DISCUSSION PAPER SERIES

IZA

IZA DP No. 9825

The Market for Paid Sick Leave

Simen Markussen
Knut Røed

March 2016

Forschungsinstitut
zur Zukunft der Arbeit
Institute for the Study
of Labor

The Market for Paid Sick Leave

Simen Markussen
Ragnar Frisch Centre for Economic Research

Knut Røed
Ragnar Frisch Centre for Economic Research
and IZA

Discussion Paper No. 9825
March 2016

IZA

P.O. Box 7240
53072 Bonn
Germany

Phone: +49-228-3894-0
Fax: +49-228-3894-180
E-mail: iza@iza.org

IZA Discussion Paper No. 9825
March 2016

ABSTRACT

The Market for Paid Sick Leave*

In many countries, general practitioners (GPs) are assigned the task of controlling the validity of their own patients' insurance claims. At the same time, they operate in a market where patients are customers free to choose their GP. Are these roles compatible? Can we trust that the gatekeeping decisions are untainted by private economic interests? Based on administrative registers from Norway with records on sick pay certification and GP-patient relationships, we present evidence to the contrary: GPs are more lenient gatekeepers the more competitive is the physician market, and a reputation for lenient gatekeeping increases the demand for their services.

JEL Classification: H55, I11, I18

Keywords: absenteeism, gatekeeping, competition, role-conflicts

Corresponding author:

Knut Røed
The Ragnar Frisch Centre for Economic Research
Gaustadalléen 21
0349 Oslo
Norway
E-mail: knut.roed@frisch.uio.no

* This paper is part of the project "Disentangling absence patterns" financed by the Norwegian Research Council (research grant no. 227103). We wish to thank Bernt Bratsberg, Magne Mogstad, and participants at the workshop on absenteeism, social insurance and retirement in Oslo February 2016 for comments/discussions.

1 Introduction

In most OECD countries, general practitioners (GPs) have been entrusted a role of huge fiscal importance: To protect the public purse from unwarranted social insurance expenditures and unnecessary (or cost-ineffective) medical treatments. In particular, they have been assigned the task of certifying sick pay and disability insurance expenditures; i.e., to decide whether or not (and for how long) a given health problem justifies insurance payouts; see, e.g. Bonato and Lusinyan (2007) and OECD (2010). This appears to be a logical and rational solution to a moral hazard problem. Given that work – at least for some employees – entails elements of disutility, there is a temptation to exaggerate or even "invent" health problems that can justify paid absence. The GP is probably the person who is best placed to objectively assess the true health condition of the patient, and thus decide on the need for a sick leave. And GPs tend to be highly respected and trusted citizens, to whom we presumably safely can entrust the difficult task of balancing the needs and desires of their own patients against the public costs and common interests.

Or can we? Deciding on the need for, e.g., paid sick leave obviously entails subjective judgment. Existing evidence has shown that there is considerable variation in the way GPs interpret and perform their gatekeeper role; see, e.g., Wilkin (1992), Grytten and Sørensen (2003), and Markussen et al. (2011; 2013). Patients are generally free to choose their GP, and can potentially substitute a lenient GP for a strict one. Moreover, in most countries, the GPs operate businesses in a competitive environment; hence, they may have financial incentives to attract and retain customers by providing (excess) access to publically paid services, treatments, and insurance payments. In essence, the GPs have been assigned the task of protecting the public (or private) insurer's purse against the customers who form the basis for their own livelihood. We normally think of a high degree of competition as desirable market characteristic, resulting in better services and lower prices. However, when some of the offered "products" are paid for by a third party (e.g., the taxpayer) more competition may also imply more waste. By combining the apparently incompatible roles of customer competition and gatekeeping, GPs may have been assigned a "mission impossible", in the sense that GPs who perform their gatekeeping role as intended by their principal may be forced out of the market.

The aim of the present paper is to examine empirically the practical consequences of this potential role conflict, in terms of the GPs' choices of gatekeeping standards and the customers' choices of GPs. The paper consists of three separate, but closely related, parts. First, we seek to identify each primary care physician's degree of gatekeeper leniency at each point in time. This is done month by month based on observed absence certification for workers on the patient list of the physician, after controlling for customer composition by means of worker fixed effects. Identification of the physicians' behavior derives from frequent movements of workers between family doctors as well as by their movements between employment (exposure to the risk of absence certification) and non-employment (no exposure). Second, we examine the extent to which workers choose family doctors with an eye to their reputation on gatekeeper leniency. This is done within the framework of a conditional logit model, where the choice set of available GPs is identified from the observed GP choices among other people in the local area, and the GPs' presumed leniency reputations are derived from past leniency indicators estimated on the basis of *other* customers only. Finally, we examine the extent to which physicians adjust their gatekeeper leniency in response to fluctuations in the demand for their services or in the cost of losing customers. This is done within the framework of a fixed effects model where the causal effects are identified solely on the basis of changes in the local competitive environment or in the physician's remuneration structure.

We are not aware of existing empirical research on the role of physician leniency for the patients' choices of GPs. Neither is there a rich literature on the extent to which GPs take personal economic motives into account in the performance of an explicitly assigned gatekeeper role. There is a substantial literature on the impacts of economic incentives on physician behavior more generally, however, showing that payment design can have a large effect on the physicians' prescription of medical treatments; see McClelland (2011) and Chandra *et al.* (2012) for recent reviews. There is also more direct evidence related to GPs referral practices. For example, based on a reform in the funding scheme for certain types of elective surgery in the UK, Dusheiko et al. (2006) show that GPs tend to have a more restrictive referral practice when money *not* spent on surgery alternatively can be channeled back to their own practice. Iversen and Lurås (2000) and Iversen (2004) show that Norwegian physicians who experience shortage of customers provide more services and thus obtain higher income *per customer* than their unconstrained colleagues, whereas Iversen and Ma (2011) show that more intense local competition between physicians leads to more diagnostic radiology referrals. For referrals to specialist treatment that substitute for own services, Godager *et al.* (2015) note that increased competition between primary care physicians has ambiguous effects on gatekeeping incentives. While high competition and/or fewer patients than desired makes it potentially costly – in terms of lost customers – to be a strict gatekeeper, the lack of patients also makes it more profitable to treat the patient within own practice. Godager *et al.* (2015) provide empirical evidence that these effects largely cancel out, and that competition among GPs has insignificant (or slightly positive) effects on GPs' referrals of patients to specialists.

For absence certification practices, no such ambiguities exist. Unless there is excess supply of patients, the physician's economic incentives unequivocally point toward going along with the wishes of their customers. However, there is to our knowledge little empirical evidence on the effect of physician incentives on absence certification. The only study we have found on this topic has been conducted by the Swedish Social Insurance Inspectorate (Inspektionen för socialförsäkringen, 2014) on the basis of a reform in 2007 whereby county-administrations were given the opportunity to scale down GP entry barriers and increase their inhabitants' freedom to choose their (absence-certifying) GP. As a result, the degree of competition increased considerably in some – but not all – counties, and this natural experiment is exploited to assess empirically the impacts of the GPs' economic incentives on their absence certification behavior. Based on a difference-in-difference identification strategy, the authors conclude that the number of certified absence spells increased by approximately 3.5 % as a direct result of the intensified competition between physicians.

In this paper, we present strong and robust empirical evidence that workers do take the physicians' gatekeeping reputation into account when selecting a new family doctor. If a physician's reputation changes such that the average worker can expect to be granted one extra day of monthly certified absence (corresponding to a movement from around the 10^{th} to the 90^{th} percentile in the estimated GP reputation-distribution), the relative probability of being chosen over each of the competing GP alternatives increases by approximately 14 %. We also present evidence indicating that many GPs take this demand curve into account by adjusting their gatekeeping practices in response to changes in their own vacancies and/or in the local competitive situation. A larger number of vacant patient slots and/or increased competition among physicians in the local residential area induce the average GP to become more lenient. Using a group of GPs on fixed wage contracts as a point of reference, we estimate that these two mechanisms together are responsible for raising the overall level of absenteeism in Norway by approximately 4 %. This is a significant, though not a huge, impact. Yet, we argue that it could have been considerably bigger had it not been for the low level of competition between family doctors in Norway.

2 Institutional setting and data

In Norway, workers receive 100 % wage compensation from the first day of sickness absence and for up to one year (up to an income ceiling of approximately NOK 530,000 ($ 62,000) p.a. (2015)).[1] The first 16 days are paid for by the employer, after which the costs are covered by general (payroll) taxation. To offset moral hazard problems, a sick leave certificate issued by a physician is normally required for all spells lasting more than three days.[2] On a typical working day, 6-7% of Norwegian employees are absent from work due to sickness, and almost 90% of these absences are certified by a physician.

Norway has, since May 2001, practiced a family (panel) doctor system, whereby each citizen is assigned a single GP. In most cases, the family doctor receives a capitation fee from the social security administration (SSA) in addition to a per-treatment-pay, which is shared between the patient and the SSA. The capitation fee is currently (2015/2016) NOK 427 (approximately $ 50) per year, and it is paid out for each individual on a family doctor's customer list, regardless of whether the individual ever shows up at the GP's office or not (approximately one third of the customers do not show up during a year). Each GP determines a desired (maximum) number of list-members, with an upper limit of 2,500. The additional per-treatment pay schedule is then regulated through negotiations between The Norwegian Medical Association and the state. A standard 15-20-minute daytime consultation is currently charged at NOK 143 ($ 17). Both the capitation fee and the per-treatment pay imply that GPs have financial incentives to keep their patients happy.

The payment system described above does not apply for all the family doctors, however. Around 4.5 % of the doctors are employed by the municipality in fixed wage contracts, independently of the number of patients and treatments. Such fixed wage contracts are typically offered in rural areas where it may be difficult to establish sufficiently profitable family doctor businesses, or where potential GP candidates are either risk averse or financially constrained.

Sickness absence certificates can in principle be issued by any authorized physician, but will normally be issued by the family doctor, except in emergency cases (outside the family doctor's opening hours and when the family doctor is busy) and when the patient is hospitalized or subject to specialist treatment. Norwegian workers are free to choose their family doctor insofar as the desired doctor has vacant patient slots. This can easily be done on a user-friendly web-portal,

[1] To translate NOK amounts to USD, we have used the exchange rate as of November 2015, with $1=NOK 8.60.

[2] Some firms have agreed to accept self-reported sickness claims for up to eight days.

but it is not possible to change family doctor more than two times per year. When a GP has reached the desired patient ceiling, it is in principle no longer possible to choose that doctor, although there are some exemptions from this rule. Persons who do not make an active choice are assigned a ”default” GP, based on their residential address.

The number of authorized family doctor positions allocated to each municipality is regulated by the Norwegian Directorate of Health. Hence, the family doctor market is not characterized by free entry, and the degree of competition is limited. As we show below, a consequence of this system is that the majority of family doctors are normally subjected to excess demand, implying that vacant patient slots are filled immediately.

The data we use in the present paper are collected from Norwegian administrative registers. They cover all workers and physicians in Norway during the period from January 2002 through December 2010, with individual level (encrypted) monthly data on person-GP-linkages. For each worker, we have monthly information on physician-certified absence, as well as on gender, age, and place of residence. For each GP, we have information on gender and age, the actual and desired number of patients, and their type of remuneration. The latter is an indicator variable denoting whether the physician is running his/her own business (denoted variable wage) or is employed by the municipality in a fixed wage contract (denoted fixed wage).

Using these data, we construct for each employee, a monthly variable indicating the number of physician-certified absence days, adjusted for grading.[3] In addition, we construct for all persons aged 18-62, a monthly indicator giving the (encrypted) identity of the chosen panel doctor.

3 Identifying and estimating GP leniency

We assume that for each GP and for each point in time, there exists a latent variable characterizing the degree of gatekeeper leniency; i.e., the readiness to certify the use of public funds in cases where there is scope for subjective judgment. Systematic differences in leniency may reflect genuine differences in the assessment of what are the “right” certification decisions (from a purely medical or from a social cost-benefit point of view), as well as differences in the willingness to deviate from these right decisions in order to satisfy the patients. The latter may (or may not) depend on the physician’s competitive position, as reflected in the demand for

[3] It is common to use graded absence certificates in Norway, implying that the worker is only partially absent from the workplace, see Markussen *et al.* (2012) for details. If, for example, a worker has a 50% absence spell for 10 days, we count this as 5 days of absence.

his/her services.[4] In this paper, leniency is identified solely on the basis of sick pay certification. We compute a monthly GP leniency indicator based on the overall certification of sick pay days for his/her employed registered customers, controlled for customer composition and calendar time. This implies that we attribute to the family physician all absences certified for his/her patients, even when they are certified by other doctors. The rationale behind this "intention to treat" strategy is that the family physician is likely to have an influence on patients' absence behavior well beyond his/her own issuing of absence certificates; e.g., through communication with colleagues at the medical center and through his/her referral practice. Moreover, the number of absence certificates issued in the family doctor's own name is heavily influenced by opening hours. Many GPs combine the family doctor business with work at hospitals or elderly care institutions 1-2 days a week or run part-time businesses for other reasons, and by attributing all absence certificates to a worker's family doctor we ensure that variations in availability are not falsely interpreted as variations in gatekeeping leniency.

Table 1. Workers and GPs in Norway. January 2002 - December 2010

Workers	
Number of workers	2,480,466
Number of monthly worker observations	63.95
Average monthly number of prescribed absence days	1.79
Percent with positive number	9.04
Average number conditional on positive	19.81
Physicians	
Number of physicians	6,782
Average number of months in practice during our data window	62.5
Worker-physician relationships	
Average GP turnover (percent of workers shifting GP during a month)	1.06
Average patient list worker turnover (percent of employed list members that were not employed or list-members in previous month)	3.72

For each family physician in Norway, we estimate the degree of gatekeeper leniency – month by month – on the basis of the employed list-members' sick-leave certificates. To account for

[4] While we focus on the influence of self-interest here, there are also other reasons why physicians deviate from a principal's understanding of the appropriate gatekeeping practices. Many physicians view themselves as advocates of their patients; see, e.g., Schwartz (2002). The World Medical Association's Code of Medical Ethics, states that "a physician shall owe his/her patients complete loyalty and all the scientific resources available to him/her" (World Medical Association, 2006), and the Charter of Medical Professionalism states patient welfare as the first fundamental principle: "Market forces, social pressures, and administrative exigencies musts not compromise this principle" (Medical Professionalism Project, 2002, p. 244).

patient list composition, we control for worker fixed effects as well as worker age (the latter on an annual basis). The exercise is based on a total number of 2.5 million workers over a period of nine years, yielding as much as 158 million monthly observations divided between around 6,800 physicians; see descriptive statistics in Table 1.[5]

Let y_{it} be the number of grade-adjusted sick pay days prescribed for worker *i* in calendar month *t*.[6] We set up the following regression equation:

$$y_{it} = \alpha_i + \delta_a + \mu^*_{jt} + e_{it} \ , \qquad (1)$$

where α_i is a worker fixed effect, δ_a is an age fixed effect, μ^*_{jt} is a physician-by-month fixed effect, and e_{it} is a residual. We may think of the worker fixed effects as representing the *demand* for sick-leave certificates, whereas the physician-by-month fixed effects represent the *supply*. It may appear a bit optimistic to seek identification of both 2.4 million worker fixed effects and almost 43,000 physician-by-month fixed effects. However, as shown Table 1, there are considerable movements of workers between physicians, making such identification feasible. Every month, approximately 1 % of the workers shift family doctor, and for each physician the monthly turnover rate of employed patients (i.e., patients used to identify their sick leave certification practice) is close to 4 %.

We estimate Equation (1) with a least squares algorithm developed by Gaure (2013) and take out the resultant $\hat{\mu}^*_{jt}$. As an indicator for physician *j*'s degree of gatekeeper leniency in month *t*, we then define

$$\hat{\mu}_{jt} = \hat{\mu}^*_{jt} - \sum_j \omega_{jt} \hat{\mu}^*_{jt} \ , \qquad (2)$$

where ω_{jt} *is* the fraction of workers affiliated to physician *j* in month *t*. This implies that we interpret as a physician's leniency in month *t* his/her month *t* fixed effect's deviation from the

[5] Note that the physician-identity used in this paper changes in the very rare cases where a physician moves his/her practice from one municipality to another. Note also that approximately 20 % of the 6,782 physicians used to estimate GP leniency are not "genuine" family doctors; i.e., we cannot identify a specific person connected to the practice. This happens, for example, when a patient list is operated by more than one physician and in cases where the regular GP has a temporary substitute. Although we are not going to use the estimated physician leniency indicators for these "non-GPs" later in this paper, we include them in the analysis here, since their customers help identifying the worker-fixed effects and thus indirectly the leniency of other physicians.

[6] Note that this is not the same as the number of certified sick pay days in the same month, as we attribute all absence days to the month they were *certified*. As the typical length of a single certification period is around 2-3 weeks (19 days on average), these periods will often stretch into the next month.

contemporary weighted mean of physician fixed effects. Hence, we disregard any common time trends in physician leniency. The reason for this is that we cannot separately identify common time trends in physician leniency from other time developments, related to, e.g., the population's health or work norms and fluctuations in epidemics.

In practice, we would expect most physicians to have a rather consistent practice style over time, implying that we could have estimated a fixed leniency indicator $\hat{\mu}_j$ rather the time-varying indicators $\hat{\mu}_{jt}$. There is obviously a lot of stochastic noise entering into our monthly indicators, and the monthly jumps tend to be implausibly large. Indeed, decomposing the overall variation in $\hat{\mu}_{jt}$ into its between- and within-physician components, we find that the between-physician component accounts for only approximately 38 % of the variance. As we return to in the next sections, there are two reasons why we have chosen to estimate the apparently noisy separate monthly indicators. The first is that it makes it possible to ensure that persons who are in the position of choosing a new panel doctor have not themselves contributed to the computation of the same doctor's leniency reputation. And the second reason is that we intend to examine the mechanisms that cause GPs to change their degree of gatekeeper leniency.

To illustrate the overall difference in gatekeeper leniency between GPs, Figure 1, panel (a), presents the distribution of the time-averaged leniency indicators for each physician (weighted by the overall number customer-months). It shows that as we move from the strictest to the most lenient family doctor, the expected number of prescribed absence days during a month increases by around 2, ceteris paribus. Given that the average number of prescribed absence days is around 1.8 (see Table 1), this is a considerable degree of variation.

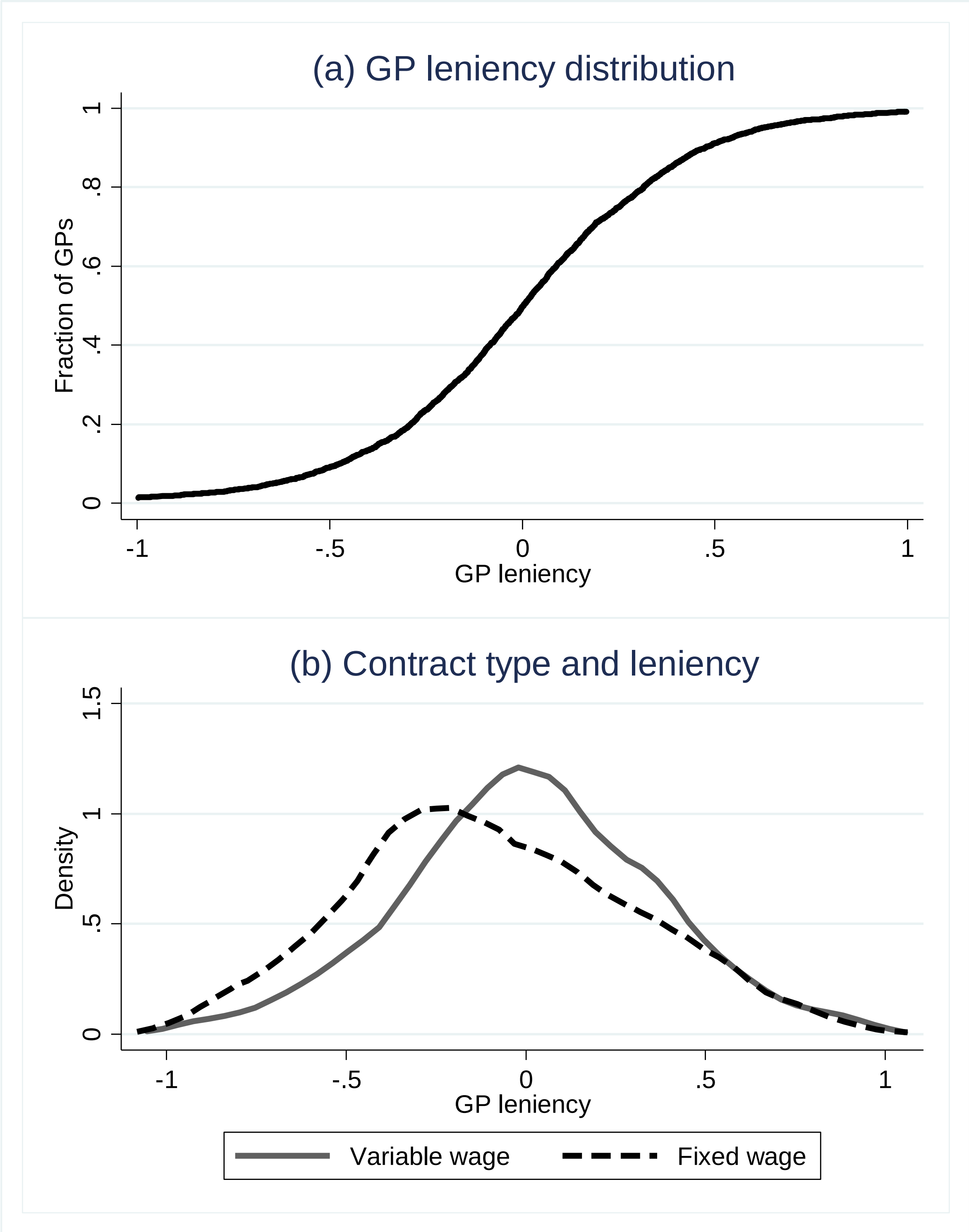

Figure 1. The distribution of time-averaged leniency indicators for the physicians in Norway.
Note: The distribution function in panel (a) is based on the weighted distribution of the time-averaged leniency indicators (with number of employed customers used as weights). The density plots in panel (b) are based on the same weighted averages, with the exception that physicians who shift between variable and fixed wage contracts during the data period are treated as two different physicians.

As described in the previous section, whereas most of the family doctors in Norway run their own business, and hence have earnings that depend on the number of customers and consultations/treatments, approximately 4.5 % of the family doctors are on fixed wage contracts. To illustrate the possible consequences that private incentives may have for the physicians' gatekeeping behaviors, Figure 1, panel (b), presents density plots for the estimated monthly leniency indicator separately for fixed wage physicians and physicians running their own business (hereafter referred to as variable wage physicians). Although there is a considerable variation in estimated gatekeeping practices within each of these groups, it seems evident that variable wage physicians tend to be a bit more lenient. The (unweighted) difference in average leniency between the two physician groups is 0.12, suggesting that variable wage physicians certify 0.12 (7 %) more absence days per month than fixed wage physicians do, ceteris paribus.

4 The patients' choice of GP

In this section, we examine the degree to which patients take gatekeeper leniency into account when they choose their family doctor? Since GP leniency is intrinsically a latent characteristic, patients must obviously choose GP with imperfect information about their gatekeeping-practices. It seems probable, though, that local rumors exist regarding physicians' practice styles, which at least to some degree mirror their true leniency.[7] We will assume adaptive behavior in the formation of a GP's leniency reputation. More specifically, in a baseline model we assume that the local reputation corresponds to the observed degree of sick pay-prescription (adjusted for patient composition) as computed in the previous section, but averaged over the past 12 months. In separate robustness analyses below, we instead assume that reputations are based on the GPs' behaviors over the past 6 or 18 months.

The analysis in this section will be conditioned on a physician shift taking place; i.e., situations where a patient chooses a (for him/her) new GP. This implies that we do not attempt to model the reason(s) why the old panel doctor is deselected in the first place, only the choice of the new one, given that this actually happens. We also require that the "old" GP was retained for at least 12 months (6/18 in the robustness analyses), and that the shift is not a part of "mass-shift" where

[7] There is now a web-portal (www.legelisten.no) designed to help citizens choose family doctor, where customers share their views and experiences with named physicians, often with a focus on their gatekeeping practices. This portal did not exist in our data period (it was established in 2012), but similar (local) discussion groups were probably common. Note that since the data used in the present paper are encrypted (due to privacy protection considerations), we are not able to compare our estimated leniency indicators to (later) customer views expressed at the web-portal.

a whole (or a large fraction of) a list is transferred to another physician (this typically happens when a GP retires or for other reason terminates the practice).[8] This way, we zoom in on situations where we know that *active* GP choices are to be made, and we avoid by construction the potential problem that the persons making these choices have themselves contributed with data used to estimate the degree of leniency for any of the GPs in their presumed choice set.

For each of these shifts we try to identify the patient's choice set, i.e., the set of GPs that made up the menu of feasible choices. To do this, we first record for each worker the correct local area of residence.[9] We then identify the set of potential GP choices by assuming that all physicians who have at least 10 other patients from the same local area belong to the choice set. In separate robustness analyses, we instead use thresholds of 5 or 50 patients to identify the choice set.

One common reason for changing family doctor is relocation to a new local area. Since we have data on residential local area on an annual basis only, we are in some cases not able to identify with certainty the correct GP choice set. In these cases we construct choice sets corresponding to both the previous and the new address, and we then include in the analysis the choice set in which we see that a GP was actually selected (if the selected GP enters both or none of the choice sets, we drop the observation).

In total, we identify around 671,000 active choice situations satisfying our requirements. However, there are several cases where workers choose GPs that were not assumed by us to be part of their choice set (this can happen, for example, if they chose a GP close to where they work rather than close to where they live). After having dropped these choice situations from the analysis, we end up with 394,720 active GP choices that can be used in our statistical analysis. Table 2 gives an overview of the size distribution of the resultant choice sets. On average, a worker can choose between 74.5 different physicians, but the variation is large – from only two alternatives in some rural areas to as much as 321 in some big city districts.

---

[8] If more than 10 customers make exactly the same shift at exactly the same time, we interpret them as resulting from a mass-shift, and drop them from the analysis. In practice, these mass-typically involve a much larger number of customers, and if we set the threshold to 50 instead of 10, we get almost exactly the same results as those reported here.

[9] In this exercise, we use "local areas" corresponding to the statistical tracts ("delområder"), drawn up by Statistics Norway; see Statistics Norway (1999) for details. They are designed to encompass neighborhoods that naturally interact, e.g., by sharing common service/shopping center facilities. There are 1,535 local areas in Norway, and a typical local area comprises around 3,100 inhabitants.

Table 2. Cases of active choices of family doctor in Norway. January 2003 - December 2009

Total number of choice situations	394,720
Average size of choice set (number of available physicians)	74,5
smallest	2
10^{th} percentile	9
25^{th} percentile	20
median	44
75^{th} percentile	118
90^{th} percentile	172
largest	321

Note: Compared to the data used in the previous section, the data window is cut by 12 months at each end. The first 12 months is lost due to our condition of at least 12 months attachment to the previous/deselected GP. The last 12 months is lost for the reason that we need "next year's" records to check for migration to a new local area in order to identify the correct GP choice set.

Based on these choice sets, we set up a model designed to explain actual GP choices. We do this on the basis of GP characteristics within the framework of a conditional logit model (McFadden, 1974), similar to the strategy used by Santos *et al.* (2016) to study the impacts of clinical quality on the choice of family doctors in England. We assume that worker *i's* utility associated with choosing a particular GP *j* can be expressed as

$$U_{ij} = \mathbf{x_{ij}}\boldsymbol{\beta} + v_{ij}, \tag{3}$$

Where $\mathbf{x_{ij}}$ is a covariate vector that differs across the alternative physicians and v_{ij} are unobservable taste components. Provided that we assume "independence of irrelevant alternatives" (IIA) – i.e., that the relative choice probabilities for any two alternative physicians depend only on the attributes of those two – the v_{ij} are independently distributed according to the type I extreme value distribution, and we can write the choice probabilities as

$$P(d_i = j) = \frac{\exp(\mathbf{x_{ij}}\boldsymbol{\beta})}{\sum_{h \in J_i} \exp(\mathbf{x_{ih}}\boldsymbol{\beta})}. \tag{4}$$

where $d_i=j$ if physician *j* was worker *i's* chosen alternative from his/her choice set J_i.

The explanatory variable of main interest is the physicians' reputations on gatekeeper leniency, as reflected in $\hat{\mu}_{jt}$ averaged over the past 12 months; i.e., for a GP choice made in month *t* we include as a reputation indicator for physician $j \in J_i$ the variable

$$\hat{\mu}_{jt}^{S} = \tfrac{1}{S}\sum_{s=t-S}^{t-1} \hat{\mu}_{js} \,, \tag{5}$$

with $S=12$ and $\hat{\mu}_{jt}$ defined in Equation (2). In addition, $\mathbf{x_{ij}}$ contains covariates describing the geographical distance between the worker's and the physician's locations, whether the GP offers an emergency service or not, the GP's gender and age, and the existence of available patient slots at the start of the month.[10]

The estimation results are shown in Table 3, for the full dataset of all workers (column (1)) as well as for a range of sub-groups (columns (2)-(7)). As noted above, the IIA assumption implies that the relative choice probabilities for any two alternative physicians depend only on the attributes of those two physicians. Thus for any alternatives *j*,*k*, it follows from (4) that we have

$$\frac{P(d_i = j)}{P(d_i = k)} = \frac{\exp(\mathbf{x_{ij}\beta})}{\exp(\mathbf{x_{ik}\beta})} = \exp\left[(\mathbf{x_{ij}} - \mathbf{x_{ik}})\boldsymbol{\beta}\right]. \tag{6}$$

This implies that we can interpret the exponentiated coefficients as the estimated change in the relative choice probabilities associated with a marginal change in the corresponding variable. The coefficient estimate associated with the leniency variable $\hat{\mu}_{jt}^{12}$ of 0.13 in column (1) thus implies that when a physician's leniency reputation increases with one unit – implying one extra expected day of certified absence per month, ceteris paribus – the relative probability of being chosen over each of the competing GP alternatives increases by $100 \times (\exp(0.13) - 1) = 13.9$ %.

[10] Since we only have precise geographical information on the location of workers' homes, we compute the distance-to-physician variables by assuming that each GP is located in the business center of the local area in which the largest number of his/her patients live. With the aid of geographical positioning data, we compute the distance between the worker's home and each physician's office. We then define two variables; one indicator variable for "walking distance" (less than 1,500 meters) and one scalar variable indicating the number of minutes it will take to drive by car.

Table 3. Workers' choice of new GP. Estimation results for baseline model (standard errors in parentheses)

	(1) All	(2) Absent t-1	(3) Present t-1	(4) Women	(5) Men	(6) Age 18-44	(7) Age 45-66
GP leniency	0.134*** (0.004)	0.285*** (0.010)	0.100*** (0.0046)	0.136*** (0.006)	0.130*** (0.006)	0.117*** (0.005)	0.197*** (0.009)
Walking distance	0.581*** (0.006)	0.527*** (0.015)	0.587*** (0.007)	0.565*** (0.009)	0.595*** (0.009)	0.580*** (0.007)	0.548*** (0.015)
Estimated driving time if not walking distance (minutes)	-0.108*** (0.000)	-0.099*** (0.001)	-0.110*** (0.001)	-0.103*** (0.001)	-0.111*** (0.001)	-0.110*** (0.001)	-0.099*** (0.001)
Emergency service provided	0.013*** (0.004)	0.035*** (0.009)	0.009** (0.004)	0.031*** (0.006)	-0.001 (0.005)	0.017*** (0.004)	0.021** (0.009)
Same sex GP	0.664*** (0.004)	0.556*** (0.010)	0.691*** (0.004)	0.532*** (0.005)	0.737*** (0.006)	0.669*** (0.004)	0.652*** (0.009)
Female GP	-0.099*** (0.004)	-0.234*** (0.010)	-0.067*** (0.005)			-0.068*** (0.005)	-0.220*** (0.009)
GP age	-0.017*** (0.000)	-0.022*** (0.001)	-0.016*** (0.000)	-0.021*** (0.000)	-0.014*** (0.000)	-0.016*** (0.000)	-0.022*** (0.001)
GP younger	0.014* (0.008)	-0.065*** (0.018)	0.033*** (0.009)	-0.005 (0.011)	0.029*** (0.011)	0.103*** (0.014)	-0.034** (0.014)
GP older	-0.139*** (0.006)	-0.096*** (0.015)	-0.149*** (0.0069)	-0.110*** (0.009)	-0.166*** (0.009)	-0.178*** (0.007)	0.029** (0.014)
Free capacity at start of month	0.985*** (0.004)	1.018*** (0.009)	0.978*** (0.004)	0.884*** (0.006)	1.078*** (0.005)	1.005*** (0.004)	0.900*** (0.0086)
N (#choices×size of choice set)	29,742,433	5,280,516	24,461,917	14,048,568	15,693,865	24,611,473	5,130,960

Note: GP leniency is defined in Eq. (5). Walking distance is defined as less than 1,500 meters (0.9 miles). Estimated driving times are computed from "open street map". ***(**)(*) indicates statistical significance at the 1(5)(10) % levels, respectively.

Table 4. Workers' choice of new GP. Robustness (standard errors in parentheses)

	(1) Baseline	(2) >4 patients	(3) >49 patients	(4) Vacant	(5) 6 months	(6) 18 months
GP leniency	0.134*** (0.004)	0.136*** (0.004)	0.086*** (0.005)	0.190*** (0.006)	0.147*** (0.003)	0.121*** (0.005)
Walking distance	0.581*** (0.006)	0.642*** (0.006)	0.505*** (0.008)	0.627*** (0.008)	0.580*** (0.006)	0.582*** (0.007)
Estimated driving time if not walking distance (minutes)	-0.108*** (0.000)	-0.118*** (0.000)	-0.053*** (0.001)	-0.109*** (0.001)	-0.104*** (0.000)	-0.110*** (0.001)
Emergency service provided	0.013*** (0.004)	-0.0050 (0.0038)	-0.003 (0.005)	0.026*** (0.005)	0.023*** (0.004)	0.012*** (0.004)
Same sex GP	0.664*** (0.004)	0.667*** (0.004)	0.645*** (0.004)	0.674*** (0.005)	0.657*** (0.004)	0.668*** (0.004)
Female GP	-0.099*** (0.004)	-0.093*** (0.004)	-0.112*** (0.005)	0.036*** (0.006)	-0.069*** (0.004)	-0.117*** (0.005)
GP age	-0.017*** (0.000)	-0.018*** (0.000)	-0.016*** (0.000)	-0.019*** (0.000)	-0.020*** (0.000)	-0.015*** (0.000)
GP younger	0.014* (0.008)	0.008 (0.008)	0.0342*** (0.009)	-0.011 (0.010)	0.019*** (0.007)	0.012 (0.009)
GP older	-0.139*** (0.006)	-0.134*** (0.006)	-0.149*** (0.007)	-0.118*** (0.008)	-0.147*** (0.006)	-0.136*** (0.007)
Free capacity at start of month	0.985*** (0.004)	0.979*** (0.004)	0.989*** (0.004)		1.038*** (0.004)	0.954*** (0.004)
N (#choices×size of choice set)	29,742,433	44,978,235	7,177,562	10,502,604	36,747,433	24,256,401

Note: Columns (2) and (3) are based on choice sets identified on the basis of at least 5 or 50 other individuals in the local area having selected the physician in question. Column (3) is based on at least 10 persons having selected the physician (as in the baseline model), but only with physicians with vacant capacity at the start of the month included. Columns (5) and (6) are based on alternative definitions of GP reputation, based on the last 6 or 18 months, respectively.
***(**)(*) indicates statistical significance at the 1(5)(10) % levels, respectively

To put the estimated effects of leniency reputation into perspective, we report in Table 3 also the estimated impacts of other physician characteristics. Geographical proximity to the worker is clearly of great importance; and being located within walking distance of the worker raises the relative choice probability considerably. For example, substituting walking distance for a 10 minutes' drive raises the relative choice probability by a factor of $\exp(0.58+10\times0.11)=5.37$. Workers also appear to prefer doctors with emergency service. Apart from these GP practice characteristics, the results in column (1) show that workers care about the GP's gender and age. They prefer GPs of the same sex as themselves, yet with a bias toward male doctors. Workers also prefer GPs that are either of the same age as themselves or younger. Finally, the table shows that having vacant slots (free capacity) at the start of the month raises the relative probability of being chosen by a factor of around 2.7. In principle it is not possible to choose a doctor without vacant capacity (see Section 2); hence, it could be argued that GPs without vacant capacity should not belong to the choice set at all. However, due to the large patient turnover we frequently see that GPs are chosen during the course of a month despite having a full list at the start of the month. In a robustness exercise, we drop physicians without vacant capacity from the choice set.

It appears probable that workers who are on a sick-leave around the time of GP selection put more weight on a prospective family doctors leniency-reputation than workers who are not on sick-leave. Columns (2) and (3) report separate estimation results for workers with and without a sick-leave certificate issued in the month prior to the shift of GP. As expected, we find that the physicians' leniency reputations have much larger impacts on the choice of GP for currently (or recently) sick-listed workers.

It may also be of some interest to examine how the weight attached to GP leniency vary across demographic groups, and columns (4)-(7) report results for women and men and for young and old workers separately. The estimated effect of GP leniency is very similar across these groups, however, with the exception that older workers seem to put a bit more weight on GP leniency reputation than younger workers do.

While we do not see any specific challenges to identification of the models estimated above, we realize that the presented results build on a number of assumptions and definitional choices for which we have no clear evidence-based guidance. One may worry that these kinds of choices have been made on the basis of a data-mining exercise seeking to identify the most "publishable" results. To allay such concerns, we present in Table 4 some results based on alternative assumptions. Columns (2)-(4) show results based on alternative assumptions regarding the workers' choice sets. In column (2), we have assumed that all physicians that are chosen by at least five other individuals in a local area belong to a worker's choice set, instead of 10, as in the baseline model (recall that a local area on average consists of 3,100 inhabitants). This raises the average size of the choice sets from 74.5 in the baseline model to as much as 108.8. Yet the estimated coefficient on GP leniency remains unchanged. In column (3), we have instead restricted the choice sets to physicians chosen by more than 50 other individuals in the local area. This reduces the average size of the choice sets to 22.7 physicians. The estimated effect of GP leniency then becomes a bit smaller, although it is still highly significant. In column (4), we maintain the threshold of 10 customers in the local area, but remove from the choice set all physicians without vacant capacity at the start of the month. This reduces the average size of the choice set to 37.4 physicians. And the estimated effect of leniency becomes a bit larger than in the baseline model.

In columns (5) and (6), we present results based on alternative assumptions regarding the formation of GP reputation. In column (5), we have included only six months of past leniency indexes to compute average leniency; i.e., we have substituted $\hat{\mu}_{jt}^{6}$ for $\hat{\mu}_{jt}^{12}$; see Equation (5). This also implies that we raise the number of observations by as much as 20 %, since we in this model can reduce the conditioning period for having been affiliated to the same GP from 12 to 6 months. Yet, the estimated effect of GP leniency remains the same. In column (6), we instead include as much as 18 months of past leniency indexes to compute average leniency (using $\hat{\mu}_{jt}^{18}$), and consequently throw out around 20 % of the observations. Still, the estimated effect of GP leniency remains almost unchanged.

An additional concern could be that we have used a rather restrictive functional form assumption for the relationship between GP leniency and customer utility. There is of course no behavioral theory that can be called upon to justify a specific functional form in this case, and the linearity assumption must be viewed primarily as a choice motivated by convenience and interpretability. As an alternative, we have therefore estimated models where we have dummy-coded the GP-leniency indicator into 10 brackets. The result from this exercise is illustrated in Figure 2, where we have plotted the estimated coefficients for the models corresponding to columns (1)-(3) in Table 3; i.e., for all workers, for recently absent workers, and for recently present workers. For comparison, we also show the estimated linear relationships (from Table 3), and to illustrate the origins of the linear slope estimates, we present the new dummy-based estimates with circles representing the numbers of GP-months that lie behind each estimate, as well as with 95 % confidence intervals. The estimates in Figure 2 suggest indeed that the association between worker utility and GP leniency is non-linear, and that, for most workers, the relationship is relatively flat in the central part of the GP leniency distribution. However, for all worker groups, there appears to be a negative utility associated with having a very strict GP. For recently/currently absent workers there also appears to be a considerable utility gain associated with having a very lenient GP; see panel (b). It is notable that for the latter group – which clearly constitutes the worker group for which GP leniency is likely to have some immediate consequences – the relationship appears to be highly monotone, and also not very far from the linearity assumption.

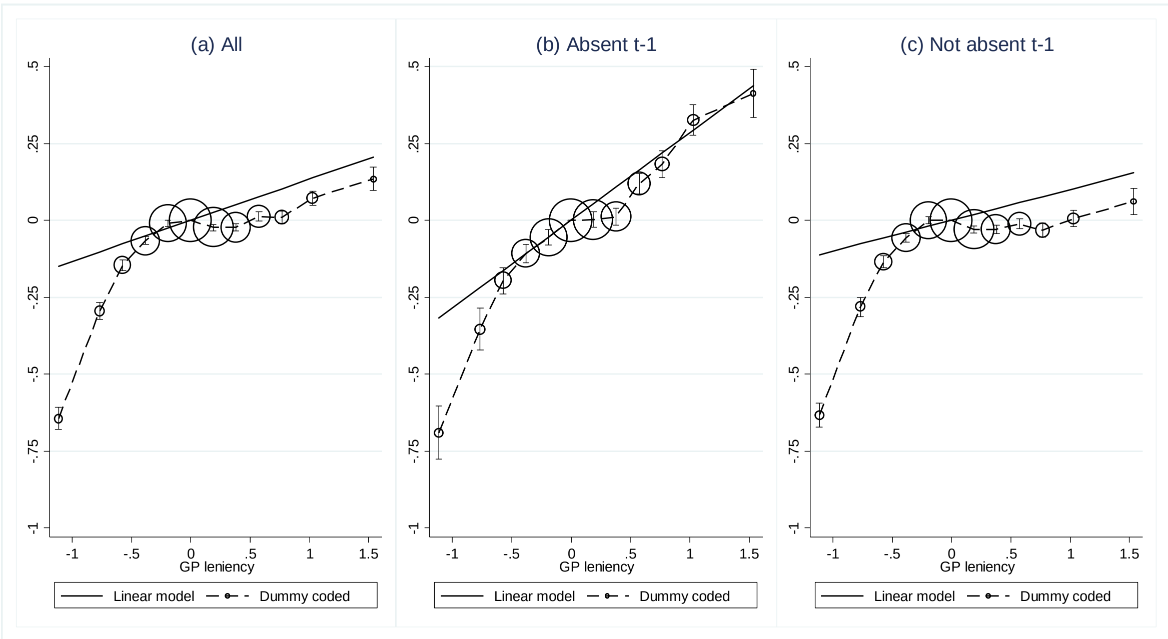

Figure 2 Non-linear estimates of the impacts of leniency
Note: The leniency indicator variables are defined such that each of the 10 bins on the horizontal axis have exactly the same length (0.2 days), except at the two ends where each category contains all leniency observations below the 1st and above the 99th percentiles in the leniency distribution, respectively.

Our reading of these exercises is that our finding that workers put considerable emphasis on a GPs leniency reputation when choosing a new family doctor is robust with respect to modeling assumptions. The relationship is strongest in the tails – and particularly in the lower tail – of the GP leniency distribution.

5 The GPs' choice of leniency

Having established that workers do take the GPs' leniency reputation into account when selecting their family doctor, the next question is whether or not the physicians respond to the demand for lenient gatekeeping practices in order to attract customers? If they do, we would expect leniency to increase in response to patient shortage or increased GP market competition. We would also expect leniency to increase when physicians move from a fixed wage contract to running their own business.

As a first descriptive examination of the physicians' leniency-responsiveness, we take a closer look at the physician markets identified in the previous section, and check whether the average leniency indicators among variable wage physicians correlate systematically with the degree of local competition. In this descriptive exercise, we include all local areas with at least 200 inhabitants divided between at least 10 physicians who each have at least 10 customers from the local area. This gives us approximately 105,000 monthly local physician-market observations. Since it is not obvious how the degree of local competition between physicians should be measured, we compute for each local area and for each calendar month, a number of alternative competition indicators; i.e.:

(a) The vacancy rate, defined as the number of vacant (unfilled) patient slots in the local area divided by the total number of registered (desired) patient slots. This is perhaps the most natural local competition indicator, as it summarizes the overall excess supply of patient slots. Note that in some cases, the vacancy rate is slightly negative, as it is possible for physicians to be temporarily overbooked.

(b) The number of vacant patient slots in the local area. This measure also summarizes overall excess supply, but in absolute numbers instead of relative to total supply as in (a).

(c) The number of GPs in the local area with at least one vacant patient slot (available physicians). This is the best indicator for the actual choice opportunities available to workers searching for a new GP.

(d) The number of physicians operating in the local area divided by the number of inhabitants. This is the standard measure of physician density.

(e) Patient turnover, defined as the fraction of patients in the local area who shifted family doctor last month. This measure captures the current tendency for patients in the local area to shift family doctor.

(f) The fraction of physicians with fixed wage contracts. In contrast to the other competition indicators, a higher fraction of fixed wage physicians indicate *less* competition, both because fixed wage contracts tend to be offered precisely *because* the GP coverage is considered too low and because fixed wage physicians do not have the same incentives to compete for patients.

Figure 3 illustrates in separate panels (a)-(f) how each of these competition indicators correlates with average estimated leniency among the local areas' variable wage physicians. For ease of exposition, we have divided the local market observations into percentiles (with approximately 1,000 market observations behind each data point). To illustrate the considerable size-differences between the local markets, we have let the sizes of the dots vary proportionally to the total number of inhabitants in the local markets included in each percentile. Finally, we have added linear bivariate regression lines, based on the size-weighted percentiles.

It is clear from Figure 3 that, regardless of the chosen competition indicator, we consistently uncover a conspicuous and robust positive relationship between the degree of local competition and average estimated physician leniency.

If the largely cross-sectional correlation patterns in Figure 3 have a causal interpretation, we would expect to find them also in a longitudinal setting; i.e., we would expect each physician's gatekeeper leniency to change in response to changes in the local competitive environment and own patient shortage. To investigate this further, we examine in more detail the individual physicians' responses to changes in their competitive situation. The data we use for this purpose is described in Table 5. They consist of panel data for 5,340 family physicians, on average observed for 71 months during our estimation period. On average, a physician is 56 customers short; i.e., he/she has 56 fewer persons on the customer-list than desired, corresponding to an average vacancy rate of 5.3 %. However, the vacancies are far from equally distributed, and in as much as 55 % of the physician-month observations, the list is full (zero vacancies).

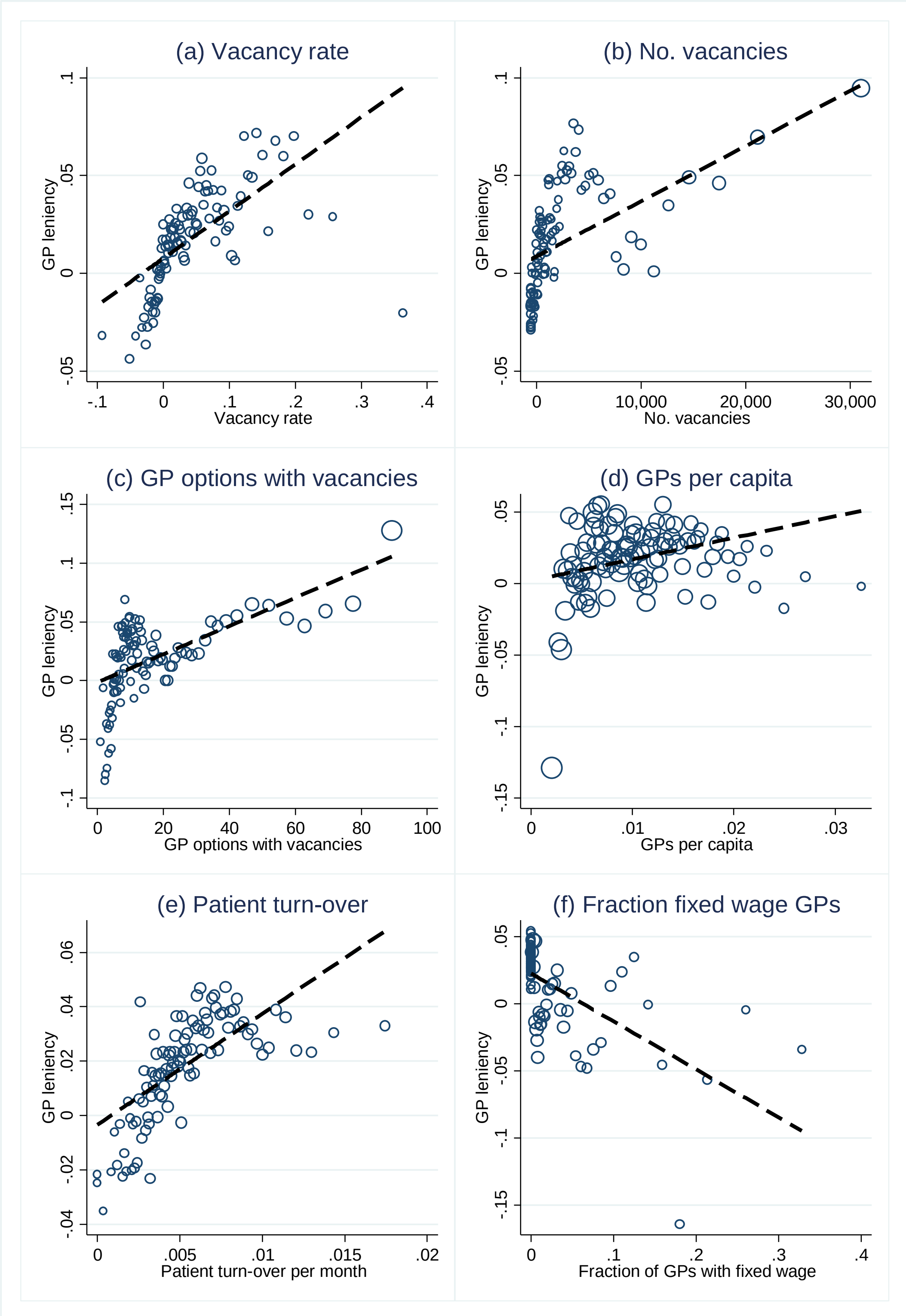

Figure 3. Average leniency indicator among variable-wage physician in local physician markets. By degree of competition.

Note: Each data point represents a percentile in the respective distributions of competition indicators. The sizes of the dots are proportional to the number of inhabitants in the local markets included in each percentile.

Table 5. GPs in Norway January 2002 - December 2010

Total number of physicians	5,340	
Average number of months represented in the dataset	71.05	
Descriptive statistics by physician month	Mean	Standard deviation
Average number of desired patients	1,410.9	399.5
Average number actual patients	1,355.0	386.4
Average number of vacancies	56.1	198.4
Fraction of GPs with full lists (no vacancies at start of month)	0.554	
Average vacancy rate (vacancies/number of desired patients)	0.053	0.175
Fraction on fixed wage contract	0.476	
Local market competition (local areas)		
Vacancy rate	0.058	0.066
No. vacancies	8,593.4	14,946.9
No. GP options with vacancies	21.13	24.16
GPs per capita	0.0007	0.0001
Patient turn-over per month	0.006	0.004
Fraction of GPs with fixed wage	0.045	0.084

Note: All numbers are weighted by the number of workers behind each GP/month measure of leniency.

To examine the determinants of gatekeeper leniency, we set up and estimate statistical models with the following structure:

$$\hat{\mu}_{jt} = \pi_j + \mathbf{z}_{\mathbf{jt-1}}\boldsymbol{\sigma} + \varepsilon_{jt}\,, \tag{7}$$

where $\mathbf{z}_{\mathbf{jt}}$ is a vector of time-varying variables that potentially influence the physicians' practice style. To ensure that the presumed endogenous responses take place strictly after the explanatory impulses, we enter all the explanatory variables with a month time lag. While we focus on the impacts of the type of wage contract and local market conditions, all estimated models also include dummy variables for calendar time (on a monthly basis) and the physician's age (on a yearly basis). In a baseline specification, we represent market conditions as the average vacancy rate among the physicians in the local market (including all physicians who have at least 10 patients from the local area, irrespective of where the physicians themselves are located). The variables are entered as deviations from their respective (global) means.

Table 6. Estimation results.
Dependent variable: Estimated GP leniency (standard errors in parentheses)

	(1) OLS	(2) Between GPs	(3) Within GP	(4) Within GP	(5) Within GP	(6) Within GP
Fixed wage (dummy)	-0.110*** (0.005)	-0.155*** (0.025)	-0.072*** (0.025)	-0.062** (0.025)	-0.059** (0.025)	-0.059** (0.026)
Variable wage (dummy) interacted with:						
Own vacancy rate					0.058*** (0.009)	0.058*** (0.009)
Local market vacancy rate	0.124*** (0.016)	0.138 (0.116)		0.228*** (0.025)	0.190*** (0.025)	0.190*** (0.025)
Fixed wage (dummy) interacted with						
Own vacancy rate						0.008 (0.033)
Local market vacancy rate						-0.078 (0.075)
N (physician-months)	379,425	379,425	379,425	379,425	379,425	379,425

Note: Local market vacancy rate is computed in a “leave-out”-fashion; i.e., without including data from own practice. All models include calendar month and years of age dummy variables. “Fixed wage” is a dummy variable indicating the GP receives a fixed wage from the municipality. “Variable wage” is a dummy variable indicating that the wage is not fixed. Columns (3) - (6) are based on models with physician fixed effects included.
***(**)(*) indicates statistical significance at the 1(5)(10) % levels, respectively.

The results are presented in Table 6. As a point of departure, columns (1) and (2) first show results from a simple ordinary least squares (OLS) regression and a “between-estimator” (using the cross section of physician averages only), where we include the fixed wage indicator and the local vacancy rate only. Columns (3)-(5) show the results from the fixed effects model, with the three key variables entered in a stepwise fashion. Column (3) show the estimated impact of having a fixed rather than a variable wage, with no controls for market situation or own vacancies. The parameter estimate of -0.072 implies that when a physician switches from a variable to a fixed wage regime, the expected number of certified absence days among his/her customers drop by -0.072 per month – corresponding to approximately 4.0 % of the average number of prescribed days; conf. Table 1. This is considerably smaller than suggested by the pure cross-sectional regression in column (2) and the OLS model in column (1), yet still significant, both from a substantive and from a statistical point of view. The foundation for identification is weak, however, as the fixed wage coefficient estimated in a model with physician fixed effects is based on the relatively few physicians who change between a fixed wage and a variable wage status (47 physicians in our data). Column (4) presents the results from a model where we have included the local vacancy rate (for variable wage physicians only). The point estimate of 0.23 implies that if the local vacancy rate increases by, say, 0.1 (10 percentage points), the average monthly number of physician-certified sick pay days prescribed by variable wage physicians in that local area increases by approximately 0.023 days.

Column (5) shows our preferred results where we also add in the physician's own vacancy rate. The estimated coefficient for this variable is considerably smaller than for the overall vacancy rate in the local market, yet highly statistically significant. Finally, column (6) adds in the local market and own vacancy rate for fixed wage physicians. Since these physicians do not have any personal economic motives for adjusting their gatekeeping practices in response to market conditions, this model specification may be viewed as a sort of placebo test. Given the low number of fixed wage physicians, the standard errors are considerably larger for these coefficients; hence statistical power probably becomes too weak to draw firm conclusions. However, the two estimated coefficients are either approximately zero or "wrongly" signed, and none of them are statistically significant.

How robust are these results? Table 7 reports estimation results for a range of alternative fixed effects models. For ease of comparison, column (1) repeats the main results from column (5) in Table 6. Columns (2)-(6) report results based on alternative formulations of the competitive pressure variables. In column (2), we have substituted the absolute *number* of own and local vacancies for the corresponding *rates.* In column (3) we have kept the own vacancy rate, but used the number of GPs with vacant patient slots as the indicator for local competition. In column (4), we have instead used the overall number of GPs per capita in the local area as the indicator for local competition. In column (5) we have used the fraction of persons who has recently shifted family doctor. And in column (6), we have used the fraction of GPs in the local area with variable wage. All these models essentially give the same answers: More competition imply more lenient gatekeepers. Finally, column (6) reports how the estimates in the baseline model change when we add in a lagged dependent variable in Equation (7).[11] This reduces the estimated coefficients slightly. The "steady state" effects (the coefficient estimate divided by one minus the estimated autoregressive parameter) remain approximately unchanged, however.

[11] Note that the transformed equation used to estimate the fixed effects model contains a residual that incorporates a covariate-adjusted individual mean (over all months), and is thus not completely exogenous with respect to the lagged dependent variable (see, e.g., Cameron and Trivedi (2005, p. 764)). Approximate consistency requires that the average residual is small relative to each period's residual, which again requires that the number of periods is large. We consider this satisfied in our case.

Table 7. Estimation results from alternative model specifications with physician fixed effects.
Dependent variable: Estimated GP leniency (standard errors in parentheses)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Fixed wage (dummy)	-0.059** (0.025)	-0.080*** (0.025)	-0.086*** (0.026)	-0.056** (0.025)	-0.066*** (0.025)	-0.079*** (0.026)	-0.046* (0.024)
Variable wage (dummy) interacted with:							
Own vacancy rate	0.058*** (0.009)		0.066*** (0.0086)	0.069*** (0.009)	0.073*** (0.0086)	0.073*** (0.0086)	0.045*** (0.0085)
Local market vacancy rate	0.190*** (0.025)						0.137*** (0.025)
Number of own vacancies (divided by 100)		0.004*** (0.001)					
Number of local market vacancies (divided by 5000)		0.008*** (0.001)					
Number of GPs with vacancies (divided by 100)			0.121*** (0.019)				
Number of GPs divided by number of inhabitants				0.115*** (0.023)			
Fraction of customers with recent GP shift					0.548** (0.241)		
Fraction of GPs with variable wage						0.134*** (0.036)	
Lagged dependent variable							0.26*** (0.0016)
N (physician-months)	379,425	379,425	379,425	379,425	379,425	379,425	379,425

Note: Where appropriate, local market characteristics are computed in a "leave-out"-fashion, i.e., without including data from own practice. All models include calendar month and years of age dummy variables. "Fixed wage" is a dummy variable indicating the GP receives a fixed wage from the municipality. "Variable wage" is a dummy variable indicating that the wage is not fixed. Physician fixed effects included in all models.
***(**)(*) indicates statistical significance at the 1(5)(10) % levels, respectively.

The message coming out of these exercises is that our main finding – that gatekeeping practices are indeed affected by market conditions and the physicians own economic incentives – is highly robust with respect to model specification.

6 The association between leniency and patient numbers

Somewhat simplified, we have found that: i) more lenient gatekeeping gives the GP more customers, and ii) more customers make the GP less lenient. What are the overall implications of these two mechanisms for the association between GP leniency and the allocation of customers? Will the lenient physicians tend to have more or less customers than the strict physicians? One way to address this question is to examine the statistical association between the predicted gatekeeper leniency and the GPs' numbers of employed customers, using all GP-months as the units of observation. Doing this, we also incorporate an element of patient choice behavior that has so far been left out of the analysis, namely the duration of customer-GP-relationships (the decision to deselect a current GP).

Let n_{jt} denote the number of employed customers held by physician j in a month t, and recall that $\hat{\mu}_{jt}$ denotes the GP's estimated leniency in the same month. Based on all the GP-month observations in our dataset, we find that the correlation between these two variables is negative $(corr(\hat{\mu}_{jt} n_{jt}) = -0.06)$. This implies that lenient physician behavior is empirically associated with fewer rather than more customers. However, in order to give this association a behavioral interpretation, we need to disentangle the within-market and across-market contributions to the correlation pattern. Across markets, we would indeed expect the association to be *positive,* as a high (low) total number of customers per GP in a market implies a low (high) degree of competition, and thus unambiguously yield more (less) lenient GP behavior (confer Figure 3). To facilitate an appropriate decomposition into within-market and across-market contributions, we assume that all GPs located within a given local area in a given month constitute a unique market.[12] We then obtain a set of non-overlapping markets and can use the law of total covariance to disentangle the within-market and the across-market contributions. Let Z_{jt} indicate the market in which GP j operates in moth t. We then have that

[12] This is obviously not correct, as we have shown in this paper that there is a considerable overlap between markets, such that many workers have GPs outside their own local area. For the purpose at hand, however, we consider it to be an acceptable approximation. Note also that we have identified each physician's locality as the local area in which the majority of his/her customers live.

$$\underbrace{\text{cov}(\hat{\mu}_{jt}, n_{jt})}_{=-4.81} = \underbrace{E\left[\text{cov}(\hat{\mu}_{jt}, n_{jt} \mid Z_{jt})\right]}_{=-6.30} + \underbrace{\text{cov}\left[E(\hat{\mu}_{jt} \mid Z_{jt}), E(n_{jt} \mid Z_{jt})\right]}_{=1.49}, \quad (8)$$

where the first expression on the right-hand-side reflects the within-market covariance between leniency and the number of customers, and the second expression reflects the across-market covariance. The numbers indicated below the expressions show the corresponding estimates computed from our data. Hence, what we find is that the negative correlation between GP leniency and customer numbers indeed arises *within* local markets only, reflecting that the GP-responses dominate the customers' choice behavior. As expected, the correlation across markets is positive.

Another way of illustrating the consequences of customer choices and leniency adjustments is to compare the unweigthed average GP leniency (taken over GP-by-month-observations) with the corresponding average weighted by the number of employed customers each month. The result of this exercise is presented in Table 8. The first row of this table reports the average of the prescribed absence days taken over all GP-months (without weighting by number of patients). The second row reports the corresponding average weighted by the number of employed customers. A comparison of the two numbers confirms that the average *chosen* physician is slightly *less* lenient than the unweighted average physician. The mechanism behind this result is further illustrated by looking at the predicted GP-contributions to absence prescription conditional on "popularity". A conspicuous finding is that the GPs with full customer lists are much less lenient than GPs who have fewer customers than desired. Taken literally, the estimates imply that had everyone been assigned the most popular physicians – those with full customer lists more than 95 % of the time – absenteeism would have been reduced by as much as 5.5 %. The obvious explanation for this is that the most lenient physicians are lenient precisely because they have difficulties recruiting a sufficient number of customers. Hence, somewhat simplified, we could say that lenient gatekeeper practice appears to be a tool used to compensate for other characteristics that entail a low probability of being selected.[13]

[13] Previous evidence from the Norwegian GP system has indicated that physicians with fewer customers than desired also are deselected 50 % more frequently than physicians with full lists are (Iversen and Lurås, 2011).

Table 8. Average prescribed absence days implied by estimated leniency. By physicians' popularity

	Implied average
Unweighted average of absence days by GP (N=379,425)	1.83 days
… weighted by number of customers each month	1.82 days (-0.74 %)
… conditional on full customer list (N=198,098)	1.82 days (-0.73 %)
… conditional on vacant customer slots (N=181,327)	1.85 days (+0.88 %)
… conditional on full lists ≤ 75 % of the time (N=246,105)	1.85 days (+1.03 %)
… conditional on full lists > 75 % of the time (N=133,320)	1.80 days (-1.91 %)
… conditional on full lists > 95 % of the time (N=48,445)	1.73 days (-5.52 %)

Note: All reported percentage changes are computed from the respective differences in estimated GP leniency only, and measured relative to the average number of absence days per month reported in the first row (1.83). This number deviates from the average number reported in Table 1 (1.79) for two reasons: i) While the average reported in Table 1 is taken over workers, the average reported here is taken over physicians (unweighted by number of workers), and ii) while Table 1 includes absence prescribed by all physicians, the present table includes absence prescribed by genuine family physicians (where a specific person can be identified); see footnote 5.

7 Concluding remarks

Does the assignment of GPs to the twin roles of running competitive businesses and guarding the public purse imply that we in practice have "let the fox mind the henhouse"? We have shown in the previous sections that workers in the process of choosing a new family doctor prefer lenient over strict gatekeepers, ceteris paribus. We have also shown that some physicians take this customer-preference for leniency into account when determining their own gatekeeping practice style. But how important are these mechanisms? What are the financial implications for the insurer (the taxpayer in our case)?

We can shed some light on the issue by using the physicians on fixed wage contracts as a point of reference. As shown in Section 3, the average difference in estimated leniency between variable and fixed wage physicians is 0.12, which corresponds quite closely to the OLS estimate in Table 6. However, based on the fixed effect (within GP) model, we have estimated that approximately half of this difference has a causal interpretation. Taken literally, this suggests that if we – as a thought experiment – gave all the GPs in Norway a fixed wage contract instead of letting them run their own businesses, we would expect the absence level in Norway to decline by approximately 0.06 days per month, or roughly 3-4 %. This is a considerable, though not huge effect.

Why do we not see even larger consequences of letting the GPs be gatekeepers for their own customers? After all, the responses identified in this paper, both on workers' GP choices and on GP behavior, appear both statistically and substantively significant. One important explanation is that the competition between physicians in Norway is weak. In large parts of the country, GPs are persistently in short supply, and as shown above, more than half of the GPs' patient lists are fully booked while the average vacancy rate is only 5 % (Table 5). And even though workers care about physician leniency, they also care about other physician characteristics. Hence, many GPs probably do not have to worry at all about insufficient demand for their services. And the most popular GPs need not be particularly lenient in order to attract the desired (or the maximum) number of customers. In fact, family doctors in Norway on average work 20 % more hours per week than the 37.5 hours defined as full time; and in a survey conducted in 2004, a third of the family doctors reported that they were subjected to an "unacceptable" level of work pressure (Aasland and Rosta, 2011).

The estimated effects are indeed larger if we focus on markets with above average competition. Combining our preferred estimates (Table 6, column 5) with the vacancy rates prevailing in the local markets with most intense physician competition (with vacancy rates around 30 %), we find that substituting fixed wage for all variable wage physician would reduce the expected absence level by approximately 7 %.

Although the findings reported in this paper hardly suggest that strict gatekeepers are necessarily driven out of business, they illustrate the potential difficulty associated with being an impartial gatekeeper for claims made by own customers. And this role conflict can extend far beyond the certification of sick pay. In most countries, family doctors play important gatekeeper roles in choices of (costly) medical treatments, in decisions regarding referrals to specialists, and in the evaluation of disability insurance claims. The fiscal consequences of this system are likely to depend on the degree of competition between GPs; the more physicians need to worry about attracting enough customers, the more costly it is to use them as the gatekeepers of social or private insurance.

References

Aasland, O. G. and Rosta, J. (2011) The working hours of general practitioners 2000 – 2008. *Tidsskrift for Den norske legeforening,* Vol. 131, 1076-1079.

Bonato, L. and Lusinyan, L. (2007) Work Absence in Europe. IMF Staff Papers, 54, 475-538.

Cameron, A. C. and Trivedi, P. K. (2005) *Microeconometrics – Methods and Applications*. Cambridge University Press.

Chandra, A., Cutler, D. and Song, Z. (2012) Who Ordered That? The Economics of Treatment Choices in Medical Care. In *Handbook of Health Economics,* Vol. 2, eds. McGuire, T. G., Pauly, M. V., and Barros. P. P., 397-432. Elsevier.

Dusheiko, M., Gravelle, H., Jacobs, R., Smith, P. (2006) The Effect of Financial Incentives on Gatekeeping Doctors: Evidence from a Natural Experiment. *Journal of Health Economics* Vol. 25, 449-478.

Gaure, S. (2013) OLS with Multiple High Dimensional Category Variables. *Computational Statistics & Data Analysis*, Vol. 66, 8–18.

Gaure, S. (2014) Correlation Bias Correction in Two-Way Fixed-Effects Linear Regression. *Stat*, Vol. 3, No 1, 379–390.

Godager, G., Iversen, T. and Ma, C.A. (2015) Competition, Gatekeeping, and Health Care Access. *Journal of Health Economics*, Vol. 39, 159–170.

Grytten, J. and Sørensen, R. (2003) Practice Variation and Physician-Specific Effects. *Journal of Health Economics,* Vol. 22, 403-418.

Inspektionen för socialförsäkringen (2014) Vårdvalets effekter på sjukskrivningarna (The effect of introducing a health care choice system on the use of sickness insurance). Rapport 2014:17.

Iversen, T. (2004) The Effects of Patient Shortage on General Practitioners' Future Income and List of Patients. *Journal of Health Economics,* Vol. 23, 673-694.

Iversen, T. and Lurås, H. (2000) Economic Motives and Professional Norms: The Case of General Medical Practice. *Journal of Economic Behavior & Organization*, Vol. 43, 447-470.

Iversen, T. and Lurås, H. (2011) Patient Switching in General Practice. *Journal of Health Economics,* Vol. 30, 894-903.

Iversen, T., Ma, C.A. (2011) Market Conditions and General Practitioners' Referrals. *International Journal of Health Care Finance and Economics*, Vol. 11, 245–265.

Markussen, S., Røed, K., Røgeberg, O.J., Gaure, S. (2011) The anatomy of absenteeism. *Journal of Health Economics,* Vol. 30, 277-292.

Markussen, S, Røed, K. and Røgeberg, O. (2013) The Changing of the Guards - Can Family Doctors Contain Worker Absenteeism? *Journal of Health Economics*, Vol. 32, No. 6, 1230–1239.

McFadden, D. L. (1974) Conditional Logit Analysis of Qualitative Choice Analysis, in P Zarembka (ed.): *Frontiers in Econometrics,* New York Academic Press, 105-42.

McClellan, M. (2011) Reforming Payments to Healthcare Providers: The Key to Slowing Healthcare Cost Growth while Improving Quality? *Journal of Economic Persp*ectives, Vol. 25, No. 2, 69-92.

Medical Professionalism Project (2002) Medical Professionalism in the New Millennium: A Physician Charter. Project of the ABIM Foundation, ACP-ASIM Foundation, and European Federation of Internal Medicine. *Annals of Internal Medicine*, Vol. 136, No. 3, 243-246.

OECD (2010) Sickness, *Disability and Work: Breaking the Barriers. A Synthesis of Findings across OECD Countries*. OECD Publishing, Paris.

Santos, R. Gravelle, H. and Propper, C. (2016) Does Quality Affect Patients' Choice of Doctor? Evidence from England. *Economic Journal*, forthcoming.

Statistics Norway. (1999) Regionale inndelinger - En oversikt over standarder i norsk offisiell statistikk. Norges Offisielle Statistikk, Oslo-Kongsvinger.

Schwartz, L. (2002) Is there an Advocate in the House? The Role of Health Care Professionals in Patient Advocacy. *Journal of Medical Ethics,* Vol. 28, 37-40.

Wilkin, D. (1992) Patterns of referral: Explaining variation, in: M. Roland and A. Coulter, (eds.): *Hospital Referrals*, Oxford General Practice Series 22, Oxford University press.

World Medical Association (2006) World Medical Association International Code of Medical Ethics.