UNIVERSITY OF OSLO
HEALTH ECONOMICS RESEARCH PROGRAMME

The interaction between patient shortage and patients` waiting time

Hilde Lurås

Tor Iversen
Center for Health Administration
University of Oslo

Working Paper 1999: 2

The interaction between patient shortage and patients' waiting time*

Tor Iversen and Hilde Lurås**

Health Economics Research Programme at the University of Oslo
HERO 1999

Center for Health Administration

* Financial support from the Norwegian Ministry of Health and Social Affairs is acknowledged.

** Health Economics Research Programme,
Center for Health Administration, University of Oslo, Rikshospitalet, The National Hospital,
N - 0027 Oslo, Norway. Telephone: 47 22868750, Fax: 47 22362562
e-mail: tor.iversen@samfunnsmed.uio.no

Health Economics Research Programme at the University of Oslo 1998 – 2003
ISSN 1501-9071, ISBN-nummer: 82-7756-026-5

Abstract

We study the interaction between patient shortage and patients' waiting time to get an appointment. From a theoretical model we predict that physicians experiencing a shortage of patients offer their patients a shorter waiting time than their unconstrained colleagues. This happens because a shorter waiting time is expected to lower the threshold for seeking care, and hence, to increase the number of patient-initiated contacts. But it also happens because a shorter waiting time can be a mean to attract new patients. The hypotheses are supported by some preliminary results from a sample of Norwegian general practitioners participating in a capitation trial.

1. Introduction

The purpose of this paper is to examine whether a shortage of patients has an influence on the waiting time a general practitioner (GP) offers his patients. The study considers a system where each GP is responsible for a personal list of patients and the payment system is a combination of capitation and fee for service. By a shortage of patients we mean that the GP has fewer patients on his list than he would prefer to have. We predict that GPs who experience a shortage of patients' offer a shorter wait than their unconstrained colleagues do. We highlight two reasons behind the prediction. Firstly, a short wait may attract many patient-initiated consultations from the existing list of patients and hence, contribute to the GP's income. Secondly, a short wait may attract new patients to the list and hence make the shortage of patients less severe.

The approach in this paper has certain similarities with Sloan and Lorant (1977) and Mueller (1985) since both in their models and in ours, the expected waiting time is a decision variable for the physician. In their model the physician is a profit maximising monopolist facing a demand curve where the money price of a visit is a decreasing function of the number of visits and the waiting time. The physician's costs are an increasing function of the number as well as the length of the visits and a decreasing function of patients' waiting time. Sloan and Lorant find that increased medical insurance coverage (demand shift) shortens the waiting time and so does an increase in the physician-population rate (supply shift). From a different set of data Mueller (1985) gets much of the same empirical results: waiting time is inversely related to both physician density and real per capita income.

While these studies examine patients' waiting time in the waiting room before the consultation, we study patients waiting time from the day they make an appointment with the doctor until the consultation takes place. We believe there are similarities in the mechanism driving the two types of waiting times. The mechanism stems from the fact that patients' demand for consultations is stochastic. Without a booking system patients experience a high capacity utilisation as a long wait in the waiting room. With a booking system the stochastic element of patient arrival in the waiting room is controlled, and a high capacity utilisation is felt as a wait from the time of reservation to the time the consultation takes place.

Our model considers a list patient system with mixed capitation and fee for service remuneration. Fees are considered to be exogenous to the individual physician. The physician is assumed to have lexicographic preferences concerning his patients' health and his own income and leisure. This assumption implies that a patient's health is never balanced against the GP's income or leisure. Details of the model are explained in section 2.

Our empirical study applies data from one of the municipalities participating in the Norwegian capitation experiment that lasted from 18 May 1993 until 18 May 1996. In autumn 1994 the GPs were asked whether they wanted additional persons on their list. GPs who answered this question positively, are considered to be rationed: they experience a shortage of patients. A more detailed description of the data can be found in section 3. In section 4 the statistical model and the empirical results are presented. We find that GPs who experience a shortage of patients, offer their patients a shorter waiting time than their unconstrained colleagues. We also find that among constrained GPs, the waiting time has a negative impact on the increase in the number of persons listed.

Section 5 contains some concluding remarks and suggestions for further research. A policy implication of our study is that the distinction between patient-initiated and physician-initiated consultation may be less sharp than often assumed in the literature.

2. A GP's optimal waiting time, number of patients and service provision

There are two types of consultations, patient-initiated and physician-initiated. The number of patient-initiated consultations per listed person, *s*, are assumed to depend upon the expected waiting time to get an appointment, *w*. A long wait is assumed to occur together with a small number of patient-initiated consultations both because some patients go to alternative providers and because some patients recover by their own:

$$s = \underline{z} + g(w) \tag{1}$$

$\underline{z}$ is an exogenous component depending on a patient's characteristics. We assume that $dg(w)/dw < 0$.

It is well known that clinical judgement vary among physicians. For instance, views may differ with respect to how often a patient with diabetes or a patient with hypertension should be called in for check-ups. Views may also differ on whether a GP who prescribes antibiotics to a patient, should call in the patient for a follow-up consultation next week, or ask the patient to contact him if his symptoms persist. The intensity of service provision will on average be higher in the first case than in the second. It is also well documented that referral rates vary more than can be explained by the characteristics of patients. Hence, the medical treatment that succeeds a patient-initiated consultation is likely to depend on the physician's

clinical judgement. A consequence of the lack of medical standards is that several practice profiles are all regarded as equally satisfactory from a medical point of view[1].

The number of physician-initiated consultations are assumed to consist of two components, $\underline{k}$ + k. $\underline{k}$ is a lower constraint on physician-initiated consultations determined by medical knowledge and professional culture. k is a component determined by each physician's discretion. It follows that practice variation is likely to occur. k is assumed to be located in the interval $[0, \overline{k}]$, where $\overline{k}$ is the upper boundary determined by identical factors as the lower boundary $\underline{k}$.

Our approach implies that economic incentives can only influence the number of consultations in the interval $[\underline{k}, \overline{k}]$. To concentrate on our main points, we shall ignore the GP's provision of other services than consultations (for instance diagnostic tests).

The aggregate demand for consultations with a GP is a stochastic variable. The demand varies from one day to the next and the variation cannot exactly be determined in advance. When all patients get an appointment at their preferred day, idle capacity will exist in periods with low demand. A decline in idle capacity will increase the probability that a patient has to wait for an appointment and will increase the expected waiting time as well. From stochastic queuing theory (see for instance Cooper (1981)) we have that expected waiting time in steady state with exponentially distributed time interval between two subsequent arrivals and between two subsequent completed services, with one server and each customer demanding one service, can be expressed by the formula:

[1] For an introduction to the literature on medical practice variation, see Andersen and Mooney (1990).

$$w = \frac{C(\frac{\eta}{\gamma})}{\gamma - \eta} \tag{2}$$

where η is the average number of customers arriving per time unit, γ is the average number of completed services per time unit when the server is busy and $C(\cdot)$ is the probability that the server is busy. For a steady state solution to exist we must have $\eta < \gamma$, otherwise the waiting time approaches infinity. In our application η is interpreted as the average number of appointments per time period, equal to $n[\underline{z} + g(w) + \underline{k} + k]$, where n is the number of persons on a GP's list of patients, and $[\underline{z} + g(w) + \underline{k} + k]$ is the total number of consultations per person during the time period. We further interpret γ as the average number of finished consultations per time unit when all available appointments are booked and hence, the GP is busy all the time. We then have that γ is equal to $\frac{T - \ell}{t}$, where T is the total time, ℓ is the GP's leisure and t is the average duration of a consultation. Eq (2) may then be transformed to:

$$w = \frac{t\, C(\frac{nt[\underline{z} + g(w) + \underline{k} + k]}{T - \ell})}{T - \ell - nt[\underline{z} + g(w) + \underline{k} + k]} \tag{3}$$

(3) determines indirectly w as a function of k, ℓ and n:

$$w = w(\underset{+}{k}, \underset{+}{n}, \underset{+}{\ell}, \underset{+}{t}) \tag{4}$$

where the sign of a partial derivative is written under the argument.

From (1) and (4) we have:

$$s = \underline{z} + g[w(k, n, \ell)] = \underline{z} + z(\underline{k}, \underline{n}, \underline{\ell}) \tag{5}$$

The physician is assumed to have lexicographic preferences concerning his patients' health, his own income and his leisure. This assumption implies that a patient's health is never balanced against the GP's income or leisure[2]. The doctor then provides services to the patient until there is no medical gain from further treatment; he will not conceal any treatment. Health services are then provided until the marginal health effect is equal to zero. This assumption simplifies the formal reasoning considerably and is also in accordance with the medical profession's own assertion. A relaxation of this assumption would imply that the effect of economic incentives is strengthened.

A GP's optimal number of patients (n), number of consultations per patient (k) and leisure (ℓ) is determined by constrained maximisation of the GP's objective function. Similarly to the number of consultations, we also introduce an exogenous constraint on the waiting time. This constraint indicates the level of accessibility that the GP, because of professional standard, feels obliged to offer. We formulate the maximisation problem with a quasilinear objective function:

$$\begin{aligned} &\underset{n,k,\ell}{Max}\ c + v(\ell) = r + qn + np[\underline{z} + z(k, n, \ell) + \underline{k} + k] + v(\ell) \\ &\text{s.t.} \\ &0 \le k \le \overline{k} \\ &0 < w(k, n, \ell, t) \le \overline{w} \\ &0 < n \le \overline{n} = h(k, n, \ell, \Omega) \end{aligned} \tag{6}$$

[2] This argument is elaborated in Iversen and Lurås (1998a).

where c is the GP's total practice income, r is a fixed wage or practice allowance, q is a capitation income per listed person and p is a fee per consultation. The marginal utility of leisure is assumed to be strictly decreasing in the amount of leisure ($v''(\ell)<0$), implying that $v(\ell)$ is strictly concave. The first and second constraint express respectively the accepted interval for the number of consultation per patient and the accepted interval for the waiting time to get an appointment according to a GP's medical standard. The third constraint expresses that a GP may experience a shortage of patient, and this constraint, $\bar{n}$, depends on the waiting time that patients observe. Other things equal, we assume that a patient prefers to be listed with a GP who offers a short wait for consultations: $\bar{n} = \bar{n}[w(k,n,\ell),\Omega] = h(\underline{k},\underline{n},\underline{\ell},\Omega)$, where Ω is a vector of a GP's characteristics, for instance gender, medical experience and personal involvement.

Necessary and sufficient (the objective function is concave) conditions for the solution of (6) are found by means of concave programming[3]:

$$q \geq \sigma - \lambda \frac{y}{n} \tag{7}$$

$$q + py[1 + z_1'(k,n,\ell)] = \rho \frac{y}{n} w_1'(k,n,\ell) + \sigma[1 - \frac{y}{n} h_1'(k,n,\ell,\Omega)] \tag{8}$$

$$npz_3'(k,n,\ell) + v'(\ell) = \rho w_3'(k,n,\ell) - \sigma h_3'(k,n,\ell,\Omega) \tag{9}$$

where λ, ρ and σ are the Lagrange parameters of the first, second and third constraint, respectively. We are now ready to study the effect of the payment system on the GP's optimal choice of the number of consultations, patients and leisure. Since the empirical part of our

[3] Details of the derivation are found in appendix A.

study applies data from a combined capitation and fee for service system, we will confine our discussion to this case. We have the following result:

<u>*Result:*</u> *With a combined capitation and fee for service payment the minimum amount of physician-initiated consultations (k = 0) with a maximum waiting time (w =* $\bar{w}$*) is expected to occur when patients are abundant. With a shortage of patients the number of physician-initiated visits is expected to increase and the waiting time is expected to decline.*

With a mixed capitation and fee for service system we have $q > 0$, $p > 0$ and r = 0. We assume first that there is no shortage of patients, i.e. $\sigma = 0$. The right hand side of (7) is then smaller or equal to zero. Since the left hand side is greater than zero, we have $\lambda = 0$ and $k = 0$. Since $1 + z_1'(k, n, \ell) > 0$ [4], the left-hand side of (8) is larger than zero. Then ρ has to be positive to satisfy (8). Accordingly we have $w = \bar{w}$.

Assume next that a shortage of patients occurs, i.e. $\sigma > 0$. If $\sigma - q < 0$, then $\lambda = 0$ and $k = 0$. If $\sigma - q = 0$, then $\lambda = 0$ and $0 < k \leq \bar{k}$. If $\sigma - q > 0$, then $\lambda > 0$ and $k = \bar{k}$. From (8) we see that $\rho \geq 0$, implying $0 < w \leq \bar{w}$.

The intuition behinds these results can be described like this: Without a shortage of patients, few physician-initiated visits are desirable for the GP because more patients then can be listed and more income from capitation can be harvested. A long wait goes together with a high utilisation of the GP's working hours and hence, a higher remuneration per hour compared

with a shorter wait. Secondly, a long wait is assumed to initiate a decline in patient-initiated visits, making it possible for the GP to have more patients listed and get more income from capitation.

With a shortage of patients the number of physician-initiated visits depends upon the strength of the rationing, expressed by σ - q. σ is the increase in utility obtained by a marginal loosening of the constraint, i.e. a marginal increase in the number of patients listed. σ - q then expresses the marginal increase in the GP's utility of providing services to an additional patient. Hence, σ - q is negative if the disutility of the time spent providing services dominates the income from the service provision. Then the number of physician-initiated consultations is not increased compared to the minimum number provided in the unconstrained case. The interpretation of σ - q > 0 should then be obvious. In the constrained case the waiting time may be lower than the maximum, $\bar{w}$. There are two reasons for this result. Firstly, a reduction of w is assumed to increase the number of patient-initiated consultations of patients already listed and hence, the GP's income. Secondly, a reduction in w may attract additional patients to the list and hence, loosening the patient constraint. If we focus on the first effect (assuming $h_1'(k,n,\ell) = 0$), we see from (8) that a necessary condition for $w < \bar{w}$ is that σ - q > 0. Adding the potential effect on additional persons listed makes it more likely that $w < \bar{w}$.

FIGURE 1

[4] Assume that $1 + z_1'(k,n,\ell) < 0$. Then the total effect of an increase in k on the number of consultations is negative. Since an increase in k and z both have an equal impact on w, a decline in w would then happen. But a decline in w is not compatible with a decline in patient initiated consultations.

We can illustrate the GP's optimal choice in the unconstrained case by means of figure 1, where indifference contours are depicted. The GP's utility increases when moving towards north-east in the diagram. The shape of an indifference contour can be explained as follows: An exogenously given n, for instance n^0, has an optimal waiting time w^0, determined by the point where an indifference curve has a tangential that corresponds to the vertical line through n^0. For a w greater than the constrained optimal w the number of consultations becomes smaller than in the constrained optimum (because of the waiting time effect on patient-initiated consultations). Hence, the number of patients has to increase for the GP to remain on the indifference curve. For a wait smaller than w^0, the number of consultations becomes larger than in w^0 (because of the waiting time effect on patient-initiated consultations). Hence, the GP has to be compensated for the extra work-load by capitation payment from an increased number of patients. The point A illustrates the GP's optimal number of patients, n^*, determined by the tangential of the horizontal line from the maximum acceptable wait, $\bar{w}$, and an indifference curve.

FIGURE 2

Figure 2 illustrates the constrained case. Since the demand curve now is located to the left of A, the unconstrained optimum A with n^* number of patient is no longer feasible. The constrained optimum B with n^{**} number of patients is characterised by a waiting time shorter than the maximum in order to attract patient-initiated consultations and new patients. If the demand for being listed with a certain GP is inelastic with respect to the waiting time (the demand curve is vertical), the optimal w becomes larger because the GP does not have to consider the negative impact of the waiting time on the number of patients listed.

We may now formulate two hypotheses to be tested in the empirical section of the paper

- To attract more patient-initiated visits and to attract additional patients to the list, a GP who experiences a shortage of patients, is expected to have a shorter waiting time than his unconstrained colleagues
- A GP, who experiences a shortage of patients, has a larger increase in the list of patients the shorter wait he offers

3. Description of the data

Data from one of the municipalities participating in the Norwegian capitation experiment are used. The experiment was launched 18 May 1993 in four municipalities, and lasted three years. All inhabitants were listed with a GP, preferably a GP of their choice. The GP's income consisted of a per capita component per enlisted person and a fee-per-item component. The capitation fee was adjusted for the age of a person on the list, and for whether the physician was a specialist in general medicine. The fee-per-item component was paid according to a fixed fee schedule and was paid partly by the National Insurance and partly by patient charges. About 50 per cent of the income of an average practice were expected to come from the capitation component, and about 50 per cent from the fee-per-item part.

Data on waiting time to get an appointment with the GP (WAIT) were collected once a year; November 1993, November 1994 and November 1995. The number of weekdays until the physician has an available appointment measures the variable. Our waiting time data include 80 GPs, and this is about 90 per cent of the GPs practising in the municipality. We have data

on waiting time at all three points of time for more than 60 per cent of the GPs. For the rest of the GPs one or two registrations are missing.

The composition of patients on a GP's list may influence the waiting time to get an appointment. For instance, of two GPs with an equal number of working hours and number of persons on their list but with different patient loads, the one with the heavier load is expected to have a longer waiting time. The reason is simply that persons on *that* list have a greater need for health care, and hence need to see the doctor more often. As an indicator of patient load we use the proportion of female patients (PROPFEM) and the proportion of patients aged seventy years and older (PROPOLD). These data were collected annually (November 1993, January 1995 and January 1996[5]).

Female physicians seem to have a practice style that differs from their male colleagues (Langwell 1982), and this might also influence on the waiting time. In the analysis we take account of the physician's gender (MALE).

A GP's medical experience will influence the way patients are treated, and also how the practice is organised. It also seems likely that the number of persons that prefer to be listed by a certain GP are influenced by the number of years a GP has been practising in the municipality. We include the number of years a GP has been practising in the municipality (PRACTIME) to take account of these aspects of the GP's practice.

[5] Changes in patient composition occur quite slowly. We therefore used the data from January 1995(1996) together with the waiting time registration in November 1994 (1995).

An important conclusion from our theoretical model is that physicians, who experience a shortage of patients, have a shorter waiting time than their unconstrained colleagues. To account for differences in practice style between the two groups of physicians, we included a dummy variable in the analysis. In autumn 1994 the GPs were asked whether they wanted additional persons on their list. The GPs, who answered this question positively, are denoted rationed. The dummy variable RATION is equal to one when a GP is rationed.

Seventeen GPs in our sample did not give us access to the information of whether they wanted additional patients on their lists. For these GPs the rationing status was determined by comparing the preferred number of patients[6] stated before the experiment started with the actual number during the experiment. This variable is available for all the GPs. We denote those GPs who had a smaller list in November 1993 than they wanted to have, and experienced an increase in the number of patients from November 1993 to January 1995, as lightly rationed. Similarly, those GPs who had a smaller list in November 1993 than they wanted to have and experienced a constant or declining number of patients from November 1993 to January 1995 is denoted as strongly rationed. For 63 GPs both kinds of rationing indicators are available. It turns out that 65 per cent of those who were lightly rationed, expressed that they wanted more patients, while only 16 per cent of the strongly rationed expressed that they wanted more patients. The reason may be that the strongly rationed have adapted to a service intensive practice style that makes them less interested in a longer list[7].

[6] Before the experiment started, all the participating GPs were asked to specify the number of persons they would like to have on their individual list.

[7] That the strongly rationed in fact have a more service intensive style of practice, is shown in Iversen and Lurås (1998).

For the 17 GPs where the information about whether they wanted more patients was missing, we therefore used light rationing as a proxy for whether a GP wanted more patients[8].

We calculated the net change in the number of persons on the GPs' individual lists from one period to the next (CHANGE). This variable is the dependent variable in analysis two, where the purpose is to examine whether a reduction in the waiting time is a mean to attract new persons to the list.

TABLE 1

Our sample consists of 206 observations. Table 1 summarises the descriptive statistics according to rationing status and for the full sample. We see that GPs in the full sample have worked about 10 years in the municipality, and 77 per cent are men. The average individual list size is 1714 persons and the average change in the list from one period to the next is 10 persons. 44 per cent of the GPs are rationed according to our rationing criterion. Persons are on average expected to wait nearly 10 days to get an appointment with their GP. 9 per cent of the persons on the individual lists were aged 70 and older and 51 per cent were women.

Comparing the characteristics according to rationing status we see that rationed GPs on average have shorter lists, experience a greater increase in the list size from one period to the next and have a shorter waiting time than their unrationed colleagues. Measured as the proportion of old and the proportion of females on the list the unrationed GPs have a heavier patient load than the rationed. We also observe that the proportion of males is higher among

[8] The empirical results in section 4 happen to be insensitive to this choice.

the rationed than among unrationed, and that the rationed GPs on average have been practising a shorter period in the municipality.

4. Empirical results

We want to estimate the impact of patient shortage on a GP's waiting time. Important considerations for our choice of estimation procedure were:

- The waiting time observations are count data (integers greater or equal to zero)
- As seen from figure 3, the waiting time distribution is right skewed. Since 9 observations are equal to zero, a logarithmic transformation implies that five per cent of the observations are lost.
- A GP is likely to have a certain practice style related to his personality, experience, the organisation of the practice, etc. Since we have three observations of each physician, our data has a panel data structure and unobserved heterogeneity is likely to create dependence between the time specific error terms for each GP.

FIGURE 3

These features of data (skewed count panel data) have similarities with the data analysed in the classical article by Hausman, Hall and Griliches (1984) and in Cameron and Trivedi (1998). We chose a similar approach starting with the Poisson regression model[9]. The Poisson density function is:

[9] It could be claimed that we should have used the exponential distribution, since the waiting time distribution is exponential when arrivals are Poisson distributed. Hence, there is a duality between waiting time distribution and

$$f(y_{it}|\mathbf{x}_{it}) = \frac{e^{-\mu_{it}}\mu_{it}^{y_{it}}}{y_{it}!}, \qquad y_{it} = 0,1,2,..... \tag{10}$$

with the mean parameter

$$E[y_{it}|\mathbf{x}_{it}] = \mu_{it} = \exp(\mathbf{x}_{\mathbf{it}}'\beta)$$

where y_{it} is the dependent variable with a subscript indicating observation no. t (t = 1,2,3) of GP number i (i = 1,2,....80) in the municipality, and $\mathbf{x_{it}}$ is a vector of independent variables and β is the related vector of parameters showing the impact of the independent variables on the mean parameter. We estimate β by means of maximum likelihood estimation[10].

The results of the estimation of the Poisson model are shown in the second column of Table 2. The effect of patient shortage has the expected sign and is statistically significant at the one per cent level: rationed GPs offer shorter waiting times than GPs experiencing no constraints on the number of persons on the list. The effect of the proportion of females on the list is positive: GPs with a large share of females have longer waiting times. This is in accordance with results from various studies, which find females to be more frequent users of physicians' services, than men are (see for instance Elstad, 1991). On the contrary and perhaps surprisingly, the proportion of old persons on the list contributes to a shorter wait. We also see from the sign of the gender variable that male GPs contributes a longer wait.

TABLE 2

event count (see Cameron and Trivedi (1998, p 106)). As seen from figure 3, the frequency distribution of our waiting times differs from the exponential distribution.

[10] The statistical software LIMDEP 7.0 is used throughout this section.

A characteristic of the Poisson distribution is that the mean is equal to the variance, i.e.: $Var[y_{it}] - E[y_{it}] = 0$. We tested for whether the difference between the variance and the mean in fact increases (decreases) with the mean (Cameron and Trivedi, 1998, p. 78). The relation turned out to be positive and significant. An overdispersion of the data is therefore present.

We corrected for the overdispersion by estimating a negative binomial model. We then have[11]:

$Var[y_{it}] = E[y_{it}]\{1 + \alpha^2 E[y_{it}]\}$

where α^2 characterises the degree of overdispersion. The negative binomial model can be derived as a mixture of the Poisson and Gamma distributions by assuming that the parameter in the Poisson distribution, θ_{it}, has a random intercept term, $\exp(\varepsilon_{it})$:

$\theta_{it} = \exp(\mathbf{x}_{\mathbf{it}}'\beta + \varepsilon_{it}) = \mu_{it} \exp(\varepsilon_{it})$

where $\exp(\varepsilon_{it})$ has a gamma distribution with mean 1.0 and variance α (see Cameron and Trivedi, 1998, p.100).

The results from the estimation of the negative binomial model are shown in column 3 of table 2 (model 2). We see that none of the signs of the estimated effects changes and the negative impact of rationing on patients waiting time is still statistically significant. The overdispersion coefficient is significantly different from zero.

Still we have not taken unobserved heterogeneity among the GPs into account. In the random effect model estimated in LIMDEP 7.0, the random effect is added to the binomial model by assuming that the overdispersion parameter is randomly distributed across groups. The results

from the estimation are shown in the fourth column of table 2. Again, we see that the signs of the estimated coefficient are unchanged and that patient shortage contributes significantly negatively to the waiting time. We also see that the random model is not rejected by the Hausman test.

To get an idea of differences in waiting times between the rationed and unrationed groups of GPs, we calculated the mean of the fitted waits for the two groups. The fitted waiting time for the rationed groups was 6.8 days, and for the unrationed group 12.0 days. Hence, a patient listed with a rationed GP may expect to wait 5.2 days or 43 per cent less than a patient listed with an unrationed GP.

A second hypothesis from the theoretical model is that rationed GPs have a larger increase in their list of patients the shorter their waiting time is. We tested if the waiting time in period $t - 1$ (lagWAIT) had any effect on the change in the list size from period $t - 1$ to period t (t = 2,3) for the group of constrained GPs. Physician specific effects were found to be significant, and we estimated a random effects model. The results are presented in table 3.

TABLE 3

The effect of waiting time on the increase in the number of persons on the list is as expected: the longer waiting time is in one period, the smaller is the increase in the list size in the next period. The magnitude of the effect is however small. One day increase in the waiting time is expected to reduce the number of enrolled patients by 2,26 in the next period. Calculations on

[11] We concentrate on this variant of the negative binomial model (NEGBIN II in Cameron and Trivedi) since this is the model used in the estimation.

basis of the results above show that an average rationed male physician (PROPOLD = 0.1, PROPFEM = 0.5) with a waiting time of one day in one period (lagWAIT) can expect 54 new patients in the next period. The increase in the list size when the waiting time is three and six days, are respectively 49 and 43 new patients in the next period.

5. Concluding remarks

In section 2 we developed two hypotheses about the relationship between patient shortage, waiting time and inflow of new patients on a GP's list:

- To attract more patient-initiated visits and to attract additional patients to the list, a GP who experiences a shortage of patients, is expected to have a shorter waiting time than his unconstrained colleagues
- A GP, who experiences a shortage of patients, has a larger increase in the list of patients the shorter wait he offers

The results from our analysis show that GPs experiencing constraints with regard to the number of persons listed offer their patients a shorter wait: The waiting time to get an appointment with an unrationed GP is 12 days while the rationed GPs offer about 5 days shorter waiting time. Although rationed GPs on average care for 230 fewer patients than their unrationed colleagues do, their shorter wait is a result of the capacity they choose to offer to care for their patients. A shorter wait stems from a lower capacity utilisation and implies that rationed GPs have a lower income per hour than their unrationed colleagues have. They

choose to work more hours for a lower income per hour rather than increasing leisure and capacity utilisation.

Our results suggest that potential patients consider physicians' waiting time when they choose the GP they want to be listed with. But we also found that GPs with their optimal number of patients on average have a longer waiting time than those GPs who want additional patients to care for. A theory by Becker (1991) may shed light on these seemingly conflicting results. Becker argues that up to a certain point excess demand signals a popular producer and hence, adds to the aggregate demand. He shows that there may well be multiple equilibria, where the least popular suppliers attract consumers by lowering the price (here the wait), while the most popular suppliers have an interest in maintaining the excess demand, because it adds to the demand facing them.

In the model we showed that GPs by regulating the waiting time can influence the patient's thresholds for seeking care and hence, the number of patient-initiated consultations. Our results imply that the distinction between patient-initiated and physician-initiated consultations may be less clear cut than often assumed in the literature. Unfortunately, we do not have data that distinguish between patient-initiated and physician-initiated consultations. We are therefore not able to explore the relation any further empirically.

Our analysis can be criticised on several points. We may not have considered all variables of relevance for the determination of the waiting time. For instance, our data do not distinguish between consultations occurring because of acute illness and other consultations. It may be that some GPs take care of acute consultations without registering the patients' wait. In that case our waiting time registrations are likely to overstate the average waiting time. However,

we have no reason to believe that this kind of bias depends on whether a GP experiences patient shortage or not.

The rationing variable used in all three periods is a measure calculated on basis of a question asked the GPs in the second period (1994). Our variable may therefore give the GPs a wrong rationing status in the first (1993) and in the second period (1995) of the analysis. The average rationed GP stated that they wanted 230 more persons on the list in 1994, and only three GPs wanted less than 100 additional persons. When we know that the actual change in list size for GPs in the municipality is quite small from one period to the next, we expect that our rationing variable is a good indicator of patient constraints in period one and three as well.

The composition of patients on the GPs list is not collected at the same point of time as the waiting time variable. Similar to changes in list size, changes in the proportion of females and the proportion of old persons on the lists are quite slow and hence, different registration periods are unlikely to give any systematic bias to the analysis.

The composition of patients on the list is a crude measure that may fail to recognise important aspects of patients' need for health care. A better measure of patients' needs may be the distribution of diagnosis on a GP's list. Variables characterising the GP are also important for explaining differences in waiting time but most of all, important for explaining why individuals choose to be listed by a certain GP. For instance, most patients prefer an individual doctor who shows empathy, and if you have a certain diagnosis you will prefer a doctor with experiences and knowledge in treating this kind of illness. Presently, we do not possess data that could be used for analysing the matching between patients and GPs.

References:

Andersen, T.F., Mooney, G., (eds.), 1990. The challenges of medical practice variation. Macmillan Press, London.

Becker, G.S., 1991. A note on restaurant prices and other examples of social influence on price, Journal of Political Economy 99, 1109-1116.

Cameron, A. C., Trivedi, P.K., 1998. Regression analysis of count data. Cambridge University Press, Cambridge.

Cooper, R. B., 1981. Basic queuing theory. North Holland Inc, New York.

Elstad, J. I., 1991. Flere leger, større bruk? Artikler om bruk av allmennlegetjenester. INAS- Rapport 1991:11. Institutt for sosialforskning, Oslo.

Hausman, J., Hall, B.H., Griliches, Z.,1984. Econometric models for count data with an application to the patent-R&D relationship, Econometrica 52, 909-938.

Iversen, T., Lurås, H., 1998. The impact of economic motives on the provision of health services in general practice. Working Paper 1998:1. Center for Health Administration, University of Oslo.

Langwell, K.M., 1982. Factors affecting the incomes of men and women physicians: further explorations, The Journal of Human Resources 17, 261 – 276.

Mueller, C.D., 1985. Waiting for physicians' services: Model and evidence, Journal of Business 58, 173-190.

Sloan, F.A., Lorant, J.H., 1977, The role of waiting time: Evidence from physicians' practices, Journal of Business 50, 486-507.

Appendix A:

Derivation of necessary and sufficient conditions

The Lagrangian is:

$$L(k,n,\ell) = g + qn + np[\underline{z} + z(k,n,\ell) + \underline{k} + k] + v(\ell)$$
$$-\lambda(k - \bar{k})$$
$$-\rho(w(k,n,\ell) - \bar{w})$$
$$-\sigma(n - h(k,n,\ell,\Omega))$$

A necessary and sufficient[12] condition for $k \geq 0$, $n > 0$ and $\ell > 0$ to solve the problem is that there are non-negative λ, ρ and μ such that:

$$\frac{\partial L(k,n,\ell)}{\partial k} \leq 0, \qquad k = 0 \qquad or \qquad \frac{\partial L(k,n,\ell)}{\partial k} = 0 \tag{A1a}$$

$$\frac{\partial L(k,n,\ell)}{\partial n} = 0 \tag{A1b}$$

$$\frac{\partial L(k,n,\ell)}{\partial \ell} = 0 \tag{A1c}$$

$$0 \leq k \leq \bar{k} \qquad \lambda = 0 \qquad if \qquad k < \bar{k} \tag{A1d}$$

$$0 < w(k,n,\ell) \leq \bar{w} \qquad \rho = 0 \qquad if \qquad w(k,n,\ell) < \bar{w} \tag{A1e}$$

$$0 < n \leq h(k,n,\ell,\Omega) \qquad \sigma = 0 \qquad if \qquad n < h(k,n,\ell,\Omega) \tag{A1f}$$

[12] The conditions are sufficient since the objective function is concave.

The explicit expressions for $\partial L/\partial k$, $\partial L/\partial n$ and $\partial L/\partial \ell$ are found from the partial derivatives of the Lagrangian:

$$\begin{aligned}
\frac{\partial L(k,n,\ell)}{\partial k} &= np[z_1'(k,n,\ell)+1]-\lambda-\rho w_1'(k,n,\ell)+\sigma h_1'(k,n,\ell,\Omega)\le 0\\
\frac{\partial L(k,n,\ell)}{\partial n} &= q+p[y+nz_2'(k,n,\ell)]-\rho w_2'(k,n,\ell)-\sigma[1-h_2'(k,n,\ell,\Omega)]=0\\
\frac{\partial L(k,n,\ell)}{\partial \ell} &= npz_3'(k,n,\ell)+v'(\ell)-\rho w_3'(k,n,\ell)+\sigma h_3'(k,n,\ell,\Omega)=0
\end{aligned} \qquad \text{(A2)}$$

where y=$\underline{z}$+z(k,n, ℓ) + $\underline{k}$+k, corresponds to the total number of consultations.

We wish to simplify (A2). By inserting from (4) into (3) and implicit differentiation with respect to *k* and *n* we get[13]:

$$w_2'(k,n,\ell)=\frac{y}{n}w_1'(k,n,\ell) \qquad \text{(A3)}$$

where $w_i'(k,n,\ell)$ is the derivative of the waiting time function wrt argument no. i.

By differentiating (5) with respect to *k* and *n* and inserting from (A3) we get:

$$z_2'(k,n,\ell)=\frac{y}{n}z_1'(k,n,\ell) \qquad \text{(A4)}$$

where $z_i'(k,n,\ell)$ is the derivative wrt argument no. i.

By inserting from (A3) and (A4) into (A2) we get:

$$np[z_1'(k,n,\ell)+1]\le \lambda+\rho w_1'(k,n,\ell)-\sigma h_1'(k,n,\ell,\Omega) \qquad \text{(A5)}$$

$$q+py[1+z_1'(k,n,\ell)]=\rho\frac{y}{n}w_1'(k,n,\ell)+\sigma[1-\frac{y}{n}h_1'(k,n,\ell,\Omega)] \qquad (8)$$

$$npz_3'(k,n,\ell) + v'(\ell) = \rho w_3'(k,n,\ell) - \sigma h_3'(k,n,\ell,\Omega) \tag{9}$$

where $\frac{y}{n} h_1'(k,n,\ell,\Omega) = h_2'(k,n,\ell,\Omega)$ follows from $\bar{n} = \bar{n}[w(k,n,\ell),\Omega] = h(\underline{k},\underline{n},\underline{\ell},\Omega)$ and (A3).

Finally, by inserting the expression for $1 + z_1'(k,n,\ell)$ from (8) into (A5) we get:

$$q \geq \sigma - \lambda \frac{y}{n} \tag{7}$$

[13] Details may be found in appendix B.

Appendix B:

Derivation of eqs. (A3) and (A4)

Inserting from (4) into (3):

$$w = \frac{t\,C\left(\frac{nt[\underline{z} + g(w,k,n,\ell) + \underline{k} + k]}{T - \ell}\right)}{T - \ell - nt[\underline{z} + g(w,k,n,\ell) + \underline{k} + k]} \qquad \text{(B1)}$$

Implicit derivation of (B1) with respect to *k* and *n* gives:

$$\frac{\partial w}{\partial k} = \frac{C'(\cdot)\,n\left(g'\frac{\partial w}{\partial k} + 1\right)\left(1 - \frac{ny}{\mu}\right) + C'(\cdot)\,n\left(g'\frac{\partial w}{\partial k} + 1\right)}{(\eta - ny)(\eta - ny)}$$

$$= \frac{n\left[C'(\cdot)\,g'\left(1 - \frac{ny}{\mu}\right) + C'(\cdot)\,g'\right]}{\left[(\eta - ny)(\eta - ny) - C'(\cdot)\,ng'\left(1 - \frac{ny}{\mu}\right) - C'(\cdot)\,ng'\right]} \qquad \text{(B2)}$$

$$\frac{\partial w}{\partial n} = \frac{C'(\cdot)\,g\left(y + ng'\frac{\partial w}{\partial n}\right)\left(1 - \frac{ny}{\mu}\right) + C'(\cdot)\,g\left(y + ng'\frac{\partial w}{\partial n}\right)}{(\eta - ny)(\eta - ny)}$$

$$= \frac{y\left[C'(\cdot)\,g'\left(1 - \frac{ny}{\mu}\right) + C'(\cdot)\,g'\right]}{\left[(\eta - ny)(\eta - ny) - C'(\cdot)\,ng'\left(1 - \frac{ny}{\mu}\right) - C'(\cdot)\,ng'\right]}$$

$$\frac{\partial w}{\partial n} = \frac{y}{n}\frac{\partial w}{\partial k}$$

From (B2) we then have (A3):

Implicit differentiating of (5) with respect to *k* and *n*:

$$\frac{\partial s}{\partial k} = g'\frac{\partial w}{\partial k} = \frac{\partial z}{\partial k} \qquad \text{(B3)}$$

$$\frac{\partial s}{\partial n} = g'\frac{\partial w}{\partial n} = \frac{\partial z}{\partial n}$$

Inserting from (A3) into (B3) gives (A4).

Table 1: Descriptive statistics – mean (standard deviation) - full sample and for the two groups according to rationing status

Variable	Definition	Rationed (N = 90)	Unrationed (N = 116)	Full sample (N = 206)
PROPOLD	The proportion of persons aged 70 and older on the list	0.09 (0.05)	0.10 (0.05)	0.09 (0.05)
PROPFEM	The proportion of females on the list	0.47 (0.10)	0.55 (0.11)	0.51 (0.11)
MALE	A dummy variable equal to one if the GP is a male	0.90	0.66	0.77
RATION	A dummy variable equal to one for GPs who want more persons on their list			0.44
WAIT	The number of weekdays to get an appointment with the GP	7.24 (7.79)	11.55 (7.95)	9.67 (8.15)
CHANGE	The change in persons on the GPs individual list from period t - 1 to period t, t = 2, 3	44.39 (62.81)	-16.29 (68.25)	10.04 (72.30)
LIST SIZE	The actual number of persons on a GPs individual list	1578 (209)	1820 (336)	1714 (311)
PRACTIME	The number of years the GP has been practising in the municipality	9.20 (6.18)	10.77 (5.32)	10.08 (5.75)

Table 2: The estimated effect (standard deviation) of independent variables on waiting times (WAIT) [14]

Variable	Model 1: Poisson	Model 2: Negative binomial	Model 2: Negative binomial with random effects
CONSTANT	0.94(0.24)**	1.38 (0.53)**	0.25 (0.58)
RATION	-0.34(0.05)**	-0.46 (0.13)**	-0.40 (0.11)**
PROPOLD	-2.80(0.55)**	-3.51 (1.56)*	-3.85 (1.27)**
PROPFEM	2.81(0.36)**	2.35 (0.80)**	2.62 (0.81)**
MALE	0.32(0.10)**	0.21 (0.28)	0.36 (0.22)
Number of observations	206	206	206
Overdispersion coefficient ($\hat{\alpha}$)		0.52 (0.06)**	
- Log likelihood	935.5	658.2	650.9
Hausman test			CHISQ(2) = 0.189 p-value = 0.91

[14] * (**) indicates that the estimated coefficient is significantly different from zero at 5 per cent (1 per cent) level.

Table 3: The estimated effect of independent variables on the change in list size for rationed GPs. (Standard deviation in parenthesis)

Variable	**Dep.variable CHANGE**
CONSTANT	58.31(92.06)
lagWAIT	-2.46(1.23)*
PROPOLD	-473.69(186.01)*
PROPFEM	48.90(144.48)
GENDER	21.08(41.33)
Number of observations	59
Hausman test	CHISQ(2) = 1.01 p-value = 0.61
R^2	0.17

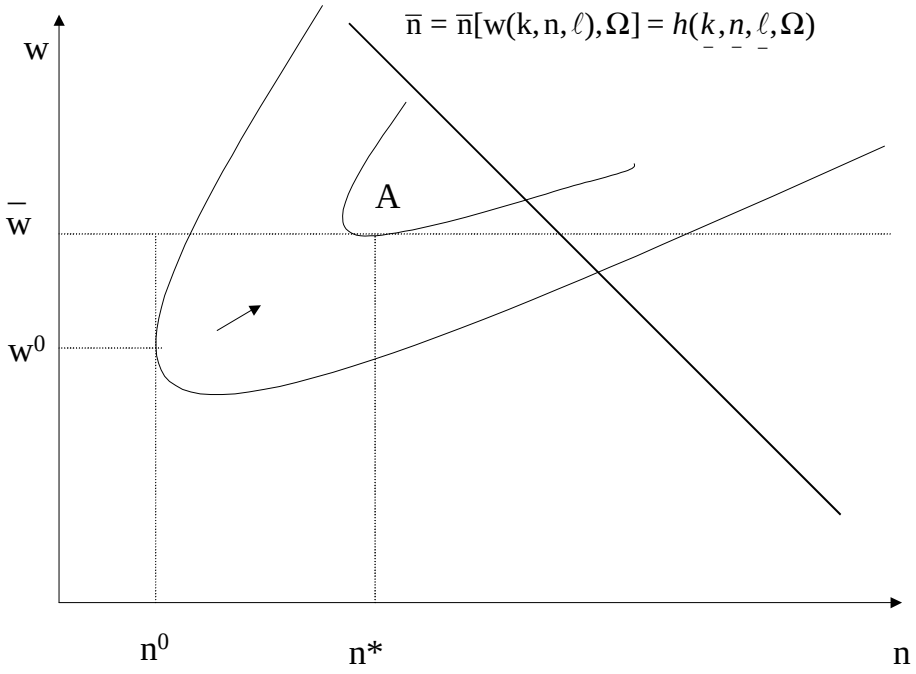
$\bar{n} = \bar{n}[w(k, n, \ell), \Omega] = h(\underline{k}, \underline{n}, \underline{\ell}, \Omega)$
w
$\underline{w}$
w^0
A
n^0
n*
n

Figure 1

Figure 2

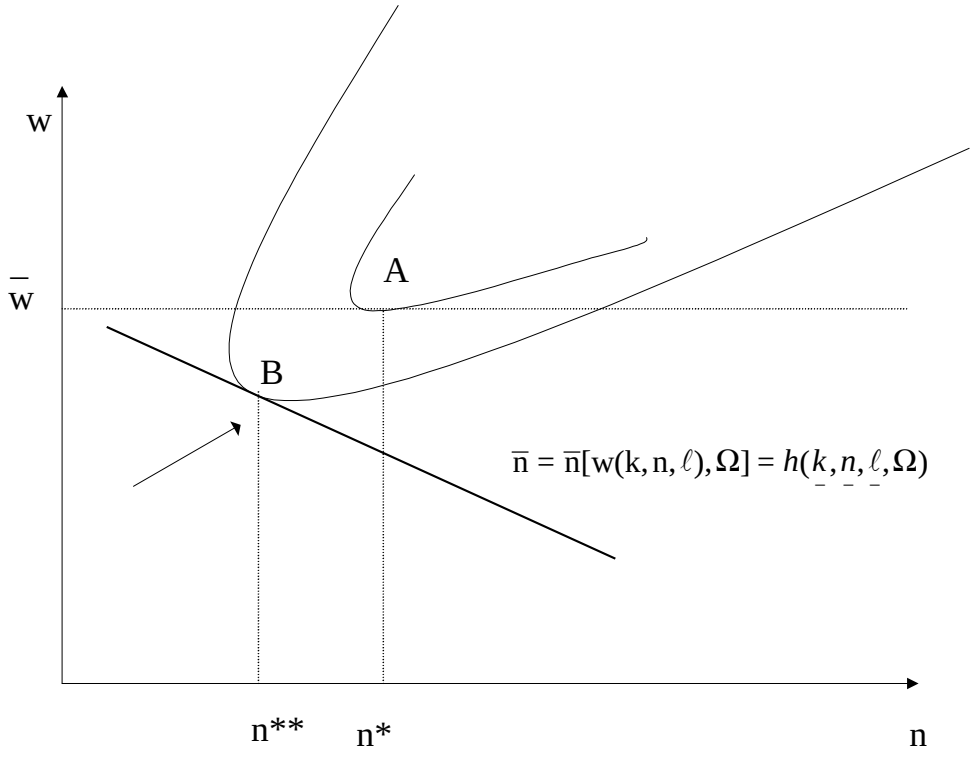
w
$\bar{w}$
A
B
$\bar{n} = \bar{n}[w(k, n, \ell), \Omega] = h(\underline{k}, \underline{n}, \underline{\ell}, \Omega)$
n**
n*
n

Figure 3: Frequency distribution of WAIT