UNIVERSITY
OF OSLO
HEALTH ECONOMICS
RESEARCH PROGRAMME

The effect of patient shortage on general practitioners' future income and list of patients

Tor Iversen
Center for Health Administration
University of Oslo

Working Paper 2003: 1

UNIVERSITAS OSLOENSIS
MDCCCXI

The effect of patient shortage on general practitioners' future income and list of patients

by Tor Iversen*

December 2002

Health Economics Research programme at the University of Oslo
HERO 2003

JEL Classification: H42; I11; I18

Keywords: Economic motives;Capitation; General practice; Patient shortage; Service provision

Author: E-mail tor.iversen@samfunnsmed.uio.no
Telephone 47 23075308/23075301 / Fax 47 23075310
Center for Health Administration, Rikshospitalet, NO-0027 Oslo, Norway

* Thanks to Sverre Grepperud and Hilde Lurås for helpful comments to a previous version of the paper. I am also grateful to the participants at the Lisbon workshop on health economics 15-16 February 2002, and in particular to Frank Windmeijer for constructive criticism, to participants at the International Health Economics Conference at Hong Kong University of Science & Technology March 11-13 2002, and seminar at the Department of Economics, University of Alicante, Spain 16 December 2002. None of them are responsible for remaining errors and shortcomings. Financial contributions from the Norwegian Ministry of Health and Social Affairs and the Research Council of Norway are acknowledged.

Health Economics Research programme at the University of Oslo
Financial support from The Research Council of Norway is acknowledged.
ISSN 1501-9071, ISBN 82-7756-104-0

ABSTRACT

The literature on supplier inducement suffers from inability to distinguish the effect of better access from the effect of patient shortage. Data from the Norwegian capitation trial in general practice give us an opportunity to make this distinction and hence, study whether service provision by physicians is income motivated.

In the capitation trial each general practitioner (GP) has a personal list of patients. The payment system is a mix of a capitation fee and a fee for service. The data set has information on patient shortage, i.e. a positive difference between a GP's preferred and actual list size, at the individual practice level. From a model of a GP's optimal choice we derive the optimal practice profile contingent on whether a GP experiences a shortage of patients or not. To what extent GPs, who experience a shortage, will undertake measures to attract patients or embark on a service intensive practice style, depends on the costs of the various measures relative to their expected benefit. The model classifies GPs into five types.

In the empirical analysis a panel of GPs is followed for five years. Hence, short-term effects due to transition to a new system should have been overcome. We show that even in the longer run, GPs who experience a shortage of patients have a higher income per listed person than their unrationed colleagues. This result is robust with regard to correction for potential selection bias based on observable and unobservable characteristics. We do not find any significant difference in income per listed person dependent on whether a rationed GP obtains an increase in the number of patients or not. A policy implication is that patient shortage is costly to the insurer because of income motivated behavior of unknown benefit to the patient.

1. Introduction

Much of the literature on supplier induced demand[1] suffers from insufficient data about the number of patients a physician is supposed to serve. Often aggregate data like physician density in a geographical area are used. But aggregate data are misleading if the variation in number of patients per physician within an area is considerable. Some physicians may then have a high level of service provision because they have many patients, while others may have a high level of service provision because many services are delivered to a small number of patients. Hence, with aggregate data we do not know whether a high level of service provision is due to many patients or to many services per patient or a combination. In this context a list patient system (capitation system) has an important advantage. Since each physician has a defined group of patients to care for, we know whether a high practice income is caused by a long list or by a high level of service provision per person listed.

Variation in service intensity among physicians is likely because the absence of clinical guidelines leaves a lot of room for individual judgement. A high level of services per listed person may well be a result of a physician's preferences and not influenced by economic motives. Hence, there is a need for an indicator of whether a high level of service provision per patient is caused by economic motives or not. One such indicator is whether a physician experiences a shortage of patients or not. Physicians with a shorter list than they would have preferred, are in a constrained optimum, implying that their combination of income and leisure are suboptimal compared to what their preferred list would have resulted in. Given that fees are rewarding, a possible response to such patient shortage is to increase the number of services delivered to patients on their list. Since this would not have been done in a situation without patient shortage, this action is said to be income motivated and not motivated by the physician's preferences for a certain practice style.

This approach to the study of economic incentives has been taken in the evaluation of the Norwegian capitation trial in general practice. Before this trial was implemented, each GP informed the National Insurance Administration about the number of persons they would like to have on their lists. The number was used to prevent that some GPs got a heavier workload than they wanted to have. But it also turned out that some GPs got a considerably shorter list of patients than they wanted to have. We therefore got an empirical distinction between those GPs who experienced a shortage of patients (rationing) and those who did not. This distinction is considered to be a fruitful one for studying the impact of economic motives on the intensity of service provision.

In Iversen and Lurås (2000) we present results from the capitation study. In short, we find that physicians who experience a shortage of patients have higher income, longer and more frequent consultations and more laboratory tests per listed person than their unconstrained colleagues. We conclude that not only preferences for a certain practice style, but also the economic motive is a factor of considerable importance for the level and composition of service provision in general practice.

A possible objection to the conclusions in Iversen and Lurås (2000) is related to the short time period of data collection and that increased service provision may be caused by an increase in the number of patient initiated consultations. For instance, patients with chronic conditions may have wanted to see their new GP to be advised regarding the treatment of their condition. Patients listed with a rationed GP may have experienced better accessibility than patients listed with a GP with his preferred list size. The difference in accessibility between the two types of GPs may have resulted in a difference in the number of consultation per listed person. The difference in accessibility may also have encouraged some people to switch to a GP with patient shortage and better accessibility. Hence, we may have discovered a temporary effect related to the introduction of the capitation system. After a while a leveling of service intensity and lists may have occurred. In the long run the effect of an initial patient shortage may therefore have disappeared.

We are able to test this objection since all four trial municipalities on a voluntarily basis decided to continue the capitation system until the system was introduced nationwide June 1st 2001. In the study presented in this paper, data for service provision and patient lists are available for five years (1994-1998). The main result is that patient shortage

[1] A review of some of the recent literature is found in Iversen & Lurås (2000).

also in the longer run has a positive effect on the provision of services per patient. On average, GPs who experienced a shortage of patients provided 15 per cent more services per patient listed than those who were not constrained. We did not find significant differences in service provision between those rationed GPs who obtained an increasing list of patients and those who did not.

The paper proceeds as follows. Section 2 models optimal decisions for a GP who experiences patient shortage. We claim that a GP who experiences a shortage of patients, may choose one of the following actions or a combination of them: improve accessibility to attract more patients, increase the intensity of health service provision per patient or enjoy more leisure. Section 3 describes the data and presents descriptive statistics. Estimation method and results are found in Section 4, and Section 5 concludes with a short discussion of the results and suggestions for further work.

2 The model

The model illustrates at the theoretical level the effects of patient shortage on a representative GP's behavior. At the outset we may think of several responses to a shortage of patients:

- providing better access to encourage more patient initiated consultations and to attract more patients to the list,
- embarking on a more service intensive practice style to harvest more income from each patient,
- contracting with the municipality on providing physician services to nursing homes, schools and various preventive services,
- reducing the working hours to enjoy more leisure,
- moving the practice to another geographical location,
- and combinations of the above alternatives

We concentrate on the alternatives covered by our data[2] and study the conditions for

[2] For instance, the option of contracting with the municipality on providing physician services to nursing homes, schools and various preventive services is omitted in this version of the model since no data that describe these

making each alternative an optimal response. Hence, we end up with a classification of optimal responses. The level of access, *s*, is measured by an index of characteristics that patients are supposed to appreciate; for instance opening hours, waiting time in the waiting room, waiting time from the booking of an ordinary consultation until the consultation takes place, the physical accessibility of the physician's office etc. The physician's cost of providing access is partly a fixed cost, *C(s)*, unrelated to the number of consultations, and partly a time cost, *t(s),* related to each consultation. The fixed cost depends on the attractiveness of the premises, and the services offered, including assortment of laboratory tests. The time cost takes account of that in a stochastic world better access also implies more spare capacity, and hence, more total time devoted to each consultation. We assume, $\mathrm{C}'(\mathrm{s})>0, \mathrm{C}''(\mathrm{s}) >0, \mathrm{t}'(\mathrm{s}) >0 \text{ and } \mathrm{t}''(\mathrm{s}) >0$, where $'$ $('')$ denotes first (second) order derivative.

As described in Scott (2000), there is no unanimity in the literature with regard to the modeling of GP behavior. Here a simple framework is chosen. The central assumption is that a professional norm prescribes that the number of services provided to a patient must be within an interval where the marginal effect of services on health is not documented to be different from zero. This interval is wider the less clinical guidelines there are. The assumption implies that the GP's income is not balanced against a patient's health, and implies that we do not need to introduce disutility from inducement found in some models. A relaxation of this assumption increases the effect of economic incentives. The assumption also implies that less practice guidelines imply more scope for economic incentives.

The physician maximizes a quasilinear utility function in monetary terms, $c+v(\ell)$, where c is income (all income is consumed) and ℓ is leisure. We assume $v'(\ell)>0$ and $v''(\ell)<0$. In this version of the model patients are homogenous and only one type of health service is provided[3]. The net income is defined as $w+qn+pn(\alpha+k)-C(s)$, where w is a fixed salary or practice allowance, q is a capitation payment per person on the physician's list, p is the fee per item of health service (or equivalently, a fee for

activities at the physician level are available.

[3] For a model that takes account of heterogeneous patients and several services for a fixed accessibility, see Iversen and Lurås (2000).

service) and n is the number of listed patients. The lowest acceptable level of service provision is denoted by α, and k is the provision of health services above that level at the physician's discretion. The definition of leisure is $T - t(s)n(\alpha + k)$ where T is the exogenous total time endowment. The practice profile is characterized by number of patients, number of services per patient and level of accessibility. The physician maximizes his constrained utility function:

$$\underset{n,k,s}{Max} \quad w + qn + pn(\alpha + k) - C(s) + v[T - t(s)n(\alpha + k)]$$

$$s.t.$$

$$(i) \quad 0 \leq k \leq \beta$$

$$(ii) \quad 0 < n \leq n^d(s, S, \omega, \Omega, N, M)$$

Constraint (i) says that the number of services provided to a patient must be within the interval, [α, α+β], where the marginal effect of services on health is not documented to be different from zero[4]. Constraint (ii) says that the number of patients is less or equal to the demand for being added to the GP's list, $n^d = n^d(s, S, \omega, \Omega, N, M)$, where

S is a vector of access offered by other GPs in the area, ω is a vector of exogenously determined characteristics of the GP as perceived by (potential) patients, and Ω is a vector that describes similar characteristics of other GPs in the area. N is the number of residents in the area and M is the number of GPs. Hence, when (ii) is non-binding the number of people who wish to be listed with the GP is greater than his preferred number. Some people who prefer to be listed with that GP are then turned away and listed with another GP. When (ii is binding, the GP is rationed in the sense that the number of people who wish to be listed with him for a low level of access (s=0) is smaller than the GP's preferred number. The GP is assumed to be able to influence the demand for being listed with him by increasing the level of accessibility in his practice; i.e. $\partial n^d/\partial s \geq 0$. One could claim that also the number of services provided should have an impact on the demand for being listed with a GP. We have not included this factor in the model, although it would strengthen our conclusion about the effect of patient shortage on service provision.

The maximization problem is analysed by means of concave programming. Necessary and sufficient conditions for k≥0, n≥0 and s≥0 to solve the problem is that there are

[4] With strict evidence based medicine and practice guidelines, we would have β=0.

non-negative λ and μ such that:

$$\frac{pn-\lambda}{t(s)n} \leq v'(\ell) \tag{1a}$$

$$\frac{q+p(\alpha+k)-\mu}{t(s)(\alpha+k)} \leq v'(\ell) \tag{1b}$$

$$\frac{\mu n_s^{'}(s;S,\omega,\Omega,N)-C'(s)}{t'(s)n(\alpha+k)} \leq v'(\ell) \tag{1c}$$

where the lhs of (1a) expresses the utility of income per time unit from an additional consultation, the lhs of (1b) expresses the utility of income per time unit from an additional person listed and the lhs of (1c) expresses the utility of net income per time unit from marginal accessibility offered. Hence, (1a)-(1c) state that the utility of income per time unit from these alternative sources at the margin is less or equal to the marginal utility of leisure.

Since the empirical data in this study are from a list patient system with a combination of capitation payment and fee for service we assume w=0, q>0 and p>0[5].

We further assume that the optimal number of patients is strictly positive such that (1b) is fulfilled with equality.

We can now distinguish between 5 types of GPs, according to level of access and volume of health services per patient. Table 1 exhibits the classification. The formal condition that each type satisfies is derived from (1a)-(1c) in the appendix.

[5] Hence, we shall not study the effect of various payment schemes on physician behavior, as in Ferrall et al. (1998).

Table 1 Classification of GPs according to optimal choice of health service provision, and level of access.

		Health service provision (k)	
		$k=0$	$0<k\leq\beta$
Level of access (s)	$s=0$	1, 5	3
	$s>0$	2	4

Type 1: The unrationed($k=s=0$)

This GP experiences an excess demand of people who prefer to be listed with him. Since the capitation fee is positive, there is always more rewarding to add new people to the list than to increase the level of service provision to those already listed[6]. Hence, $k=0$ and the minimum service intensity is offered. Also, $s=0$ since high accessibility is unnecessary because of other characteristics appreciated by patients.

The remaining four types have in common that they do not achieve their optimal number of patients with $s=0$; hence $\mu>0$, i.e. and they are constrained with regard to the number of patients.

Type 2: Improved access ($s>0$, $k=0$)

For this type of GP the net income per time unit of a marginal increase in *s* and hence, *n* is equal to the marginal utility of leisure. Furthermore, the income per time unit of providing health services beyond the minimum is lower than the marginal utility of leisure. Typically, this type of GP experiences a high return of improved access in terms of more patients listed.

[6] In Iversen and Lurås (2000) we show that this may not be true with more than one type of service, if relative fees deviate too much from relative costs of providing them.

Type 3: High service intensity ($s=0, 0<k\leq\beta$)

This implies that income per time unit from health service provision is greater than the marginal utility of leisure because the intensity of service provision is constrained by the professional norm. The net income per time unit of providing better access is smaller than the marginal utility of leisure; for instance because of a small effect of accessibility on the number of people who want to be listed.

Type 4: Improved access and high service intensity ($s>0, 0<k\leq\beta$)

This type of GP is between type 2 and type 3 regarding the return from improved access in terms of an increased list of patients. The income per time unit of providing health services is greater or equal to the income per time unit (at the margin) from providing better access, which is equal to the marginal utility of leisure.

Type 5: More leisure ($k=s=\tau=0$)

Since this GP is rationed, he would have preferred less leisure, more patients and higher income, but the income from all kinds of additional efforts is too small to compensate for the loss of leisure. In table 1, type 5 is located in the same cell as type 1, since both types have s=k=0. But contrary to type 1, type 5 is interested in having a longer list for s=0. Hence, he experiences a kind of forced leisure.

From the demand function, $n^d = n^d(s, S, \omega, \Omega, N, M)$, we have that demand for a GP depends upon the accessibility offered by the GP relative to other GPs, the perceived exogenous characteristics relative to other GPs and the total number of residents the GPs compete for. Hence, there is an interdependency of optimal choices of GPs within a geographical market. This interdependency could be modeled as a Nash-equilibrium in a non-cooperative game. We shall, however, not go further into this issue here.

3 Description of the data

Data are from the Norwegian capitation trial that lasted from 1993 to 1996, and preceded the system introduced nationwide from June 1st 2001. The trial included four municipalities with a total of 250,000 inhabitants and 150 GPs. All inhabitants were listed with a GP, preferably a GP of their own choice. All GPs in the municipalities participated in the trial. Hence, the problem of self-selection should be considerably less than in systems where GPs select the kind of system they wish to join. The GPs were not allowed to selectively refuse to list a person, but they were permitted to close their list for new patients when the list became too long.

In Norway, primary care is the responsibility of the municipalities, which constitute the lowest level of government. Some GPs are municipal employees on a fixed salary, while the majority is privately practicing and contracting with the municipality. The financing of general practice is split between the state (the National Insurance Scheme), the municipalities and the patients. Before the trial started, the municipalities paid an input based practice allowance to the privately practicing GPs depending on opening hours, number of auxiliaries etc, while the state paid a fee for service component according to a fixed schedule negotiated between the state and the Norwegian Medical Association. In the capitation trial previously employed GPs became privately practicing. Compared with the former system, the practice allowance was replaced with a capitation component that depended on the number of patients a GP had on his personal list. Only minor changes were done to the fee for service schedule. All GPs in our study were privately practicing during the trial. The patients pay a fixed fee per consultation with an annual ceiling. If the ceiling is reached, the fee is refunded from the National Insurance. Patients' copayment in the trial was similar to the rest of the country.

When the capitation trial ended, all four municipalities decided to continue the list patient system with the same organization and payment system for GPs as during the trial. Hence, we got the opportunity of having data for a longer period than the trial's duration.

Data on annual income from fee for service (including copayment) are collected for the five subsequent years 1994-1998. For each year the annual income from fee for service per listed person (*Inpercap*) is calculated for each GP. The GPs in one of the municipalities did not report data on copayment after the trial ended. Hence, for GPs in this municipality (*Municip2),* we have income data for two years. The GPs in the other three municipalities (*Municip1, Municip3, Municip4)* provided complete income data for all five years. In the empirical analysis the income from fees per listed person is used as an indicator of the intensity of a GP's service provision.

The characteristics of a list is described by the number of people on the list (*List size*), the proportion of women (*Propfem*) and the proportion aged 70 and older (*Propold*). Several studies have suggested that women and old people are more frequent users of health services than other groups. Hence, the variables describing the composition of the list are included to account for variations in workload because of variation in patients' need. These data were collected at November 1st 1993, January 1st 1995, January 1st 1996, January 1st 1997, January 1st 1998 and January 1st 1999.

Before the capitation trial was initiated, each GP stated the number of people he would like to have on his list (*Prelistsize*). Since GPs' preferred workload vary, the preferred number of people on the list is likely to vary substantially between GPs. For each GP we compare preferred and actual list size and obtain an indicator of whether a GP is constrained regarding the number of people listed. We denote GPs with a shorter list than preferred, as rationed (*Ration*), corresponding to the GP types 2-5 in the classification scheme in Section 2.

For one of the municipalities we have accessibility data measured by the time from the booking of an appointment to a consultation takes place. In Iversen and Lurås (2002) we show that a patient listed with a rationed GP may expect to wait 3.7 days or almost 40 per cent less than a patient listed with an unrationed GP. We also find that among the rationed GPs a reduction in waiting time had the effect of increasing the number of listed persons in the next period. Hence, we find empirical support for the hypothesis that improved accessibility attracts additional patients, as suggested in Section 2.

Unfortunately, we have no data on accessibility for the other municipalities. Hence, we cannot study whether improved access affects list size positively. We therefore classify the rationed GPs according to whether their lists in fact increase or decline. We have done this classification in two alternative ways described in detail in Section 4.

Descriptive statistics of the variables according to rationing status is presented in Table 2.

Table 2 Descriptive statistics - mean (standard deviation) of the variables in all periods

Variable	Definition	*Ration*= 0 (sufficient patients) (167)	*Ration*= 1 (desired< actual) (322)	All GPs (489)
Inpercap	Annual income from fees and patient charges per listed person in NOK	NOK 216 (NOK 79)	NOK 266 (NOK 88)	NOK 249 (NOK 88)
Prelistsize	A GP's statement of preferred list size before the capitation experiment started	1694 (476)	1978 (385)	1881 (440)
List size	The actual number of persons on a GP's individual list	1784 (413)	1669 (356)	1708 (380)
Propold	The proportion of persons aged 70 and older on the list	0.083 (0.059)	0.112 (0.061)	0.102 (0.062)
Propfem	The proportion of females on the list	0.531 (0.106)	0.496 (0.105)	0.508 (0.107)
Female	A dummy variable equal to one if the physician is a female	0.371	0.193	0.254
Age	The GP's age in years	42.5 (6.8)	43.6 (6.2)	43.2 (6.4)
Salaried	A dummy variable equal to one if the physician was a salaried community physician prior to the trial.	0.425	0.323	0.358

The data set includes 489 observations of 109 GPs[7], and represents accordingly an unbalanced panel. In Table 2 the GPs are split according to whether they were rationed at the outset of the trial (*Ration*). A fairly large proportion of the GPs turns out to be rationed. The average annual income from fees and patient charges per listed person is NOK 249[8]. The amount is considerably higher (NOK 266) for the rationed than for the unrationed GPs (NOK 216). The rationed GPs also wanted considerably longer lists than their unrationed colleagues did. But the actual list size is on average longer for the

[7] The remaining 41 GPs (150 – 109) were omitted because they provided incomplete income data, were not practicing at the start of the trial or had a list (less than 500 people) that indicated that general practice was not their main work.

unrationed than for the rationed. The proportion of old people on the lists is on average 10.2 per cent. The proportion is somewhat higher among the rationed than among the unrationed. The proportion of women on the lists is somewhat higher among the unrationed than among the rationed GPs. This is probably related to the presence of a higher proportion of female GPs (*Female*) among the unrationed than among the rationed. We also see that among the unrationed a higher proportion was employed on fixed salary (*Salaried*) before the trial started. We expect former salaried GPs to have lower income from fees, since they are supposed to be less familiar with the fee schedule than their colleagues with a private practice prior to the trial.

The figures 1 and 2 show the list size and income per listed person according to periods of observation and rationing status. There is a considerable decline in the mean list size of the unrationed and a small increase of the mean list size of the rationed. Those among the rationed who have a longer list in 1999 than in 1993 (*Ration_A*), have on average 118 more people on the list in 1999 than in 1993. Those who experienced a decline (*Ration_B*), have on average a reduction of 80 people listed. On the other hand, we see from Figure 2 that the group *Ration_B* has a greater increase in the income per listed person than the group *Ration_A*.

[8] 7.5 NOK is approximately 1 USD.

Figure 1 Average list size according to rationing status

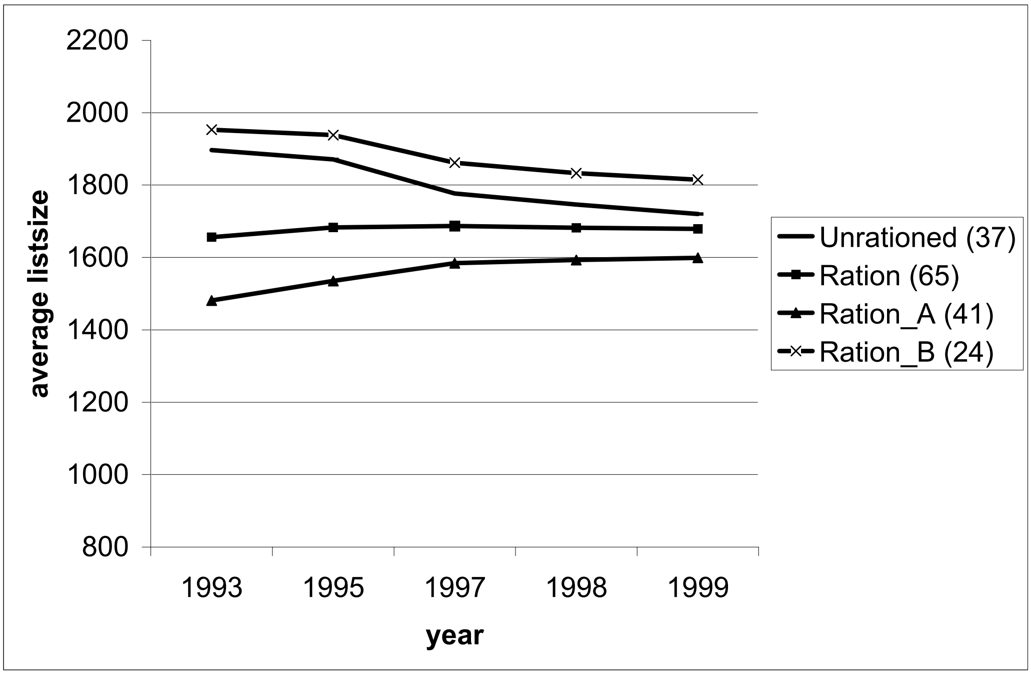
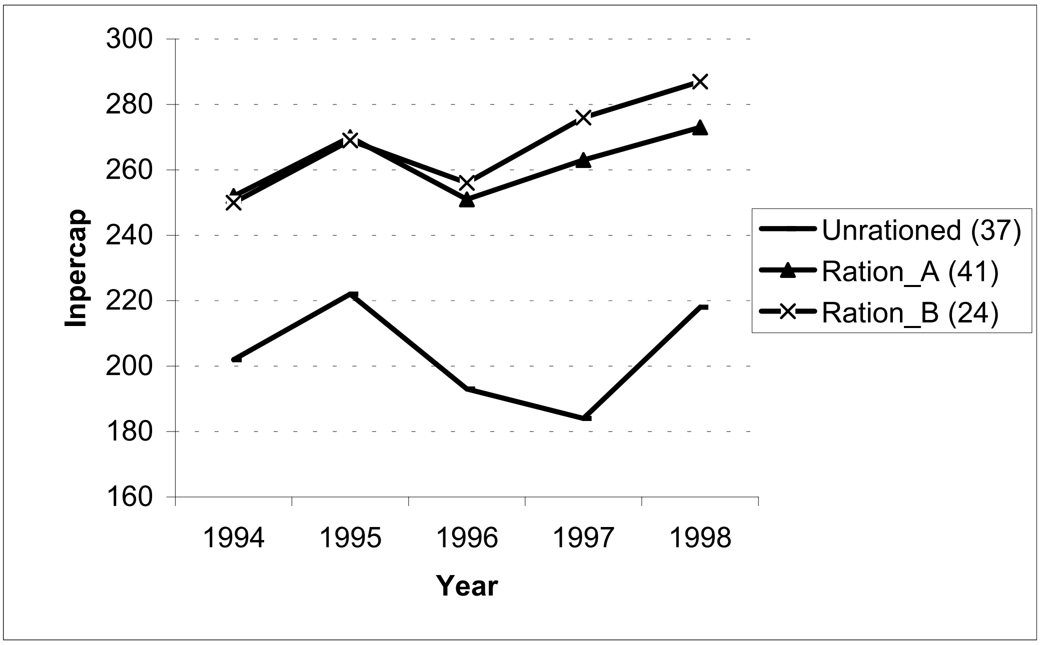


Figure 2 Average income per listed patient (Inpercap) according to rationing status



4 Estimation and results

Data on GPs' service provision are observed in five periods and each GP belongs to a specific municipality. Our data therefore have both a panel data structure and a hierarchical structure. We take account of the hierarchical structure of the data by introducing dummies for municipalities, *Municip2, Municip3* and *Municip4*, where the GP density is highest in the reference municipality (*Municip1*) and lowest in *Municip4*. Since each GP has an individual practice style related to his personality, medical experience, the organisation of his practice etc., we expect heterogeneity not accounted for by our explanatory variables. We tested for unobserved heterogeneity by estimating a model with a dummy for each GP (fixed effects) and a H_0 saying that all estimated individual effects are equal. The F-test rejects H_0 with a level of statistical significance of less than 1 per cent.

Unobserved heterogeneity among GPs can be handled either by means of a fixed effects model or by a random effects model. In the simplest version of the random effects model we have a random intercept and all other coefficients being fixed:

$$y_{it} = a_i + \mathbf{bx}_{it} + v_{it} \qquad \text{where } a_i = a + u_i \qquad (2)$$

y_{it} is the dependent variable with a subscript indicating observation number *t* of GP number *i* and $\mathbf{x_{it}}$ is a vector of explanatory variables with a similar subscript, **b** is a vector of coefficients and a_i is the intercept of GP number *i.* The term *'a'* is the constant part of the intercept, u_i is the random individual intercept and v_{it} is a classical disturbance. The two components are assumed to be independent of each other, with $E(u_i)=E(v_{it})=0$, $E(v_{it}^2)=\sigma_v^2, E(u_i^2)=\sigma_u^2$. Hence, we have that $Corr(v_{it}+u_i, v_{is}+u_i)=\rho_A=\frac{\sigma_u^2}{\sigma_v^2+\sigma_u^2}$, showing the correlation coefficient between two observations of the same GP. If the individual random effects are independent of the regressors, the random effects model is the more efficient estimation method since fewer parameters need to be estimated compared to the fixed effects model. Whether the random effects model is appropriate, is tested for by the Hausman chi-quared statistic. We did this test by including the explanatory variables that vary over time: *Propold,* *Propfem* and a time trend, *Time,* as regressors[9]. The null hypothesis is not rejected, and we continue with the random effect model.

The main result from the panel data regression is shown in Table 3.

Table 3: The estimated effect of a shortage of patients on income per listed person (INPERCAP) (standard errors in parenthesis)[10]

	Model 1		Model 2	
Propfem	132.4	(87.6)	160.2*	(43.0)
Propold	327.3*	(112.5)	300.0*	(59.7)
Time	9.3*	(1.9)	10.4*	(2.5)
Municip2	-57.1	(32.2)	-59.2*	(26.1)
Municip3	-76.0*	(27.0)	-77.6*	(20.9)
Municip4	-109.9*	(28.7)	-112.7*	(22.4)
Salaried	16.6	(14.4)	18.5*	(8.5)
Ration	33.5*	(13.0)	32.2*	(7.7)
Female	-30.3	(23.9)	-38.1*	(13.7)
Constant	178.3*	(49.2)	168.9*	(31.9)
$\hat{\rho}_A$	0.51		0.42	
R^2	0.28		0.29	
Number of observations	489		489	

In Table 3, the estimation method used distinguishes Model 1 from Model 2[11]. In Model 1 the coefficients are estimated using generalized least squares (GLS) while Model 2 applies ordinary least squares (OLS) with corrected standard error estimates. We see that the estimated correlation coefficient between several observations of a GP, $\hat{\rho}_A$, is considerable. We also see that the magnitude of the estimated coefficients is hardly influenced by the estimation method applied. The standard errors are in general smaller in Model 2 than in Model 1, resulting in all estimated coefficients being statistically

[9] The rationing variable, *Ration*, had to be left out, since it is independent of time and therefore cannot be distinguished from the individual specific fixed effect.

[10] '* ' indicates that the estimated parameter is significantly different from zero at the five per cent level with a two tailed test.

[11] The models are estimated with the statistical software Limdep 7.0.

significant in Model 2. As expected, the proportion of females on the list and the proportion of old people both contribute to a high level of service provision per person listed. We also see a positive time trend reflecting that the dependent variable is measured in nominal units and that the level of fees has increased over time. From the estimated coefficients of the municipality dummies we see that a decline in the GP density contributes to a lower level of service provision per listed person. This is as expected, since a higher GP density in a municipality improves patients' access to GPs.

The estimated effect of patient shortage (*Ration*) tells us that that rationing implies an expected increase in income per listed person of about NOK 33, which amounts to 15 per cent of the per patient income of an unrationed GP.

From Table 2 we saw that the rationed GPs differ from the unrationed in particular with regard to preferred list size, the proportion of female GPs and the proportion who were salaried prior to the trial. We therefore suspect that rationing is not a random event. Then the characteristics that determine whether a GP is rationed may also explain why he has a high intensity of service provision. There may for instance be a particular type of GP who both has a taste for many patients and for giving them much treatment. We then have a similar problem to the selection problem in the evaluation of social programmes, recently reviewed by Blundell and Costa Dias (2000). In our case the rationed GPs may be considered as the treatment group and the income per listed person as the outcome variable. We then have:

$$y_i = \begin{cases} y_{i0} = \mathbf{b}\mathbf{X}_i + v_{i0} & if\ T_i = 0 \\ y_{i1} = \mathbf{b}\mathbf{X}_i + \gamma + v_{i1} & if\ T_i = 1 \end{cases}$$

where T=1 denotes *ration* (treatment) and T=0 otherwise, **b** and **X** are vectors of explanatory variables, as described earlier. To simplify the exposition, we have dropped the time indicator. The expected effect of rationing for individual *i* is:

$$E(y_{i1}) - E(y_{i0}) = \gamma + E(v_i \mid T_i = 1) - E(v_i \mid T_i = 0)$$

We are interested in separating the systematic effect of rationing (γ) from the unobserved factors. If we also had observed the outcome for the rationed had they not been rationed, we would have $E(v_i \mid T_i = 1) - E(v_i \mid T_i = 0) = 0$, and there would be no evaluation problem. We could find the effect of rationing by simply comparing the value of the outcome variable in the rationed situation with the value of the outcome variable in the unrationed situation. This is basically what is done in Table 3. The randomized experiment approximates this solution by randomly assigning the eligible population to the treatment group and to the control group, respectively.

When randomization is not an option, selection bias may occur, and hence, we may have $E(v_i \mid T_i = 1) - E(v_i \mid T_i = 0) \neq 0$. We then try to model the factors that contribute to whether GP *i* is rationed:

$$\begin{aligned} &T_i^* = \alpha' \mathbf{Z}_i + w_i \\ &T_i = 1 \, if \, T_i^* > 0, \quad 0 \, otherwise \end{aligned} \tag{3}$$

where $\mathbf{Z}_i$ is a vector of explanatory variables, α' a vector of coefficients, and $w_i \sim IID(0, \sigma_w^2)$ is the stochastic term.

Several statistical methods are suggested and used to correct for selection bias. In this study we consider two: The Heckman (1979) two step estimator and the propensity score matching method (Rosenbaum and Rubin, 1983, 1984).

In the two step estimator, *w* and *v*, are assumed to be jointly normally distributed. We first estimate by a probit model the propensity of being rationed contingent on a choice of variables, **Z**; i.e. Prob(*Ration*=1 | **Z**), where GP's *Age, Age^2, Female, Prelistsize* and *Salaried* are included in **Z** as explanatory variables. We then estimate (2) taken into account the binormal distribution of *v* and *w*: $(v_{it}, w_i) \sim bivariate\ normal\,[0, 0, \sigma_v^2, \sigma_w^2, \rho]$. Table 4 exhibits the results:

Table 4 The estimated effect (standard deviation) of independent variables on income per listed person (INPERCAP) (standard errors in parenthesis)

Variable	Probit	Model 3: Random effects with selection correction
Prelistsize	0.892* (0.351)	
(1000 persons)		
Age	-0.578 (0.406)	
Age^2	0.001 (0.001)	
Female	-0.544 (0.310)	-19.5 (27.3)
Salaried	0.418 (0.322)	14.9 (14.6)
Ration		75.4 (53.0)
Propfem		123.3 (88.5)
Propold		308.8** (115.3)
Municip2		-63.1 (33.2)
Municip3		-84.3** (29.0)
Municip4		-121.2** (32.1)
Time		9.4** (1.8)
Constant		163.5** (52.7)
$\hat{\lambda}$		-27.4 (33.5)
$\hat{\rho}_A$		0.52
Number of observations	109	489
- 2 Log likelihood	129.2	
R^2		0.29

The probit model predicts 72 per cent of the observations correctly. We see that the coefficient of the selection term ($\hat{\lambda}$) in model 3 is statistically insignificant, indicating that the effect of rationing may be estimated without taking selection on unobservable characteristics into account, as in Table 3.

The matching propensity score method assumes that selection takes place on observable characteristics, but not on unobservables. Each of the 71 rationed individuals is matched with an unrationed individual with an approximately equal probability of being rationed. The sample of unrationed GPs then shrinks from 38 to 25. On average the difference in the probability of being rationed between the rationed and the matched control was 0.008. The effect of *Ration* is then estimated as the average difference in adjusted *Inpercap* between the 71 rationed and their 25 matched controls[12]. The adjustment is done by means of linear regression as shown in Table 3, but with *Ration* omitted as explanatory variable. As shown in Table 5, we both consider differences in adjusted *Inpercap* in single years and the average for the whole period. We present results of matching with nearest neighbor (the matched control with the closest propensity score), since matching with an average of a larger number of matches, does not make much difference for the result. The estimated average effect was NOK 28.7 and hence, quite close to the effect (NOK 33.5) we found in Table 3.

Table 5 Difference between rationed and matched unrationed in regression adjusted from fees and patient charges per listed person

Dependent variable		Difference between rationed and matched unrationed (nearest neigbour)
Inpercap94	Adjusted income from fees and patient charges per listed person in NOK in 1994	17.9 (10.6)
Inpercap95	Regression adjusted income from fees and patient charges per listed person in NOK in 1995	17.5 (10.8)
Inpercap96	Regression adjusted income from fees and patient charges per listed person in NOK in 1996	24.0* (10.4)
Inpercap97	Regression adjusted income from fees and patient charges per listed person in NOK in 1997	25.6* (11.0)
Inpercap98	Regression adjusted income from fees and patient charges per listed person in NOK in 1998	30.2* (12.2)
AverInpercap	Regression adjusted average annual income from fees and patient charges per listed person in NOK 1994-1998	28.7* (11.7)

The positive effect of rationing on service provision therefore survives the correction for

[12] The estimation was done by means of the program, psmatch, authored by Barbara Sianesi in the software Stata 7.0.

selection bias by propensity score matching. Hence, we conclude that the higher level of service provision is related to rationing in itself. Potential shortcomings should however be mentioned. Since not all unrationed are used as controls (in this case 13 unrationed GPs are not used), there is a loss of information due to non-overlapping support of X. On the other hand, some of the unrationed are used more than one time as a matched control. Another problem is that although a rather small difference in the propensity of being rationed between the rationed and the non-rationed was obtained, we cannot be sure that selection on unobservables is absent, as assumed by the model.

So far we have not considered the difference in service intensity that may stem from variation in change in list size over time among the rationed. Again, we may face a selection problem since unobservables (for instance, accessibility is unobservable due to lack of data) that affect the increase in list size may also affect the level of service intensity. We correct for this potential bias by estimating a Difference-in-Differences estimator[13]:

$$y_{it} = a + \eta_i + \theta_t + \tau x_{it} + u_{it}$$

where y_{it} is the outcome variable, a is a constant term, η_i is the individual specific fixed effect, θ_t is the common time specific fixed effect, x_{it} is a dummy equal to one if individual i experiences an increase in the list in period t and u_{it} is a temporary individual specific effect with zero expectation and constant variance. If we now take the difference in outcome during the treatment period for each individual and then take the difference between the treated and non treated, a, η_i and θ_t cancel out, and we are left with τ as the expected treatment effect.

We have made two alternative specifications for the increase in the number of people listed. In the first alternative (Model A) GP no. i has $x_{it}=1$ if he experiences an increase in list size from period $t-1$ to period t and 0 otherwise. But it may also be that what matters for the outcome variable is the total change in list size from 1993 to period t. Hence, in this alternative (Model B) $x_{it}=1$ if the GP experiences an increase in list size from period 1 (1993) to period t and 0 otherwise.

[13] See Blundell and Costa Dias (2000) for a more detailed description of the method.

Table 6: The estimated effect of an increased list of patients on income per person listed (INPERCAP) (standard errors in parenthesis) among the rationed GPs

	Model A	Model B
Increase in no. patients listed	-3.7 (9.1)	-17.4 (12.0)
Constant	273.5* (4.4)	280.5* (6.5)
Individual specific effect	Yes	Yes
Common time specific effect	Yes	Yes
R^2	0.72	0.72
Number of observations	304	304

From Table 6 we see that the effect of an increase in the number of listed patients is insignificant regardless of the choice of specification. Hence, we are not able to show an effect of an increase in the number of patients on intensity of service provision among the GPs who were rationed at the outset of the trial.

The results from the empirical analyses in this section may be summed up like this: Rationing regarding the number of patients listed seems to increase the intensity of service provision per person listed. Whether a rationed GP experiences an increase in the list or not, has no significant effect on the intensity of service provision.

5 Concluding remarks

With a list patient system each GP's number of patients is known. Hence, we know whether a high volume of services delivered by a GP is caused by a long list of patients or by many services to each patient listed. This is crucial information for the study of economic incentives in health care, and is the empirical basis for the present study. From an analysis of a GP's optimal practice profile a classification according to intensity of health service provision and accessibility is derived. In particular, alternative responses to patient shortage are explored. We derive the conditions for each of the following actions or a combination of them: improving accessibility to attract more patients, increasing the intensity of health service provision per patient or enjoying more leisure.

Data from the four municipalities that participated in the Norwegian capitation trial 1993-1996 are used in the empirical analysis. Since the municipalities on a voluntary basis decided to continue the list patient system after the trial ended, we had access to data describing a panel of GPs in five consecutive years. As expected we find a positive effect of GP density at the municipal level on the intensity of health service provision. We also find a positive effect of patient shortage on service provision per listed person at the GP level. A main result from the regressions is that GPs who experience a shortage of patients are expected to have about 15 per cent more income per patient listed than their unrationed colleagues. The result survives the correction for potential selection bias. Whether a rationed GP experiences an increase in the list of patients or not, had no significant effect on the intensity of service provision. In Table 1 this behavior corresponds to the cells 3 and 4.

The available data for this study were not appropriate to take all the dimensions of the theoretical classification into account. In particular, we had no data at hand that showed the offered accessibility. Hence, we had to estimate the effect of an increase in the list size on service intensity directly. But we cannot rule out that an increased number of patients listed are due to personal characteristics of the GP gradually spread by word of mouth.

With data from the two years 1994 and 1995, Iversen and Lurås (2000) find that GPs who experienced a shortage of patients provided more services per person listed than their unrationed colleagues. An objection against this result has been that the estimated effect may be temporary and related to the introduction of a new system. Over time lists may adjust and the variation in service intensity may decline. The present study shows that the results in Iversen and Lurås (2000) is not just a temporary phenomenon, but persist five years after the system was introduced. In fact the estimated effect was 5 per cent greater when estimated with data for five years.

A policy implication of this study is that patient shortage is costly to the insurer because of income motivated behavior of unknown benefit to the patient. This income motivated behavior is fed by the fee for service component of the payment system. An alternative would be to drop the fee for service component and let the payment system be based on the capitation fee only. The GPs would then compete for patients without considering the income from services per se. Services delivered would then be a means to attract patients to the list and hence, to generate capitation income. The problem is of course that under a pure capitation system not all patients are equally attractive because of variation in need for services. A risk adjustment component would then be required to prevent risk selection by the GPs. It is well known from the literature that a risk adjusted capitation system is hard to construct in practice. The present study therefore demonstrates the classical trade off between selection and inefficiency in health care.

We have so far only considered optimal choice from the physician's perspective, and have not analyzed the relation to optimal service provision from the society's point of view. An important challenge for future research is therefore to gain more knowledge of the optimal mix of capitation and fee for service in general practice from the society's perspective.

Appendix

Type 1: The unrationed

This type is characterized by μ=0:

$$\frac{pn-\lambda}{t(s)n} \leq v'(\ell) \tag{1a}$$

$$\frac{q+p(\alpha+k)}{t(s)(\alpha+k)} = v'(\ell) \tag{A1b}$$

$$\frac{-C'(s)}{t'(s)n(\alpha+k)} < v'(\ell) \tag{A1c}$$

From (A1c) we have s=0. Eq (A1b) inserted into (1a) yields $-\frac{q}{\alpha+k}-\frac{\lambda}{n} \leq 0$, implying strict inequality and hence, λ=0 and k=0. Since there is a positive capitation fee, there will always be more rewarding to add new people to list than to increase the level of service provision to those already listed.

The other types have in common that they do not achieve their optimal number of patients with s=0; hence μ>0. We distinguish between eight types:

Type 2: Improved access

This type is characterized by s>0 and k=0, and hence:

$$\frac{pn}{t(s)n} < v'(\ell) \tag{A2a}$$

$$\frac{q+p(\alpha+k)-\mu}{t(s)(\alpha+k)} = v'(\ell) \tag{A2b}$$

$$\frac{\mu n_s'(s;S,\omega,\Omega,N)-C'(s)}{t'(s)n(\alpha+k)} = v'(\ell) \tag{A2c}$$

Inserting for μ from (A2b) into (A2c) yields:

$$\frac{[q+p(\alpha+k)]n_s'(s;S,\omega,\Omega,N)-C'(s)}{t'(s)n(\alpha+k)+t(s)n_s'(s;S,\omega,\Omega,N)(\alpha+k)} = v'(\ell) \tag{A2e}$$

describing as the optimum condition that income per time unit from a marginal increase in s is equal to the marginal utility of leisure. From (A2a) we have that the income per time unit of providing higher service intensity is lower than the income per time unit from providing better accessibility.

Type 3: High service intensity

For this type we have 0<k≤β and s=0 and hence:

$$\frac{pn-\lambda}{t(s)n} = v'(\ell) \tag{A3a}$$

$$\frac{[q+p(\alpha+k)]n_s^{'}(s;S,\omega,\Omega,N)-C'(s)}{t'(s)n(\alpha+k)+t(s)n_s^{'}(s;S,\omega,\Omega,N)(\alpha+k)} < v'(\ell) \tag{A3e}$$

From (A3a) and (A3e) we conclude that the income per time unit from health service provision is greater than the marginal utility of leisure because the intensity of service provision is constrained by the professional norm. The income per time unit of providing better access is smaller than the marginal utility of leisure; for instance because of a small effect of accessibility on the number of people who want to be listed.

Type 4: Improved access and high service intensity

This type is characterized by s>0 and 0<k≤β:

$$\frac{pn-\lambda}{t(s)n} = v'(\ell) \tag{A3a}$$

$$\frac{[q+p(\alpha+k)]n_s^{'}(s;S,\omega,\Omega,N)-C'(s)}{t'(s)n(\alpha+k)+t(s)n_s^{'}(s;S,\omega,\Omega,N)(\alpha+k)} = v'(\ell) \tag{A2e}$$

For this type of GP the income per time unit of providing health services is greater than the income per time unit from providing better accessibility which is equal to the marginal utility of leisure.

$$\frac{[q+p(\alpha+k)]n_s^{'}(s;S,\omega,\Omega,N)-C'(s)}{t'(s)n(\alpha+k)+t(s)n_s^{'}(s;S,\omega,\Omega,N)(\alpha+k)} < v'(\ell) \tag{A3e}$$

In this case the wage rate is higher than both the income per time unit of providing better accessibility and from providing health services.

Type 5: More leisure

We then have τ=k=s=0:

$$\frac{pn}{t(s)n} < v'(\ell) \qquad \text{(A2a)}$$

$$\frac{[q+p(\alpha+k)]n_s^{'}(s;S,\omega,\Omega,N)-C'(s)}{t'(s)n(\alpha+k)+t(s)n_s^{'}(s;S,\omega,\Omega,N)(\alpha+k)} < v'(\ell) \qquad \text{(A3e)}$$

Since this type of GP is rationed, he would have preferred more patients and a higher income but the income from service provision only is not great enough to compensate for loss of leisure, hence μ<q.

References:

Blundell, R., Costa Dias, M., 2000. Evaluation methods for non-experimental data. Fiscal Studies 21, 327-468.

Ferrall, C., Gregory, A. W., Tholl, W. G., 1998. Endogenous work hours and practice patterns of Canadian physicians. Canadian Journal of Economics XXXI, 1-27.

Heckman, J., 1979. Sample selection bias as a specification error. Econometrica 47, 153-161.

Iversen, T., Lurås, H., 2002. Waiting time as a competitive device: an example from general medical practice. International Journal of Health Care Finance and Economics 2, 189-204.

Iversen, T. and Lurås, H., 2000. Economic motives and professional norms: The case of general medical practice. Journal of Economic Behavior and Organization 43, 447-471.

Rosenbaum, P. and Rubin, D.B., 1983. The central role of the propensity score in observational studies for causal effects. Biometrika 70, 41-55.

Rosenbaum, P. and Rubin, D.B., 1984. Reducing bias in observational studies using subclassification on the propensity score. Journal of the American Statistical Association 79, 516-524.

Scott, A., 2000. Economics of general practice. In A. J. Culyer and J. P. Newhouse: Handbook of Health Economics, Volume 1 (Elsevier Science, Amsterdam) 1175-1200.