Does quality influence choice of general practitioner?

An analysis of matched doctor-patient panel data

Erik Biørn
Department of Economics, University of Oslo

Geir Godager
Institute of Health Management and Health Economics, University of Oslo

Working paper 2008: 3

HERO

Does quality influence choice of general practitioner?

An analysis of matched doctor-patient panel data

Erik Biørn
Department of Economics, University of Oslo

Geir Godager
Institute of Health Management and Health Economics, University of Oslo

Health Economics Research Programme at the University of Oslo
HERO 2008

Health Economics Research Programme at the University of Oslo.
Financial support from The Research Council of Norway is acknowledged.
ISSN 1501-9071 (print version.), ISSN 1890-1735 (online), ISBN 82-7756-186-5

DOES QUALITY INFLUENCE CHOICE OF GENERAL PRACTITIONER? AN ANALYSIS OF MATCHED DOCTOR-PATIENT PANEL DATA

ERIK BIØRN
Department of Economics, University of Oslo,
P.O. Box 1095 Blindern, 0317 Oslo, Norway
E-mail: erik.biorn@econ.uio.no

GEIR GODAGER
Institute of Health Management and Health Economics, University of Oslo,
P.O. Box 1089 Blindern, 0317 Oslo, Norway
E-mail: g.c.godager@medisin.uio.no

ABSTRACT: The impact of quality on the demand facing health care providers has important implications for the industrial organization of health care markets. In this paper we study the consumers' choice of general practitioner (GP) assuming they are unable to observe the true quality of GP services. A panel data set for 484 Norwegian GPs, with summary information on their patient stocks, renders the opportunity to identify and measure the impact of GP quality on the demand, accounting for patient health heterogeneity in several ways. We apply modeling and estimation procedures involving latent structural variables, *inter alia*, a LISREL type of model, is used. The patient excess mortality rate at the GP level is one indicator of the quality. We estimate the effect of this quality variable on the demand for each GP's services. Our results, obtained from two different econometric model versions, indicate that GP quality has a clear positive effect on demand.

KEYWORDS: GP services. Health care quality. Health care demand. Latent variables. LISREL. Panel data. Norway

JEL CLASSIFICATION: C23, C33, D83, H51, H75, I11, I18

ACKNOWLEDGEMENTS: We are grateful for comments from Tor Iversen, Karin Monstad, Ismo Linnosmaa, Terje Skjerpen and participants at the 27th Nordic Health Economists Study Group Meeting, Copenhagen, August 2006, a Workshop on Health Economics, Oslo, August 2006, and the 6th iHEA World Congress in Health Economics, Copenhagen, July 2007. Financial support from the Norwegian Ministry of Health and Care Services and the Research Council of Norway is acknowledged. Some of the data applied are provided by the National Insurance Administration and Statistics Norway and have been prepared for research purposes by The Norwegian Social Science Data Services. The authors alone are responsible for all analyses and interpretations.

1 Introduction

Asymmetric information between physicians and their patients is a basic characteristic of the market for health care services. In the words of Arrow (1963):

> *"...medical knowledge is so complicated, the information possessed by the physician as to the consequences and possibilities of treatment is necessarily very much greater than that of the patient, or at least so it is believed by both parties".*

Patients are therefore often considered to be poor judges of service quality. However, those who have repeated encounters with the same health care provider, will accumulate information on services and treatment outcomes, thus narrowing the information gap. The market for general practitioners' services is characterized by durable doctor-patient relations that may improve the patients' quality assessment. The aim of this paper is to investigate empirically whether the demand facing a general practitioner (GP) responds to the quality of the provided services.

General background

The impact of quality on the demand facing health care providers has important implications for the organization of health care markets. There is a growing literature on competition and quality in such markets, from which an important result is that the effect of stronger competition on quality depends crucially on the relative sizes of the price elasticity and the quality elasticity of demand. More competition may bring about reductions in quality if the quality elasticity is small compared to the price elasticity (Dranove and Satterthwaite, 2000, Gaynor, 2006). Further, the impact of quality on the demand facing health care providers has important implications for the optimal design of payment systems. A familiar result is that a retrospective payment scheme in the form of cost reimbursement is likely to pursue the goal of quality provision while giving weak incentives to provide cost reducing efforts. Conversely, prospective payment schemes tend to strengthen the incentives for cost reduction, while weakening the incentives for providing quality. A combination of payment mechanisms is thus likely to perform better than payment systems employing only one parameter. However, if quality affects demand, a first-best solution can, in theory, be obtained under a pure prospective payment scheme (Ma, 1994). This suggests that the effect of quality on demand – and information on its numerical size – is a key factor determining the optimal calibration of the parameters in the payment system: If the market punishes providers who are skimping on quality, the payment system can put more weight on the parameters that encourage cost reducing efforts.

Relation to literature

A conventional empirical approach when seeking to assess the effect of quality on demand for health services is to estimate the effect of provider characteristics on individual consumers' choice of provider, applying different models for individuals' discrete choice. An influential paper in this tradition is Luft *et al.* (1990). They specifically study the effect of quality indicators such as death and complication rates, teaching status of hospital, and out of state admissions on patients' choice of hospital, using logit models, and find positive effects for several of the applied indicators. Using similar quality indicators and methods, Burns and Wholey (1992) extend the framework by including in their logit models characteristics of the admitting physician. They find that quality affects demand positively, and that characteristics of the admitting physician are important determinants of patients' hospital choice. More recently, Howard (2005), applying a mixed logit model on data on kidney transplantations, estimated the effect of the deviation from expected failure rate on probabilities of hospital choice. The results indicate that hospitals with a higher than expected failure rate have smaller probabilities for being chosen. A different empirical strategy is followed in Chirikos (1992), in estimating, by linear regression, the effect of individual hospitals' *quality spending* on their market shares. The results support the hypothesis that increased provider quality affects demand positively.

The present paper adds to the literature in several ways. First, no previous empirical studies seem to have considered the demand effects of quality in the market for general practitioners. Second, in the current literature the relationship between demand and various indicators of quality, such as mortality rates, failure rates or hospital type, and other independent variables, are estimated separately. The present paper contributes to the literature by simultaneously estimating the relationship between demand and several quality indicators, applying linear structural equation modeling (LISREL) and estimation methods. Within this framework we acknowledge both the multidimensional aspect of the quality concept, and that it may be considered as more appropriate to interpret outcome measures such as *mortality rates* or *failure rates* as functions of quality, rather than as measuring quality itself. Third, our econometric model has a wider field of application as it provides a method to separate the effect of quality on outcome measures from the effect of patient health.

Setting of the study

In June 2001 a regular GP scheme was introduced in Norwegian general practice, making the GPs responsible for the provision of primary care services to the persons listed at their practice. Prior to the reform the health authorities gathered the information needed to assign one GP to each Norwegian inhabitant. All inhabitants were asked to rank their three most preferred GPs in a form, and all GPs were asked to report the maximum number of patients they would like to take care of. An algorithm was designed to utilize this information and obtain a one-to-one match between inhabitants and GPs.

Our data set has a panel format with the GP as the observation unit, but for some variables only one observation per GP exists. The data stem from The Norwegian General Practitioners Database, covering all Norwegian GPs, supplemented by measures of the

GP density in each municipality and of age-gender specific mortality rates. Among the variables recorded are the number of persons who ranked each GP as most preferred when returning the entry form, the number of mortalities among each GP's listed patients during a six-month period, and the proportion of the listed persons who switch to other GPs in later periods. For a stratified sample of GPs, relating to 14 municipalities, from this official GP database the data set has been extended to also include the median income and wealth of the listed persons and the proportion of them who have not finished high-school. In the analysis, we interpret the number of first-rankings and the proportion of listed persons who switch to other GPs, as indicators of the demand facing each GP. Our main hypothesis is that there exists a latent stochastic variable, denoted as *GP quality*, which, when heterogeneity related to the health status of the listed persons and other observed heterogeneity have been accounted for, is positively related to the demand facing each individual GP and negatively related to the recorded excess mortality of the GP's listed patients. We find empirical support to this hypothesis.

Two kinds of models are considered: a Panel Data model with latent heterogeneity related to perceived GP quality and a multi-equation *LISREL type of model*, including both GP quality and the health of the stock of persons on the GP's list as latent variables, both of which are assumed to affect demand as well as other observed variables. For some variables, including the proportion of persons switching and the excess mortality, we have data in the panel data format. This is profitable for quantifying the latent heterogeneity and its consequences.

The rest of the paper proceeds as follows. The modeling of the demand in the market for GPs is discussed in the following two sections. In Section 2, we present a theoretical argument supporting the view that the expected demand facing each individual GP can be a function of quality, even if the true quality is unobserved to his/her potential patients. The discussion motivates testable predictions and hypotheses to be examined in the paper. In Section 3, we present the two econometric models. The data are described in Section 4, while estimation and test results are presented in Section 5. In Section 6 we discuss the results and conclude.

2 A model of patients' quality perceptions and demand for GPs

Quality and demand

In order to model the consumers' choice of GP when they are unable to observe the true quality of GP services, we first show that if the errors in the quality assessment of potential patients have certain properties, the demand for GPs will depend on quality even when the latter is unobservable. Consumers are imperfectly informed about the quality of GPs and we therefore distinguish between *true* and *perceived* quality. We assume that the only criterion for selecting a GP is the perceived quality of the services provided. Quality of health care services is a complex entity that is not easily represented by a scalar measure. In this model, however, quality can (in principle) be quantified and completely described by a number on a finite scale. One may thus think of quality as an input factor in the GP's 'health production function'. While predetermined abilities, such as individual talent, obviously influence quality, the GP also has discretionary power to influence quality by the exertion of 'quality generating efforts', such as concentration. The individual GP's quality of services is determined by his/her abilities and preferences and we assume that the latter are both time invariant, implying that quality also is time invariant.

The consumer's information set is comprised by quantitative information on the quality of each available GP. This information, however, is 'contaminated' by stochastic errors. The quality of GPs as perceived by a consumer may be higher or lower than the true quality, and two consumers are likely to have different beliefs regarding the quality of the same GP. The stochastic properties of what we may think of as measurement errors drive the matching of consumers and GPs in the model. Let μ_j denote the true quality of GP j ($j = 1, \ldots, M$), while q_{ij} denotes the quality of GP j as *perceived* by consumer i ($i = 1, \ldots, N$). We assume that q_{ij} is normally distributed with $\mathsf{E}(q_{ij}) = \mu_j$ and $\mathrm{var}(q_{ij}) = \sigma_j^2$, *i.e.*, the distribution of the perceived quality differs between GPs. We thus allow for the possibility that the population of consumers may have more accurate information about GPs who have been active in the market for a long time (low σ_j^2) compared to GPs who have established their practice recently (high σ_j^2). We let $u_{ij} = q_{ij} - \mu_j$ and assume that the MN u_{ij}s are uncorrelated both across GPs and over consumers. Altogether, we can therefore state our assumptions as

$$(1) \qquad \begin{aligned} &q_{ij} = \mu_j + u_{ij}, && i, l = 1, \ldots, N, \\ &u_{ij} \sim \mathsf{N}(0, \sigma_j^2), \quad \mathsf{E}(u_{ij} u_{lk}) = 0, \quad i \neq l \text{ or } j \neq k, && j, k = 1, \ldots, M. \end{aligned}$$

We simply assume that consumer i considers perceived quality q_{ij} as indicating the true quality of GP j. Since the normal distribution has an infinite support, the distribution of *perceived* qualities associated with the GPs with the highest and the lowest *true* quality overlap. This ensures that even if $\mu_j < \mu_k$, there is a positive probability that $q_{ij} > q_{ik}$, so that any GP has a strictly positive probability of being selected by any consumer. This implication seems reasonable if the differences in true quality is not too large.

Matching GPs and consumers

We may think of the matching of GPs and consumers as a lottery. A draw is a realization of $q_{i1}, q_{i2}, \ldots, q_{iM}$, that is, the realizations of the beliefs of the quality of each and one of the GPs for consumer i. There are thus N independent drawings performed in the market, one for each consumer. Let $\phi(q_{ij}; \mu_j, \sigma_j^2)$ denote the density function of q_{ij}, the perceived quality of GP j, which according to (1) is distributed as $\mathsf{N}(\mu_j, \sigma_j^2)$. To simplify notation we let $\Delta_{ijk} = q_{ij} - q_{ik}$ be the difference between consumer i's perceived quality of GPs j and k. It follows from (1) that $\Delta_{ijk} \sim \mathsf{N}(\nu_{jk}, \theta_{jk}^2)$, with density function $\phi(\Delta_{ijk}; \nu_{jk}, \theta_{jk}^2)$, where

$$(2) \qquad \nu_{jk} = \mu_j - \mu_k, \qquad \theta_{jk}^2 = \sigma_j^2 + \sigma_k^2, \qquad j, k = 1, \ldots, M.$$

Let $\mathcal{A}_{jk}$ denote the event that $\Delta_{ijk} > 0$ for an arbitrary consumer, i. Then

$$(3) \qquad P(\mathcal{A}_{jk}) = P(\Delta_{ijk} > 0) = \int_0^\infty \phi(\Delta_{ijk}; \nu_{jk}, \theta_{jk}^2) d\Delta_{ijk} \equiv p_{jk}$$

is the probability that GP j has a higher perceived quality than GP k. Since the draws are assumed to be independent, the event that GP j has the highest perceived quality in a random draw can thus be expressed as

$$\mathcal{B}_j = \bigcap_{\substack{k=1 \\ k \neq j}}^{M} \mathcal{A}_{jk}, \quad j = 1, \ldots, M.$$

The probability of this event can be expressed as:

$$(4) \qquad P(\mathcal{B}_j) = P(\Delta_{ijk} > 0;\ \forall k \neq j) = \prod_{\substack{k=1 \\ k \neq j}}^{M} p_{jk} \equiv \pi_j, \qquad j = 1, \ldots, M.$$

The probability that GP j has the highest perceived quality in a draw, π_j, is a function of $\mu_1, \ldots, \mu_M; \sigma_1^2, \ldots, \sigma_M^2$. Since (3) implies that $\partial p_{jk}/\partial \nu_{jk} > 0 \ \forall j \neq k$, it follows from (2) and (4) that $\partial \pi_j/\partial \mu_j > 0$ and $\partial \pi_j/\partial \mu_k < 0, \ k \neq j$. Hence, for all GPs, the probability of being selected by any consumer is an increasing function of the *true* quality. When consumers select the GP with the highest *perceived* quality, the expected demand facing GP j is $\pi_j N$. The expected demand facing any GP therefore changes in proportion to the number of consumers in the market. Or stated otherwise, the probability of being selected by a random consumer can be interpreted as the GP's expected market share. The *ex-post* market share converges to the probability of being selected by a random consumer as the number of drawings increases.

From this model we can make the following *predictions*:

[P1] GPs with high quality of services have a higher probability of being selected by a randomly chosen consumer than a GP whose services are of lower quality.

[P2] The selection probabilities π_j are independent of the number of consumers, N. For a given population of M GPs, expected demand for the services of any of them, is a linear function of N.

This, rather simple, model implies that the consumers are unable to affect the precision of their own quality assessment, reflected by the assumption that perceived quality of GP j has the same variance, σ_j^2, for all consumers. The model could be generalized to allow for consumer heterogeneity in the sense that some are more skillful or eager in gathering and processing information in the market than others. This could have been accounted for by replacing σ_j^2 by σ_{ij}^2, where $\sigma_{ij}^2 < \sigma_{hj}^2$ if consumer i has taken efforts to become better informed about GP j's quality than has consumer h. A prediction from such an extended model may be that high-quality GPs tend to have a higher proportion of skilled or eager consumers on their lists than the low-quality GPs (Godager, 2008). The possible existence of such a selection mechanism is important since consumers who are skilled or eager in collecting information, may have a health status and a death probability different from those not so skilled or eager.

The crucial question then becomes: which groups of consumers, according to observable characteristics, devote most attention and efforts in searching for the best GP? On the one hand, less healthy consumers, with a high expected mortality rate, may be thought to be *particularly concerned* about their choice of GP and as a result be more willing to collect information than the average consumer. This may contribute to increasing the average mortality rate among the patients listed with high-quality GPs. On the other hand, consumers who are more healthy and resourceful and have low expected mortality may be *particularly able* to collect and process such information. This may contribute to the outcome of the selection mechanism being reversed, *i.e.*, lowering the average mortality rates of the persons listed with high-quality GPs. Consequently, from *a priori* reasoning it is not obvious that the outcome of (observed or unobserved) patient heterogeneity will be neither that high-quality GPs attract patients with an average health status which differs from that of the low-quality GPs, nor if there is a difference, in which direction it will go. If a mechanism systematically selecting patients with different expected mortality rates for GPs of different professional quality is at work, and heterogeneity in health status among listed patients is not taken care of in our modelling, we are likely to face severe difficulties when trying to estimate the impact on demand of GP service quality. The models to be described below have different degree of sophistication and are not equally well designed to meet this challenge. We address this issue in more detail in sections 3 and 5.

In elaborating the theory element above, we, for simplicity, have considered the consumers' mean perceived quality of GP j, μ_j, as non-stochastic. This interpretation is provisional and intended to be valid only in a *conditional* sense. When, in the following, this theory element will be embedded in an econometric model involving both latent and observed variables related to GP quality, this variable will change its status and become a latent, stochastic variable.

3 Econometric models

Motivation

In order to represent, and hopefully quantify, how the demand for GP services responds to GP quality and other relevant variables – as motivated by the theoretical argument put forth in the previous section – two kinds of models will be considered. The first, *Model A*, is a *two-equation Panel Data model accounting for latent unit-specific heterogeneity.* We associate the latter with, *inter alia*, perceived GP quality. The second, *Model B*, is a more complex, multi-equation model of the *LISREL type.* It includes not only GP quality among its latent variables, but also the initial health status of the persons entered on the GPs' lists. This extension serves to control for the fact that GP quality and observed GP heterogeneity may interact with observed and latent heterogeneity of the listed persons in multiple ways when determining demand as observed in the market.

Model A: Two-equation panel data random effects regression model

Assume that, in a certain district, at time t, there are M_t GPs, indexed by, $j = 1, \dots, M_t$, and N_t patients, indexed by $i = 1, \dots, N_t$. As before, we let μ_j denote the true quality of GP j, unobserved both to the consumers and the health administrators, and now treated as stochastic. Let further y_{1jt} and y_{2jt} denote two observable variables, which may be considered indicators of μ_j at time t. The interpretation adopted in Model A is that y_{1jt} is the demand facing GP j, and y_{2jt} is the excess death rate of persons on the list of this GP at time t. We specifically measure demand only by the number of consumers ranking the GP as the most strongly preferred prior to the implementation of the regular GP reform, and it is observed in period $t = 1$ only. The variables assumed to explain (y_{1j1}, y_{2jt}) are quality and observable variables, of which some vary across both GPs and time periods, denoted as two-dimensional variables, and some are GP-specific.

We specify

$$
\begin{aligned}
y_{1j1} &= \boldsymbol{x}_{1j1}\boldsymbol{\beta}_1 + \boldsymbol{z}_{1j}\boldsymbol{\gamma}_1 + \alpha_{1j} + u_{1j1}, \quad j = 1, \dots, M_1, \\
y_{2jt} &= \boldsymbol{x}_{2jt}\boldsymbol{\beta}_2 + \boldsymbol{z}_{2j}\boldsymbol{\gamma}_2 + \alpha_{2j} + u_{2jt}, \quad j = 1, \dots, M_t;\ t = 1, \dots, T,
\end{aligned} \tag{5}
$$

$$
\left(\begin{bmatrix} u_{1j1} \\ u_{2jt} \end{bmatrix} \middle| \begin{bmatrix} \boldsymbol{x}_{1j1}, \boldsymbol{z}_{1j}, \alpha_{1j} \\ \boldsymbol{x}_{2jt}, \boldsymbol{z}_{2j}, \alpha_{2j} \end{bmatrix} \right) \sim \mathsf{IID}(\mathbf{0}, \boldsymbol{\Sigma}), \quad \mathbf{0} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \ \boldsymbol{\Sigma} = \begin{bmatrix} \sigma_{u1u1} & \sigma_{u1u2} \\ \sigma_{u1u2} & \sigma_{u2u2} \end{bmatrix}, \tag{6}
$$

where $(\boldsymbol{x}_{1j1}, \boldsymbol{x}_{2jt})$ and $(\boldsymbol{z}_{1j}, \boldsymbol{z}_{2j})$ are the row vectors of two-dimensional and GP-specific variables, respectively, $\boldsymbol{\beta}_1, \boldsymbol{\gamma}_1, \boldsymbol{\beta}_2, \boldsymbol{\gamma}_2$ are column vectors of coefficients, and $(\alpha_{1j}, \alpha_{2j})$ are stochastic latent variables relating to the GP j's quality, the latter assumed to affect patients' demand as well as their mortality. A crucial part of the model are the equations which connect these latent variables with the latent quality μ_j. We consider two ways of formalizing this relationship stochastically, denoted as Versions 1 and 2. In both versions, parallel with the extended scope of the model, the statistical status of μ_j will be changed from being a deterministic expectation, interpreted conditionally, to being a latent stochastic variable, the distribution of which is specified as part of the econometric panel data model.

LATENT HETEROGENEITY. VERSION 1: We first specify

$$
\begin{aligned}
\alpha_{1j} &= \lambda_1 \mu_j + \varepsilon_{1j}, \\
\alpha_{2j} &= \lambda_2 \mu_j + \varepsilon_{2j},
\end{aligned} \tag{7}
$$

$$
\left(\begin{bmatrix} \mu_j \\ \varepsilon_{1j} \\ \varepsilon_{2j} \end{bmatrix} \middle| \begin{bmatrix} \boldsymbol{x}_{1j1}, \boldsymbol{z}_{1j} \\ \boldsymbol{x}_{2jt}, \boldsymbol{z}_{2j} \end{bmatrix} \right) \sim \mathsf{IID}(\mathbf{0}, \boldsymbol{\Omega}), \quad \mathbf{0} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \ \boldsymbol{\Omega} = \begin{bmatrix} \sigma_\mu^2 & 0 & 0 \\ 0 & \sigma_{\varepsilon1\varepsilon1} & \sigma_{\varepsilon1\varepsilon2} \\ 0 & \sigma_{\varepsilon2\varepsilon1} & \sigma_{\varepsilon2\varepsilon2} \end{bmatrix}, \tag{8}
$$

where we expect $\lambda_1 > 0$, $\lambda_2 < 0$, and $\sigma_{\varepsilon1\varepsilon2} = \sigma_{\varepsilon2\varepsilon1} < 0$. When μ_j is low, *i.e.*, when GP j is a low-quality doctor, then his/her patients will have a higher mortality rate than can be explained by $(\boldsymbol{x}_{2jt}, \boldsymbol{z}_{2j})$, and he/she will meet a lower demand than can be explained by $(\boldsymbol{x}_{1j1}, \boldsymbol{z}_{1j})$. Equations (5) and (7) define a four-equation system of structural equations explaining $(y_{1j1}, y_{2jt}, \alpha_{1j}, \alpha_{2j})$ by $(\boldsymbol{x}_{1j1}, \boldsymbol{x}_{2jt}, \boldsymbol{z}_{1j}, \boldsymbol{z}_{2j}, \mu_j)$ and noise terms. Inserting (7) into (5) yields the reduced form

$$
\begin{aligned}
y_{1j1} &= \boldsymbol{x}_{1j1}\boldsymbol{\beta}_1 + \boldsymbol{z}_{1j}\boldsymbol{\gamma}_1 + \lambda_1 \mu_j + \varepsilon_{1j} + u_{1j1}, \quad && j = 1, \dots, M_1, \\
y_{2jt} &= \boldsymbol{x}_{2jt}\boldsymbol{\beta}_2 + \boldsymbol{z}_{2j}\boldsymbol{\gamma}_2 + \lambda_2 \mu_j + \varepsilon_{2j} + u_{2jt}, \quad && j = 1, \dots, M_t; \ t = 1, \dots, T,
\end{aligned} \tag{9}
$$

LATENT HETEROGENEITY. VERSION 2: The alternative version is

$$
\alpha_{1j} = \lambda \alpha_{2j} + \varepsilon_j, \tag{10}
$$

$$
\left(\begin{bmatrix} \alpha_{2j} \\ \varepsilon_j \end{bmatrix} \middle| \begin{bmatrix} \boldsymbol{x}_{1j1}, \boldsymbol{z}_{1j} \\ \boldsymbol{x}_{2jt}, \boldsymbol{z}_{2j} \end{bmatrix} \right) \sim \mathsf{IID} \left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_{\alpha2}^2 & 0 \\ 0 & \sigma_\varepsilon^2 \end{bmatrix} \right), \tag{11}
$$

where we expect $\lambda < 0$. Equations (5) and (10) define a three-equation system of structural equations which explains $(y_{1j1}, y_{2jt}, \alpha_{1j})$ by $(\boldsymbol{x}_{1j1}, \boldsymbol{x}_{2jt}, \boldsymbol{z}_{1j}, \boldsymbol{z}_{2j}, \alpha_{2j})$ and noise terms. Inserting (10) into (5) we get, instead of (9), the reduced form

$$
\begin{aligned}
y_{1j1} &= \boldsymbol{x}_{1j1}\boldsymbol{\beta}_1 + \boldsymbol{z}_{1j}\boldsymbol{\gamma}_1 + \lambda \alpha_{2j} + \varepsilon_j + u_{1j1}, \quad && j = 1, \dots, M_1, \\
y_{2jt} &= \boldsymbol{x}_{2jt}\boldsymbol{\beta}_2 + \boldsymbol{z}_{2j}\boldsymbol{\gamma}_2 + \alpha_{2j} + u_{2jt}, \quad && j = 1, \dots, M_t; \ t = 1, \dots, T.
\end{aligned} \tag{12}
$$

The latter equations, with $\lambda = \lambda_1/\lambda_2$ and $\varepsilon_j = \varepsilon_{1j} - \lambda\varepsilon_{2j}$, could, of course, alternatively have been derived from (5) and (7). However, when (8) holds, (12) is not a reduced form, since α_{2j} is correlated with ε_{2j} and therefore with the composite disturbance in the cross section equation in (12), $\varepsilon_j + u_{1j1}$.

The basic differences between the two model versions can be explained as follows: First, it follows from (7) and (8) that Version 1 implies

$$
\mathsf{E} \left(\begin{bmatrix} \alpha_{1j}^2 & \alpha_{1j}\alpha_{2j} \\ \alpha_{2j}\alpha_{1j} & \alpha_{2j}^2 \end{bmatrix} \middle| \begin{bmatrix} \boldsymbol{x}_{1j1}, \boldsymbol{z}_{1j} \\ \boldsymbol{x}_{2jt}, \boldsymbol{z}_{2j} \end{bmatrix} \right) = \begin{bmatrix} \lambda_1^2 \sigma_\mu^2 + \sigma_{\varepsilon1\varepsilon1} & \lambda_1\lambda_2 \sigma_\mu^2 + \sigma_{\varepsilon1\varepsilon2} \\ \lambda_2\lambda_1 \sigma_\mu^2 + \sigma_{\varepsilon2\varepsilon1} & \lambda_2^2 \sigma_\mu^2 + \sigma_{\varepsilon2\varepsilon2} \end{bmatrix}, \tag{13}
$$

and $\mathrm{cov}(\alpha_{ij}, \varepsilon_{kj}) = \sigma_{\varepsilon i \varepsilon k}$ $(i = 1,2; \ k = 1,2)$, which violate (10)–(11). Second, while Version 1 treats latent quality μ_j as a symmetric 'causal factor' for y_{1j1} and y_{2jt}, Version 2, by treating α_{2j} as the latent causal factor, introduces an asymmetry in the way quality affects latent GP-specific heterogeneity in the two equations in (12).

The empirical implementation of Model A, to be presented in Section 5, relies on Version 2, in that estimation is done sequentially and a predicted value of α_{2j} obtained from the second equation in (12), the excess mortality equation, serves as a proxy for GP quality in the first equation, the demand equation. The estimators used in Section 5 may thus be consistent in Version 2, but inconsistent in Version 1.

Model B: LISREL model with GP quality and patient health latent

Model A gives a rather restrictive, uni-directional description of how demand for GP services is related to GP quality. A LISREL model, *i.e.*, a linear multi-equation structural model with both manifest and latent structural variables may be a better solution to the problem of modeling sample separation. Model B, now to be described, belongs to this class. See Goldberger (1972), Jöreskog (1977), Aigner *et al.* (1984, Sections 4 and 5), and Jöreskog *et al.* (2000) for further discussion of LISREL models.

Again, we exploit the panel design of our data set, with the GP as the observational unit, containing GP-specific time-series for some variables, including patient-switching and mortality rates, as well as GP-specific and patient specific time invariant variables. We let t be the time index and suppress the GP subscript. Boldface and slim letters denote matrices/vectors and scalars, respectively. Model B has three categories of variables: observable (manifest) structural variables, latent structural variables, and error/noise variables. In the baseline version of the model, the categorization of the variables – corresponding to the standard notation for latent and manifest variables in the LISREL documentation – is as follows:

Observable (manifest) structural variables:

- y_1: Number of persons wanting to be entered on list initially, in period 1 (scalar)
- y_{2t}: Number of persons switching to another GP in period t (scalar)
- $\boldsymbol{x}_1$: Observed GP characteristics initially, in period 1 [(6×1)-vector]
- $\boldsymbol{x}_2$: Observed patient characteristics initially, in period 1 [(3×1)-vector]
- x_{3t}: Excess mortality of patient stock in period t (scalar)
- $\boldsymbol{x}_4$: Other time-invariant GP-characteristics unrelated to GP quality [(2×1)-vector]
- $\boldsymbol{y}_2 \equiv [y_{21}, \dots, y_{2T}]'$
- $\boldsymbol{x}_3 \equiv [x_{31}, \dots, x_{3T}]'$

Latent structural variables:

- η_1: Demand directed towards GP (latent, time-invariant scalar)
- ξ_1: GP quality (latent, time-invariant scalar)
- ξ_2: Patient health (latent, time-invariant scalar)
- $\boldsymbol{\xi}_3 \equiv \boldsymbol{x}_4$: Technical redefinition[1]

[1] This redefinition is motivated by the fact that LISREL does not allow x variables to affect the η variables directly in cases where the model also include ξ variables.

Error/noise variables:

ζ_1: Disturbance in demand function

$\varepsilon_1, \varepsilon_{2t}$: Errors in the measurement equations for demand

$\boldsymbol{\delta}_1$: Errors in equations relating GP quality to GP characteristics [(6×1)-vector]

$\boldsymbol{\delta}_2$: Errors in equations relating patient health to patient characteristics. [(3×1)-vector]

δ_{3t}: Errors in equations relating patient health and GP quality to excess mortality (scalar)

$\boldsymbol{\delta}_3 \equiv [\delta_{31}, \ldots, \delta_{3T}]'$

$\boldsymbol{\varepsilon}_2 \equiv [\varepsilon_{21}, \ldots, \varepsilon_{2T}]'$

A basic hypothesis of the *baseline version* of Model B is that *GP quality, ξ_1, and patient health status, ξ_2, both time invariant scalars, are exogenous to the rest of the system.* The quality variable ξ_1 corresponds to the variable μ_j in Model A, Version 1. Time invariance and exogeneity are also assumed for the time invariant GP characteristics, $\boldsymbol{x}_4 = \boldsymbol{\xi}_3$, in the model represented by the gender and the country of origin of the GP; see below. These four variables are considered as determined from outside, inherent in the GP and in the patient, and hence are not subject to feedback from the rest of the system. This is an important assumption, which, for at least ξ_1 and ξ_2, may be questioned. To some extent it will be modified later on (Section 5), in examining the robustness of the primary conclusions concerning the link between GP quality and patient demand to changes in basic assumptions. These genuinely exogenous variables are, in the baseline model, indicated by observable 'counterparts', which, by assumption, become endogenous.

The baseline model has four elements: (i) a demand function for GP services expressed in terms of latent variables, (ii) measurement equations indicating this latent demand, (iii) measurement equations indicating GP quality and health status of listed persons, and (iv) distributional assumptions for the latent exogenous variables and the error terms.

First, the baseline version of the *demand function*, relating latent demand (endogenous) to GP quality (exogenous), and latent health status and other characteristics of the listed persons (all exogenous), is:

$$\eta_1 = \Gamma_{11}\xi_1 + \Gamma_{12}\xi_2 + \boldsymbol{\Gamma}_{13}\boldsymbol{\xi}_3 + \zeta_1 = \begin{bmatrix} \Gamma_{11} & \Gamma_{12} & \boldsymbol{\Gamma}_{13} \end{bmatrix} \begin{bmatrix} \xi_1 \\ \xi_2 \\ \boldsymbol{\xi}_3 \end{bmatrix} + \zeta_1. \tag{14}$$

We can interpret $\Gamma_{11}, \Gamma_{12}, \boldsymbol{\Gamma}_{13}$ as (vectors of) *structural coefficients* and ζ_1 as a disturbance.

Second, the baseline version of the *measurement system for latent demand* is

$$\begin{bmatrix} y_1 \\ \boldsymbol{y}_2 \end{bmatrix} = \begin{bmatrix} \Lambda_{Y11} \\ \boldsymbol{\Lambda}_{Y21} \end{bmatrix} \eta_1 + \begin{bmatrix} \varepsilon_1 \\ \boldsymbol{\varepsilon}_2 \end{bmatrix}. \tag{15}$$

This subsystem expresses that $y_1, \boldsymbol{y}_{21}, \ldots, \boldsymbol{y}_{2T}$ are treated as $T+1$ observable indicators of the latent demand for GP services. Technically, in factor-analytic terminology, we can interpret Λ_{Y11} and $\mathbf{\Lambda}_{Y21}$ as *factor loadings* for, respectively, the number of persons wanting to be on the list initially (positive loading) and the number of persons switching to another GP in a later period (negative loading), on latent demand. In standard regression terminology, we can interpret Λ_{Y11} and $\mathbf{\Lambda}_{Y21}$ as the marginal effects of the latent variables on the corresponding observable variables. The error terms $(\varepsilon_1, \boldsymbol{\varepsilon}_2)$ may contain, *inter alia*, measurement errors. Third, the baseline version of the *measurement system for GP quality and patient health* is specified as

$$(16)\qquad \begin{bmatrix} \boldsymbol{x}_1 \\ \boldsymbol{x}_2 \\ \boldsymbol{x}_3 \\ \boldsymbol{x}_4 \end{bmatrix} = \begin{bmatrix} \mathbf{\Lambda}_{X11} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{\Lambda}_{X22} & \mathbf{0} \\ \mathbf{\Lambda}_{X31} & \mathbf{\Lambda}_{X32} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \boldsymbol{I} \end{bmatrix} \begin{bmatrix} \xi_1 \\ \xi_2 \\ \boldsymbol{\xi}_3 \end{bmatrix} + \begin{bmatrix} \boldsymbol{\delta}_1 \\ \boldsymbol{\delta}_2 \\ \boldsymbol{\delta}_3 \\ \mathbf{0} \end{bmatrix}.$$

This subsystem expresses, that the vector of observed GP characteristics, $\boldsymbol{x}_1$, is related to latent GP quality, that the vector of observed patient characteristics is related to latent patient health, and that the T vector of excess mortalities, $\boldsymbol{x}_3$, is related to both GP quality and patient health. Technically, in factor-analytic terminology, $\mathbf{\Lambda}_{X11}, \mathbf{\Lambda}_{X31}$ can be interpreted as, respectively, factor loadings for GP characteristics and excess patient mortality on latent GP quality. Likewise, $\mathbf{\Lambda}_{X22}, \mathbf{\Lambda}_{X32}$ can be interpreted as factor loadings for, respectively, patient characteristics and excess patient mortality on patient health. The error terms $(\boldsymbol{\delta}_1, \boldsymbol{\delta}_2, \boldsymbol{\delta}_3)$ may, contain, *inter alia*, measurement errors. The fourth equation in this sub-system simply states $\boldsymbol{x}_4 = \boldsymbol{\xi}_3$, which implies, *inter alia*, that these variables, representing observed heterogeneity of the GPs, are assumed to be error-free.

Fourth, the process determining the latent exogenous variables $\xi_1, \xi_2, \boldsymbol{\xi}_3$ is modeled in terms of their first-order and second-order moments as follows:

$$(17)\qquad \mathsf{E}\begin{bmatrix} \xi_1 \\ \xi_2 \\ \boldsymbol{\xi}_3 \end{bmatrix} = \begin{bmatrix} \mu_{\xi1} \\ \mu_{\xi2} \\ \boldsymbol{\mu}_{\xi3} \end{bmatrix}, \qquad \mathsf{V}\begin{bmatrix} \xi_1 \\ \xi_2 \\ \boldsymbol{\xi}_3 \end{bmatrix} = \begin{bmatrix} \Phi_{11} & \Phi_{12} & \mathbf{\Phi}_{13} \\ \Phi_{21} & \Phi_{22} & \mathbf{\Phi}_{23} \\ \mathbf{\Phi}_{31} & \mathbf{\Phi}_{32} & \mathbf{\Phi}_{33} \end{bmatrix},$$

while the distributions of the error and noise terms are assumed to satisfy

$$(18)\qquad \mathsf{E}[\zeta_1] = 0, \qquad \mathsf{V}[\zeta_1] = \Psi_{11},$$

$$(19)\qquad \mathsf{E}\begin{bmatrix} \boldsymbol{\delta}_1 \\ \boldsymbol{\delta}_2 \\ \boldsymbol{\delta}_3 \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \qquad \mathsf{V}\begin{bmatrix} \boldsymbol{\delta}_1 \\ \boldsymbol{\delta}_2 \\ \boldsymbol{\delta}_3 \end{bmatrix} = \begin{bmatrix} \mathbf{\Theta}_{\delta11} & \mathbf{\Theta}_{\delta12} & \mathbf{\Theta}_{\delta13} \\ \mathbf{\Theta}_{\delta21} & \mathbf{\Theta}_{\delta22} & \mathbf{\Theta}_{\delta23} \\ \mathbf{\Theta}_{\delta31} & \mathbf{\Theta}_{\delta32} & \mathbf{\Theta}_{\delta33} \end{bmatrix},$$

$$(20)\qquad \mathsf{E}\begin{bmatrix} \varepsilon_1 \\ \boldsymbol{\varepsilon}_2 \end{bmatrix} = \begin{bmatrix} 0 \\ \mathbf{0} \end{bmatrix}, \qquad \mathsf{V}\begin{bmatrix} \varepsilon_1 \\ \boldsymbol{\varepsilon}_2 \end{bmatrix} = \begin{bmatrix} \Theta_{\varepsilon11} & \mathbf{\Theta}_{\varepsilon12} \\ \mathbf{\Theta}_{\varepsilon21} & \mathbf{\Theta}_{\varepsilon22} \end{bmatrix},$$

$$(21)\qquad \begin{bmatrix} \xi_1 \\ \xi_2 \end{bmatrix} \perp \zeta_1 \perp \begin{bmatrix} \boldsymbol{\delta}_1 \\ \boldsymbol{\delta}_2 \end{bmatrix} \perp \begin{bmatrix} \varepsilon_1 \\ \boldsymbol{\varepsilon}_2 \end{bmatrix}.$$

The final assumption, (21), where $\perp$ denotes orthogonal, is crucial for the modeling of causality and non-causality in Model B. It expresses, *inter alia*, the assumed exogeneity for GP quality and patient health. Its essence is that these variables, being modeled by (17), remain unaffected by the perturbations in the demand equation disturbances, and the errors in the measurement systems for demand (endogenous) and latent GP quality and latent patient health (exogenous). Since arguments may be raised that this model disregards a possible effect of GP quality on the listed patients' initial health status, we will in addition consider a modified version, Model C, in which this potential link is modeled and hence may be tested for.

4 Data

Data sources and data design

Prior to the introduction of the regular GP scheme in June 2001, the health authorities gathered the information needed to assign GPs to the entire Norwegian population. All inhabitants were asked to rank their three most preferred GPs in an entry form. The GPs were asked to report the maximum number of patients they would like to take care of. The health authorities utilized this information as an input in an algorithm allocating inhabitants to GPs. Most people got listed with the GP whom they had consulted prior to the reform (Lurås, *et al.*, 2003).

Our data stem from The Norwegian General Practitioners Database supplemented by a measure of the GP density, as calculated from the number of contracted GPs in each municipality in June 2001, as well as aggregate age/gender specific mortality rates. The latter are calculated by means of aggregate mortality rates constructed by Statistics Norway. The Norwegian General Practitioners Database contains information on all Norwegian GPs, and the variables describing the individual GPs practice is provided by the National Insurance Administration (NIA) every six month. The database is administered by the Norwegian Social Science Data Services, who merge the information reported by NIA with socio-demographic variables as income, wealth and marital status, registered by statistics Norway. For GPs practicing in 14 municipalities, sampled by stratification, the database also includes characteristics for the patients who were listed in the GP's practice in June 2001, such as the median income and median wealth, and the proportion who have not finished high-school. For each GP we know the number of persons who ranked the GP at the top when returning the entry form, in this paper to be given the interpretation as an indicator of the *demand* facing the GP. After the reform was implemented, the GP database is updated at regular intervals to give the number of persons who are *actually listed* in the practice. After excluding observations with key variables missing, our unbalanced panel data set consists of a sample of 484 GPs observed up to 7 six-month periods.[2] The pattern of observation is described in Table 1, from which we see that 441, or 91 %, of the GPs are observed in all 7 periods.

Table 1: Pattern of observations

Response pattern	No. of GPs	Freq., %	Cum. freq., %
1111111	441	91.12	91.12
11111..	11	2.27	93.39
111111.	11	2.27	95.66
1111...	8	1.65	97.31
111....	7	1.45	98.76
11.....	5	1.03	99.79
11...11	1	0.21	100.00
..	484	100.00	..

Table 2 lists and defines the variables applied in this paper, Table 3 gives overall descriptive statistics for the variables, and Table 4 gives descriptive statistics of the GP-specific means of the time varying variables. Descriptive statistics for variables at the level of the municipality are given in Table 5. We distinguish between variables observed at the GP level and variables which are observed at the municipality level and hence are common to all GPs practising in the same municipality.

The symbols used for the observable variables in the exposition of Models A and B above, $(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z})$, have their empirical counterparts among the the variables in Table 2. This correspondence is given below (the GP subscript, for simplicity, suppressed):

Model A:
$y'_1 = [DEMAND]$, $y'_{2t} = [ACTMORT_t]$, $\boldsymbol{x}'_1$ is empty, $\boldsymbol{x}'_{2t} = [EXPMORT_t]$

$$\boldsymbol{z}'_1 = \begin{bmatrix} GPDENS \\ MARRIEDGP \\ SPECGEN \\ SPECCOM \\ SPECOTH \\ \widehat{ALPHA} \\ IMMIGRGP \\ FEMALEGP \\ AGEGP \\ AGEGPSQ \end{bmatrix}, \quad \boldsymbol{z}'_2 = \begin{bmatrix} CENTRAL \\ LESSCENT \\ LEASTCENT \\ LOSUBMIT \\ LOEDUC \\ PINCOME \\ PWEALTH \\ SPECGEN \\ SPECCOM \\ SPECOTH \\ FEMALEGP \\ AGEGP \\ AGEGPSQ \end{bmatrix}$$

[2]The GPs from the municipality Tromsø, 44 in total, were excluded from the sample. Here, the regular GP scheme was implemented already in 1993 and very few inhabitants returned the entry form.

Table 2: Variable definitions

Variable	Definition, Type of variable	Formula
DEAD	No. of dead persons on GP's list	
EXPDEAD	No. of persons on GP's list expected to die per year	Expected mortality rates based on age distribution of persons on list and population age-specific mortality rates
ACTMORT	Actual no. of mortalities per 1000 persons listed	= DEAD/LISTSIZE
EXPMORT	Expected no. of mortalities per 1000 persons listed	= EXPDEAD/LISTSIZE
EXCMORT	Excess mortality relative to list size	= ACTMORT – EXPMORT
LISTSIZE	GP's actual no. of patients	
DEMAND	No. of persons ranking this GP as most preferred when returning entry form	
DEMAND1	Demand for this GP normalized against GP density in municipality	= DEMAND * GPDENSITY
AGEGP	Age of GP, January 2002	
LEAKRATE	Share of patients switching to another GP.	= no. of persons leaving/LISTSIZE
LOLEAK	log(LEAKRATE/(1-LEAKRATE))	
FEMALEGP	Dummy variable	=1 if GP is female
MARRIEDGP	Dummy variable	=1 if GP is married
IMMIGRGP	Dummy variable	=1 if GP is non-Scandinavian citizen
SALARY	Dummy variable	=1 if GP is remunerated by a fixed salary scheme
SPECGEN	Dummy variable	=1 if GP is a specialist in general practice
SPECCOM	Dummy variable	=1 if GP is a specialist in community medicine
SPECOTH	Dummy variable	=1 if GP is a specialist in other kind of medicine
LEASTCENT	Dummy variable	=1 if practice in Least central municipality
LESSCENT	Dummy variable	=1 if practice in Less central municipality
CENTRAL	Dummy variable	=1 if practice in Central municipality
MOSTCENT	Dummy variable	=1 if practice in Most central municipality
PINCOME	Median income (NOK 1000) of persons assigned to this GP in 2001	
PWEALTH	Median wealth (NOK 1000) of persons assigned to this GP in 2001	
PFORMSUB	Share of persons returning forms in 2001 among those assigned to this GP in 2001	
LOSUBMIT	log(PFORMSUB/(1-PFORMSUB))	
PEDUC	Share of persons without finished high-school among those assigned to this GP in 2001	
LOEDUC	log(PEDUC/(1-PEDUC))	
GPDENSITY	No. of GPs per 1000 inhabitans in municipality	

Model B:

$$\boldsymbol{y}_1' = [\, DEMAND \,], \quad \boldsymbol{y}_2' = \begin{bmatrix} LOLEAK_1 \\ LOLEAK_2 \\ LOLEAK_3 \\ LOLEAK_4 \\ LOLEAK_5 \\ LOLEAK_6 \\ LOLEAK_7 \end{bmatrix},$$

$$\boldsymbol{x}_1' = \begin{bmatrix} SPECGEN \\ SPECCOM \\ SPECOTH \\ SALARY \\ MARRIED \\ AGEGP \end{bmatrix}, \quad \boldsymbol{x}_2' = \begin{bmatrix} LOSUBMIT \\ LOEDUC \\ PINCOME \end{bmatrix}, \quad \boldsymbol{x}_3' = \begin{bmatrix} EXCMORT_1 \\ EXCMORT_2 \\ EXCMORT_3 \\ EXCMORT_4 \\ EXCMORT_5 \\ EXCMORT_6 \\ EXCMORT_7 \end{bmatrix}, \quad \boldsymbol{x}_4' = \begin{bmatrix} FEMALEGP \\ IMMIGRGP \end{bmatrix}$$

Variables at the GP level, including patient stock characteristics

The variables collected at the GP level and related to the mortality of the persons on the GP's list are DEAD, EXPDEAD, ACTMORT, EXPMORT, and EXCMORT. The number of individuals leaving the list and the number of mortalities on each individual GP's list during a six-month period is registered in the GP database, except for the year 2002, where this information is registered for the whole calendar year only. We have allocated the mortalities and the switches in this year on the two half-years, according to the list sizes in the two half-years. DEAD denotes the number of mortalities during a period, and ACTMORT measures mortality per thousand listed patients.

GPs with a relatively high proportion of elderly people on their lists are presumably recorded with a relatively high mortality rate in any period. In order to compensate for this heterogeneity when measuring excess mortality, we proceeded as follows: Among the information registered in the GP database is the number of listed males and females belonging to each of the age categories 0–7, 8–19, 20–29, 30–39, 40–49, 50–59, 60–69, 70–79 years, and 80 years and above. By applying the gender and the age specific death probabilities (Statistics Norway, 2005a) and the age distribution in Norway (Statistics Norway, 2005b), we can for each GP calculate the expected number of mortalities, EXPDEAD, and the expected per thousand mortality rate, EXPMORT, *i.e.*, EXPDEAD per thousand listed persons. This enables us to calculate EXCMORT: the difference between the actual and the expected mortality rates at the GP level, henceforth to be referred to as the GP-specific *excess mortality rate*

From Table 3 we see that the overall mean of the actual number of mortalities during a six-month period (DEAD) is 5.63, and from Table 4 that its *GP-specific mean* ranges from 0 to 22 mortalities. The overall mean of the expected number of mortalities (EXPDEAD) is 5.12, with range from 0.05 to 23.9. By combining the aggregate death probabilities and the age-gender distribution of listed patients we have obtained a two-dimensional variable, with a mean value not far from the actual mean number of deaths according to the mortality statistics included in the General Practitioners Database.

Table 3: Global descriptive statistics

Variable	Obs	Mean	St. Dev	Skewness	Kurtosis	Min	Max
DEAD	3260	5.6254	4.2566	1.2850	5.3638	0	29
EXPDEAD	3275	5.1252	3.2353	1.1384	4.6549	0.0503	23.9037
ACTMORT	3251	4.7629	3.5683	1.9031	11.9554	0	37.9669
EXPMORT	3256	4.2390	2.1982	0.8033	3.3628	0.3781	13.6722
EXCMORT	3251	0.5235	2.6810	2.2873	17.5835	-8.4804	29.4960
DEMAND	3275	825.9289	520.5296	1.1275	4.8392	21	3152
DEMAND1	3275	716.0950	447.6848	1.1181	4.8671	18.3694	2757.1620
LEAKRATE	3251	0.0290	0.0269	6.2202	108.6469	0	0.6660
AGEGP	3275	47.0293	7.5861	0.0610	2.8225	28	70
FEMALEGP	3275	0.3600	0.4801	0.5833	1.3403	0	1
MARRIEDGP	3275	0.6889	0.4630	-0.8159	1.6656	0	1
IMMIGRGP	3275	0.0504	0.2188	4.1111	17.9015	0	1
SALARY	3275	0.0544	0.2267	3.9315	16.4564	0	1
LISTSIZE	3256	1211.0820	401.3513	0.1665	3.4662	123	2687
SPECGEN	3275	0.5597	0.4965	-0.2405	1.0578	0	1
SPECCOM	3275	0.0696	0.2545	3.3821	12.4389	0	1
SPECOTH	3275	0.0345	0.1825	5.1008	27.0180	0	1
LEASTCENT	3275	0.0256	0.1581	6.0012	37.0144	0	1
LESSCENT	3275	0.0116	0.1071	9.1212	84.1960	0	1
CENTRAL	3275	0.0403	0.1967	4.6747	22.8526	0	1
MOSTCENT	3275	0.9224	0.2675	-3.1588	10.9778	0	1
PINCOME	3275	196.5151	23.7784	0.2245	2.7626	130.9460	261.7615
PWEALTH	3275	63.4333	79.2646	1.2259	4.9701	-94.0780	467.3245
PFORMSUB	3275	0.7320	0.2642	-0.7781	2.5387	0.0554	1.0000
PEDUC	3275	0.4446	0.1524	-0.2344	2.4003	0.0213	0.8090

Table 4: Descriptive statistics for GP-specific means, No. of GPs: 484

Variable	Mean	St. Dev	Skew	Kurt	Min	Max	Between variation as share of total, %
DEAD	5.5612	3.6936	1.1570	4.8702	0.0000	22.0816	75.3
EXPDEAD	5.0638	3.2286	1.2459	5.2236	0.3427	22.6007	99.6
ACTMORT	4.7216	2.9549	1.7391	11.2484	0.0000	28.1421	68.6
EXPMORT	4.2065	2.1513	0.8130	3.3556	0.4489	12.4530	95.8
EXCMORT	0.5151	1.8582	3.3390	30.4057	-4.1782	20.3758	48.0
LEAKRATE	0.0308	0.0255	4.4032	36.3194	0.0007	0.2737	90.2
LISTSIZE	1201.6160	401.2103	0.1835	3.4536	152.7857	2620.0000	99.9 %

The overall mean number of mortalities per thousand listed persons is 4.76. Its GP-specific means range from zero to 28 deaths per thousand. EXPMORT has an overall mean of 4.24 deaths per thousand, and the GP-specific means ranges from 0.45 to 12.45. An important variable in the analyses is the excess mortality rate, EXCMORT. As explained, a positive (negative) value means that the mortality rate at the GP level is higher (lower) than expected from the age and gender distribution of the persons on each GP's list. We note that the overall mean of EXCMORT is positive. The reason for this could be, on the one hand that the mortality tables are constructed from cross-sectional variation in mortalities during a period of only one year, on the other hand that life expectancy is known to be lower than the national average in the municipality Oslo (Statistics Norway, 2006a & 2006b), which is the location of 426 of the 484 GPs in the data set. A third explanation may be that the number of deaths in the first period is somewhat overestimated.[3]

We denote the number of consumers ranking a specific GP as the most preferred GP as DEMAND directed towards this GP. This variable is time-invariant, as the matching of GPs and patients has been undertaken only once in connection with the implementation of the regular GP reform in 2001. The average GP was preferred by 826 inhabitants, but there is a lot of variation. The most popular GP in our data set was preferred by 3152 inhabitants, whereas some GPs were not preferred by any inhabitants. Being requested by a large number of inhabitants in a municipality with a high GP density is not equivalent to be strongly requested in a municipality where the GP density is low. We have considered two ways of controlling for differences in GP density across municipalities. The primary one is to include a measure of GP density as an additional explanatory variable representing observed heterogeneity. The secondary one is to weight DEMAND by a measure of GP density. The specific measure of the GP density applied here is the number of GPs per thousand inhabitants in the municipality (GPDENSITY); see the next sub-section for an elaboration. The specific measure of weighted demand we used is the variables DEMAND1, obtained as the product of DEMAND and GPDENSITY. Taking DEMAND1 as the relevant demand variable in our analysis implies that a given number of first rankings is interpreted as a higher demand in a municipality with a high GP density than in a municipality where GP density is low. Prediction **[P2]** in Section 2 provides the rationale for transforming the demand variable in this way.

We have information on the number of persons leaving a GP's list in order to enroll on a competitor's list.[4] We refer to the proportion of listed persons leaving the list and switching to other GPs in the municipality as LEAKRATE, and interpret this variable as a time-varying indicator of the demand. Its overall mean is 3 %, its GP-specific means vary from close to zero to 27 %, and the between GP variation accounts for as much as 90 % of the total variation.

Our data set also reports the GP's age (AGEGP), gender (FEMALEGP), and marital status (MARRIEDGP) as well as the GP's birth country. We see that the average GP is

[3]The *period-specific means* of ACTMORT seem to be higher in the first period than in the later periods. This is not unexpected as the first period is one month longer than the other ones, a fact we have adjusted for simply by multiplying the number of mortalities in the first period with $\frac{6}{7}$. We suspect that mortalities in the period April to June 2001 may also have been registered, although with a lag. In estimating the models, we have therefore alternatively applied an adjustment factor of $\frac{6}{9}$, which would have been correct if mortalities from April to June 2001 were indeed included among those registered for the second half-year 2001. The main results are not affected by this modification of the adjustment factor.

[4]Patients leaving the list because they migrate or move to another municipality are excluded from these numbers. We thus interpret LEAKRATE as the proportion of the listed persons who switch because they actually prefer another GP.

47 years old, that 36 % of the GPs are females and that 69 % are married. We have constructed a binary variable, IMMIGRGP, equal to 1 if the GP is born in a non-Scandinavian country. About 5 % of the GPs in our sample have this property. The variable denoted SALARY is a binary variable equal to one if the GP is remunerated by means of a fixed salary contract when practicing as a GP, and we see that 5 % of the GPs have this kind of contract.

The number of patients *actually* listed in the practice at the beginning and end of each period is registered in the GP database. To take account of within-period changes of this variable, we construct the average of the numbers recorded at the beginning and at the end of each period, giving the variable LISTSIZE. Its overall mean is 1211 persons, while its GP-specific means range from 153 to 2620.

Our data set also reports whether the GP is a specialist in general medicine (dummy variable SPECGEN), in community medicine (dummy variable SPECCOM) or in another medical field (dummy variable SPECOTH) – all of which are time-varying dummies, but the within-GP variation is small. Overall, 56 % of the GPs are specialists in general medicine, 7 % are specialists in community medicine. and 3 % are specialists in an other field.

Our GP-level data also contain the following information on the patients who were listed in the practice in June 2001: the median net income and median net wealth among the listed patients older than 30 years, the proportion of listed patients who are older than 30 and have not finished high school and the proportion of the listed patients who submitted the entry form signalling GP preferences. By construction, these variables are uni-dimensional, as this information is not updated after the implementation of the regular General Practitioner Scheme. The income and wealth variables, measured in 1.000 NOK, PINCOME and PWEALTH, have overall means 196.5 and 63.4, respectively. Not unexpectedly, they vary considerably: the GP whose listed patients are on average richest, have a median income twice that of the GP whose listed patients have the lowest median income. The corresponding median wealth, PWEALTH extends from -94.2 to 467.3. In the LISREL analysis, after some trial runs, we decided to exclude PWEALTH from the variable list, in order to ensure convergence. We suspect that the reason for this is that income and wealth are highly correlated.

Finally, PFORMSUB denotes the proportion of the listed patients who submitted the entry form prior to the implementation of the regular General Practitioner Scheme. This variable varies from nearly zero to one, indicating that some GPs were not assigned any patients who submitted the entry form, while other GPs were assigned only patients who submitted the form. The mean of this variable is 0.73, indicating that the average GP have a list where 73 % of the patients submitted the entry form. We denote by PEDUC the proportion of the listed patients who are older than 30 and have not finished high school. We see that the average GP have 45 % of the listed patients characterized by not having finished high school. This share also varies considerably, from 2 % to 81 %.

Variables at the level of the municipality

Some important variables reported are specific to the municipality in which the practice of the GP is localized. Statistics Norway has constructed an indicator of *centrality*, placing each Norwegian municipality in one of 4 centrality categories. This indicator captures, *inter alia*, the population density and the distance to the nearest city of a certain size. We refer to these categories, in an order of increasing centrality, as least central (LEASTCENT), less central (LESSCENT), central (CENTRAL) and most central (MOSTCENT). In our sample six municipalities are categorized as least central municipalities, one as being less central, three municipalities as being central, while three municipalities are categorized as most central. A description of the GP density measure, GPDENSITY, and the number of GPs within each municipality are given in Table 5. We see that the range of GPDENSITY is from 0.55 to 1.57 GPs per thousand inhabitants.

TABLE 5: DESCRIPTIVE STATISTICS FOR VARIABLES AT THE MUNICIPAL LEVEL

Municipality	No. of GPs	GPDENSITY
Frogn	7	0.69
Oslo	426	0.87
Stor-Elvdal	2	0.68
Søndre Land	5	0.98
Notodden	10	0.82
Tvedestrand	5	0.84
Vindafjord	4	0.83
Os	12	0.86
Jølster	2	1.01
Ulstein	6	0.91
Overhalla	2	0.55
Beiarn	1	1.57
Porsanger	2	0.69
Total	484	

5 Estimation and test results

Our data set include several variables expected to be related to the individual quality of the GP, primarily the demand variable DEMAND and excess mortality variable EXCMORT. It is important to note that information on the number of patient mortalities at the level of the GP is not publicly available in Norway. It is thus highly unlikely that individuals' choice of GP is directly related to these numbers. We derived from our theoretical model in Section 2, the prediction, [**P1**], that a positive relationship exists between the quality of the individual GP and the demand facing the GP, even when individual consumers may have incorrect perceptions of GP quality. How perceptions are formed is unknown and we make no attempt to open this "black box", as an enquiry into the formation of human perceptions is beyond the scope of this paper. We let consumers be heterogeneous with regard to their preferences, the information they possess, and the way they process information. Consumers may choose the same GP for different reasons, or different GPs for the same reason. We expect however that the GPs' appearance, experiences from earlier consultations with available GPs, advice from relatives and friends and even rumor to enter the "black box" as inputs in the formation of individuals' quality perceptions.

As explained in Sections 3 and 4, the statistical modeling of the 'causality chain' giving rise to this relationship is rather different in the two econometric models we consider, Model A and Model B. In addition, mainly as a robustness check of our main conclusion, results obtained from a third model, Model C – essentially a modification of Model B in one important respect – will be briefly considered at the end.

The estimation procedure for **Model A**, Version 2, represented by Equation (12), is, as explained in Section 3, a stepwise procedure. Using in both steps modules in the STATA 9 software, we estimate in the first step the effect of the variables representing observed heterogeneity and other assumed exogenous variables on the mortality rates and extract the predicted value of the random effect for each GP in the sample, ALPHAHAT.[5] In estimating this mortality equation, *i.e.*, the second equation of (12), we allow for the possibility that the residuals are not independent within municipalities, and report robust standard errors. In the second step, this prediction, treated as an exogenous variable, is inserted in the demand equation, *i.e.*, the first equation of (12). [6]

For the multi-equation model, **Model B**, we apply the Maximum Likelihood (ML) procedure in the LISREL 8.80 software. [7] The actual number of linear equations to be simultaneously estimated is 25. In this model both the quality of the individual GPs and the unobserved aggregate health status of the listed patients occur as latent exogenous variables, as explained in Section 3.

Results for Model A

The mortality equation

The dependent variable in this equation is ACTMORT. Observable heterogeneity is controlled for in various ways. First, to control for differences in the age and gender distribution of the GPs' listed patients, we include the expected mortality rates EXPMORT as an explanatory variable. Second, to account for heterogeneity between municipali-

[5]See Hsiao (2003, Section 6.2.2.c) and Lee and Griffiths (1979) on the prediction of random effects from panel data.

[6]An even simpler alternative also considered is a single-equation model where the excess mortality rate is inserted directly as a quality indicator in the demand equation – in a sense merging the two equations in (12) into one equation. The underlying assumption is that the GP specific level of patient excess mortality is exogenous. This approach, however, is defective to the extent that the GPs have an inhomogeneous patient stock with respect to the average health status, which will induce a bias in the coefficient estimate of the quality variable. The results from this 'single-equation version' of Model A is presented in Appendix A, Table A.1

[7]The Covariances and Means to be analyzed are estimated by the EM procedure, as there are some missing observations due to the unbalanced panel data.

ties of different centrality, we include the centrality dummies LEASTCENT, LESSCENT and CENTRAL. Third, to account for observable GP heterogeneity we include AGEGP, three GP speciality dummies as well as FEMALEGP. Fourth, we include variables describing the listed patients, with intention to control for the possibility that the average health status of patients varies between GPs. Since there is evidence in the medical literature that life expectancy and health status is related to education, income and wealth (Lantz *et al.*, 1998, Papas *et al.*, 1993), we include PEDUC, PINCOME and PWEALTH as proxies for the average health status of the persons listed with each GP. Fifth, as discussed in Section 2, still another kind of heterogeneity may also occur: individuals who chose not to submit the entry form stating their preferences for certain GPs, may have an average health status different from those who returned the entry form. We take account of this by including PFORMSUB as an explanatory variable. Since the range of PEDUC and PFORMSUB is restricted to the $(0, 1)$ interval, we transform them by the *log-odds* using the formula $\ln(\frac{x}{1-x})$ in order to extend their range to $(-\infty, +\infty)$ which gives a better balance with the unbounded range of the other explanatory variables.

TABLE 6: MODEL A. MORTALITY EQUATION, GLS ESTIMATES

No. of obs.=3251. No. of GPs=484. Obs. per GP.: min=1, mean=6.7, max=7

Regressor	Estimate	Std.Err.
EXPMORT	1.2900	0.0383**
CENTRAL	-0.6365	0.1656**
LESSCENT	-0.3897	0.0434**
LEASTCENT	-0.2359	0.2578
LOSUBMIT	-0.0496	0.0025**
LOEDUC	-0.0468	0.0185*
PINCOME	-0.0097	0.0015**
PWEALTH	-0.0094	0.0006**
SPECGEN	0.0156	0.0323
SPECCOM	0.2164	0.0536**
SPECOTH	-0.6432	0.0717**
FEMALEGP	-0.1720	0.0422**
AGEGP	-0.0547	0.0498
AGEGPSQ	0.0003	0.0005
CONST	4.0038	1.0395*
$\sigma_{\alpha 2}$	1.3602	
σ_{u2}	2.1076	
ρ	0.2940	
R^2		
within	0.0578	
between	0.7104	
overall	0.5071	
Wald chi2(11)	165868	
p-value	0.0000	

*) Significantly $\neq 0$ at the 5 % level (two-tailed test)

**) Significantly $\neq 0$ at the 1 % level (two-tailed test)

The results from the mortality equation are presented in Table 6. All of the estimated coefficients except LEASTCENT, SPECGEN, AGEGP and AGEGPSQ are statistically significant, and the overall R^2 is rather high: 0.5071. The coefficient of EXPMORT is positive, as expected, since high expected mortality should have a positive effect on actual mortality. Further, GPs having their practice in a central or less central municipality have a significantly lower patient mortality rate than GPs in most central municipalities. The negative estimated coefficient of LOSUBMIT indicates that GPs who were assigned a high proportion of the persons who expressed their GP preferences in advance, have lower patient mortality than GPs who obtained a low proportion of patients who actively selected their GP. The coefficient on the education variable is negative, which is not in accordance with intuition saying that GPs with a high proportion of low-educated patients have a higher mortality rate and may be attributed to education being correlated with income and wealth.[8] The estimated coefficients of income and wealth have the expected negative signs. Since these variables are measured in 1000 NOK, an increase in PINCOME and PWEALTH of NOK 100.000 would be accompanied by a reduction in the mortality rates of 0.97 and 0.94 deaths per thousand, respectively. We see that, *ceteris paribus*, the patient mortality rate of GPs who are specialists in community medicine is significantly higher and that of GPs who are specialists in a field other than general medicine and community medicine is significantly lower than the patient mortality rate of other GPs. Finally, female GPs have, *ceteris paribus*, a lower mortality rate of their patient stock than male GPs.

Statistics describing the predicted values of the GP specific heterogeneity variables, *i.e.*, of the α_{2j}s obtained from the second equation of (12), denoted as ALPHAHAT, are given in Table 7. According to our interpretation of Model A, α_{2j} represents a latent variable that is linearly related to quality, confer Equations (7) and (10). A histogram is given in Figure 1. Its form is not very far from a bell-shape, although with an outlier at the right end, equal to 13.11 deaths per thousand.

We next proceed to consider the results for the demand equations, in which the ALPHAHAT predictions are among its explanatory variables.

TABLE 7: MODEL A: PREDICTED QUALITY INDICATOR ALPHAHAT. DESCRIPTIVE STATISTICS

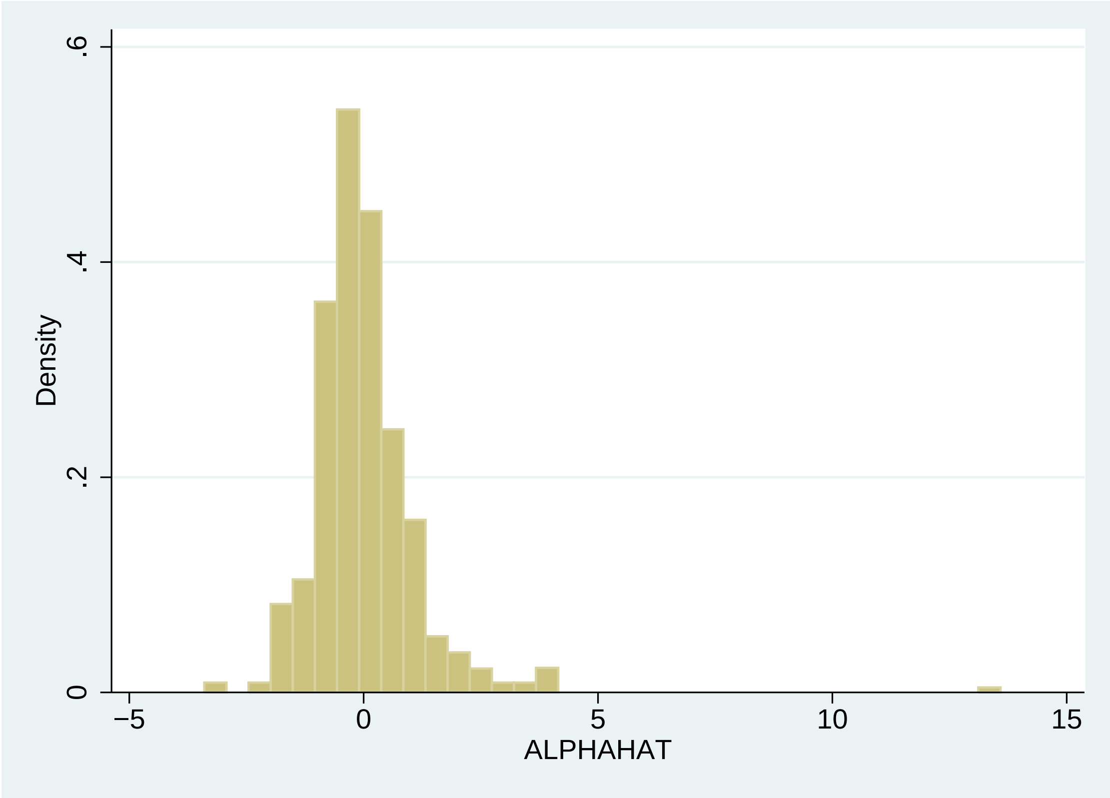
VARIABLE	MEAN	ST. DEV	SKEW	KURT	MIN	MAX
ALPHAHAT	0.0000	1.1546	3.5900	37.5611	-3.3982	13.1143

The demand equation

The estimation results for the demand equation, in the two versions explained above, are presented in Table 8. In panel A, to which we will give most attention, the dependent variable is DEMAND. Supplementary results, to investigate the sensitivity of the findings,

[8]This intuition is supported by supplementary regressions in which income and wealth were excluded as regressors, giving a significantly positive coefficient of LOEDUC.

Figure 1: Histogram of ALPHAHAT



when applying the GP density weighted measure of demand, DEMAND1, are given in panel B. In panel A, the GP density is included as a explanatory variable instead.

Explanatory variables included in both versions of the equation are the specialization dummies, SPECGEN and SPECOTH, which may be interpreted by the market as observable quality indicators. We also include the specialization dummy SPECCOM. We expect this variable to have a negative effect on demand. One reason for this may be that GPs who are specialists in community medicine are known to participate more frequently in the community health service than GPs who do not have this specialization (Godager and Lurås, 2007), and as a result, they may supply fewer business hours per week and hence appear less attractive for patients. To take account of observable GP heterogeneity we in addition include dummy variables MARRIEDGP, IMMIGRGP and FEMALEGP, as well as AGEGP as explanatory variables. Because the demand variable is time invariant, cf. the first equation of (12), weighted between-GP estimation is used, the weighting being motivated from the differing number of observations behind the GP-specific means. The GP density has a statistically significant effect on demand, and its coefficient has the expected negative sign (Table 8, Panel A). Second, the estimated coefficient of ALPHAHAT is negative and statistically significant – supporting the intuition that increased

Table 8: Model A: Demand equation: Between GP estimates

No. of obs.=3251. No. of GPs=484. Obs. per GP.: min=1, mean=6.7, max=7

A. Demand measure: No. of first rankings, DEMAND

Regressor	Estimate	Std.Err.
GPDENS1	-950.3971	445.4716*
MARRIEDGP	144.8451	48.7506**
SPECGEN	158.4471	54.4290**
SPECCOM	-156.025	90.4873
SPECOTH	-159.9245	125.6506
ALPHAHAT	-66.018	19.1214**
IMMIGRGP	-114.8588	104.0845
FEMALEGP	-89.3546	48.5764
AGEGP	82.8435	29.0487**
AGEGPSQ	-0.8065	0.3023**
CONST	-545.2106	824.6051
R^2		
within	0.0000	
between	0.1403	
overall	0.1372	
$F(10, 473)$	7.72	
p-value	0.0000	

B. Demand measure: No. of GP density weighted first rankings, DEMAND1

Regressor	Estimate	Std.Err.
MARRIEDGP	124.3429	42.1185**
SPECGEN	140.2252	46.9876**
SPECCOM	-136.4362	78.0651
SPECOTH	-134.8392	108.5775
ALPHAHAT	-56.6003	16.5214**
IMMIGRGP	-95.0004	89.9192
FEMALEGP	-75.4430	41.9200
AGEGP	72.4525	24.9295**
AGEGPSQ	-0.7066	0.2595**
CONST	-1204.6690	586.5191*
R^2		
within	0.0000	
between	0.1298	
overall	0.1265	
$F(9, 474)$	7.86	
p-value	0.0000	

∗) Significantly $\neq 0$ at the 5 % level (two-tailed test)
∗∗) Significantly $\neq 0$ at the 1 % level (two-tailed test)

quality induces increased demand facing the GP. Furthermore, the estimated effect of MARRIEDGP and SPECGEN indicate that being married and being a specialist in general medicine contribute, *ceteris paribus*, to a higher market demand. While FEMALEGP is not statistically significant, AGEGP comes out with a positive and statistically significant effect. The latter results may be explained by the fact that a GP's age is correlated with the number of years in GP practice, and that a GP who has been practicing for a long time may be included in the choice set of a larger proportion of the consumers in the market. This mechanism, however, is not explicitly accounted for in our theoretical model. It would have been possible to capture it by introducing heterogeneous groups of consumers with different choice sets within the same market, and this can be done simply by furnishing the parameter N with a group subscript.

As to the signs of the effects as well as their significance, the results in Panel B are very similar to those in Panel A. Being married and being a specialist in general medicine are both estimated to have a positive effect according to this model as well, and again, AGEGP comes out with a significantly positive coefficient.

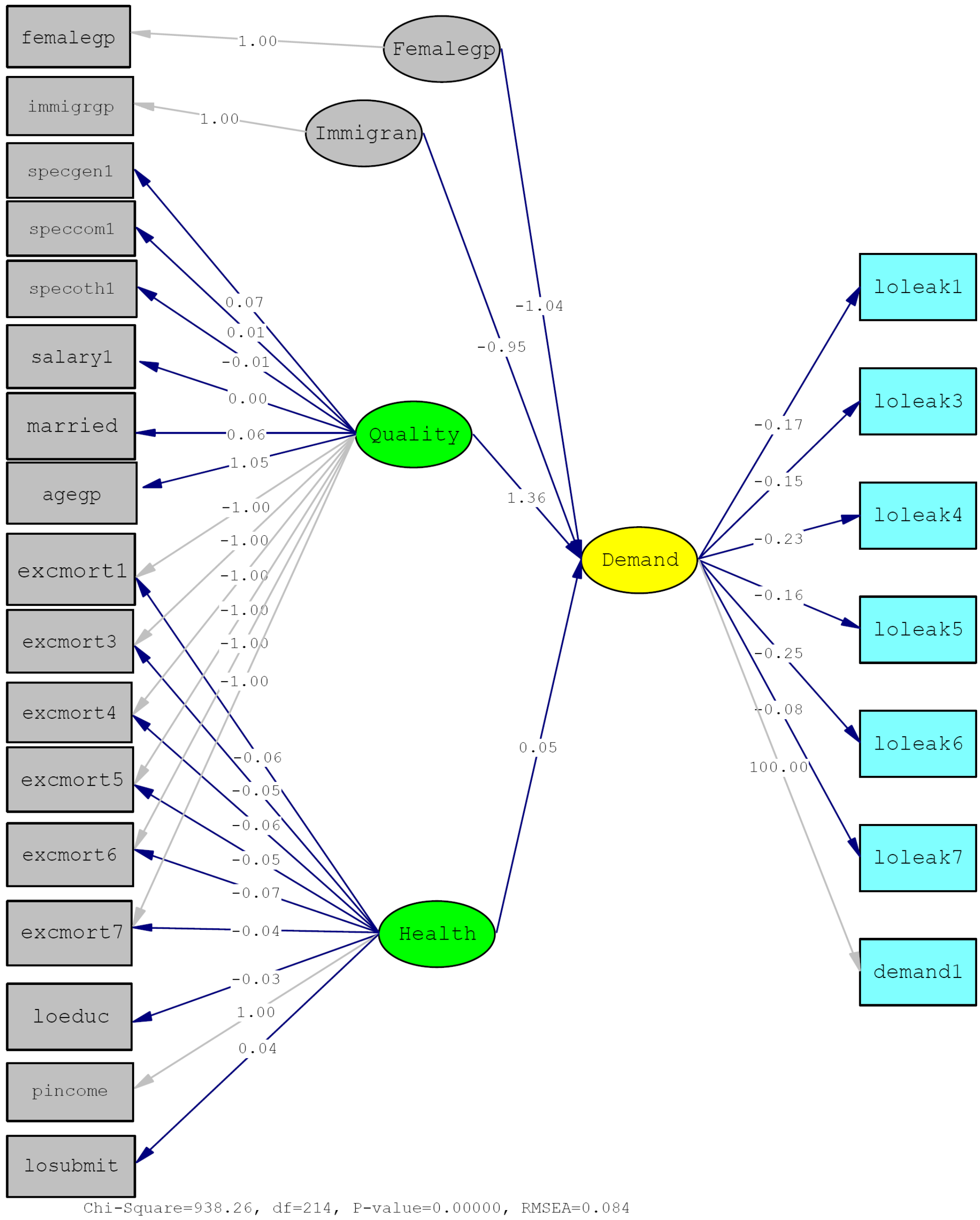
Results for Model B

The path diagram produced by the LISREL program, given in Figure 2, is a useful starting point for the description of the approach. As explained in Section 3, the LISREL model consists of two parts: the measurement equations and the structural equations, representing the relationship between the latent variables. In the sequel, we will follow the conventional LISREL notation, letting latent variables be indicated by names having capitalized first letter, whereas observable variables are written without capitalized first letters.[9] The measurement model specifies how our unobserved, latent variables *Quality*, *Health* and *Demand* are indicated by observed variables; cf. equations (16) and (15).

By modeling the demand facing the individual GP as a latent variable we are able to utilize information on the rate at which patients leave the GP's list in order to join a competitor's list. This approach thus takes into account how the demand facing the GPs has developed after the introduction of the General Practitioner Scheme. In our case the measurement model consists of two parts. The measurement equations for the exogenous latent variables, in the LISREL notation in Section 3 referred to as ξ-variables, are henceforth referred to as the *X-measurement model*. The measurement equations for the dependent latent variable, in LISREL notation referred to as an η-variable, is henceforth referred to as the *Y-measurement model*. When interpreting the approach and the results below, it should be recalled that the panel structure of the data – including the repeated observations of patients leaving the GP's list as well as of the excess mortality rates at the level of the GP – is essential for obtaining the inference we want to make, as the three latent variables in focus on Model B, *Quality*, *Health* and *Demand*, are all time

[9]We denote the variables Femalegp and Immigrant *as if* they were latent. See note 1.

Figure 2: LISREL Path diagram, Model B



invariant.[10] In total, Model B comprises 25 equations that are simultaneously estimated by the Maximum Likelihood (ML) method. We start by presenting the results from

[10] As explained in Section 4, the observations from period 2 and 3 are linearly dependent by construction. Therefore observations from period 2 are excluded from the data set when estimating Model B.

estimation of the structural equation before presenting the results from the estimation of the equations in the two measurement models.

The structural model (demand equation)

The results from estimation of the structural equation, corresponding to Equation (14) in Section 3, are given in Table 9. Our first major finding is that the estimated effect

TABLE 9: MODEL B: EQUATION FOR LATENT DEMAND. COEFFICIENT ESTIMATES

No. of obs.=484.

Regressor	Estimate	Std.Err.
Quality	1.3583	0.1617**
Health	0.0456	0.0177**
Femalegp	-1.0422	0.3881**
Immigrant	-0.9545	0.8356

*) Significantly ≠ 0 at the 5 % level (two-tailed test)
**) Significantly ≠ 0 at the 1 % level (two-tailed test)

of *Quality* on *Demand* is statistically significant and has the expected sign. Since the measurement scale of *Demand* is the number of first rankings measured in hundreds, the interpretation of the coefficient estimate of 1.36 obtained is that a marginal increase in *Quality*, equivalent to a marginal reduction in the excess mortality rates, increases *Demand* by 136 persons. Second, we find that the estimated coefficient of *Health* also is statistically significant with a positive sign, which implies that the more healthy a GP's patients are, for a given level of quality, the higher demand will be experienced. Also, the estimated coefficient of Femalegp is negative and statistically significant.

We remember from Section 2 that heterogeneity of patients ability or willingness to collect and process available quality information may induce selection mechanisms resulting in systematic differences in morbidity and mortality between GPs with different levels of quality. We have no prior knowledge of the direction in which this selection mechanism may go. However, failing to control for differing aggregate health status of listed patients is likely to result in a simultaneity bias and/or an excluded variable bias when estimating the effect of quality on demand. Our approach separates the effect of quality on excess mortalities from the effect of health at the level of the GP through the exclusion restrictions imposed on the measurement equations for Quality and Health: six variables describing the GP are included in the measurement equations for *Quality*, but excluded from the measurement equations for *Health*, while three variables are included in the measurement equations for *Health* but excluded from the measurement equations for *Quality*.

The estimated covariance matrix of the latent variables is given in Table 10. Our results indicate that there is a negative association between *Quality* and *Health*. What we observe is thus consistent with a situation where a selection mechanism exists such that

Table 10: Model B: Variance and correlation matrix of latent variables

Variances along the main diagonal, correlation coefficients below the diagonal. No. of obs.=484.

	Demand	Quality	Health
Demand	15.6585		
Quality	0.4358	2.4438	
Health	0.0123	-0.4064	460.2335

GPs with low quality of services are endowed with a patient stock with a better health, as compared to GPs with higher quality of services. One may argue that Model B does not reveal the effect of the GPs' quality on the initial health state of listed patients, as it is set up to measure the effect of quality when controlling for initial health status of patients showing between GP variation. To address the latter issue, and for the purpose of conducting a robustness check of our main findings, we have additionally estimated an alternative LISREL model where all latent variables enter as η variables, *i.e.*, as formally endogenous – *Quality* and *Health* being exogenous latent variables in Model B. Such a model setup allows the estimation of the marginal effect of GP's quality on the initial state of health. The results from this model, denoted as *Model C*, are reported in Appendix A. The results confirm the results from Model B, that latent quality affects latent demand positively. The numerical size of the effect is somewhat smaller, however. The most important single result from estimation of Model C is that quality is found not to have significant effect on the aggregate health status of the GP's listed patients.

The X-measurement model

The left hand side of Figure 2 describes the relations in the X-measurement model. Arrows indicate the relation between the observable variables and the ξ variables, and the reported numbers on each arrow corresponds to the estimated or fixed *'factor loading'*.Since latent variables by definition do not have a scale of measurement, the scale of the latent variables are defined by fixing one or more of the factor loadings. In order to make the results from LISREL estimation comparable to the results of the previous model we have scaled the ξ variable *Quality* to be measured in (negative) units of *per thousand excess mortality rates*. By fixing the factor loading of *pincome* on *Health*, we have scaled the ξ variable *Health* to be measured in units of *thousand Norwegian kroner*. Note that the arrows representing the fixed factor loadings are shaded compared to the massive arrows representing estimated factor loadings. Although formally treated as latent variables, the ξ-variables *Femalegp* and *Immigrgp* are identical to their observable counterparts *femalegp* and *immigrgp*. This is done by fixing the factor loadings to unity and fixing the error terms in the two corresponding regression equations to zero. The interpretation is that the GP's gender and country of birth is expected to affect the demand directly, without being indicators of either quality or health status of listed patients. The excess mortality rate from each period is entered as indicators of GP quality and as indicators of the aggregate measure of health status of listed patients. Other variables included as indicators of quality are dummy variables indicating GP specialization, a dummy variable indicating whether or not the GP is remunerated by means of a fixed salary, a dummy variable indicating whether or not the GP is married and the GP's age. The proportion of listed individuals without finished high school and the proportion of listed individuals who submitted the entry form in 2001 are included as indicators of the aggregate health measure. The results from estimation of the equations in the X-measurement model, corresponding to Equation (16) in Section 3, are given in Table 11. The estimated factor loading of *Quality* on *married*, *agegp* and *specgen1* are

Table 11: Model B: X-measurement equations

No. of obs.=484.

ξ variable	Quality		Health	
Dep var	Estimate	Std.Err.	Estimate	Std.Err.
excmort1	-1	(Fixed)	-0.0614	0.0127**
excmort3	-1	(Fixed)	-0.0498	0.0121**
excmort4	-1	(Fixed)	-0.0581	0.0123**
excmort5	-1	(Fixed)	-0.0458	0.0120**
excmort6	-1	(Fixed)	-0.0673	0.0127**
excmort7	-1	(Fixed)	-0.0445	0.0119**
loeduc	0	(Fixed)	-0.0283	0.0025**
losubmit	0	(Fixed)	0.0374	0.0158*
pincome	0	(Fixed)	1	(Fixed)
femalegp	0	(Fixed)	0	(Fixed)
immigrgp	0	(Fixed)	0	(Fixed)
married	0.0558	0.0161**	0	(Fixed)
agegp	1.0471	0.2692**	0	(Fixed)
salary1	0.0033	0.0079	0	(Fixed)
specgen1	0.0658	0.0174**	0	(Fixed)
specoth1	-0.0056	0.0064	0	(Fixed)
speccom1	0.0056	0.0086	0	(Fixed)

*) Significantly $\neq 0$ at the 5 % level (two-tailed test)

**) Significantly $\neq 0$ at the 1 % level (two-tailed test)

statistically significant, and the interpretation is that married GPs, GPs that are specialists in general medicine is associated with higher quality and that there is a positive relation between GP age and provided quality. We see that all the factor loadings of *Health* is statistically significant, and we see that excess mortality rates are negatively related to *Health*. We see that *Health* is positively related to the share of individuals who submitted the entry form in 2001, and that our latent aggregate health measure is negatively related to the proportion of listed individuals with short schooling, as expected. The estimated covariance matrix of error terms in the X-measurement model is given in Table 12. We have added some restrictions on the correlation of error terms from different x-regressions. With some exceptions error terms from different regressions are uncorrelated. Error terms in regressions on the quality indicators related to GP

Table 12: Model B: Error covariance matrix of X-measurement equations

	excmort1	excmort3	excmort4	excmort5	excmort6	excmort7
excmort1	0.7120					
excmort3	0.1371	0.6403				
excmort4	–	0.0727	0.6520			
excmort5	–	–	-0.0077	0.6397		
excmort6	–	–	–	-0.1060	0.6625	
excmort7	–	–	–	–	0.0046	0.6026

	loeduc	pincome	losubmit	salary1	married	agegp	specgen1	specoth1	speccom1
loeduc	0.2931								
pincome	–	0.1951							
losubmit	–	–	0.9866						
salary1	–	–	–	0.9995					
married	–	–	–	–	0.9648				
agegp	–	–	–	–	–	0.9550			
specgen1	–	–	–	–	–	–	0.9575		
specoth1	–	–	–	–	–	–	-0.0043	0.9979	
speccom1	–	–	–	–	–	–	0.1463	-0.0082	0.9988

specialization, *specgen1*, *specoth1* and *speccom1* are allowed to be correlated. Further, error terms on in the excess mortality rates regression from period t are allowed to be correlated with error terms in period $t-1$.

The Y-measurement model

The right hand side of Figure 2 describes the relations in the Y-measurement model. Here the observable variable *demand1* enters the model as an indicator of the η variable *Demand*. The variable *demand1* is identical to the GP density weighted number of first rankings applied in estimation of Model A. The factor loading of *Demand* on *demand1* is fixed to 100 and the η variable *Demand* is thus measured in units of *hundred first rankings*.

Table 13: Model B: Y-measurement equations

No. of obs.=484.

η variable	Demand	
Dep var	Estimate	Std.Err.
loleak1	-0.1671	0.0165**
loleak3	-0.1542	0.0167**
loleak4	-0.2337	0.0759**
loleak5	-0.1600	0.0465**
loleak6	-0.2541	0.0899*
loleak7	-0.0850	0.0494
demand1	100	(Fixed)

*) Significantly $\neq 0$ at the 5 % level (two-tailed test)

**) Significantly $\neq 0$ at the 1 % level (two-tailed test)

The log-odds-ratio of the proportion of listed persons leaving the list in period t, $loleak_t$, are entered as indicators of *Demand*. The results from estimation of the equations in the Y-measurement model, corresponding to Equation (15) in Section 3, are given in Table 13. We see that all the estimated factor loadings are statistically significant, and that the factor loadings of the $loleak_t$ variables and the factor loading of *Demand* on *demand1* are of opposite sign, as one should expect. The estimated covariance matrix of error terms in the Y-measurement model is given in Table 14. No restrictions are specified with regard to the structure of this matrix.

Table 14: Model B: Error covariance matrix of Y-measurement equations

	loleak1	loleak3	loleak4	loleak5	loleak6	loleak7	demand1
loleak1	0.4024						
loleak3	0.2469	0.4318					
loleak4	0.1819	0.2454	0.9014				
loleak5	0.2058	0.2489	0.5517	0.8818			
loleak6	0.1312	0.1983	0.3027	0.4874	0.9123		
loleak7	0.2182	0.2768	0.1667	0.2079	0.1671	0.9673	
demand1	-0.0235	0.0049	-0.0101	-0.0394	0.0359	-0.0858	0.2182

6 Discussion and conclusion

Our analysis supports the hypothesis that the demand responds to the individual quality of the GP. Even though our two different approaches to quantifying this effect rely on different assumptions and different methods, it may be argued that our two sets of results pointing in the same direction, contribute to strengthening this conclusion. Our results indicate that a marginal quality increase equivalent to a reduction of the mortality rate of one per thousand – being the implicit measurement scale of our quality indicator – increases the demand facing the GP by 57 persons according to Model A and by 136 persons according to Model B. In interpreting this finding one should recall that the empirical results for Model A (presented in Section 5), rely on Version 2 from Section 3, which imposes an asymmetry in the way quality affects latent GP-specific heterogeneity in the two equations. The estimators used under Model A, although enjoying consistency under Version 2, are inconsistent if the less restrictive Version 1 is valid.

The latter has been our primary motivation for also modeling latent heterogeneity within a LISREL framework. Our LISREL model, Model B, may be considered less restrictive than Model A and also enables us to address more appropriately the issue that the aggregate health status of listed patients, also considered an unobservable variable, is likely to be related to the mortality rates at the level of the GP.

We believe that this econometric modeling tool has a wider application in assessing the impact of quality of health care providers on demand for health services than the one presented here. Our LISREL model separates the effect of quality on outcome measures from the effect of patient health at the provider level through the exclusion restrictions imposed on the measurement equations for the latent variables Quality and Health. An idea for future research is to apply this modelling and estimation strategy to matched hospital-patient data, in order to measure the impact of hospital quality on demand. An important question from the point of view of policy implications is: how *strong* is the effect of quality on demand? In the context of regression models answers to such questions are often provided by means of appropriate elasticities. Due to the fact that the mean value of latent variables are undefined, the elasticities of interest, such as *Quality elasticity of demand* are also undefined. However, to get a 'metric' for assessing the magnitude of the effect we may compare the effect of quality on demand with the effect of the observable GP specific variables like the GP gender dummy and the dummy for whether or not the GP is born in a non-scandinavian country. In case of model B, this can be achieved by utilizing the standardized solution of our LISREL estimation in which all effects are scaled in terms of their standard deviation. The standardized solution thus obtained indicates that the effect of quality on demand is more than four times the effect of the GP gender dummy and ten times the effect of non-Scandinavian GP dummy. It is also illuminating to interpret the latter effect in relation to the standard deviation of the Quality variable σ_Q (confer Table 10). We then find that an increase in quality equivalent to a reduction of the patient mortality rate by one per thousand corresponds to a change in Quality equal to $0.6397\sigma_Q$. An increase of this order of magnitude results in an increase in Demand corresponding to 16% relative to the global mean of DEMAND, given in table Table 3.

References

Arrow, K.E. 1963. Uncertainty and the welfare economics of medical care. *American Economic Review* **53**, 941-973.

Aigner, D.J., Hsiao, C., Kapteyn, A., and Wansbeek, T. (1984): Latent Variable Models in Econometrics, in Griliches, Z. and M.D. Intriligator (eds.): *Handbook of econometrics, vol. II*, Amsterdam: North-Holland, 1321-1393.

Burns, L.R. and Wholey, D.R. (1992): The Impact of Physician Characteristics in Conditional Choice Models for Hospital Care. *Journal of Health Economics* **11**, 43-62.

Chirikos, Thomas N.(1992): Quality Competition in Local Hospital Markets: Some Econometric Evidence from the Period 1982-1988. *Social Science & Medicine* **34**, 1011-1021.

Dranove, D. and Satterthwaite, M.A., (2000): The Industrial Organization Of Health Care Markets, in A.J. Culyer and J.P. Newhouse (eds.): *Handbook of Health Economics, vol.1.* Amsterdam: Elsevier Science, pp. 1093-1139.

Gaynor, M. (2006): What Do We Know About Competition And Quality In Health Care Markets? NBER working paper No. w12301. National Bureau of Economic Research.

Godager, G.(2008): Consumer Choice and Patient Selection in Health Care: Evidence From the Reform in Norwegian General Practice. Unpublished working Paper.

Godager, G., and Lurås, H. (2007): Dual Job Holding General Practitioners: The Effect of Patient Shortage. Unpublished working Paper.

Goldberger, A.S. (1972): Structural Equation Methods in the Social Sciences. *Econometrica* **40**, 979-1001

Howard, D.H. (2005): Quality and Consumer Choice in Healthcare: Evidence from Kidney Transplantation. *Topics in Economic Analysis & Policy.*

Hsiao, C. (2003): *Analysis of Panel Data*, Second edition. Cambridge: Cambridge University Press.

Jöreskog, K.G. (1977): Structural Equation Models in the Social Sciences: Specification, Estimation and Testing, in P.R. Krishnaiah (ed.): *Applications of statistics.* Amsterdam: North Holland, pp. 265-287.

Jöreskog, K.G., D. Sörbom, D., du Toit, D. and du Toit. M. (2000): *LISREL 8: New Statistical Features.* Lincolnwood, IL: Scientific Software International.

Lantz *et al.* (1998): Socioeconomic Factors, Health Behaviors, and Mortality. Results From a Nationally Representative Prospective Study of US Adults. *The Journal of the American Medical Association* **279**,1703-1708

Lee, L.F., and Griffiths, W.E. (1979): The Prior Likelihood and Best Unbiased Prediction in Stochastic Coefficient Linear Models. Working Papers in Econometrics and Applied Statistics No. 1. Department of Economic Statistics, University of New England.

Luft, H.S. *et al.* (1990): Does Quality Influence Choice of Hospital? *Journal of the American Medical Association* **263**, 2899-2906.

Lurås, H., Paulsen, B. and Finnvold, J.E. (2003): De fleste fikk den de ville ha. (In Norwegian.) *Samfunnsspeilet* 05/2003.

Ma, C.T.A. (1994): Health Care Payment Systems: Cost and Quality Incentives. *Journal of Economics and Management Strategy* **3**, 93-112.

Papas G., *et al.* (1993): The Increasing Disparity in Mortality Between Socioeconomic Groups in the United States, 1960 and 1986. *The New England Journal of Medicine* **329**, 103-109.

Statistics Norway (2005a): *Dødelighetstabeller 2004.* (In Norwegian.) Oslo: Statistics Norway. `http://www.ssb.no/emner/02/02/10/dode/arkiv/tab-2005-04-28-05.html`.

Statistics Norway (2005b): *Folkemengde, etter kjønn og ettårig alder.* (In Norwegian.) Oslo: Statistics Norway. `http//:statbank.ssb.no/statistikkbanken`.

Statistics Norway (2006a): *Indeks for levekårsproblemer og delindekser for enkeltindikatorer* `http://www.ssb.no/vis/emner/03/hjulet/tabell-04.html`

Statistics Norway (2006b): *Alders- og kjønnsstandardisert dødelighet. Gjennomsnitt for årene 2001-2005. Dødelighet per 100 000 innbyggere. Konfidensintervall for dødelighetsraten* `http://www.ssb.no/emner/03/hjulet/tabell-08.1.html`

Appendix A: Excess mortality exogenous in demand equation

The estimation results when excess mortality rates at the level of the GP is entered as an exogenous explanatory variable are presented in Table A.1, panel A based on the dependent variable DEMAND, panel B on the dependent variable DEMAND1. The only difference between the results presented in Table A.1 and the results presented in Table 8 is that ALPHAHAT is replaced by EXCMORT. The latter variable is now assumed to be a valid proxy for quality, and is assumed to be uncorrelated with the error components. We see that the estimated effect of GP density is statistically significant and has the expected sign. We see that being married and being a specialist in general medicine have a positive and statistically significant effect on demand. We also see that AGEGP has a statistically significant effect. The effect of EXCMORT is negative and statistically significant. An interpretation may be that a marginal increase in quality, measured in terms of a reduction in the excess mortality rates, increase the demand facing the GP. The magnitude of the estimated coefficient indicates that a marginal quality increase equivalent to a reduction of the mortality rate of one per thousand, increases the demand facing the GP by 90 persons (Panel A). We note that the estimated coefficients on EXCMORT are larger in absolute value as compared to the coefficients on ALPHAHAT in Table 8. An interpretation of this result, while recalling that the dimension of the coefficients are the same, is that by representing quality by EXCMORT, as in Table A.1, we disregard the variation in health status between GPs. We cannot assess whether high excess mortality is a result of low quality or bad health status of the listed patients.

Table A.1: Model A: Between GP estimates of demand equation.
No. of obs.=3251. No. of GPs=484. Obs. per GP.: min=1, mean=6.7, max=7

A. Demand measure: No. of first rankings, DEMAND

Regressor	Estimate	Std.Err.
GPDENS1	-922.5879	424.9770*
MARRIEDGP	113.9111	46.7327*
SPECGEN	150.8321	51.9366**
SPECCOM	-156.9416	86.3357
SPECOTH	-182.0114	119.8388
EXCMORT	-90.6113	11.7378**
IMMIGRGP	-119.6318	99.2157
FEMALEGP	-133.5571	46.6809**
AGEGP	72.0700	27.7608**
AGEGPSQ	-0.7207	0.2887*
CONST	-167.7292	788.7412
R^2		
within	0.0000	
between	0.2172	
overall	0.1408	
$F(10, 473)$	13.12	
p-value	0.0000	

B. GP density-weighted no. of first rankings, DEMAND1

Regressor	Estimate	Std.Err.
MARRIEDGP	97.3718	40.3584*
SPECGEN	133.7817	44.8165**
SPECCOM	-137.5563	74.4515
SPECOTH	-153.7605	103.5101
EXCMORT	-78.5167	10.1394**
IMMIGRGP	-99.4600	85.6771
FEMALEGP	-113.5944	40.2677**
AGEGP	62.9205	23.8148**
AGEGPSQ	-0.6302	0.2478*
CONST	-851.8743	561.5582*
R^2		
within	0.0000	
between	0.2084	
overall	0.1323	
$F(9, 474)$	13.87	
p-value	0.0000	

$*$) Significantly $\neq 0$ at the 5 % level (two-tailed test)
$**$) Significantly $\neq 0$ at the 1 % level (two-tailed test)

Appendix B: Alternative LISREL model. Model C

Figure B.1: LISREL Path diagram, Model C

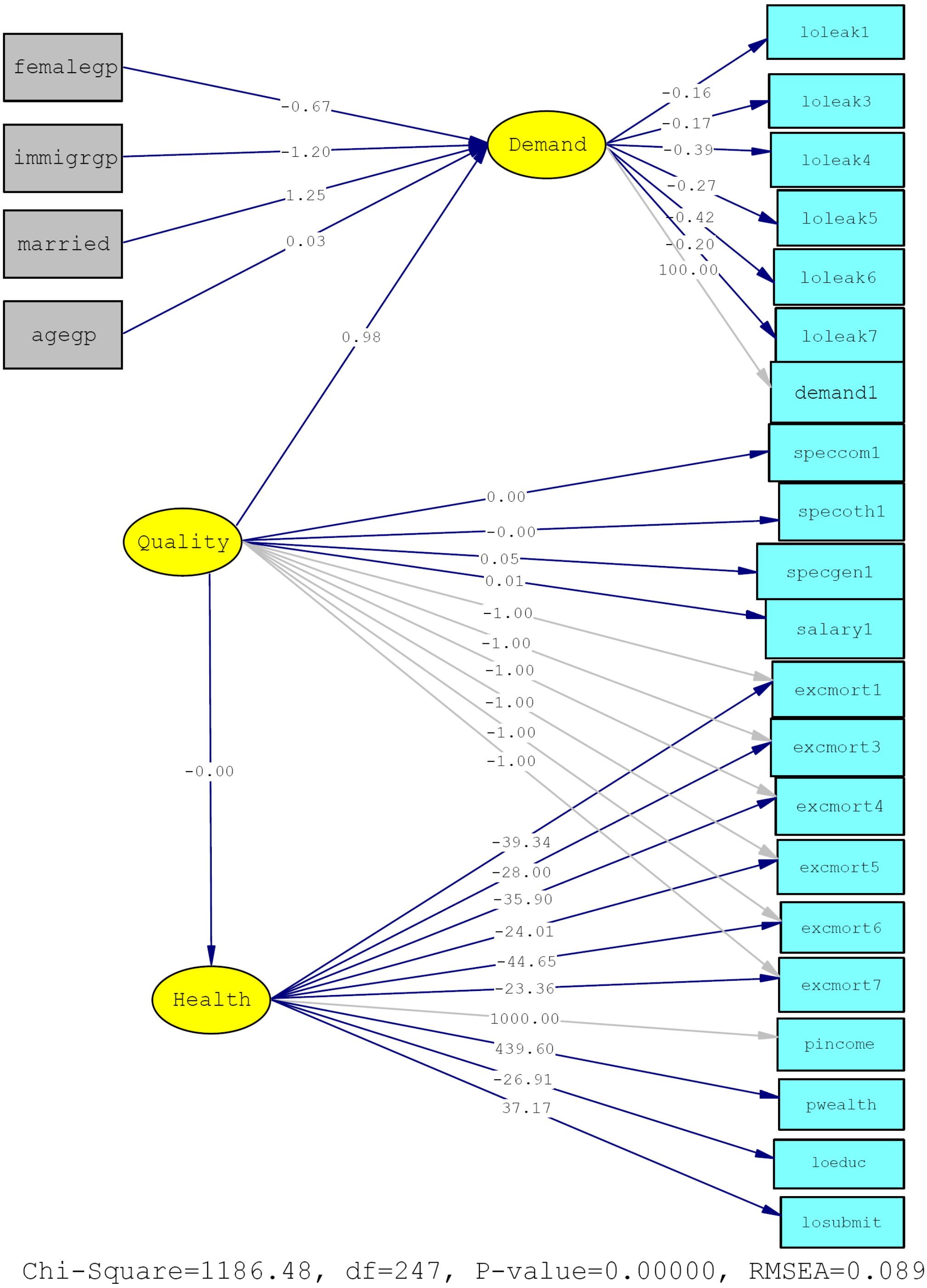


Table B.1: Model C: Equations for latent dependent variables.

Coefficient estimates. No. of obs.=484.

	Regressors									
	Quality		immigrant		married		agegp		femalegp	
Dep var	Est.	Std.Err.	Est.	Std.Err.	Est.	Std.Err.	Est.	Std.Err.	Est.	Std.Err.
Demand	0.9759	0.1487**	-1.2020	0.7935	1.2539	0.3748**	0.0293	0.0226	-0.6711	0.3733
Health	-0.0019	0.0019	0	(Fixed)	0	(Fixed)	0	(Fixed)	0	(Fixed)

*) Significantly $\neq 0$ at the 5 % level (two-tailed test)
**) Significantly $\neq 0$ at the 1 % level (two-tailed test)

Table B.2: Model C: Variance and correlation matrix of latent variables

Variances along the main diagonal, correlation coefficients below the diagonal. No. of obs.=484.

	Quality	Demand	Health
Quality	2.1382		
Demand	0.3426	17.3484	
Health	-0.1268	-0.0434	0.0005

Table B.3: Model C: Y-measurement equation

No. of obs.=484.

	Regressors					
	Quality		Demand		Health	
DEP VAR	Estimate	Std.Err.	Estimate	Std.Err.	Estimate	Std.Err.
demand1	0	(Fixed)	100.0000	(Fixed)	0	(Fixed)
loleak1	0	(Fixed)	-0.1576	(0.0178)**	0	(Fixed)
loleak3	0	(Fixed)	-0.1650	(0.0186)**	0	(Fixed)
loleak4	0	(Fixed)	-0.3930	(0.0598)**	0	(Fixed)
loleak5	0	(Fixed)	-0.2673	(0.0392)**	0	(Fixed)
loleak6	0	(Fixed)	-0.4151	(0.0663)**	0	(Fixed)
loleak7	0	(Fixed)	-0.2038	(0.0339)**	0	(Fixed)
speccom1	0.0016	(0.0092)	0	(Fixed)	0	(Fixed)
specoth1	-0.0046	(0.0069)	0	(Fixed)	0	(Fixed)
specgen1	0.0452	(0.0181)**	0	(Fixed)	0	(Fixed)
salary1	0.0055	(0.0085)	0	(Fixed)	0	(Fixed)
excmort1	-1.0000	(Fixed)	0	(Fixed)	-39.3361	(9.9107)**
excmort3	-1.0000	(Fixed)	0	(Fixed)	-28.0050	(9.1767)**
excmort4	-1.0000	(Fixed)	0	(Fixed)	-35.8985	(9.4614)**
excmort5	-1.0000	(Fixed)	0	(Fixed)	-24.0063	(9.1388)**
excmort6	-1.0000	(Fixed)	0	(Fixed)	-44.6527	(9.8519)**
excmort7	-1.0000	(Fixed)	0	(Fixed)	-23.3650	(9.0128)**
pincome	0	(Fixed)	0	(Fixed)	1000.0000	(Fixed)
pwealth	0	(Fixed)	0	(Fixed)	439.5965	(174.5656)**
loeduc	0	(Fixed)	0	(Fixed)	-26.9137	(2.5591)**
losubmit	0	(Fixed)	0	(Fixed)	37.1699	(15.3537)**

*) Significantly $\neq 0$ at the 5 % level (two-tailed test)
**) Significantly $\neq 0$ at the 1 % level (two-tailed test)

Table B.4: Model C: Error covariance matrix in subsystem for Y-measurement

	demand1	loleak1	loleak3	loleak4	loleak5	loleak6	loleak7
demand1	21999.3356						
loleak1	9.7168	0.2958					
loleak3	51.4310	0.1182	0.1773				
loleak4	283.0998	–	0.0333	5.9740			
loleak5	174.9324	–	–	1.5528	2.0378		
loleak6	339.9960	–	–	–	1.4497	8.5124	
loleak7	153.7280	–	–	–	–	-0.1123	2.7219

	excmort1	excmort3	excmort4	excmort5	excmort6	excmort7
excmort1	6.2133					
excmort3	1.0092	3.9033				
excmort4	–	0.3908	4.4066			
excmort5	–	–	-0.1208	3.8241		
excmort6	–	–	–	-0.8207	5.1367	
excmort7	–	–	–	–	-0.0872	3.2198

	speccom1	specoth1	specgen1	salary1	pincome	pwealth	loeduc	losubmit
speccom1	0.0635							
specoth1	-0.0005	0.0358						
specgen1	0.0191	-0.0009	0.2444					
salary1	–	–	–	0.0544				
pincome	–	–		–	87.6548			
pwealth	–	–	–	–	–	6131.2983		
loeduc	–	–	–	–	–	–	0.1696	
losubmit	–	–	–	–	–	–	–	47.5383