berry, Steven T. 1994. Estimating Discrete-Choice Models of Product Differentiation. *The RAND Journal of Economics*, **25**(2), 242–262.

Bloom, Nicholas, Propper, Carol, Seiler, Stephan, & Van Reenen, John. 2015. The impact of competition on management quality: evidence from public hospitals. *The Review of Economic Studies*, **82**(2), 457–489.

Chandra, Amitabh, & Staiger, Douglas. 2016. Sources of Inefficiency in Healthcare and Education. *American Economic Review*, **106**(5), 383–387.

Chandra, Amitabh, Finkelstein, Amy, Sacarny, Adam, & Syverson, Chad. 2016. Productivity dispersion in medicine and manufacturing. *American Economic Review*, **106**(5), 99–103.

Christensen, Laurits R, & Greene, William H. 1976. Economies of scale in US electric power generation. *Journal of political Economy*, **84**(4, Part 1), 655–676.

Crawford, Gregory S, Shcherbakov, Oleksandr, & Shum, Matthew. 2019. Quality Overprovision in Cable Television Markets. *American Economic Review*, **109**(3), 956–995.

Dalton, Christina M, Gowrisankaran, Gautam, & Town, Robert. 2019. *Salience, myopia, and complex dynamic incentives: Evidence from Medicare Part D.* Tech. rept. National Bureau of Economic Research.

Einav, Liran, Finkelstein, Amy, & Mahoney, Neale. 2018. Provider Incentives and Healthcare Costs: Evidence From Long-Term Care Hospitals. *Econometrica*, **86**(6), 2161–2219.

Eliason, Paul. 2017. *Market Power and Quality: Congestion and Spatial Competition in the Dialysis Industry.* Tech. rept. Working Paper.

Eliason, Paul J, Grieco, Paul L E, McDevitt, Ryan C, & Roberts, James W. 2018. Strategic patient discharge: The case of long-term care hospitals. *American Economic Review*, **108**(11), 3232–3265.

Evans, David S, & Heckman, James J. 1984. A Test for Subadditivity of the Cost Function with an Application to the Bell System. *The American Economic Review*, **74**(4), 615–623.

Fan, Ying. 2013. Ownership Consolidation and Product Characteristics: A Study of the US Daily Newspaper Market. *American Economic Review*, **103**(5), 1598–1628.

Fleitas, Sebastián. 2018. Who benefits when inertia is reduced? Competition, quality and returns to skill in health care markets.

Fordham, Richard, Skinner, Jane, Wang, Xia, Nolan, John, Group, Exeter Primary Outcome Study, & Others. 2012. The economic benefit of hip replacement: a 5-year follow-up of costs and outcomes in the Exeter Primary Outcomes Study. *BMJ open*, **2**(3), e000752.

Gaynor, Martin. 2006. *What Do We Know About Competition and Quality in Health Care Markets?* Working Paper 12301. National Bureau of Economic Research.

Gaynor, Martin, & Town, Robert J. 2011. *Competition in Health Care Markets.* Working Paper 17208. National Bureau of Economic Research.

Gaynor, Martin, Moreno-Serra, Rodrigo, & Propper, Carol. 2013. Death by Market Power: Reform, Competition, and Patient Outcomes in the National Health Service. *American Economic Journal: Economic Policy*, **5**(4), 134–166.

Gaynor, Martin, Ho, Kate, & Town, Robert. 2014 (jan). *The Industrial Organization of Health Care Markets.* Working Paper 19800. National Bureau of Economic Research.

Gaynor, Martin, Propper, Carol, & Seiler, Stephan. 2016. Free to Choose? Reform, Choice, and Consideration Sets in the English National Health Service. *American Economic Review*, **106**(11), 3521–3557.

Grieco, Paul L E, & McDevitt, Ryan C. 2016. Productivity and quality in health care: Evidence from the dialysis industry. *The Review of Economic Studies*, **84**(3), 1071–1105.

Gutacker, Nils, Siciliani, Luigi, Moscelli, Giuseppe, & Gravelle, Hugh. 2016. Choice of hospital: Which type of quality matters? *Journal of health economics*, **50**, 230–246.

Hackmann, Martin B. 2019. Incentivizing Better Quality of Care: The Role of Medicaid and Competition in the Nursing Home Industry. *American Economic Review*, **109**(5), 1684–1716.

Ho, Kate, Hogan, Joseph, & Scott Morton, Fiona. 2017. The impact of consumer inattention on insurer pricing in the Medicare Part D program. *The RAND Journal of Economics*, **48**(4), 877–905.

Kelly, Elaine, & Stoye, George. 2016. *New joints: private providers and rising demand in the English National Health Service.* Tech. rept. IFS Working Papers.

Lancsar, Emily, Gu, Yuanyuan, Gyrd-Hansen, Dorte, Butler, Jim, Ratcliffe, Julie, Bulfone, Liliana, & Donaldson, Cam. 2020a. The relative value of different QALY types. *Journal of Health Economics*, 102303.

Lancsar, Emily, Gu, Yuanyuan, Gyrd-Hansen, Dorte, Butler, Jim, Ratcliffe, Julie, Bulfone, Liliana, & Donaldson, Cam. 2020b. The relative value of different QALY types. *Journal of Health Economics*, 102303.

Oakley, Ben, Nightingale, Jessica, Moran, Christopher G, & Moppett, Iain K. 2017. Does achieving the best practice tariff improve outcomes in hip fracture patients? An observational cohort study. *BMJ open*, **7**(2), e014190.

Olley, G Steven, & Pakes, Ariel. 1992. *The dynamics of productivity in the telecommunications equipment industry.* Tech. rept. National Bureau of Economic Research.

Pennington, Mark, Baker, Rachel, Brouwer, Werner, Mason, Helen, Hansen, Dorte Gyrd, Robinson, Angela, Donaldson, Cam, & Team, EuroVaQ. 2015. Comparing WTP values of different types of QALY gain elicited from the general public. *Health economics*, **24**(3), 280–293.

Propper, Carol. 1995. The disutility of time spent on the United Kingdom's National Health Service waiting lists. *Journal of Human Resources*, 677–700.

Santos, Rita, Gravelle, Hugh, & Propper, Carol. 2017. Does quality affect patients' choice of doctor? Evidence from England. *The Economic Journal*, **127**(600), 445–494.

Small, Kenneth A, & Rosen, Harvey S. 1981. Applied welfare economics with discrete choice models. *Econometrica: Journal of the Econometric Society*, 105–130.

Spence, Michael;. 1975. Monopoly, quality, and regulation.

White, Lawrence. 1972. Quality Variation When Prices Are Regulated. *Bell Journal of Economics*, **3**(2), 425–436.

8 Appendix A - Comparing an increase in uniform prices and quality-based prices (rewards)

In this session I compare graphically and with the help of some delta algebra the welfare effects of uniform prices *v.* quality-based prices (rewards). The effectiveness of quality-based prices depends on the correlation between cost efficiency and competition as well as on the curvature and slope of the cost functions. In this section I want to show how the comparative performance of the two price regimes depends on these different factors. For the purpose of this exposition I abstract away from the presence of measurement error in quality.

There two main effects that arise from the use of quality-based prices. Direct effects that arise from targeting directly with rewards only the hospitals providing higher quality. Indirect effects, instead, arise from additional competition created by these prices. The hospital providing higher quality would be incentivized to increase quality, spurring competition from the other hospitals that are not rewarded. Some patients would change provider and if they create more welfare with the new provider, either because of externalities or lower costs, this would be welfare enhancing. Additionally, this would create an "un-rewarded" increase in quality.

To better understand these two mechanisms at play I show the market outcome of an equal increase in either the size of rewards τ or the size of uniform prices ϕ and compare their effect on welfare in two cases. The first case, that would cover the direct effects, is one where there only local monopolies in the NHS. The second case would cover the indirect effects and would show the effect of liberalizing a market with more than one hospital, letting patients the freedom of choice instead of limiting them to one hospital option.

$$W = \sum_{j=L,H} CS_j(z_j^*, q_j(z_j^*)) + \sum_{j=L,H} \Psi_{Ext}(z_j^*) + \sum_{j=L,H} v_j(z_j^*)$$

$$- \sum_{j=L,H} C_j(z_j^*, q_j(z_j^*)) - \lambda \sum_{j=L,H} [(1+\phi)\bar{p}q_j(z_j^*) + \tau\bar{p}q_j(z_j^*)\mathbb{1}(z_j > \bar{z})]$$

In the first case, I consider the case of two local monopolies with different MR_z and MC_z. Hospital H, defined such that $\tau \to z_H^* > \bar{z}\ \forall\tau \in T$, support τ and Hospital L, defined such that $\tau \to z_L^* < \bar{z}\ \forall\tau \in T$, support τ. Firstly, I increase τ by $\Delta\tau$ while keeping $\phi = 0$, then I increase ϕ by $\Delta\phi$ while keeping $\tau = 0$. Crucially, $\Delta\tau = \Delta\phi$. Hospital H is affected by $\Delta\tau$ and $\Delta\phi$ in the same way.

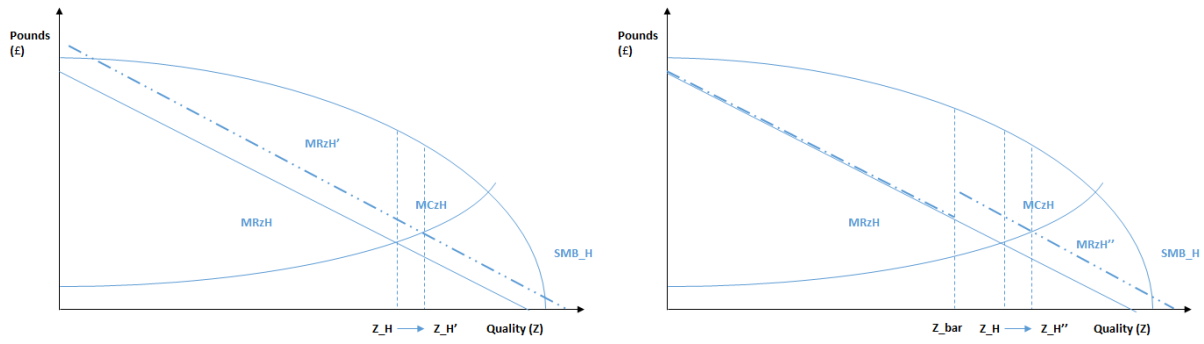

Figure 8: ϕ (uniform) increase for H τ (reward) increase for H

Then:

$$\frac{\Delta W}{\Delta \phi} > \frac{\Delta W}{\Delta \tau} \quad \rightarrow \quad \frac{\Delta W_L}{\Delta \phi} + \frac{\Delta W_H}{\Delta \phi} > \frac{\Delta W_L}{\Delta \tau} + \frac{\Delta W_H}{\Delta \tau} \quad \rightarrow \quad \frac{\Delta W_L}{\Delta \phi} > 0$$

$$\overbrace{\underbrace{\left(\frac{\Delta CS_L}{\Delta \phi} + \frac{\Delta v_L}{\Delta \phi} + \frac{\Delta \Psi_{Ext}(z_L^*)}{\Delta \phi}\right)}_{SMB_L \text{ w/ stock of patients}} + \underbrace{\frac{\Delta CS_L}{\Delta q_L}\frac{\Delta q_L}{\Delta z_L^*}}_{\text{Extra patients from outside op.}}}^{\text{Social Marginal Benefit} \times \frac{\Delta z_L^*}{\Delta \phi}} \Big) \frac{\Delta z_L^*}{\Delta \phi} > \overbrace{\left(\frac{\Delta C_L}{\Delta z_L^*} + \frac{\Delta C_L}{\Delta q_L}\frac{\Delta q_L}{\Delta z_L^*}\right)\frac{\Delta z_L^*}{\Delta \phi} + \underbrace{\left(\lambda \bar{p} q_L + \lambda \bar{p}\frac{\Delta q_L}{\Delta z_L^*}\frac{\Delta z_L^*}{\Delta \phi}\right)}_{\text{Additional gov exp}}}^{\Delta \text{ Social Marginal Cost}}$$

$$\underbrace{\left(\frac{\Delta CS_L}{\Delta \phi} + \frac{\Delta v_L}{\Delta \phi} + \frac{\Delta \Psi_{Ext}(z_L^*)}{\Delta \phi}\right)}_{SMB_L \text{ w/ stock of patients}} - \frac{\Delta C_L}{\Delta z_L^*}\Big)\frac{\Delta z_L^*}{\Delta \phi} + \underbrace{\left(\frac{\Delta CS_L}{\Delta q_L} - \frac{\Delta C_L}{\Delta q_L}\right)\frac{\Delta q_L}{\Delta z_L^*}\frac{\Delta z_L^*}{\Delta \phi}}_{\text{Extra patients from outside op.}} - \underbrace{\left(\lambda \bar{p} q_L + \lambda \bar{p}\frac{\Delta q_L}{\Delta z_L^*}\frac{\Delta z_L^*}{\Delta \phi}\right)}_{\text{Additional gov exp}} > 0$$

All the action is then with the lower quality hospital L, uniform prices deliver a greater welfare effect than the quality-based rewards if the additional quality delivered by L is creating net positive welfare. Graphically, this is the case if the shaded area in the graph (given by the by the social marginal benefit minus hospital costs) below is greater than the additional government expenditure needed to fund the price increase.

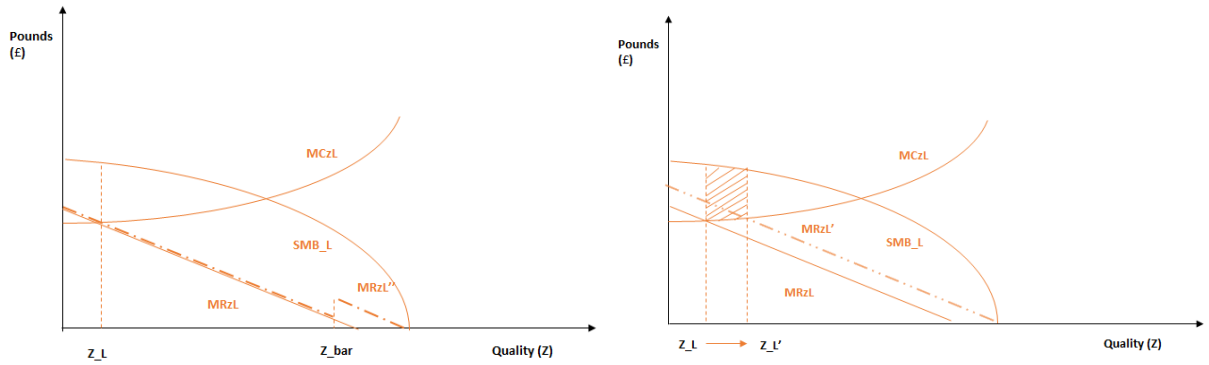

Figure 10: τ (reward) increase for L ϕ (uniform) increase for L

In the second case I show the effect of competition. Let's think of a situation where the previous local monopolies existed as monopolies because the NHS was not allowing patients in one county to go to the hospital in the other county. A liberalization of patient choice would allow for competition. Hospital L would now being able to attract less patients (also from the outside option) while hospital H would be able to attract more. The reason for this is that H has higher quality and some patients will switch to H (assuming that if there are other characteristics these are the same between the two hospitals before and after the liberalization). L creates now less welfare, while H creates more welfare. If the welfare created by H is greater than the loss by L then the re-allocation is welfare improving (graphically the difference between the two shaded areas minus the additional expenditure

created by the new patients that join H from the outside option). The patients who switch do so because they would get more utility from hospital H, but the welfare difference depends on the costs, the non-profit motives and externalities created by H compared to L. This analysis abstracts from the case of differences in other characteristics. If L has other characteristics which are different (e.g. waiting time) from H, it may also attract a net inflow patients after the liberalization. In that case, H would be able to attract less patients and would create less welfare, however the difference of the shaded areas would still give the difference in welfare (gross of the additional government expenditure for the new patients joining L from the outside option).

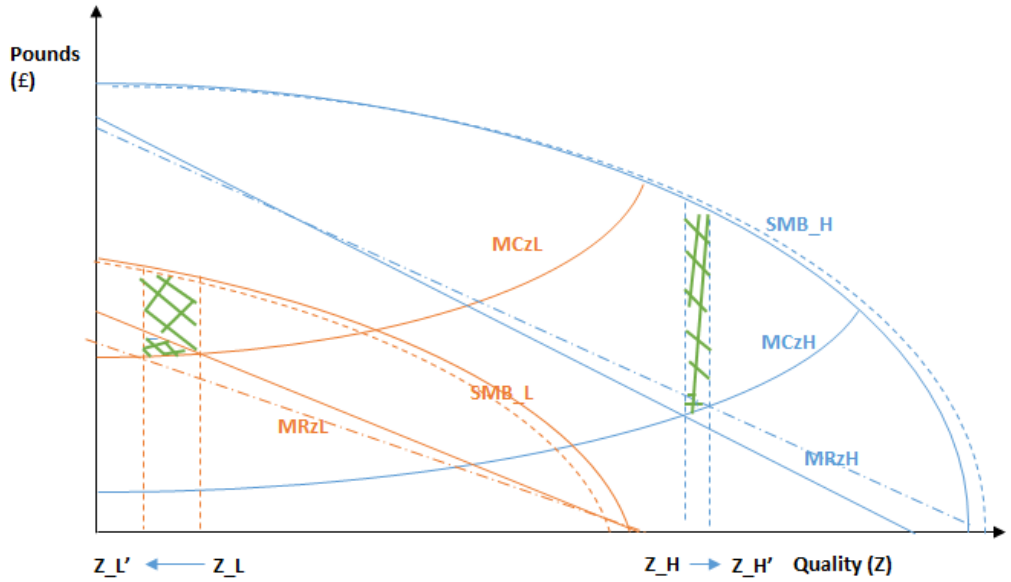

Figure 11: Effect of liberalization (no other hospital characteristic except quality)

Combining the two effects, welfare is affected by additional benefits from competition, as described at the beginning of this section. In a competitive environment Hospital H would increase quality further causing hospital L to match in part this quality increase (quality is a strategic complement). As L would not receive a reward, this would lead to an increase in quality for which the government would not have to pay. Additionally, this would lead to an additional re-allocation towards H, and, if the costs of H are lower or H generates more non-profit motives or externalities, this would have an additional positive effect on welfare.[61]

9 Appendix B - Production function and cost function duality

From the problem of the hospital choosing its inputs I can derive the cost function modeled in section 5. Consider the following maximization problem where the production function $F(K, L, M, \Omega) = K^{\beta_K} L^{\beta_L} M^{\beta_M} \Omega$. I omit for simplicity of notation the subscripts j, t

[61] A uniform price increase may also lead to a re-allocation toward one or the other provider depending on the MR_z and MC_z curves.

for hospital j and time t. Each hospital transforms inputs K,L,M in two outputs q, z according to the transformation function $T(q,z) = F(K,L,M,\Omega)$. K is capital, L is labor, M is material, C is cost, r, w, p are input prices and Ω is productivity.[62]

$$\max_{K,L,M} K^{\beta_K} L^{\beta_L} M^{\beta_M} \Omega$$
$$s.t.$$
$$C = rK + wL + pM$$

Let's define the lagrangian of the problem:

$$\mathcal{L} : K^{\beta_K} L^{\beta_L} M^{\beta_M} \Omega + \lambda_c (C - rK - wL - pM)$$

Setting the F.O.C.'s to zero:

$$\frac{\partial \mathcal{L}}{\partial K} : \beta_K \frac{K^{\beta_K}}{K} L^{\beta_L} M^{\beta_M} \Omega + \lambda_c r = 0 \tag{8}$$
$$\frac{\partial \mathcal{L}}{\partial L} : \beta_L K^{\beta_K} \frac{L^{\beta_L}}{L} M^{\beta_M} \Omega + \lambda_c w = 0 \tag{9}$$
$$\frac{\partial \mathcal{L}}{\partial M} : \beta_L K^{\beta_K} L^{\beta_L} \frac{M^{\beta_M}}{M} \Omega + \lambda_c p = 0 \tag{10}$$
$$\frac{\partial \mathcal{L}}{\partial \lambda_c} : C - rK - wL - pM = 0 \tag{11}$$

Re-arranging and dividing (5) by (6) and (5) by (7), I obtain:

$$\frac{(5)}{(6)} : \frac{r}{w} = \frac{\beta_K}{\beta_L} \frac{L}{K} \iff K = \frac{w}{r} \frac{\beta_K}{\beta_L} L \tag{12}$$
$$\frac{(5)}{(7)} : \frac{r}{p} = \frac{\beta_K}{\beta_M} \frac{M}{K} \iff K = \frac{p}{r} \frac{\beta_K}{\beta_M} M \tag{13}$$

Equating the above:

$$M = \frac{w}{p} \frac{\beta_M}{\beta_L} L \tag{14}$$

Substituting in $T(q,z) = K^{\beta_K} L^{\beta_L} M^{\beta_M} \Omega$:

[62] An equivalent formulation would include senior labor and junior labor and would not change anything for the estimation of the cost function. A similar reasoning would apply to different types of prostheses. However, while all levels of seniority of doctors are available and chosen by hospitals, hospitals choose only one type of prostheses. This will influence which input prices appear in the estimating equation, I discuss this further in the main body of the paper.

$$T(q,z) = \left(\frac{w}{r}\frac{\beta_K}{\beta_L}L\right)^{\beta_K} L^{\beta_L}\left(\frac{w}{p}\frac{\beta_M}{\beta_L}L\right)^{\beta_M}\Omega \tag{15}$$

$$\frac{T(q,z)}{\Omega} = \left(\frac{w}{r}\frac{\beta_K}{\beta_L}\right)^{\beta_K}\left(\frac{w}{p}\frac{\beta_M}{\beta_L}\right)^{\beta_M} L^{\beta_K+\beta_L+\beta_M} \tag{16}$$

$$L = \left(\frac{w}{r}\frac{\beta_K}{\beta_L}\right)^{-\frac{\beta_K}{\beta_K+\beta_L+\beta_M}}\left(\frac{w}{p}\frac{\beta_M}{\beta_L}\right)^{-\frac{\beta_M}{\beta_K+\beta_L+\beta_M}}\left(\frac{T(q,z)}{\Omega}\right)^{\frac{1}{\beta_K+\beta_L+\beta_M}} \tag{17}$$

Substituting (16) into (11) and (13):

$$K = \left(\frac{w}{r}\frac{\beta_K}{\beta_L}\right)^{\frac{\beta_L+\beta_M}{\beta_K+\beta_L+\beta_M}}\left(\frac{w}{p}\frac{\beta_M}{\beta_L}\right)^{-\frac{\beta_M}{\beta_K+\beta_L+\beta_M}}\left(\frac{T(q,z)}{\Omega}\right)^{\frac{1}{\beta_K+\beta_L+\beta_M}} \tag{18}$$

$$M = \left(\frac{w}{p}\frac{\beta_M}{\beta_L}\right)^{\frac{\beta_L+\beta_K}{\beta_K+\beta_L+\beta_M}}\left(\frac{w}{r}\frac{\beta_K}{\beta_L}\right)^{-\frac{\beta_K}{\beta_K+\beta_L+\beta_M}}\left(\frac{T(q,z)}{\Omega}\right)^{\frac{1}{\beta_K+\beta_L+\beta_M}} \tag{19}$$

Substituting (16), (17), and (18) in $C = rK + wL + pM$:

$$C = \left(\frac{1}{\Omega}\right)^{\frac{1}{\beta_K+\beta_L+\beta_M}}[r\left(\frac{w}{r}\frac{\beta_K}{\beta_L}\right)^{\frac{\beta_L+\beta_M}{\beta_K+\beta_L+\beta_M}}\left(\frac{w}{p}\frac{\beta_M}{\beta_L}\right)^{-\frac{\beta_M}{\beta_K+\beta_L+\beta_M}}$$
$$+w\left(\frac{w}{r}\frac{\beta_K}{\beta_L}\right)^{-\frac{\beta_K}{\beta_K+\beta_L+\beta_M}}\left(\frac{w}{p}\frac{\beta_M}{\beta_L}\right)^{-\frac{\beta_M}{\beta_K+\beta_L+\beta_M}}$$

$$+p\left(\frac{w}{p}\frac{\beta_M}{\beta_L}\right)^{\frac{\beta_L+\beta_K}{\beta_K+\beta_L+\beta_M}}\left(\frac{w}{r}\frac{\beta_K}{\beta_L}\right)^{-\frac{\beta_K}{\beta_K+\beta_L+\beta_M}}]T(q,z)^{\frac{1}{\beta_K+\beta_L+\beta_M}}$$

$$C = \left(\frac{1}{\Omega}\right)^{\frac{1}{\beta_K+\beta_L+\beta_M}}[r\left(\frac{w}{r}\frac{\beta_K}{\beta_L}\right)^{\frac{\beta_L+\beta_M}{\beta_K+\beta_L+\beta_M}}\left(\frac{w}{p}\frac{\beta_M}{\beta_L}\right)^{-\frac{\beta_M}{\beta_K+\beta_L+\beta_M}}\left(\frac{T(q,z)}{\Omega}\right)^{\frac{1}{\beta_K+\beta_L+\beta_M}}$$
$$+w\left(\frac{w}{r}\frac{\beta_K}{\beta_L}\right)^{-\frac{\beta_K}{\beta_K+\beta_L+\beta_M}}\left(\frac{w}{p}\frac{\beta_M}{\beta_L}\right)^{-\frac{\beta_M}{\beta_K+\beta_L+\beta_M}}$$

$$+p\left(\frac{w}{p}\frac{\beta_M}{\beta_L}\right)^{\frac{\beta_L+\beta_K}{\beta_K+\beta_L+\beta_M}}\left(\frac{w}{r}\frac{\beta_K}{\beta_L}\right)^{-\frac{\beta_K}{\beta_K+\beta_L+\beta_M}}]T(q,z)^{\frac{1}{\beta_K+\beta_L+\beta_M}}$$

Re-arranging, where $\sum_b \beta_b = \beta_K + \beta_L + \beta_M$:

$$C = \left(\frac{1}{\Omega}\right)^{\frac{1}{\sum_b \beta_b}}[w^{\frac{\beta_L}{\sum_b \beta_b}} r^{\frac{\beta_K}{\sum_b \beta_b}} p^{\frac{\beta_M}{\sum_b \beta_b}}(\frac{\beta_M^{\frac{\beta_L+\beta_K}{\sum_b \beta_b}}}{\beta_L^{\frac{\beta_L}{\sum_b \beta_b}}+\beta_K^{\frac{\beta_K}{\sum_b \beta_b}}}$$
$$+\frac{\beta_K^{\frac{\beta_L+\beta_M}{\sum_b \beta_b}}}{\beta_L^{\frac{\beta_L}{\sum_b \beta_b}}+\beta_M^{\frac{\beta_M}{\sum_b \beta_b}}}+\frac{\beta_L^{\frac{\beta_K+\beta_M}{\sum_b \beta_b}}}{\beta_K^{\frac{\beta_K}{\sum_b \beta_b}}+\beta_M^{\frac{\beta_M}{\sum_b \beta_b}}})]T(q,z)^{\frac{1}{\sum_b \beta_b}}$$

By taking the logs of the expression above I obtain the equation presented in section 5, where g and h are functions of input prices and $\beta's$ respectively, and $log(h(\beta_L, \beta_K, \beta_M))$ is included in the constant term α_0:

$$log(C) = log(T(q,z)^{\frac{1}{\sum_b \beta_b}}) + log(g(Input\ prices)) + log(h(\beta_L, \beta_K, \beta_M)) + log\left(\frac{1}{\Omega}\right)^{\frac{1}{\sum_b \beta_b}}$$

If $T(q,z) = qz^{\tilde{\alpha}_{z1}} exp^{\tilde{\alpha}_{q2}(log(q))^2} exp^{\tilde{\alpha}_{z2}(log(z))^2}$ then $\alpha_{q1} = \frac{1}{\sum_b \beta_b}$ and the equation above corresponds to equation (5). Equation (5) can also be seen as a second order Taylor expansion of an unknown function of q, z where the derivative of the interaction term of q, z is assumed to be zero. To retrieve productivity the only assumption necessary is that q multiplies all other elements in $T(q,z)$ and its exponent is normalized to 1. In this way $\alpha_{q1} = \frac{1}{\sum_b \beta_b}$ and it is possible to back out Ω knowing ω from equation (5), because $\omega = log\left(\frac{1}{\Omega}\right)^{\frac{1}{\sum_b \beta_b}}$.

10 Appendix C - Qualitative analysis

As part of this study, I have also contacted orthopedic doctors in charge of rehabilitation and surgery to test some of my assumptions. My intent was to make sure I correctly understood the decisions of doctors and patients as well as the main facts regarding hip replacement surgeries. Here, I highlight two points that are useful in the following analysis.

First, the contacted doctors agreed that the most important element in a successful hip replacement is the quality of the surgery and not the rehabilitation afterwards. Therefore, it is not surprising that the NHS decided to create quality-based prices for the payment for the surgery.

Secondly, in light of these interviews with doctors and reading documents from regulators, hospitals and medical scholars, I individuated the following sources of quality: better prostheses, more and better staff (doctors and also nurses) and better processes. In line with Hackmann (2019) and Grieco & McDevitt (2016) it is also true that more staff per patient will lead to higher quality, but it is not the only factor, e.g. better trained doctors and nurses will also lead to better outcomes as well as a better organization. In particular, doctors with more seniority can lead to better outcomes because, usually, the surgical techniques needed are perfected over decades of experience. In contrast, costs can be higher not only because of higher quality, but also because of inefficiencies. They can be bad personnel organization, late diagnosis of a problem due to lower level of monitoring of patients' condition, excessively long stays due either to mis-coordination or preventable complications.

11 Appendix D - Quality choices additional material

Comparative statics: In Figure 8 the MVQ or Marginal Revenue of Quality (MR(z)') is shifted upwards (to MR(z)") or downwards (to MR(z)) by a higher or lower regulated price $\bar{p}$ and by the intensity of the non-profit motives $\frac{\partial v}{\partial z_j}$.

Hospital choice

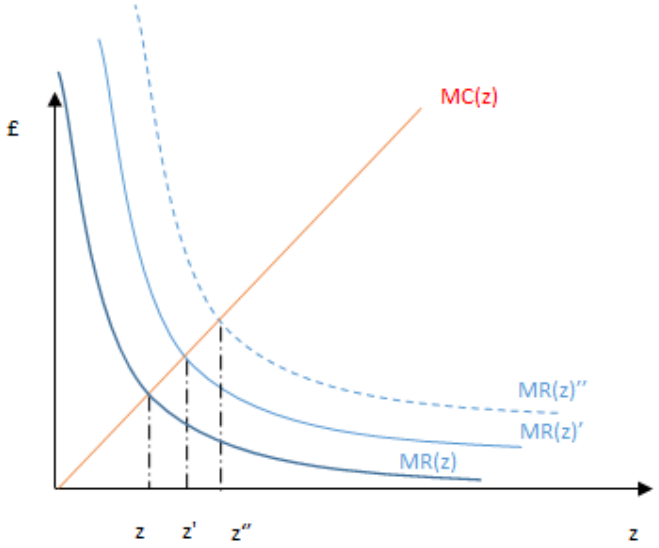

Figure 12: Marginal value of quality and marginal cost of quality z

12 Appendix E - Counterfactual welfare decomposition

In this section I present the a decomposition of the welfare effects of the different price reforms studied in my counterfactuals. It can be seen that the role of externalities and non-profit motives is important. The reason for this is the discrepancy between the QALY value attached to hip replacement surgeries and the preference parameters for quality of the patients. This can be due to lack of information about the future benefits of the surgeries or behavioral biases. Additionally, as discussed in the main body of the paper, improved mobility is bound to reduce government expenditure for patients receiving surgeries as well as positively affect family and personal decisions about work and consumption.

Table 7: Total welfare under different reform scenarios. The cost of raising public funds $\lambda = 0.3$. The last column "ΔTax Cost" to see distortionary cost of taxation.

	Δwelfare in GBP (£)					
	ΔCS	$\Delta\Psi(z)$	$\Delta v(z)$	ΔGov. Exp.	ΔTax Cost	ΔWelfare
Panel A: Uniform price increases						
$+7\%\bar{p}$	7.6m	5.0m	3.9m	21.6m	6.5m	3.9m
$+15\%\bar{p}$	16.1m	10.5m	8.4m	46.8m	14.1m	7.7m
Panel B: Rewards for high quality						
$+10\%\bar{p}$ above 1 std from median z	12.3m	8.1m	6.2m	21.0m	6.3m	8.9m
$+20\%\bar{p}$ above 1 std from median z	22.9m	14.9m	11.7m	45.9m	13.7m	14.4m
Panel C: Punishments for low quality						
$-10\%\bar{p}$ below 1 std from median z	-15.1m	-10.0m	-9.0m	-23.4m	-7.0 m	-13.7m

13 Appendix F - Demand estimation including distance from centroids

Distance is considered to be very important in patients decisions. If I had data on patients addresses or on hospitals market shares for people living in each Census tract ("MSOA") I would use a Berry logit model (Berry (1994)) to retrieve the parameters for distance. Unfortunately, I do not have this data, but I do have data on the number of people (I focus on 55 years and older) living in each tract and on the number of patients going to each hospital j. I can use this information to construct the equation below (as Elickson et al. (2019)).

$$q_{jt} = \sum_{n \in M_j} \psi_n pop_{nt} s_{njt} + \xi_{jt} \quad \text{where} \quad s_{njt} = \frac{e^{\delta_{njt}}}{1 + \sum_{u \in J_n} e^{\delta_{nut}}} \tag{20}$$

In the equation we have the following additional elements. J_n is the set of hospitals in the choice set of individuals living in Census tract n (one of 12 geographical suddivisions of England).[63] M_j is the set of Census tracts included in the catchment area of hospital j, this set is also defined using the 12 geographical regions.[64] ψ_n is instead a national prevalence rate for hisp replacements among people aged 55 and older, which is used as an approximation for the actual tract-specific prevalence rate. pop_{nt} is the population of people aged 55 and older of tract n at time t. ξ_{jt} captures firm specific error term which can be interpreted as measurement error or as an unexpected demand shock.[65] Finally, I can write the formula for the market share of hospital j in tract n in the way spelled out above based on Berry (1994).

Now I can use the formula in (20) instead of the usual Berry logit formula to retrieve the parameters of interest. In particular, I use non-linear least squares to estimate equation (3) minimizing the difference between the observed q_{jt} and the predicted equivalent.

[63] We perform robustness checks with 5km and 10km radiuses.
[64] We actually differentiate between urban catchment areas and rural catchment areas.
[65] This shock will be kept constant in the subsequent counterfactual analysis.

Table 8: Demand estimates following Ellickson et al. (2020)

Dependent variable: from logit model		
	(1)	(2)
quality ($\widetilde{z}$)	0.26	0.09
	(0.13)	(0.04)
distance	-1.92	-0.08
	(0.85)	(0.00)
percentage wait more than 120 days	0.59	-0.02
	(0.86)	(0.34)
General reputation - survey	0.31	0.08
	(0.13)	(0.01)
Cons	-23.9	-10.5
	(9.97)	(1.15)
Year fixed effects	Yes	Yes
Market fixed effects	No	Yes
No. obs	579,539	579,539

Note: standard errors in parentheses.

14 Appendix G - Demand estimation with heterogeneity in parameters

In this section I show the results for two groups of patients: patients with and without co-morbidities. I show the results of the counterfactuals for the two categories of patients, particularly looking at the consumer surplus.

Table 9: Percentage of the gap covered with consumer welfare under social optimum

		+20m	+50m	+80m	+110m	+140m	+170m
Panel A: Patients without complexity	Higher $\bar{p}$	6	13	20	28	36	43
	Rewards above median	9	17	26	34	42	50
Panel B:Patients with complexity	Higher $\bar{p}$	6	13	20	29	36	44
	Rewards above median	9	18	27	35	42	51

Note: Monetary equivalent of utility based on Propper (1990)

15 Appendix H - Cost estimation following ACF (2015)

As additional robustness check I estimated the cost function following Ackerberg *et al.* (2015), but adapted to a cost function.

Schematically, the procedure follows the same logic as in a production function:

- **1st stage:** Estimate the regression above and $\widehat{C}$
- Retrieve $\hat{\omega(\theta)} = \widehat{C} - C(\alpha' s_{guessed})$ with $C(\alpha' s_{guessed})$ I mean the part of the cost function with the alpha's parameters -i.e., terms of quality and quantity. The initial guess is relatively close to OLS.
- Assume an AR(1) process for ω -i.e., $\omega_t = \rho\omega_{t-1} + \xi_t$- to retrieve the structural error ξ_t. This is the only dynamic aspect of the model, it is nevertheless important, as persistency in productivity implies that hospitals cannot increase quality as they want by simply hiring more or better staff. They are limited by their productivity in the previous period. Crucially, I assume hospitals do not control productivity.
- **2nd stage:** Construct the moment conditions $\mathbb{E}[\xi_t \mathbf{h_{t-1}}]$ and $\mathbb{E}[\xi_t \mathbf{s_t}]$ between the structural error ξ and the instruments $\mathbf{s_t}$, $\mathbf{h_{t-1}}$: $q_{t-1}, \widehat{z_{t-1}}$ (and further lagged terms and interactions included in $\mathbf{h_{t-1}}$), as well as quality shifters.

Notably, given that I assume that the measurement errors are not serially correlated using instruments also allows me to control for biases due to measurement errors.

Table 10: Cost function estimates

Dependent variable: total cost					
Const	q	z	q^2	z^2	zq
97.333	0.999	-15.081	0.006	3.040	-0.018
(61.889)	(0.140)	(9.053)	(0.007)	(1.814)	(0.047)

Note: block bootstrap standard deviation in parenthesis

Table 11: Reform analysis. (1) hospital and year fixed effects, (2) hospital and time-smaller areas fixed effects, (3) hospital and time-larger areas fixed effects. Clustered standard errors at the hospital level.

	Dependent variable: quality		
	(1)	(2)	(3)
γ_0^p	-0.02	-0.018	-0.02
	(0.009)	(0.009)	(0.009)
γ_1^p	0.1208	0.1189	0.1208
	(0.006)	(0.006)	(0.006)
Const.	0.2624	0.2788	0.2739
	(0.006)	(0.015)	(0.009)
R^2	0.87	0.93	0.87
Obs.	749	749	749

16 Appendix I -Evidence from pilot reform

In this section, I show evidence from a pilot 2014 reform by the NHS that was designed to punish (with lower regulated prices) hospitals providing quality below a certain threshold (three standard deviation below the national average). In particular, I show that post-reform the average quality provided declined -compared to the improvement in the control-, even if only to a small extent.

To estimate the effect of the reform I use a difference-in-differences strategy. I use primary hip and knee replacements ('primary' as in first intervention for the patient) as treatment group, because the NHS reform targets these surgeries and as control group hip and knee replacement revisions, not affected by the reform. These surgeries are typically performed years after the first intervention, or "primary hip replacement", and are meant to replace the previous prosthesis affected by wear-and-tear with a new one.

$$Quality_{jkt} = \gamma_0^p \mathbb{1}Reform_{2014,t} * \mathbb{1}Primary + \gamma_1^p \mathbb{1}Primary + H_j + \eta_t^{area} + \epsilon_{jkt}$$

The same doctor and staffs may perform both types of procedures, so there may be positive spillovers between the two types of surgeries. For this reason, the results should be interpreted more as a lower bound than the actual effect of the reform.

The effect of the reform is very small, around a 5% reduction in quality with respect to the baseline. In this analysis I cannot determine the welfare effects that are behind these results or understand the possible drawbacks and benefits of different possible quality-based prices. To this end I develop a structural model and simulate the welfare effects of different designs of quality-based prices. In my simulations I can also individuate which hospitals are driving the net effects and through which mechanism.

17 Appendix J - Reforms with lower non-profit motives

In the table below I present the results from a series of counterfactuals setting lower values for the non-profit motives. The percentage reduction reduces their importance as well as their heterogeneity.

The increased effectiveness of the quality-based rewards is driven by the fact that non-profit motives are less heterogeneous in the different scenarios in panel A. In the baseline scenario non-profit motives compensate for differences in market power and productivity. As their importance declines, quality based prices can partially compensate for these differences, while uniform prices cannot perform equally well.

Table 12: % welfare gap to social optimum for different levels of non-profit motives

	$100\%\frac{\partial v}{\partial z}$	$90\%\frac{\partial v}{\partial z}$	$75\%\frac{\partial v}{\partial z}$	$50\%\frac{\partial v}{\partial z}$	$25\%\frac{\partial v}{\partial z}$	$10\%\frac{\partial v}{\partial z}$
Panel A: With non-profit motives heterogeneity						
+7% $\bar{p}$ uniform price increase	6.1%	6.2%	6.3%	6.6%	6.8%	6.9%
+10%$\bar{p}$ above 1 st. dev. from mean z	13.6%	14.4%	16.0%	16.7%	17.2%	18.0%

Note: cost of public funds are assumed 30% of raised funds (Poterba, 1996), marginal externality valued at £150