UNIVERSITY OF GOTHENBURG
SCHOOL OF BUSINESS, ECONOMICS AND LAW

Medical Mark-Up

A Check-Up on the Quality-Competition in Primary Care

Michael Erlandsson **Markus Ljungqvist**

Abstract

We show that quality-competition among private firms in the Swedish primary care have worsened during the last 15 years. As a result, public funds turn into profits, rather than quality care for patients. Our interpretation is that providers have not increased their (cost of) quality at the same rate as the compensation has increased. We measure the intensity of quality-competition with markups, which we retrieve from a production function, as De Loecker and Warzynski (2012). Since this is the first application of this methodology to estimate quality competition, we adapt the method by including quality in the Ackerberg, Caves, and Frazer (2015) two-stage estimator to control for time-variant unobserved firm quality and productivity.

Supervisor: Johan Stennek
Master's Thesis in Economics (30 ECTS), Spring 2022

Graduate School
School of Business, Economics and Law
University of Gothenburg, Sweden

Preface

We want to dedicate this thesis in "loving" memory of the many drafts and the hundreds of pages laid to rest which enabled us to produce this thesis. On a more sincere note, we want to thank our supervisor Johan Stennek, for his enthusiasm, attention to detail and ability to ask interesting questions. His intricate understanding of competition shaped the theoretical foundation of this thesis – quality competition. We also want to thank Florin Maican for spending his free-time to help us with our research proposal, which later became this thesis. In particular, his extensive knowledge of empirical applications in industrial organization guided us towards our empirical framework. It was the vigorous marking of our thesis in red pen, by both Johan and Florin, that enabled us to develop into better economists and writers. We are forever grateful for their extraordinary commitment to our growth, which we will carry with us in our future challenges!

Contents

List of Figures

List of Tables

1 Introduction

We provide evidence that quality-competition among private firms active in Swedish primary care have worsened over the last 15 years. Specifically, we show that private firms earn higher markups over time. In 2006, private firms hardly earned a markup while in 2020, the reimbursement rate was roughly 30% greater than the firms' marginal costs. This result indicates that public funds, at an increasing rate, turn in to private profits instead of qualitative care for the citizens.

The deregulation of public healthcare sectors is becoming ever more widespread due to, among other things, the perceived lack of efficiency in the provision of welfare services (Gaynor, Ho, and Town, 2015). Thus, governments allow private firms to enter healthcare markets and receive reimbursements for their services. The idea is that, in order to attract patients, providers must raise quality and thereby engage in quality-competition (Vengberg, Fredriksson, and Winblad, 2019). This may be a solution to the fact that quality is observable, but unreliable to measure and therefore impossible to enforce by contract or regulation.

It appears reasonable to assume that patients in health care, at least partially, select caregivers based on their qualities. Quality is, for example, choosing a convenient location, conducting more thorough appointments or extending opening hours. Nordgren and Ahgren (2010) found that these types of factors are important when patients choose their provider. If all patients switched to the firm with the highest quality, quality-competition would be extremely intense. Then, firms would offer as much quality as they can afford and profits would be minimal.

That Sweden relies on quality-competition to limit the profitability of private firms active in the deregulated primary care is clearly articulated in a statement by the Swedish Competition Authority (*Dnr 710/2016* 2017). What makes Swedish primary care unique, from an international perspective, is that there are no formal limitations on profitability of private firms (Välfärdsutredningen, 2016). The possibility of earning rents from public funds in primary care, which previously went exclusively to public non-profit providers, has been a source of controversy. The debate is to a large extent based on ideology and theory. Our aim is to provide some hard evidence.

We first theoretically show that if patients' choice of provider is insensitive

to quality, firms will earn a markup[1] since they are able to set the marginal cost lower than the reimbursement. Second, we study the intensity of quality-competition among private for-profit firms in the Swedish primary care market by applying the methodology of De Loecker and Warzynski (2012). Firms in the Swedish primary care sector do not set prices. Instead, private care-providers compete by raising their (cost of) quality to attract patients.

The De Loecker and Warzynski (2012) methodology relies on an estimation of a trans-log production function to calculate annual firm-level markups, utilizing widely available accounting data. This is possible since the marginal cost can be expressed as the ratio of wages and the marginal product of labor, which is obtained from an estimated production function. The estimation relies on previous methods developed in the production function estimation literature which control for productivity differences across firms, which vary over time. Specifically, we apply the two-stage estimator of Ackerberg, Caves, and Frazer (2015) which is a correction of Olley and Pakes (1996) and Levinsohn and Petrin (2003) that allow for adjustment costs of labor.

Our application of the methodology to the primary care market requires certain modifications to the estimator. Most notably, we consider the case when (to us) unobserved quality, in addition to productivity, is as a source of firm heterogeneity which we control for. We base our adaption of the methodology on the work of Doraszelski and Jaumandreu (2013) and Maican and Orth (2017) who incorporates factors which affect firms' ability to predict their future productivity. More precisely, we incorporate a proxy for average firm quality, which firm take into account when predicting their future quality and productivity.

In short, we contribute to the literature with an innovative application of the methodology where we measure the intensity of quality-competition by evaluating annual firm level markups. To our knowledge, there are no similar applications of this methodology to evaluate quality-competition, nor have quality-competition been evaluated in primary care over time.

We find that the markups of private firms active in the Swedish primary care market have increased significantly between 2005 and 2012. The markups stabilized at higher levels between 2012 and 2019, and then reached an all-time high in 2020. We find the most plausible reason for worsened quality-competition to be that the reimbursement have increased at a greater rate than that of the (cost of) quality. The reason might be that consumers are less sensitive to changes in quality at higher levels of quality. We suggest that future policy reforms target patients' sensitivity to changes in quality, rather than increases in reimbursement.

[1] We define markups as the ratio of the administered price and the marginal cost.

2 Literature Review

Previous literature evaluates quality-competition in the Swedish primary care in three main ways. These are accounting profit, market concentration and policy reform studies. Our study is different in that we directly measure quality-competition using estimated markups.

The Swedish Competition Authority (2014) obtained accounting data for Swedish primary care centers in 2012. They evaluated the accounting profits for these centers and found that approximately 40% of primary care centers did not make an accounting profit. The issue with these types of studies, is the assumption that accounting costs reflects the marginal costs. This is problematic, since accounting costs includes costs that are not related to the service provision. Instead, our thesis use estimated marginal costs, which alleviate this concern.

Another type of study evaluates the effect of market concentration on quality. The idea is that markets with more active firms should have more intense competition, which should raise quality. These types of studies often use aggregated measures of quality, such as ambulatory care sensitive cases, to measure quality. Dietrichson, Ellegard, and Kjellsson (2020) found that care center concentration only had a modest effect on quality of care in the Swedish primary care. This is a valid approach, under the assumption that firms are homogenous. Our study is different in that we do not assume firm homogeneity, since we measure quality-competition directly with firm-level markups.

Policy reforms, which aim to enhance competition in healthcare, are a common source of variation for studies of quality-competition (Handel and Ho, 2021). The idea is that by enabling firms to compete for patients, there should be competition. There are several studies conducted of this type, which conclude that higher degrees of patient choice associates with higher levels of quality (Gaynor, Ho, and Town, 2015; Gaynor, Moreno-Serra, and Propper, 2013; Gaynor, Propper, and Seiler, 2016). For our research scope, this approach is problematic since it relies on the assumption that the ability to compete implies competition. Specifically, we are interested in evaluating the intensity of competition, which is affected by several factors, rather than the effect of a policy reform.

One factor which inhibits patients' ability to choose the provider with the highest quality, and is therefore a source of market power, is informational frictions.

Anell et al. (2021) found that there are informational frictions in the Swedish primary care which prevents patients from changing provider. This indicates that regardless of firm concentration, the quality-competition might not work as intended.

We use an alternative approach, compared to previous studies, to evaluate the intensity of quality-competition. This study fills the presented gap in the literature by contributing with an innovative method to evaluate how well quality-competition functions, and thereby if public funds turn into private profits instead of qualitative care.

3 Primary Care in Sweden

In Sweden, 85%[1] of healthcare spending stems from public funding and 14% is funded by out-of-pocket spending, which is close to the EU average of 15%[2] (OECD, Health Systems, and Policies, 2021). The healthcare expenditures have risen with approximately 50% in the last 10 year, as illustrated in figure 3.1 (OECD, 2019). Alongside the increasing expenditure of the last 10 years, the Swedish healthcare sector underwent large deregulation.

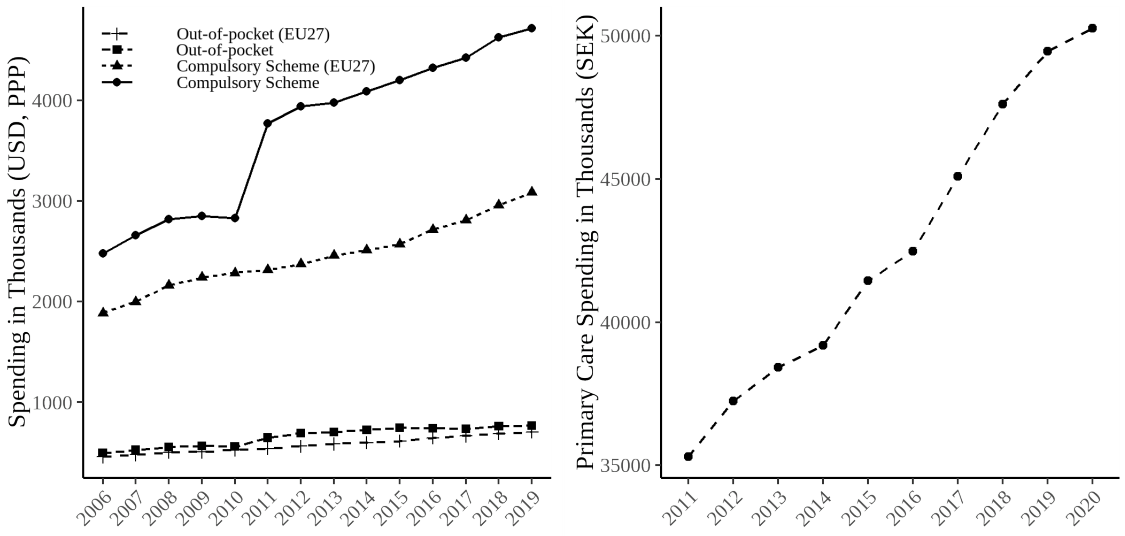

Figure 3.1: Healthcare Spending

The different roles of primary care in Sweden is to provide basic medical treatment, rehabilitation, preventative care and act as gatekeepers for hospitals (Anell, 2011). Despite the largely decentralized system, all the 21 regions provide primary care with similar structure, such as using a capitation[3] and per-visit reimbursement (Swedish Agency for Health and Care Services Analysis, 2015).

Private providers make up a large share of the primary care sector, relative to other health care sectors in Sweden. In 2017, privately owned centers made up 43% of all centers and 45% of all doctor's visits were to private providers (Swedish

[1] Only 1% of spending allocates to voluntary or private insurance schemes.

[2] The data is from 2019

[3] Capitation based systems means that a provider gets reimbursed for listed patients rather than per treatment basis.

Agency for Health and Care Services Analysis, 2015). Figure 3.2 shows that visits to private providers[4] increase steadily, while visits to public providers decrease slightly over time.

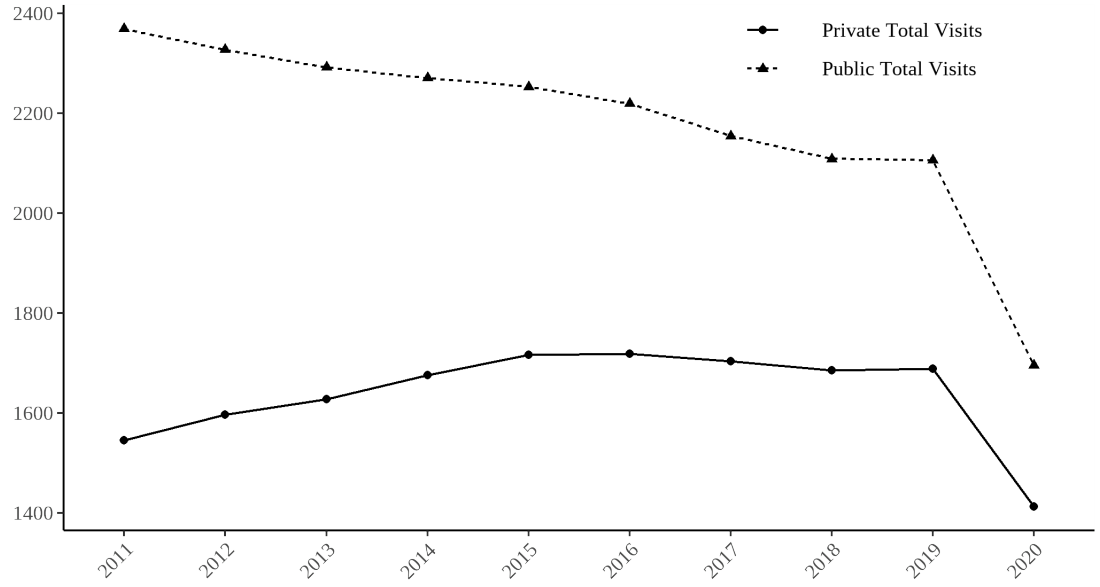

Figure 3.2: Total Visits to Public and Private Providers over Time in Primary Care

The increased private presence in primary care stems from the introduction of the Act on System of Choice in the Public Sector (henceforth, LOV), which became mandatory on a national scale in 2010. However, LOV was introduced by some regions prior to it becoming mandatory, due to the decentralized nature of the system. The introduction of LOV allowed patients to freely choose among providers and change provider an unlimited amount of times. By allowing patients to choose providers, providers attract patients by increasing the quality of their care.

In order to incentivize the entry of private firms on the market, private firms are able to earn a profit. However, prior to entry, a firm must fulfill regional requirements[5] to ensure a minimum level of quality (Anell, 2011). From an international perspective, these requirements for entry can be viewed as relatively liberal (Anell, 2015). Since the first introduction of patient choice in Halland[6] in 2007, the number of private primary care centers have rapidly increased, as illustrated in figure 3.3.

[4] Visits are to all professional categories within primary care

[5] Some regions require a level of capital, opening hours and educated staff. There is however variation between regions in these requirements.

[6] A region in Sweden.

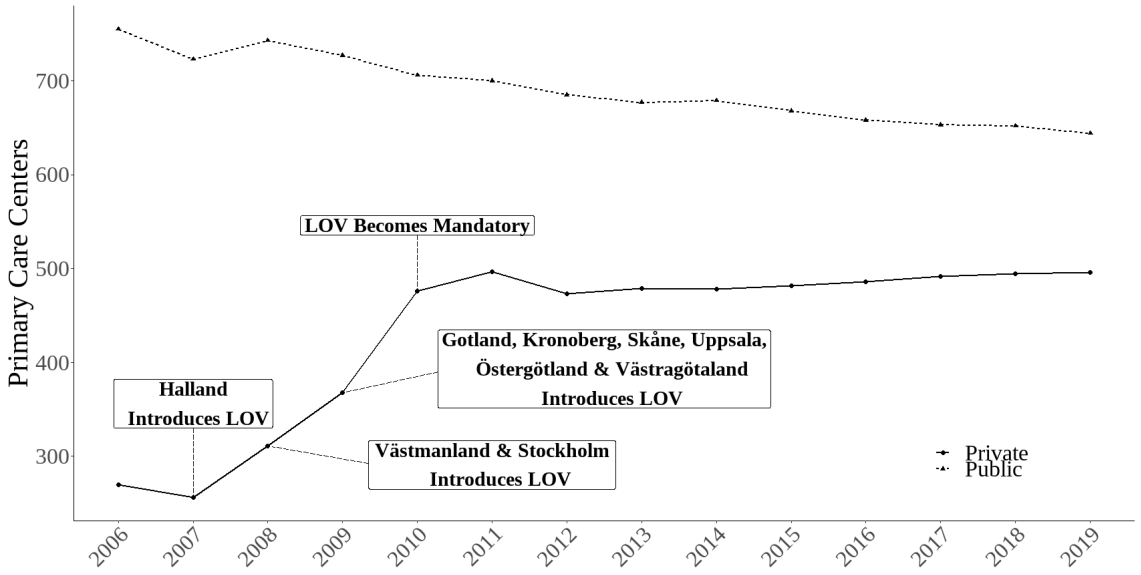

Figure 3.3: LOV Introduction and Number of Primary Care Centers over Time

One of the main motivations for patient choice was to increase quality by quality-competition (Anell, 2011). Since patients do not bear the direct cost of a visit, they should instead choose the provider with the highest quality. Providers should therefore raise quality in order to attract patients.

Since 2009, all regions in Sweden measures the perceived quality in primary care (Swedish Association of Local Authorities and Regions, 2022). Currently, the regions evaluate perceived quality through the national patient survey, which measures quality in 7 dimensions. The motivation behind the dimensions is that they represent important and sought after aspects within primary care, by both patients and providers. An example of an indicator is continuity, which is an important factor for both patients choosing a provider, and to lower emergency visits. The other dimensions are availability, compassion, information, knowledge, engagement as well as overall impressions. One important aspect to note here is that the survey uses patients' subjective opinion to measure quality. Perceived quality might not capture the actual medical quality, since there is a certain limitation to patients' ability to accurately evaluate medical quality of care.

There are, however, also "objective" measures of the performance of the primary care system, such as the rate of Ambulatory Care Sensitive Conditions (ACSC). These are cases, such as diabetes or heart failure, which can be prevented by proper primary care[7] (Rosano et al., 2013). If the rate of ACSC is lower, it indicates a higher quality primary care.

Figure 3.4 shows the rate of ACSC was relatively steady prior to 2011, ranging between 1900 and 2000 cases per 100000 inhabitants. In 2012, the rate of ACSC started to decline and reached an all-time low in 2020. In this respect, the quality

[7] There are several definitions of ACSC, we follow the definition by SKR (2018)

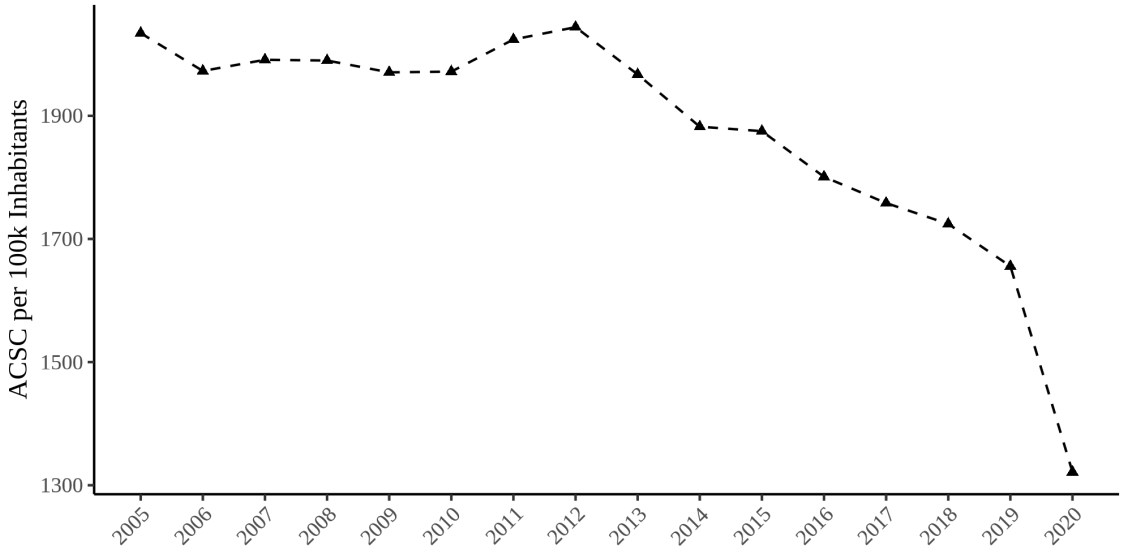

Figure 3.4: Evolution of Ambulatory Care Sensitive Cases (ACSC) in Sweden

of primary care is greater than ever.

However, figure 3.1 illustrates that the expenditure on healthcare and primary care is also greater than ever. The record low rate of ACSC do therefore not imply that quality-competition have reached an all-time high. It is entirely possible that firm quality have not risen in proportion with the reimbursement. This is the topic of the present paper.

4 Quality-Competition

Private firms in the Swedish primary care receive reimbursements set by the regional authorities. The reimbursement also means that patients do not directly bear the cost of their treatment. Patients will therefore decide which provider to visit based on other factors than price, such as quality.

The regions are unable to reliably measure quality, beyond ensuring some minimal thresholds, therefore the reimbursement is independent of quality. However, firms still have an incentive to increase their quality of care in order to attract more patients. Our definition of quality is wide, and do not only include clinical quality. Quality could, for example, be to make a more thorough appointment and cover additional issues in the same visit. Another example is to hire more physicians than absolutely necessary in order to decrease waiting times. Quality could also be as simple as refurbishing the care center by painting the walls or choosing a better location.

While quality always comes at a cost, it is also what motivates patients to choose one provider over the other. The firm must therefore balance the potential of attracting additional patients by raising the quality with its cost, in order to maximize profit. If patients do not react to an increase in quality, firms have no incentive to increase quality.

Consider a primary care market with a single firm and patients who will need to make visits to said firm to satisfy their healthcare demand. This market resembles the primary care which existed prior to the patient choice reform, with only public provision. The patient is perfectly insensitive to changes in quality, since the single active firm is the only choice. Therefore, the firm has no incentive to set quality at higher levels, and quality will be held at a relatively low levels unless there is some intrinsic motivation to set higher quality.

Now, consider that an additional firm enters the market. If all patients select the provider with the highest quality, independent of how small the differences are, the firms will set quality at the highest possible level. In other words, the firms will raise their quality of care until the marginal cost of treating the patients is equal to the reimbursement. This is "Bertrand competition" in quality.

Finally, consider that the patients do not have perfect information, and that switching as well as travel costs are present. Then the patients would be less sensitive

to changes in quality, which leads patients to not only focus on quality when choosing a caregiver. The insensitivity to quality changes will make it profitable for firms to lower quality so that the marginal cost of serving a patient is lower than the reimbursement. Should the patient be perfectly insensitive to changes in quality, the patients will choose a random firm. This implies that both firms will set their quality as in a monopoly setting. The intuition behind this is that increasing quality will not attract the patients, making the quality-competition non-existent.

More formally, consider a firm (i) active in the Swedish primary care market at year (t) that produces a quantity of enrollments and visits (Q_{it}). The firm then employs labor (L_{it}), which is variable. Besides labor, the production requires capital (K_{it}) which is fixed in the short-run. Furthermore, the firm has a level of productivity (ρ_{it}), that affects the firm's input decisions. The firm also determine a level of quality (ν_{it}) which affect the input decisions as well. Thus, we characterize the general production technology as:

$$\nu_{it} \cdot Q_{it} = \rho_{it} \cdot f(K_{it}, L_{it}). \tag{4.1}$$

The left-hand side of 4.1 ($\nu_{it} \cdot Q_{it}$) can, to a certain extent, be interpreted as the total amount of QALYs produced. In this case, ν_{it} is interpreted as the amount of QALY produced for each visit.

As previously stated, quality always comes at a cost and is therefore a shifter of the production function in our framework. For example, by spending more time with each patient (and thereby increasing quality), the production requires more capital and labor.

Note that changes in quality has its own associated marginal cost, which is different from the marginal cost of quantity. Formally, these the two marginal costs are related:

Lemma. *Assume the production technology to be characterized by 4.1. Then, a firm's marginal cost of quality is related to its marginal cost of quantity.*[1]*:*

$$MC_\nu = \frac{MC_Q}{\nu_{it}} \cdot Q_{it} \tag{4.2}$$

An important aspect of this market is that even though producers have different quality, they do provide similar products. Patients will choose between a provider, and it's competitors, based on their qualities. For simplicity, we let a firm's demand depend on the average quality of the firms in the market ($\bar{\nu}_t$) and its own quality (ν_{it}):

$$Q_{it} = D_{it}(\nu_{it}, \bar{\nu}_t). \tag{4.3}$$

The average firm quality ($\bar{\nu}_t$) affect the patient's choice, since the quality of

[1] See A.1 in Appendix A for proof.

the single firm will be evaluated in the context of other firms. If the quality of the market increases, but the quality of the firm do not change, the firm will attract fewer customers than previously.

Each caregiver sets its quality to maximize profit. We represent the firm's decision with the following profit-maximization problem:

$$\Pi_{it}(\nu_{it}, \bar{\nu}_t) = P_t \cdot D_{it}(\nu_{it}, \bar{\nu}_t) - C_{it}(D_{it}(\nu_{it}, \bar{\nu}_t), \nu_{it}). \tag{4.4}$$

Since patients choice of provider depend on firms' quality (ν_{it}), firms will compete against each other by increasing their quality and thereby marginal costs. If patients are very sensitive to changes in quality, all patients will select the provider with the highest quality. Firms will then increase their quality until the marginal cost of serving the patient is equal to the reimbursement, $P_t(\cdot) = MC_Q(\cdot)$. If firms are able to profit, the patients are to a certain degree insensitive to changes in quality. Thus, competition crucially depends on the quality elasticity of demand, which we define by $\tau_{it} = \frac{\partial Q_{it}(\nu_{it}, z_{it})}{\partial \nu_{it}} \cdot \frac{\nu_{it}}{Q}$. Formally:

Proposition. *A firm's ability to retain markup is determined by how sensitive patients are to changes in provider quality*[2]*:*

$$\frac{P_t}{MC_Q} = 1 + \tau_{it}^{-1}. \tag{4.5}$$

If patients in the primary care market are highly sensitive to changes in quality, firms will have low market power. However, if patients in the primary care market are highly insensitive to changes in quality, then the market power of the firm will be high.

Patient's sensitivity to changes in quality is affected by, for example, how easy it is to change provider. High levels of switching and travel costs or information frictions makes it more difficult to change provider. Thus, in markets with high information friction or travel costs, the consumers will be less sensitive to changes in quality. In this sense, what determines firms' ability to lower their marginal costs below the reimbursement is factors such as travel costs and information frictions, which is captured in the quality elasticity of demand.

[2] See A.2 in Appendix A for proof.

5 Empirical Framework

5.1 Estimation and Identification

We estimate a trans-log production function to retrieve annual markups for each firm. We base our methodology on the De Loecker and Warzynski (2012) framework. By estimating a trans-log production function, we are able to retrieve firm-level output elasticity of labor. By combining this with accounting data of the cost share of labor, we are able to solve for the ratio between the reimbursement and marginal cost, i.e., the markup[1].

When estimating production functions, researchers in the productivity literature assumes that productivity is the only unobserved factor which affect input decision. Productivity is, among other things, good managerial ability which enables the firm to hire less personnel while maintaining the same output. However, we consider the case when quality, in addition to productivity, is unobservable and time-variant. Quality could, for example, be continuity of care, which requires the firm to hire additional personnel to spend more time with patients. In this respect, both productivity and quality are shifters of the production function.

These unobservable factors are an issue when estimating production functions with OLS, since issues of endogeneity will bias the estimator (Marschak and Andrews, 1944). Furthermore, since these unobserved factors vary over time they cannot be controlled for by firm fixed effects, originally suggested by Hoch (1955), which only absorb time-invariant firm heterogeneity.

Olley and Pakes (1996) suggested an estimation in which time-variant unobserved heterogeneity, in the form of productivity, is controlled for by an investment demand function. Levinsohn and Petrin (2003) altered the estimation, by replacing the investment demand function with an intermediate input demand function to control for productivity. Finally, the correction suggested by Ackerberg, Caves, and Frazer (2015) accounts for adjustment costs in labor, which solve the functional dependency problem of the two previous estimators. Therefore, we apply the Ackerberg, Caves, and Frazer (2015) estimator.

The Ackerberg, Caves, and Frazer (2015) estimator build on the assumption that firms are, to a certain extent, able to predict their productivity in the next pe-

[1] see A.3 in Appendix A for derivation

riod. In these applications, firms are assumed to only account for their current state of productivity when predicting their future productivity. In the frontier of the productivity literature, econometricians instead assume that firms take other factors in to account when predicting their productivity (De Loecker and Syverson, 2021). For example, Doraszelski and Jaumandreu (2013) incorporates investments in research and development in firms' prediction. There are also more sophisticated models, such as Maican and Orth (2017), where the authors incorporate factors such as firm entry in the firms' prediction of future productivity. By incorporating these factors, firms' productivity instead follows an endogenous Markov process.

Since we incorporate quality in the production function, in addition to productivity, we cannot assume that productivity and quality follows an exogenous Markov process due to quality-competition. Firms active in markets where the average firm quality is higher must set their quality at a higher level in order to attract patients. Firms will therefore evaluate the average firm quality in the current period in order to predict their quality in the next period, which they base their current input decisions on.

We assume, as Maican and Orth (2017), that firms take their predicted productivity into account when deciding how much labor to employ. However, we extend the assumption, so that firms also incorporate their quality in the hiring decision. By inverting a labor demand function, we are therefore able to proxy for the firms' quality and productivity in order to obtain the firms' predicted output. This is the first of the two stages in the Ackerberg, Caves, and Frazer (2015) two-stage estimation.

The second stage use the predicted output to estimate the firms' predicted productivity and quality. However, firms are unable to perfectly predict their future quality and productivity. Shocks, such as advancements in technology or policy reforms, will affect firms' actual productivity and quality. By retrieving the predicted quality and productivity, we are therefore able to obtain the unpredicted shocks as well.

We use the correlation between estimated shocks to productivity and quality with the inputs to form the moment conditions of a Generalized Methods of Moment estimation. Specifically, we assume that capital cannot be adjusted in the short run in reaction to shocks in productivity and quality. Additionally, we assume that labor is able to adjust in reaction to these shocks and is therefore considered variable. We therefore use the lag of labor as an instrument for the current labor when estimating the second stage, which is standard in this literature.

More formally, we consider the following trans-log production function for firms active in the Swedish primary care market:

$$q_{it} = \beta_l l_{it} + \beta_{ll} l_{it}^2 + \beta_k k_{it} + \beta_{kk} k_{it}^2 + \beta_{lk} l_{it} k_{it} + \omega_{it} + \epsilon_{it}. \tag{5.1}$$

We denote the log of quantity for firm i at time t as q_{it}. The log of labor is l_{it} and the log of capital is k_{it}. To a certain extent, we can view both ω_{it} and ϵ_{it} as a part of the unobserved error term. The key difference is that ϵ_{it} is exogenous and uncorrelated with the input choices, while ω_{it} is endogenous and affect input choices. We model ω_{it} as the log-ratio of productivity (ρ_{it}) and quality (ν_{it}), i.e $\omega_{it} = ln\left(\frac{\rho_{it}}{\nu_{it}}\right)$.

In order to control for ω_{it} we apply, as mentioned, the Ackerberg, Caves, and Frazer (2015) two-step estimator. The purpose of the first step is to obtain an estimate of $\hat{q}_{it}$, which we interpret as the output the firms expect to obtain using its production technology. In order to control for ω_{it} in the first stage, we use an inverted labor demand function, as suggested by Maican and Orth (2017) with their application in the service sector. The firms are assumed to have knowledge of ω_{it} and base their hiring decisions on it, alongside other factors. By inverting the labor demand function, we are able to approximate the log ratio of productivity and quality (ω_{it}) by the observable variables of the inverted labor demand function[2]. More precisely, we consider the following inverted labor demand function to approximate ω_{it}:

$$\omega_{it} = l_t^{-1}(l_{it}, k_{it}, w_{it}) = l_t^{-1}(\cdot), \tag{5.2}$$

where we denote the log of wages as w_{it}. The functional form of the inverted labor demand function ($l_t^{-1}(\cdot)$) is unknown. Therefore, we approximate it with a third degree polynomial expansion[3]. Specifically, we run the following OLS regression at the first stage in order to obtain the firms' expected output, $\hat{q}_{it}$:

$$q_{it} = \phi_t(l_{it}, k_{it}, w_{it}) + \epsilon_{it}. \tag{5.3}$$

We then use the estimates in 5.3 to calculate the predicted output $\hat{q}_{it}$, which is given by $\hat{q}_{it} = q_{it} - \hat{\epsilon}_{it}$. This is an important step, since firms make input choices based on predicted output and not the actual output, which is unknown at the time of making input decisions. Furthermore, this step enables us to calculate ω_{it} as a function of our estimates, by rearranging the production function in 5.1 as:

$$\omega_{it}(\beta) = \hat{q}_{it} - \beta_l l_{it} + \beta_{ll} l_{it}^2 + \beta_k k_{it} + \beta_{kk} k_{it}^2 + \beta_{lk} l_{it} k_{it}. \tag{5.4}$$

Above concludes the last part of the first stage in our estimation. The second stage is a Generalized Method of Moments estimation of 5.1. We base the moment conditions of the estimation on timing and behavior assumption of firms' prediction of their future quality and productivity. Before proceeding with the second stage, we will clarify these assumptions.

First, firms observe their productivity and quality in the current period, ω_{it-1}.

[2] Assuming the sign of each partial derivative of the labor demand function have the same sign, i.e., strict monotonicity.

[3] This approach is standard in the literature, see De Loecker and Syverson (2021).

Our addition is that firms also observe the average firm quality of the market $\bar{\nu}_{t-1}$, in order to predict what level their productivity and quality will be in the next period (t). Thus, the firms' prediction follows an endogenous Markov process, not an exogenous Markov process[4]. Second, firms make the prediction, which we estimate in equation 5.5 using OLS, and makes input decisions based on their prediction. Third, when the firms are in the future period (t), there will be exogenous and random shocks (ξ_{it})[5] to the firms' prediction. These shocks will make the firms' input decisions suboptimal, since the firms' predicted quality and productivity is not equal to the actual outcome. The firms will therefore adjust the variable inputs, while the fixed inputs will remain the same as decided in $t-1$.

We now proceed with the second stage of the estimation procedure. First, we estimate the firms' predicted productivity and quality using OLS:

$$\omega_{it} = \gamma_0 + \gamma_1 \omega_{it-1} + \gamma_2 \omega_{it-1}^2 + + \gamma_3 \omega_{it-1}^3 + \gamma_4 \bar{\nu}_{t-1} + \xi_{it}. \tag{5.5}$$

Firms are unable to perfectly predict their future productivity and quality, due to shocks (ξ_{it}). By estimating 5.5 we obtain an estimate of ξ_{it}, and as mentioned, we use these shocks to formulate our moment conditions in the Generalized Methods of Moment estimation.

Returning to our timing assumptions, we assume capital to be fixed in the short run. Capital is therefore unable to react to the shocks to productivity and quality, ξ_{it}. Labor is, however, variable and will be able to react to shocks in productivity and quality. We therefore use the lag of labor as an instrument for current labor when forming our moment conditions[6]. The lag of labor is highly correlated with current labor and, due to our timing assumption, uncorrelated with shocks to productivity and quality. Thus, the moment conditions formed for the Generalized Methods of Moments estimation are:

$$E\left(\xi_{it}(\boldsymbol{\beta}) \begin{pmatrix} l_{it-1} \\ k_{it} \\ l_{it-1}^2 \\ k_{it}^2 \\ l_{it-1}k_{it} \end{pmatrix}\right) = 0. \tag{5.6}$$

Lastly, we use the estimates from the second stage of the estimation procedure to calculate the annual markups of each firm[7]. Due to our timing assumption, we calculate the labors cost share (α_{it}) using the predicted output $(\hat{q}_{it})$, instead of the actual output. By taking the partial derivative of 5.1 and multiplying with the

[4] The difference between an endogenous and an exogenous Markov process is best illustrated by altering 5.5 to the following exogenous form: $\omega_{it} = \gamma_0 + \gamma_1 \omega_{it-1} + \gamma_2 \omega_{it-1}^2 + \gamma_3 \omega_{it-1}^3 + \xi_{it}$

[5] Such shocks can be adaptation of new technology or sick workers and more.

[6] The standard in the literature is to use the lag of the variable input, labor in our case, as an instrument De Loecker and Syverson (2021).

[7] Derived in A.3 of appendix A

inverse of the labors cost share (α_{it}^{-1}), we obtain the markup for each firm over time:

$$\frac{\partial q_{it}}{\partial l_{it}} \cdot \alpha^{-1} = \frac{P_t}{MC_Q}. \tag{5.7}$$

5.2 Data

We gather our main data-set from Retriever Business, where we select all firms within the Swedish Standard Industrial Classification code 86.211 "general primary care medical practice activities with doctors". The dataset contains yearly accounting data for all firms.

In table 5.1, we illustrate the descriptive statistics of the relevant variables for our purposes. We also include notation and how it is measured in the table. The table includes the mean, quantiles, median and standard deviation to illustrate the distribution and variation in the variables of interest.

Table 5.1: Descriptive Statistics

Measure	Notation	Mean	Median	Q25	Q75	SD
Net Sales*	Q	126493	20458	8652	38796	770029
N. of Employees	L	133	16	7	31	829
Total Assets*	K	67572	6326	3211	12159	519843
Personnel Costs*	WL	75747	9746	3936	19662	462660
ACSC**	$\bar{\nu}$	1812	1875	1724	1972	204
Input Share Labor	α	0.5141	0.5136	0.4643	0.5621	0.0978

* In 1000 SEK, ** Cases per 100 000 inhabitants

Output, Q_{it}, is hard to measure in healthcare markets since patients demand health and not doctors visits. Therefore, we measure output as net sales, which is one of the most accurately measured variables in accounting (De Loecker and Syverson, 2021). Since regional authorities administer prices, firms cannot directly influence them.

Labor (L) is measured as number of employees. Wages, W_{it} is measured by dividing the personnel costs, $W_{it}L_{it}$ with labor. Capital, K, is measured with total fixed assets. The input share of labor, α is defined as net personnel cost over net sales.

We also use data on Ambulatory Care Sensitive Conditions (ACSC)[8] from the Swedish inpatient register. We use the ACSC per 100 000 inhabitants to proxy the average firm quality, $\bar{\nu}_t$. The rate of ACSC is considered an objective measure of primary care quality and is often used in empirical applications to evaluate the performance of primary care (Rosano et al., 2013). A lower rate of ACSC indicate

[8] ACSC is a set of conditions which could be prevented with proper primary care. We consider the Swedish definition of ACSC, defined by SKR (2018)

a higher quality primary care system since these are hospitalizations which could have be prevented by the primary care.

The two data-sets are originally in a wide format, which we transform to panel data format and merge. We omit observation where net sales is missing, firms with less than three employees, capital less than 1000 SEK and personnel cost less than 1000 SEK annually. This is done since our methodology requires non-zero values when minimizing the objective function. Additionally, the estimation procedure also requires us to introduce lagged variables, which further impose restrictions on the data-set. Our final data set contains 2456 observations of 359 firm spanning between 2005 and 2020. Notice that the data-set is unbalanced, since a balanced data-set would introduce selection bias because the observed firms would only be those who existed during the entire period.

Since there is no R software package for our implementation of the production function estimation, we alter the source code of Rovigatti (2017) to our needs. Specifically, our alteration allow the inclusion of control variables in the endogenous Markov process of the unobserved firm variables. Furthermore, we calculate block bootstrapped standard error of the Generalized Methods of Moments estimator. Each block consists of 100 firms randomly drawn, with replacement, 1000 times. With each sample, we estimate our model specification and calculate the parameters. This creates a distribution of estimates, from which we calculate the standard error of the estimates. Finally, we generate all LaTeX tables with the package of Hlavac (2018).

6 Results & Analysis

6.1 Production Function

We estimate the production function: $q_{it} = \beta_l l_{it} + \beta_{ll} l_{it}^2 + \beta_k k_{it} + \beta_{kk} k_{it}^2 + \beta_{kl} k_{it} l_{it} + \omega_{it} + \epsilon_{it}$, where ω_{it} denotes the log ratio of productivity and quality. Recall that this ratio is unobserved by the econometrican and varies over time for each firm and affect the firm's input decision (unobserved time-variant heterogeneity). We control for the unobserved time-variant heterogeneity with the Ackerberg, Caves, and Frazer (2015) estimator. Specifically, we include the rate of ACSC to proxy for the average firm quality in model (1).

We present two specifications in this section. First, we present model (2), where we assume that the unobserved heterogeneity evolves according to an exogenous Markov process, i.e., it only depends on the previous period's state. If the rate of ACSC is not included, the estimates of the production function, and therefore markups, should be biased. Second, model (1), where we assume that the unobserved heterogeneity evolves according to an endogenous Markov process, where it depends on the previous state and the previous period's rate of ACSC.

Since we estimate a trans-log production function, the marginal effect of the inputs will depend on the level of capital and labor a firm has. Specifically, we obtain the marginal effect of capital and the marginal effect of labor by taking the partial derivative of the production function with respect to labor and the partial derivative with respect to capital:

$$\frac{\partial q_{it}}{\partial l_{it}} = \beta_l + 2 \cdot \beta_{ll} \cdot l_{it} + \beta_{lk} \cdot k_{it} \tag{6.1}$$

$$\frac{\partial q_{it}}{\partial k_{it}} = \beta_k + 2 \cdot \beta_{kk} \cdot k_{it} + \beta_{lk} \cdot l_{it} \tag{6.2}$$

Because the production function is in logs, we interpret the marginal effects as output elasticities, i.e., a 1% increase in labor will lead to a $(\partial q_{it}/\partial l_{it})$% increase of output and a 1% increase of capital will lead to a $(\partial q_{it}/\partial k_{it})$% of output.

In table 6.1, we illustrate the estimated production function, when we assume that quality and productivity follows an endogenous Markov process and when we assume an exogenous Markov process.

Table 6.1: Regression Results

	Endogenous* (1)	Exogenous* (2)
β_l	1.1270	1.1255
	(0.5438)	(0.5462)
β_{ll}	-0.0364	-0.0338
	(0.0966)	(0.1017)
β_k	-0.0072	-0.0087
	(0.0980)	(0.1007)
β_{kk}	0.0397	0.0636
	(0.5261)	(0.5563)
β_{kl}	0.0123	0.0112
	(0.0366)	(0.0384)

Note: We illustrate the distribution of the block bootstrapped standard errors of model (1) in Figure B.1.
* Block bootstrapped standard errors in parentheses.

In order to compare the performance of the models, we set labor and capital to the mean value and calculate the output elasticity at the mean. The output elasticity at the mean for our primary model, (1) is 0.99 for labor and 0.79 for capital. Compared to model (2), the output elasticity at the mean is 1.03 for labor and 1.15 for capital. Primary care is a service sector, and we should therefore expect the output elasticity of labor to be higher than that of capital. Model (2) estimate an output elasticity of capital which is higher than labor, which is dubious since labor should be more important in healthcare sectors. By controlling for endogeneity in quality and productivity, we are able to estimate a lower value for the output elasticity of capital.

6.2 Markups

We use the estimates of the production function to calculate the markups, as in equation 5.7. In Figure 6.1 we present the mean markup, weighted by the firm's market share[1]. The figure shows that markups follow a rising trend, with an initial increase followed by a period of markups between 1.1 and 1.25. This increasing trend show that the quality-competition among private firms active in the Swedish primary care market have decreased.

One interesting note is that the median markup, in Figure B.3 of Appendix B, is roughly 50 percentage points greater than the weighted mean markup. This indicates that larger firms might have lower markups than smaller firms. This

[1] We calculate the weighted mean as $\sum s_{it} \cdot \mu_{it}$, where s_{it} represents a firm's turnover at time t as a share of total turnover at time t.

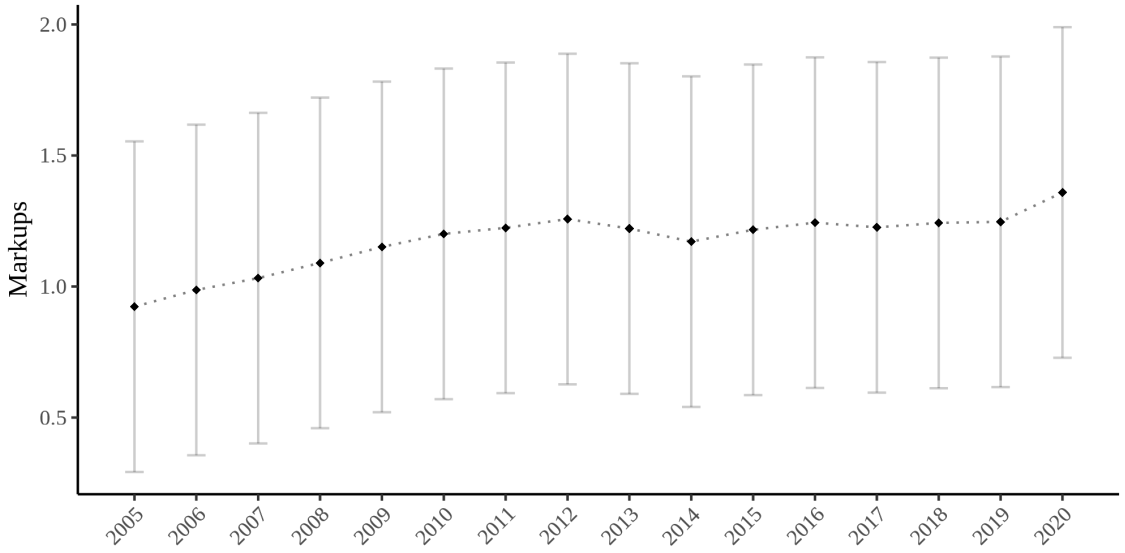

Figure 6.1: Turnover Weighted Mean Markups over Time
Error bars indicate one standard deviation from the mean

might be counterintuitive for some, since measures of market concentration (such as Herfindahl-Hirschman index) imply that firm size is as a source of market power. This type of heterogeneity in markups will have policy implications. It is therefore of interest to further evaluate the association between firm size and markup in the Swedish primary care market.

First, We group the firms by the quantiles of their turnover in Figure B.2 in Appendix B. The figure illustrates that firms which belong to the 0th quantile, of turnover, have markups which are about 100-150% greater than firms who belong to the 100th quantile. Furthermore, the 25th, 50th and 75th quantile ranges between the markups of the 0th and 100th quantile of turnover. This provides further evidence that larger firms, measured in turnover, have lower markups on average. However, the 100th and 75th quantile have an increasing trend, whereas the other are either decreasing or relatively stable.

Second, to further the investigation of the firm size's association with markups, we run a within-estimator regression. Specifically, we run the following regression:

$$ln(P_t/MC_{it}) = \alpha_0 + \alpha_1 ln(share) + F_i + F_t + \epsilon_{it}. \tag{6.3}$$

In regression 6.3 we control for both firm F_i and time F_t fixed effects to isolate the effect of the market share. The effect of firm size is statistically significant from 0, at the 1% level (see table B.1). This implies that a 1% increase in market share associates with a 0.065% decrease in markup. This indicates that larger firms, measured in market share, have a smaller markup on average.

Finally, we show the distribution of markups for each year in Figure B.4 of Appendix B. It seems as if the markup of the earlier years (2005-2008) is slightly

more scattered, than the markups of the years after. There are few, if any, outliers in markups for each year. This indicates that there is less heterogeneity in firms' markups in recent years.

6.3 Robustness

We show that the increasing trend in markups, which is our main result, are robust to changes in model specification. Specifically, we compare the results of four models. First, our primary model (1) in which we assume that productivity and quality follows an endogenous Markov process. Second, we present model (2) in which we assume that productivity and quality follows an exogenous Markov process. Third, in model (3) we apply time and firm fixed effects with the within estimator, which assumes that productivity and quality is constant over time for each firm. Fourth, model (4) is a OLS regression, which assumes that productivity and quality do not affect the input decisions.

In Figure 6.2 we present the estimated output elasticity of labor for each of the four model specifications. The estimated output elasticity of labor varies between the

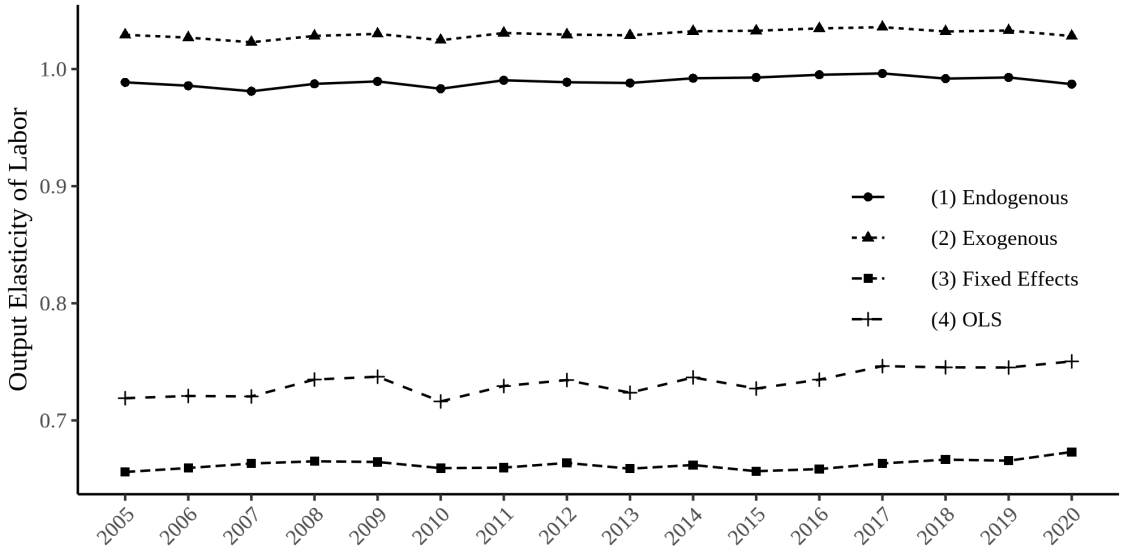

Figure 6.2: Mean Output Elasticity of Labor with Four Estimation Procedures

different model specification. However, the output elasticity is relatively stable over time for each of the estimators. The trend of the estimated output elasticity of labor does not seem to be as dependent on the model specification, while specification highly influences the level of output elasticity. There is a small difference in the estimated output elasticity of (1) and (2), which are at a higher level, than (3) and (4), which are closer to each other.

We calculate the markups as $P_t/MC_{it} = \partial q_{it}/\partial l_{it} \cdot \alpha_{it}^{-1}$. The labors cost share, α_{it} is observed from the data and is used to calculate the markups in (3) and (4).

In (1) and (2) we correct the labors cost share with the estimated output from the first stage of the Ackerberg, Caves, and Frazer (2015) estimator. The purpose of this correction is to adjust the output for exogenous shocks. This will therefore be a source of variation in the estimated markups between the different estimators, which is why we compare these two in figure 6.3.

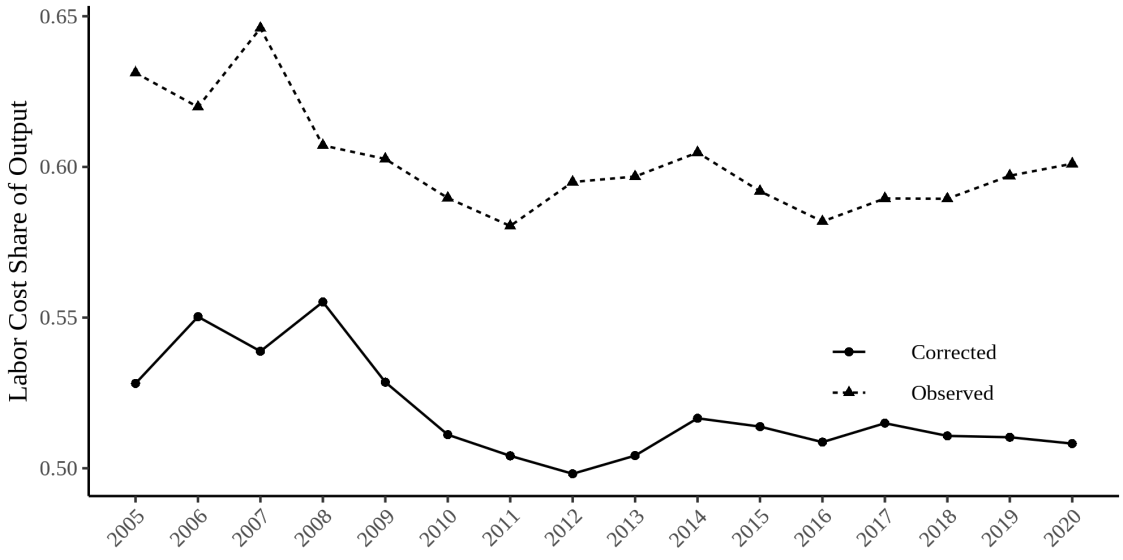

Figure 6.3: Labor Cost Share of Output

The corrected labor cost share of output is roughly 10 percentage points higher than the observed cost share. However, both follow a very similar pattern, which decreases slightly over time. The combined variation from the output elasticity of labor and the labors cost share of output make up the variation in markups, as seen in Figure 6.4. Figure 6.4 illustrates that all 4 estimators show an

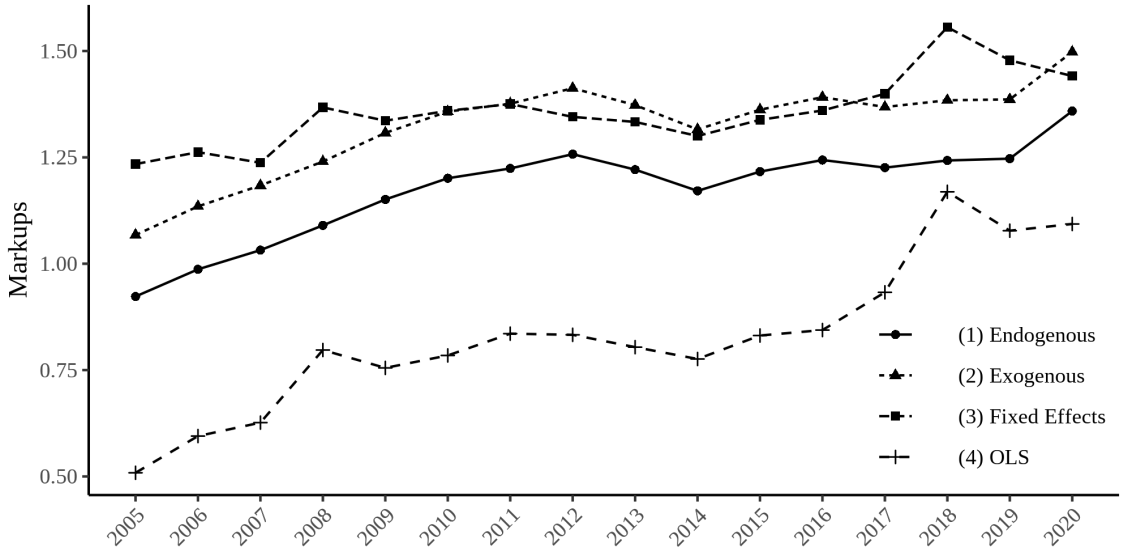

Figure 6.4: Multiple Turnover Weighted Mean Markups

increasing trend in estimated markups, although at different levels. Model (4) estimates markups which ranges between 0.5 and 1, while our main model (1) estimates markups ranging slightly below 1 to 1.25. Model (3) and model (2) perform very similarly and overlap each other in some years. However, the markups in (1), (2) and (3) do not differ much.

7 Discussion

We find that markups are increasing in the Swedish primary care market. Here, we investigate three possible explanations of why the markups are rising. These are decreasing quality, increases in productivity and increases in reimbursement.

Rising markups could be interpreted as firms providing lower quality care than previously. Markups consist of the reimbursement and the marginal cost. If markups increase, it could be because quality, and therefore marginal cost, have decreased. Testimonials from primary care providers show that providers engage in strategic behavior (Zaremba, 2013). Certain providers split treatments, which could have been a single appointment, into several appointments in order to increase profit. There is also a disbelief among physicians in the healthcare system's ability to deliver qualitative care (Swedish Agency for Health and Care Services Analysis, 2017). One potential source of the disbelief might be that providers engage in strategic behavior to maximize profits. However, we find it unlikely that the quality of care have decreased due to the strategic behavior since the rate of ACSC have decreased.

Another possible explanation for the rising trend of markup, is that productivity might have increased. The productivity of firms affect their markup since an increase in productivity, ceteris paribus, decrease the marginal cost. Regional authorities and primary care managers tend to favor measures such as resource-use and volume of production, rather than measures of quality, to evaluate the performance of primary care (Arvidsson, Dahlin, and Anell, 2021). The focus on productivity, as well as technical innovation, could mean that primary care providers have increased their productivity over the last 15 years. However, we are not able to measure the productivity of firms, and therefore we cannot confirm this explanation.

Finally, one explanation might be that expenditure on the Swedish primary care have increased during the last 15 years. While quality might also have increased, the reimbursement might have grown in a greater rate than the (cost of) quality. We are partially able to verify this scenario, since the rate of ACSC is decreasing and expenditure on primary care is increasing. Thus, we find this scenario the most plausible.

Firms' ability to turn an increase in reimbursement into a profit crucially depends on the patients' sensitivity to quality changes. High markups imply that patients are less sensitive to changes in quality. If we assume that quality has

increased, although at a lower rate than the reimbursement, one possible explanation of the rising markups is that patients are less sensitive to changes in quality at higher levels of markups. This would imply that it would be more beneficial to make investments to decrease information frictions, rather than increase reimbursement further.

There is great heterogeneity in firms' markup. One concrete example we find is that firms with a larger market share on average have lower markups. Since there is heterogeneity in firm markups, there is heterogeneity in the quality elasticity of demand. Effective policy reforms should therefore target firms where patients are less sensitive to changes in quality. It is therefore of interest to know where patients are less sensitive to changes in quality and other firm or market characteristics which influences patients sensitivity to changes in quality. Future research should further evaluate the heterogeneity of both markups and quality elasticity of demand. For example, it would be interesting to evaluate if firms active in rural areas have a greater or lower markup.

However, we are unable to infer if the level of markup is reasonable. The share weighted markups were around 1.3 in 2020, which imply that the reimbursement is around 30% greater than the marginal cost of the private firms active in the Swedish primary care. The markup should cover additional costs, beyond the marginal cost, such as fixed costs or startup costs. Differences between industries in these costs makes us unable to compare these markups with findings from the previous literature. Cross-industry comparison with sectors that are able to set their own prices and have vastly different services or products is problematic, since individual sectors' characteristics will influence the level of markups. This is, to our knowledge, the first study which estimates markups to evaluate quality-competition in a healthcare market. Future studies could compare markups between different healthcare sectors to evaluate if the level of markups are "justified".

To conclude, this study relies on accounting data of private firms active in the Swedish primary care. It would be interesting to conduct a similar study using data at the center level, with more detailed input, wage, and enrollment data. Our application use the bare minimum data, therefore we consider the Swedish primary care market a national market. With availability of center data, it would be possible to evaluate differences in markup between regions. If the data included coordinates for the centers, then distance between centers could be used in order to define even more relevant markets. Data at the center level is not available today, since all regions collect different data at the center level. It is of great interest to collect this type of data in order to evaluate the public spending on primary care.

8 Concluding Remarks

We find that quality-competition have worsened during the last 15 years. We present three explanations of why quality-competition have worsened, where the most plausible explanation is that the reimbursement have increased at a more rapid rate than the (cost of) quality.

The increasing expenditure on primary care should to a greater extent target information frictions, rather than increases in reimbursement. The great heterogeneity of markups suggests that it would be beneficial for reforms to target firms with especially high markups. For example, we find that smaller firms earn greater markups, on average. Policy reforms could therefore target smaller firm, since they earn higher markups. However, we do not evaluate all characteristics of firms with high market power. It is therefore of great interest for future research to evaluate which additional firm characteristics, and markets, associate with higher markups.

Our unique application of the De Loecker and Warzynski (2012) framework opens up a new avenue to evaluate quality-competition. Our work can be viewed as a first step towards a new approach to evaluate quality-competition, and could potentially be employed in sectors besides primary care. One example of an interesting application could be in the Swedish school market, since it also relies on quality-competition to limit profitability.

References

Ackerberg, Kevin Caves, and Garth Frazer (2015). "Identification Properties of Recent Production Function Estimators". eng. In: *Econometrica* 83(6), pp. 2411–2451. ISSN: 0012-9682.

Anell, Anders (2011). "Choice and privatisation in Swedish primary care". eng. In: *Health economics, policy and law* 6(4), pp. 549–569. ISSN: 1744-1331.

Anell, Anders (2015). "The Public–Private Pendulum — Patient Choice and Equity in Sweden". eng. In: *The New England journal of medicine* 372(1), pp. 1–4. ISSN: 0028-4793.

Anell, Anders et al. (2021). "Information, switching costs, and consumer choice: Evidence from two randomised field experiments in Swedish primary health care & nbsp". eng. In: *Journal of public economics* 196. ISSN: 0047-2727.

Arvidsson, Eva, Sofia Dahlin, and Anders Anell (2021). "Conditions and barriers for quality improvement work: a qualitative study of how professionals and health centre managers experience audit and feedback practices in Swedish primary care". eng. In: *BMC Family Practice* 22(1), pp. 1–113. ISSN: 1471-2296.

De Loecker, Jan and Chad Syverson (2021). "An industrial organization perspective on productivity". In: *Handbook of Industrial Organization*. Vol. 4. 1. Elsevier, pp. 141–223.

De Loecker, Jan and Frederic Warzynski (2012). "Markups and Firm-Level Export Status". eng. In: *The American economic review* 102(6), pp. 2437–2471. ISSN: 0002-8282.

Dietrichson, J., L. M. Ellegard, and Gustav Kjellsson (2020). "Patient choice, entry, and the quality of primary care: Evidence from Swedish reforms". eng. In: *Health Economics, 2020, Vol. 29, Iss. 6, pp. 716-.730* 29(6), pp. 716–730.

Dnr 710/2016 (Feb. 2017). Swedish Competition Authority.

Doraszelski, U. and J. Jaumandreu (2013). "R&D and Productivity: Estimating Endogenous Productivity". eng. In: *The Review of economic studies* 80(4 (285)), pp. 1338–1383. ISSN: 0034-6527.

Gaynor, Martin, Kate Ho, and Robert J Town (2015). "The Industrial Organization of Health-Care Markets". eng. In: *Journal of economic literature* 53(2), pp. 235–284. ISSN: 0022-0515.

Gaynor, Martin, Rodrigo Moreno-Serra, and Carol Propper (2013). "Death by Market Power: Reform, Competition, and Patient Outcomes in the National Health Service". eng. In: *American economic journal. Economic policy* 5(4), pp. 134–166. ISSN: 1945-7731.

Gaynor, Martin, Carol Propper, and Stephan Seiler (2016). "Free to Choose? Reform, Choice, and Consideration Sets in the English National Health Service". eng. In: *The American economic review* 106(11), pp. 3521–3557. ISSN: 0002-8282.

Handel, Ben and Kate Ho (2021). "Chapter 16 - The industrial organization of health care markets". In: *Handbook of Industrial Organization, Volume 5*. Ed. by Kate Ho, Ali Hortaçsu, and Alessandro Lizzeri. Vol. 5. Handbook of Industrial Organization 1. Elsevier, pp. 521–614. DOI: https://doi.org/10.1016/bs.hesind.2021.11.016. URL: https://www.sciencedirect.com/science/article/pii/S1573448X21000169.

Hlavac, Marek (2018). *stargazer: Well-Formatted Regression and Summary Statistics Tables*. Bratislava, Slovakia. URL: https://CRAN.R-project.org/package=stargazer.

Hoch, Irving (1955). "Estimation of Production Function Parameters and Testing for Efficiency". eng. In: *Econometrica* 23, p. 325. ISSN: 0012-9682.

Levinsohn, James and Amil Petrin (2003). "Estimating Production Functions Using Inputs to Control for Unobservables". eng. In: *The Review of economic studies* 70(2), pp. 317–341. ISSN: 0034-6527.

Maican, Florin and Matilda Orth (2017). "Productivity Dynamics and the Role of 'Big-Box' Entrants in Retailing: PRODUCTIVITY AND THE ROLE OF 'BIG-BOX' ENTRANTS IN RETAILING". eng. In: *The Journal of industrial economics* 65(2), pp. 397–438. ISSN: 0022-1821.

Marschak, Jacob and William Andrews (1944). "Random Simultaneous Equations and the Theory of Production: Introduction". eng. In: *Econometrica* 12(3, 4), p. 143. ISSN: 0012-9682.

Nordgren, Lars and Bengt Ahgren (Jan. 2010). *Val av primärvård: resultat från en brukarundersökning baserad på invånarepaneler*.

OECD (2019). *Sweden: Country Health Profile 2019*, p. 24. URL: https://www.oecd-ilibrary.org/content/publication/2dcb7ca6-en.

OECD, European Observatory on Health Systems, and Policies (2021). *Sweden: Country Health Profile 2021*, p. 24. DOI: https://doi.org/https://doi.org/10.1787/b9027e42-en. URL: https://www.oecd-ilibrary.org/content/publication/b9027e42-en.

Olley, G. Steven and Ariel Pakes (1996). "The Dynamics of Productivity in the Telecommunications Equipment Industry". eng. In: *Econometrica* 64(6), pp. 1263–1297. ISSN: 0012-9682.

Rosano, Aldo et al. (2013). "The relationship between avoidable hospitalization and accessibility to primary care: a systematic review". eng. In: *European journal of public health* 23(3), pp. 356–360. ISSN: 1101-1262.

Rovigatti, Gabriele (2017). "Production function estimation in r: The Prodest Package". In: *Journal of Open Source Software* 2(18), p. 371.

SKR (2018). *Indikatorer för sammanhållen vård och omsorg*.

Swedish Agency for Health and Care Services Analysis (2015). *Vården ur primärvårdsläkarnas perspektiv*, p. 150. DOI: https://doi.org/https://doi.org/10.1787/2dcb7ca6-en. URL: https://www.vardanalys.se/wp-content/uploads/2015/12/Rapport-2015-9-V%5C%C3%5C%A5rden-ur-prim%5C%C3%5C%A4rv%5C%C3%5C%A5rdsl%5C%C3%5C%A4karnas-perspektiv.pdf.

Swedish Agency for Health and Care Services Analysis (2017). *Primärvården i belysning*, p. 125. DOI: https://doi.org/https://doi.org/10.1787/2dcb7ca6-en. URL: https://www.vardanalys.se/wp-content/uploads/2017/12/PM-2017-5-Prim%5C%C3%5C%A4rv%5C%C3%5C%A5rden-i-belysning.pdf.

Swedish Association of Local Authorities and Regions (2022). *Detta är Nationell Patientenkät*. URL: https://patientenkat.se/nationellpatientenkat/omnationellpatientenkat.44342.html (visited on 05/09/2022).

Swedish Competition Authority (2014). "Etablering och konkurrens bland vårdcentraler (KKV rapport 2014:2". In.

Välfärdsutredningen (2016). *Ordning och reda i välfärden: betänkande*. swe. Statens offentliga utredningar, 2016:78. Wolters Kluwer: Stockholm. ISBN: 9789138245224.

Vengberg, Sofie, Mio Fredriksson, and Ulrika Winblad (2019). "Patient choice and provider competition – Quality enhancing drivers in primary care?" eng. In: *Social science & medicine (1982)* 226, pp. 217–224. ISSN: 0277-9536.

Zaremba, Maciej (2013). *Patientens pris : ett reportage om den svenska sjukvården och marknaden*. swe. ISBN: 9789187347153.

A Appendix

A.1 Proof Lemma

In order to obtain the marginal cost of producing an additional unit, we rewrite equation 4.1 as:

$$Q_{it}(\cdot) = \frac{\rho_{it}}{\nu_{it}} \cdot f(K_{it}, L_{it}). \tag{A.1}$$

First, we obtain the marginal product of labor for quantity, by taking the partial derivative of A.1 with respect to L :

$$\frac{\partial Q_{it}(\cdot)}{\partial L_{it}} = \frac{\rho_{it}}{\nu_{it}} \cdot f_L(K_{it}, L_{it}). \tag{A.2}$$

Additionally, we need to obtain the partial derivative of quality with respect to labor to be able to relate quality to quantity:

$$\frac{\partial \nu_{it}(\cdot)}{\partial L_{it}} = \frac{\rho_{it}}{Q_{it}} \cdot f_L(K_{it}, L_{it}). \tag{A.3}$$

Finally, we multiply both of the partial derivatives, A.2 and A.3, with the associated cost of labor, wages (W), to obtain the marginal costs:

$$MC_Q = W_{it} \cdot \nu_{it} \cdot \frac{\rho}{f_L(K_{it}, L_{it})};\ MC_\nu = W_{it} \cdot Q_{it} \cdot \frac{\rho}{f_L(K_{it}, L_{it})}. \tag{A.4}$$

A.2 Proof of Proposition

Firms active in quasi-markets face the following maximization problem:

$$\Pi_{it}(\nu_{it}) = P_t \cdot D_{it}(\nu_{it}, \bar{\nu}_t) - C_{it}(D_{it}(\nu_{it}, \bar{\nu}_t), \nu_{it}). \tag{A.5}$$

Taking the partial derivative of the profit function with respect to quality to obtain the first order condition:

$$P_t \cdot \frac{\partial D_{it}(\nu_{it}, \bar{\nu}_t)}{\partial \nu_{it}} - MC_Q \cdot \frac{\partial D_{it}(\nu_{it}, \bar{\nu}_t)}{\partial \nu_{it}} - MC_\nu = 0. \tag{A.6}$$

We rewrite A.6 as:

$$P_t - MC_Q = MC_\nu \cdot \frac{1}{\frac{\partial D_{it}(\nu_{it},\bar{\nu}_t)}{\partial \nu_{it}}}. \tag{A.7}$$

Rearranging the right-hand side of A.7 by inserting our lemma (4.2):

$$MC_\nu \cdot \frac{1}{\frac{\partial D_{it}(\nu_{it},\bar{\nu}_t)}{\partial \nu_{it}}} = \frac{D_{it}}{\nu_{it}} \cdot MC_Q \cdot \frac{1}{\frac{\partial D_{it}(\nu_{it},\bar{\nu}_t)}{\partial \nu_{it}}}. \tag{A.8}$$

By inserting A.7, we obtain:

$$\frac{P_t - MC_Q}{MC_Q} = \frac{1}{\frac{\partial D_{it}(\nu_{it},\bar{\nu}_t)}{\partial \nu_{it}} \cdot \frac{\nu_{it}}{Q_{it}}}. \tag{A.9}$$

Denoting the quality elasticity of demand as $\tau_{it} = \frac{\partial D_{it}(\nu_{it},\bar{\nu}_t)}{\partial \nu_{it}} \cdot \frac{\nu_{it}}{Q}$ and defining markups as P_t/MC_Q we obtain:

$$\frac{P_t}{MC_Q} - \frac{MC_Q}{MC_Q} = \tau_{it}^{-1}. \tag{A.10}$$

Rearranging A.10 for simplicity as below:

$$\frac{P_t}{MC_Q} = 1 + \tau_{it}^{-1}. \tag{A.11}$$

A.3 Markups in terms of Output Elasticity of Labor

This derivation is from the work of De Loecker and Warzynski (2012). Consider the general production function of A.2. In addition, assume that the provider minimizes cost of production:

$$\mathcal{L}_{it} = W_{it}L_{it} + R_{it}K_{it} + \lambda_{it}\left[Q_{it} - Q_{it}(\cdot)\right]. \tag{A.12}$$

We denote rent of capital as R_{it}. Since we assume cost minimization, λ_{it} is interpreted as the cost of relaxing the budget with one unit. This implies that λ_{it} is synonymous with marginal cost of quantity, $\lambda_{it} \equiv MC_Q$. We now take the partial derivative of the equation A.12 with respect to L_{it} in order to express markups as a function of output elasticity:

$$\frac{\partial \mathcal{L}_{it}}{\partial L_{it}} = W_{it} - \lambda_{it}\left[\frac{\partial Q_{it}(\cdot)}{\partial L_{it}}\right] = 0. \tag{A.13}$$

A.13 is then multiplied with L_{it}/Q_{it},

$$\frac{\partial Q}{\partial L} \cdot \frac{L_{it}}{Q_{it}} = \frac{1}{\lambda_{it}} \frac{W_{it}L_{it}}{Q_{it}}. \tag{A.14}$$

In order to express equation A.14 in terms of markup, P_t/MC_Q, both sides of equation A.14 is multiplied with P_t/P_t:

$$\frac{\partial Q}{\partial L} \cdot \frac{L_{it}}{Q_{it}} \cdot \frac{P_t}{P_t} = \frac{P_t}{\lambda} \cdot \frac{W_{it}L_{it}}{P_tQ_{it}}. \tag{A.15}$$

Since $MC_Q \equiv \lambda_{it}$, we are able to rewrite A.15 is as:

$$\frac{\partial Q_{it}}{\partial L_{it}} \cdot \frac{L_{it}}{Q_{it}} = \frac{P_t}{MC_Q} \cdot \frac{W_{it}L_{it}}{P_tQ_{it}}. \tag{A.16}$$

In equation A.16, $W_{it}L_{it}$ is the total expenditure on labor which and P_tQ_{it} is turnover. Both of these are observable from accounting data, which implies that $\alpha_{it} = \frac{W_{it}L_{it}}{P_tQ_{it}}$ is observable. Rewriting equation A.15 as:

$$\frac{P_t}{MC_Q} = \left[\frac{\partial Q_{it}}{\partial L_{it}} \cdot \frac{L_{it}}{Q_{it}}\right] \cdot \frac{1}{\alpha_{it}}, \tag{A.17}$$

we see that by estimating the output elasticity of labor and dividing it with the labor cost share we obtain an estimate of markups.

B Appendix

Table B.1: Regression on Markup

	Dependent variable:
	$\log(\mu)$
log(share)	-0.065^{***}
	(0.006)
Observations	2,456
R^2	0.062
Adjusted R^2	-0.106
F Statistic	138.125^{***} (df = 1; 2081)
Note:	*p<0.1; **p<0.05; ***p<0.01

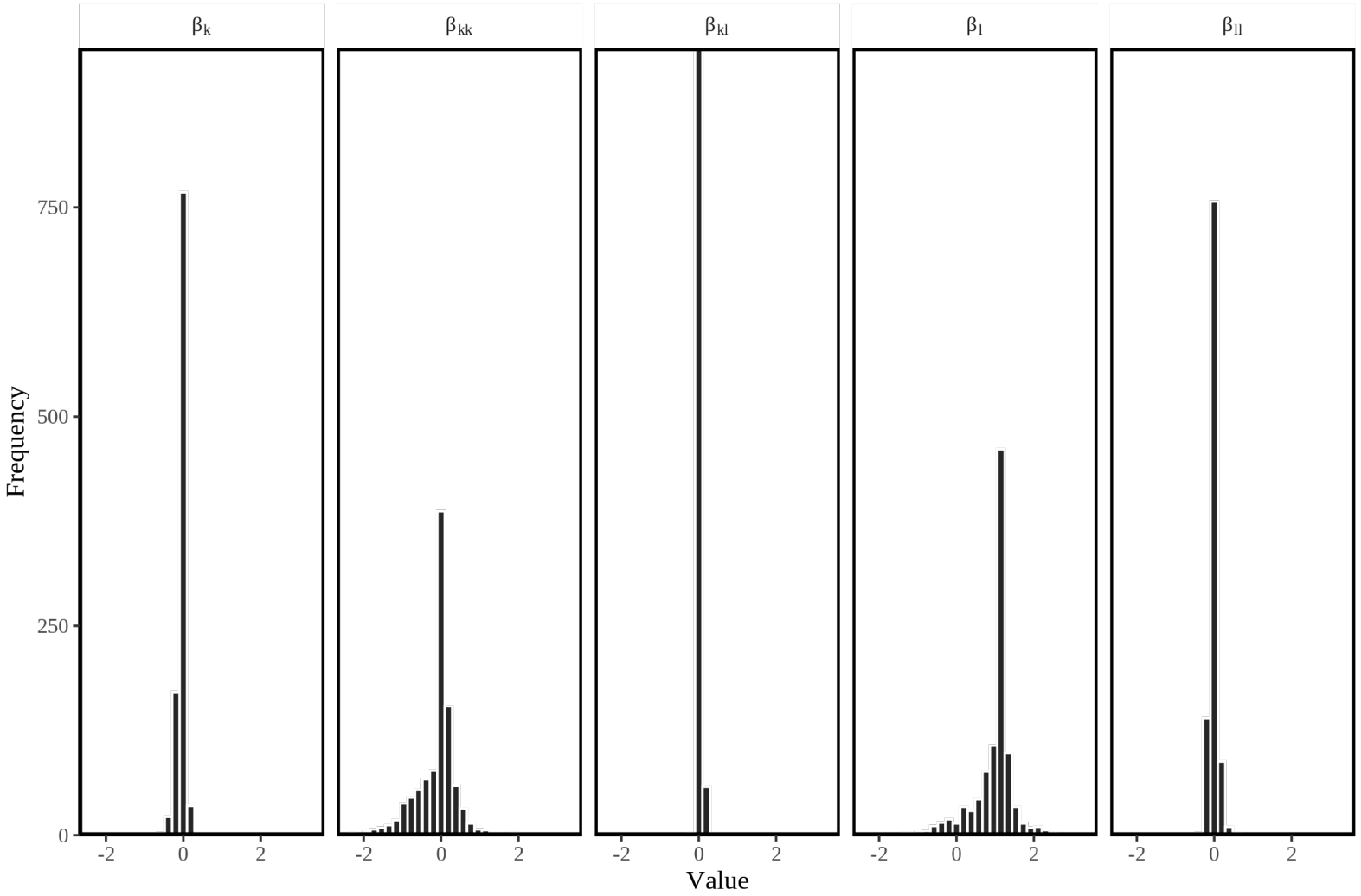

Figure B.1: Distribution of Bootstrapped Estimates

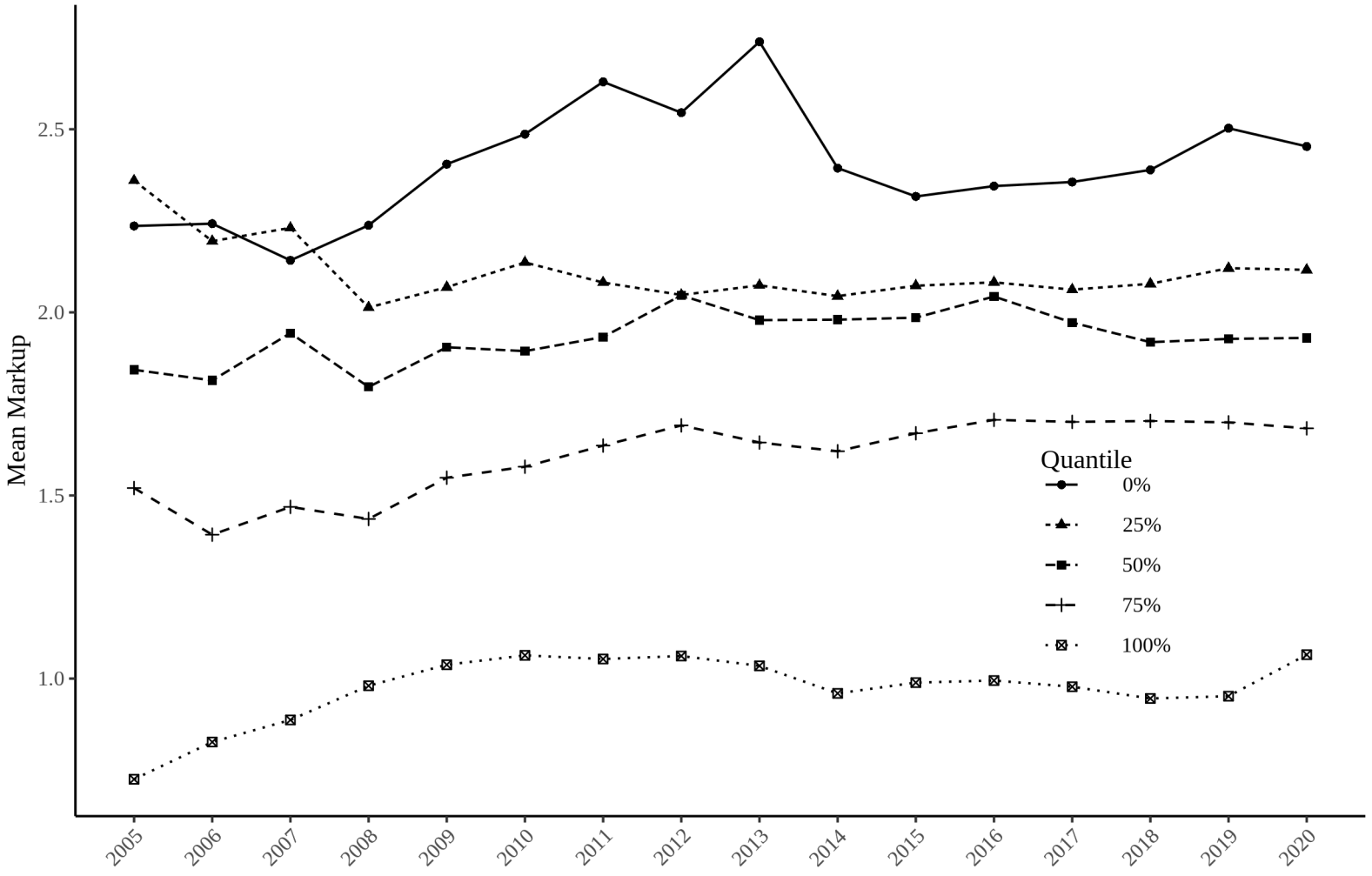

Figure B.2: Evolution of Markups, by Quantile of Turnover

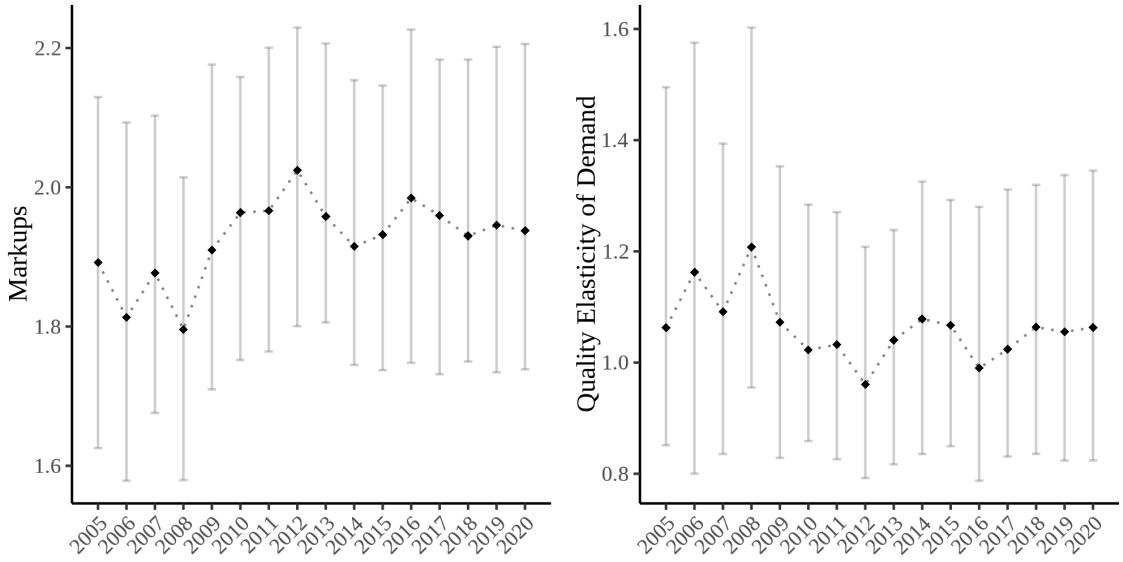

Figure B.3: Median Markups and Quality Elasticity over Time
Error bars indicate 25th and 75th quantile

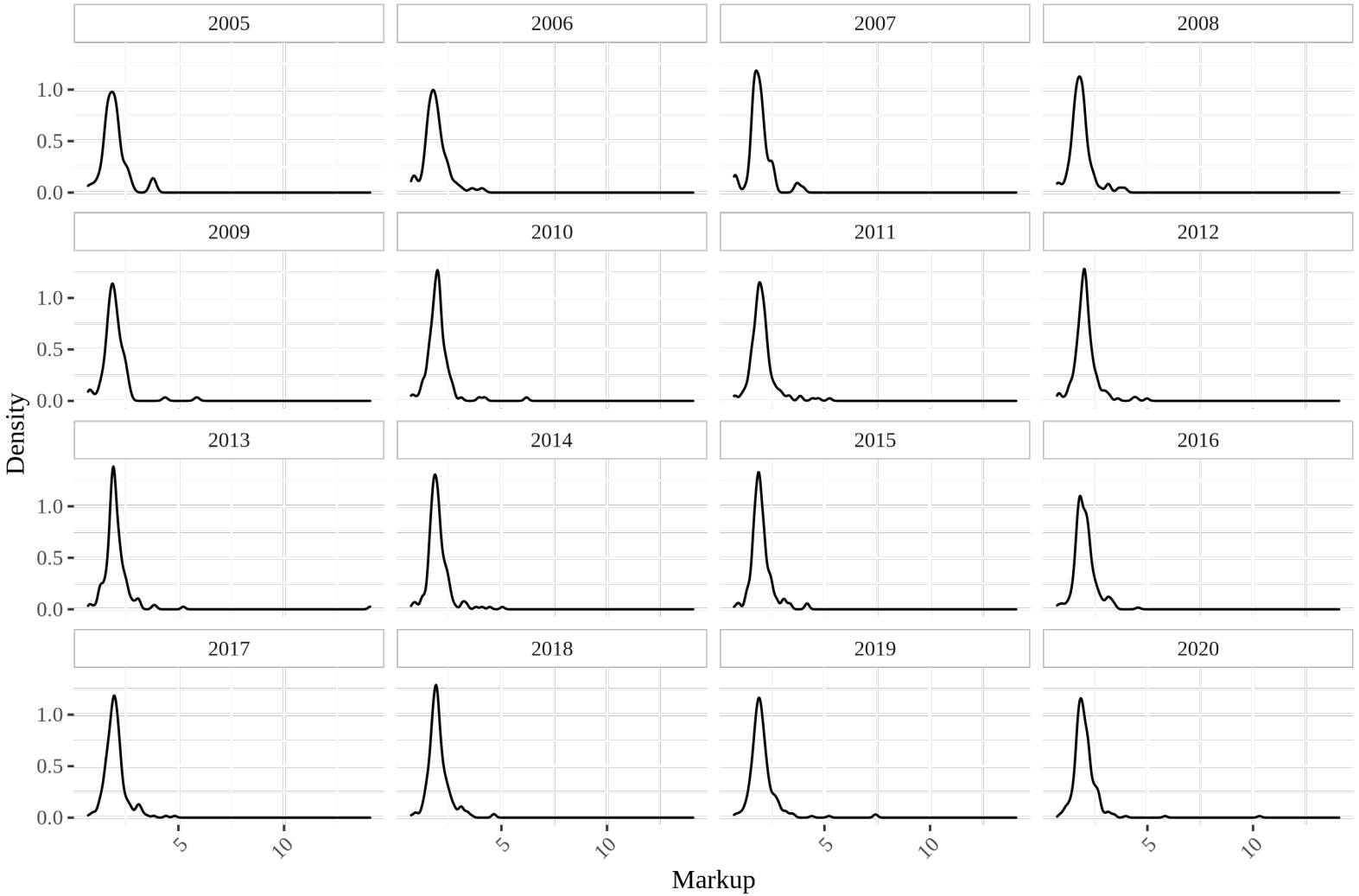

Figure B.4: Distribution of Estimated Markups by Year