Centre for Household, Income, Labour and Demographic economics

Working Papers

Rationing the Public Provision of Healthcare in the Presence of Private Supplements: Evidence from the Italian NHS

D. Fabbri and C. Monfardini

ChilD n. 21/2006

e-mail: de-child@unito.it
Web site: http://www.child-centre.it

RATIONING THE PUBLIC PROVISION OF HEALTHCARE IN THE PRESENCE OF PRIVATE SUPPLEMENTS:

EVIDENCE FROM THE ITALIAN NHS

Daniele Fabbri*

Department of Economics, University of Bologna, Italy and CHILD

Chiara Monfardini

Department of Economics, University of Bologna, Italy and CHILD

This version: November, 2006

ABSTRACT

In this paper we assess the relative effectiveness of user charges and administrative waiting times as a tool for rationing public healthcare in Italy. We measure demand elasticities by estimating a simultaneous equation model of GP primary care visits, public specialist consultations and private specialist consultations, as if they were part of an incomplete system of demand. We find that own price elasticity of the demand for public specialist consultation is about -0.3, while administrative waiting time plays a less important role. No substitution exists between the demand for public and private specialists, so that user charges act as a net deterrent for over-consumption. The public provision of healthcare does not induce the wealthy to opt out. Moreover our evidence suggests that user charges and waiting lists do not serve redistributive purposes.

Keywords: healthcare demand elasticities, user charges, waiting lists, multivariate count data model

JEL: C34, C35, C51, D12, I11

Acknowledgments: We are grateful to Thomas Buchmueller, Miguel Gouveia, Rowena Pecchenino, Federico Perali and partecipants to IIPF-2004, IHEA-2005 conference, for helpful suggestions. We are especially indebted to Pravin Trivedi for stimulating discussions and helpful comments. Usual disclaimers apply.

* Corresponding author: Daniele Fabbri, Dipartimento di Scienze Economiche, Università di Bologna, Piazza Scaravilli 2 - 40126 Bologna, Italy. E-mail: dfabbri@economia.unibo.it

1 INTRODUCTION

Increasing healthcare expenditure is a major policy issue in most developed countries. Despite an attenuation thanks to the lesser disability of marginal survivors, population aging and technological advances will almost certainly push healthcare expenditure far above existing ratios to GDP (Cutler and Sheiner, 1998). In this turmoil, public reforms are repeatedly urged to cap spiralling expenditure. Over the past two decades demand-side cost-sharing for medical care has been widely adopted for this purpose in OECD countries (Docteur and Oxley, 2003). Most such measures have reduced eligibility and improved the targeting of public healthcare programs, while also raising user charges.

Regardless of differences due to the regulation of health insurance markets and supply-side cost-sharing arrangements, in many countries demand-side cost-sharing comes on top of rationing by means of waiting lists.[1] This is the rule in all public healthcare programs. The literature on waiting lists and waiting times [surveyed by Cullis et al. (2000)] is abundant, but quantitative evidence on the effect of administrative waiting times on demand is scant. The empirical studies are based on aggregate data, refer to the British NHS, and focus on elective surgery [see Martin and Smith (1999), Gravelle et al. (2000, 2002)]. This literature suffers from two major limitations. First, it ignores the role of patient co-payment, which is assumed to be negligible. Second, it fails to allow for the possibility of substitution of private for public healthcare, even though Cutler and Gruber (1996) and Besley et al. (1999) provide evidence that modifying the extent and access to public provision triggers adjustment in markets for supplements and substitutes.

This paper offers the first empirical assessment of the relative and joint effectiveness of user charges and administrative waiting time in curbing the demand for public physicians' care, while accounting for the presence of imperfect substitutes available in the market. We conduct a microeconometric analysis of the demand for physicians' care in Italy and evaluate demand elasticities, gaining insight into how much rationing is attainable through waiting lists and increased co-payments. We work on a

[1] According to Siciliani and Hurst (2004) waiting times for elective surgery are a "serious health policy issue in 12 countries involved in the OECD Waiting Time project". Schoen et al. (2005) report that among the countries surveyed in the 2005 Commonwealth Fund International Health Policy Survey, long waits for specialist appointments are a relevant issue in Australia, Canada, UK and New Zealand. The persistence of waiting lists in OECD countries other than the US, led some [see Anderson, et. al. (2005)] to consider them as a possible explanation for the US health spending differential.

large dataset coming from a national household survey conducted every 5 years by the Italian National Institute of Statistics (ISTAT), using the survey conducted between July 1999 and June 2000 (the most recent available). The demand for physician care in Italy takes three forms: primary care and public and private specialist consultations. Each of these components is proxied by the number of visits sought by the individual in the four weeks before the interview. By exploiting geographical variations in waiting times and fees, we can measure own and cross elasticities for each demand component.

Our approach is close to the literature that examined the impact of non-monetary factors on the demand for healthcare [see Acton (1975), Colle and Grossman (1978), Goldman and Grossman (1978), Coffey (1983)]. In those contributions monetary prices, jointly with physical waiting time, distance and travel costs, were properly considered in the estimation of a single demand equation performed on individual data. We extend the approach to the joint estimation of multiple healthcare consumption equations as if they were part of an incomplete system of demand (Hausman, 1981, Epstein, 1982, LaFrance, 1985). Our empirical strategy relies on the joint modelling of the three visit counts, in the spirit of a seemingly unrelated regression model approach, with a complete vector of prices and waiting times for the three alternatives included among the regressors of each equation.

The paper is organised as follows. The next section provides concise institutional background on the market for physician care in Italy. Section 3 presents the incomplete system of demand we estimate in our econometric analysis. Section 4 presents the econometric model. Section 5 describes the data; major results are presented in section 6 and discussed in section 7. Section 8 concludes.

2 SOME BACKGROUND ON PHYSICIAN CARE IN THE ITALIAN NHS

The Italian National Health Service, founded in 1978, is a universal system with comprehensive insurance and uniform healthcare for the entire population. It is financed mainly by general taxation. However, depending on a citizen's income, age and health condition, co-payments are also charged for drugs, out-patient treatment, some diagnostic and laboratory tests, and medical appliances. Every year funds are transferred from the central government to each Regional Health Authority (RHA) according to a capitation rule and then reallocated among approximately 200 Local Health Authorities (LHA). Within its budget, each LHAs are responsible for financing the healthcare consumption of the "enrolled" population, and is also partly responsible for healthcare production.

Primary care is provided by GPs, who have the status of independent, self-employed physicians, working for an LHA under a public contract. They are paid according to a capitation fee, determined by the number of people registered with each doctor. Under the contract GPs are expected to provide most primary care. They also act as gatekeepers for access to secondary services whose provision is refunded by the NHS, such as diagnostic checks, hospital admissions, and specialist visits. Although primary care GPs are given financial incentives to share clinic premises with their colleagues, they usually work in single practices. People may choose any physician, among those under contract for the LHA they reside in, as long as the physician's list has not reached the ceiling of 1800 enrollees.

The Italian NHS plays a major role in the market for specialist consultations, where public, closely regulated and mainly salaried specialists compete with private, less strictly regulated ones. Specialized NHS out-patient services, including visits, diagnostics and treatment, are provided either by the LHA or by accredited public and private facilities with which the LHA has agreements and contracts. People can access NHS specialist care only with their GP's authorization and referral. Once the GP has prescribed the visit or the treatment, the individual is free to choose any NHS-accredited provider, even one outside his LHA. A co-payment is required as an additional source of financing for the provision of specialist out-patient care and as a way of curbing consumption. The particular amount of co-payment is discretionary for each RHA up to a ceiling determined by national law. The ceilings are well below the market clearing level, so that queues of patients form, and in this way supply is rationed.

Because of waiting lists, co-payments and unsatisfactory quality, many patients seek care outside the NHS, resorting to the private market for specialist care. This market is quite well developed.[2] Private specialists are subject to an authorization based on minimum standard requirements, which turn out to be very loose indeed. Fees, quality and most other relevant features of these medical practices are then subject mainly to market forces.[3] It is generally true that for the kind of specialist visits we consider here, i.e. excluding out-patient treatment, the private alternative to NHS supply is of better quality and commensurably higher-priced.

In Italy the demand for medical consultations, in an episode of illness, is therefore divided among three main alternatives, which may be substitutes or complements depending on "local" circumstances. The pattern of substitution/complementarity is

[2] Almost 20% of specialist diagnostic procedures are performed by private providers.
[3] Some of the rules are laid down by professional self-regulatory associations, but again these are not particularly stringent.

unpredictable, a-priori, and is thus a task for empirical analysis. Since it strongly influences the effect of public user charges, it is crucial to consider GP visits, public and private specialist visits as if they were part of an (incomplete) demand system.

3 AN INCOMPLETE SYSTEM OF DEMAND MODEL FOR VISITS

Given this framework, let us sketch a model for the demand for physician care. Given the persistence of waiting lists as a major rationing device, we rely on a consumer behaviour representation as proposed by Lindsay and Feigenbaum (1984) and revisited by Martin and Smith (1999). The view taken in these works is that waiting time can be considered, in an equilibrium framework, as hassle cost reaching a level just high enough to align the demand for services with the supply. Administrative waiting time, unlike physical waiting time, clears the market by reducing the individual valuation of the expected health gains from a given treatment.

The model is as follows. Individuals consume a bundle of consumption goods $\boldsymbol{X}=(\boldsymbol{x}', x_{m+1})'$, with $\boldsymbol{X} \in R^{m+1}$. We set the good $m+1$ as the numeraire to normalize the other prices and individual income. The commodity vector $\boldsymbol{X}$ is then priced according to a conform vector $\boldsymbol{Q}=(\boldsymbol{q}', 1)'$. For convenience, we let the numeraire stand for the bundle of food goods, so that $\boldsymbol{q}$ comprises all the other goods.

In case of need, the patient can demand one of the three different types of physician visit in his local market: GP primary care and public and private specialist consultations. In the local market he observes prices and waiting times to access public and private specialist consultations in the immediate past. Since user charges are adjusted infrequently, observed waiting times to obtain a public specialist's visits are those that clear the local market at the given prices.[4] Let $\boldsymbol{y}=(y_{GP}, y_{PUB}, y_{PRI})'$ be the vector comprising the number of visits the individual chooses to consume. GP visits are free, public specialist visits are priced according to f^*_{PUB} and private specialists according to f^*_{PRI}. The vector of prices is therefore given by: $\boldsymbol{P^*}=(0, f^*_{PUB}, f^*_{PRI})'$. Administrative waiting times that correspondingly clear the market in a long-run equilibrium are at the level $\boldsymbol{W^*}=(0, w^*_{PUB}, w^*_{PRI})'$.[5] We assume that preferences are represented by a utility function $\boldsymbol{U}(\boldsymbol{y}, \boldsymbol{x}; \boldsymbol{z}, \boldsymbol{W^*})$ with standard properties that is conditioned on a vector $\boldsymbol{z}$ of individual factors including health status variables and the

[4] We do not consider prioritization schemes, which in fact actually play only a minor role here given the low therapeutic value of physician consultations. Moreover, we do not attach the same interpretation to observed waiting time in the market for private consultations.

[5] GP visits are rationed exclusively by physical waiting time.

vector of administrative waiting times $\boldsymbol{W}$.[6] In the short run, having normalized prices and income by the price of the consumption good $m+1$, our consumer solves the following problem:

$$\max_{x,y}[U(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{z}, \boldsymbol{W^*}): \boldsymbol{q'x} + x_{m+1} + \boldsymbol{P^{*\prime}y} = \boldsymbol{M}, \boldsymbol{x}, \boldsymbol{y} \geq 0]$$

where $\boldsymbol{M}$ is the available income. Due to data limitations we cannot observe the individual consumption of goods $\boldsymbol{x}$, so the ordinary demand system that we estimate is the following:

$$y_j=h^j(\boldsymbol{P^*}, \boldsymbol{q}, \boldsymbol{M}; \boldsymbol{z}, \boldsymbol{W^*}), \qquad j=GP, PUB, PRI. \tag{1}$$

In addition to (1) the individual demands goods x_l according to the demand functions $x_l=h^l(\boldsymbol{P^*}, \boldsymbol{q}, \boldsymbol{M}; \boldsymbol{z}, \boldsymbol{W^*})$, $l=1, ..., m+1$, but these are not observed. Given that in our case $m > 0$, we have to rely on the estimation of the incomplete demand system (1).

4 THE ECONOMETRIC MODEL

We specify the empirical counterpart of (1) as a multivariate count data model for the three types of doctor consultation (*GP*, *PUB*, *PRI*). For the same individual, these different measures of healthcare utilization are likely to be jointly dependent and their interrelation can be described in an analogous way to the seemingly unrelated regression model, leading to more efficient estimates with respect to univariate regressions.

In the recent econometric literature, many multivariate count data models have been developed (for an overview see Cameron and Trivedi, 2005). Generally speaking, correlated counts are generated by introducing in the marginal models some latent heterogeneity factor. Similarly to the univariate case, a stream of modelling avoids specific parametric assumptions for the heterogeneity terms and treats the problem semiparametrically through the adoption of finite mixtures[7]. We follow here the discrete factor method approach of Mroz (1999) and specify a discrete distribution for an heterogeneity term which is common to all the equations of the model but appears with equation specific factor loadings. This last feature keeps the modelling quite simple, but

[6] Conditioning on waiting times is consistent with the idea of decay function reducing the value of a treatment, in Lindsay and Feigenbaum (1984).

[7] The semiparametric approach based on finite mixtures to control for unobserved heterogeneity in econometrics dates back to the proposal of Heckman and Singer (1984) for duration models. Deb and Trivedi (1997, 2002) propose its use for modeling univariate counts in a single process approach.

introduces some flexibility in the correlation pattern of unobserved heterogeneity components across outcomes, that are not constrained to be perfectly correlated.

Denote with y_{ji} the number of visits of type j consumed by the individual i, j=*GP, PUB, PRI*. Let $x_i = (P_i', W_i', q_i, z_i')'$ be the vector of exogenous regressors including the vector of visit prices $P_i = (0, f_{PUB_i}, f_{PRI_i})'$, the vector of waiting times $W_i = (0, w_{PUB_i}, w_{PRI_i})'$, the vector of other goods' prices q_i, and a vector z_i of covariates comprising all relevant factors like individual income, education, age, health status, etc.. As a first step we consider the three counts $(y_{GP_i}, y_{PUB_i}, y_{PRI_i})$ as driven by independent Negative Binomial processes. This specification allows for overdispersion in each count, a pattern we observe in the unconditional mean and variance of our sample. The conditional density for the count variable j relative to individual i is given by:

$$f_j(y_{ji} \mid x_i) = \frac{\Gamma(\vartheta_{ji} + y_{ji})}{\Gamma(y_{ji}+1)\Gamma(\vartheta_{ji})} \left[\frac{\mu_{ji}}{\mu_{ji} + \vartheta_{ji}}\right]^{y_{ji}} \left[\frac{\vartheta_{ji}}{\mu_{ji} + \vartheta_{ji}}\right]^{\vartheta_{ji}} \quad y_{ji} = 0,1,2,..... \qquad (2)$$

where: $x_i = (p_i', z_i')'$ is the vector of exogenous regressors, $\Gamma(.)$ is the gamma function, $\vartheta_{ji} = \frac{\mu_{ji}^{k_j}}{\alpha_j}$, $\alpha_j > 0$ is an overdispersion parameter, k_j an arbitrary constant. $\mu_{ji} = \exp(\beta_j' x_i)$ represents the conditional mean of the process: $\mu_{ji} = E(y_{ji} | x_i) = \exp(\beta_j' x_i)$, where β_j is a count specific vector of unknown parameters. The conditional variance function is: $V(y_{ji} | x_{ji}) = \mu_{ji} + \alpha_j \mu_{ji}^{2-k_j}$. If $\alpha_j=0$ we get the Poisson model, in which the conditional variance equals the conditional mean and overdispersion is not allowed. We resort to the Negbin-2 model setting $k_j=0$.

As a second step we let the three processes to be simultaneously determined. Since the more frail or the more anxious individuals tend to consume more of all types of visits a pattern of positive dependence across the three counts might plausibly occur. To capture such kind of dependence we introduce the common unobserved individual heterogeneity component, v_i, in the Negbin mean function of the three processes:

$$\mu_{ji}^{(v)} = \exp(\beta_j' x_i + \theta_j v_i) \qquad (3)$$

where θ_j are unknown equation-specific factor loadings that are estimable upon some identification restrictions discussed below, implying that the unobserved components of each process conditional mean, $\theta_j v_i$, are allowed not to be perfectly correlated. Notice that when the observed regressors do not include prices, the resulting dependence structure among counts incorporates both possible complementarity and substitution effects exerted through unobserved visit's prices and the impact of unobserved individual heterogeneity. Conditioning on prices, we are able to disentangle these two effects, and can ascribe the correlation pattern entirely to individual heterogeneity. Conditionally to v_i, the joint density of the three count variables is given by:

$$f(y_i \mid x_i, v_i; \beta, \vartheta) = \prod_{j \in (GP, PuS, PrS)} f_j^{(v)}(y_{ji} \mid x_i, v_i; \beta_j, \vartheta_j) = f(y_i \mid x_i, v_i) \qquad (4)$$

where $y_i = (y_{GP_i}, y_{PuS_i}, y_{PrS_i})'$, the vectors β and ϑ collect the parameters of the mean functions of the three counts, and $f_j^{(v)}(.)$ denotes the conditional density obtained from (1) after generalizing the conditional mean with the heterogeneity component according to (2). In order to write down the likelihood function, we have to integrate out the unmeasured scalar heterogeneity term. This requires choosing a functional form for the distribution of v_i, $G(v_i)$. We assume that $G(v_i)$ may be approximated by a discrete distribution with a finite number of points of support. In the resulting discrete factor specification, the individual likelihood function has the following expression:

$$L_i(\beta, \theta, \pi, v) = \sum_{k=1}^{K} \pi_k \prod_j L_{jik}(\beta_j, \theta_j, v_k) \qquad (5)$$

with: $L_{jik}(\beta_j, \theta_j, v_k) = f_j^{(v)}(y_{ji} \mid x_i, v_k; \beta_j, \vartheta_j)$, $\prod_j L_{jik}(\beta_j, \theta_j, v_k) = f(y_i \mid x_i, v_k)$

where $v = (v_1, ..., v_k, ..., v_K)'$, v_k are the K points of support with associated probabilities π_k, and $\pi = (\pi_1, ..., \pi_k, ..., \pi_K)'$. The location of the points of support v and their associated probabilities π are estimated jointly with the other parameters of the model. The location of the discrete factor is arbitrary when the model contains an intercept. Identification requires normalising one of the support points to zero, with the remaining support points expressed as deviations from it. Moreover given that the scale of the discrete factor is not identified, one loading factor must be set to a non-zero constant (we set θ_{GP} =1).[8] Adopting a discrete distribution for the heterogeneity component amounts to assume that each observation on the triplet of counts y_i is generated from a population consisting of a finite-mixture of K distinct sub-populations (or latent classes). Each sub-population k is characterized by the joint density $f(y_i \mid x_i, v_k)$, and constitutes a proportion π_k of the overall population. Empirical evidence in the literature suggests that adequate nonparametric representation of $G(v_i)$ is reached with a low number of support points.[9] In many applied health economics studies this number is

[8] See Mroz (1999) for a discussion of identification for discrete factor simultaneous equation models.

[9] See Cutler (1995), Deb and Trivedi (1997), Hamilton and Bramley-Harker (1999).

found to be two, with the consequence that the latent groups are interpreted as “frequent users/ill”, with high average demand for medical care, and “infrequent users/healthy”, with low average demand.

To sum up, in our final specification the finite mixture allows for heterogeneity in the conditional mean functions of the component densities, while the Negbin specification allows for overdispersion within each class. This model proves to be rather convenient for analysing price and income demand elasticities. It is easy to see that when the mean function is like: $\mu = E(y \mid x) = \exp(\beta' x)$ the elasticity of the count with respect to variable x_k is given by:

$$\varepsilon_k = \frac{\partial E(y|x)}{\partial x_k} * \frac{x_k}{E(y|x)} = \beta_k x_k . \qquad (6)$$

5 DATA

We work on a dataset from the "Indagine Statistica Multiscopo sulle Famiglie: condizioni di salute e ricorso ai servizi sanitari" (ISMF), a national household survey conducted by the Italian National Statistical Institute (ISTAT) every 5 years. The last available survey was conducted from July 1999 to June 2000 on a sample of 52,332 households comprising 140,011 individuals.

The survey provides a full account of individual health, healthcare utilization, demographics, socio-economic status and other relevant economic variables such as private health insurance policies. We exploit the section of the survey on individual healthcare consumption during the four weeks before the interview, comprising the number of GP visits and specialist consultations with a public and/or a private specialist. Specialist consultations are classified into 14 different specialties. For empirical manageability we pool them into a single class. We disregard dentist visits as their cost variability is very great. To avoid the confusion due to unobserved family factors[10] affecting the consumption of all members, we restrict the analysis to the household head. After two further selections relating to the definition of some of the required explanatory variables, the final sample has 45,601 observations.

Tables 1 and **2** show the three counts in our dependent variables and sample moments of the distributions. In the span of a month, the participation rate, irrespective of the type of visit, is 27.6% (the share of individuals who consume only 1 visit is about 16%). On average, individuals consume 0.3 GP visits, 0.11 public specialist visits and 0.08 private visits. Overdispersion is clearly present: sample variance is almost twice

[10] See Deb (2001) for a clear statement of this problem.

the mean. We get better insight by conditioning on "some" positive, i.e. analysing those individuals that consumed at least one visit in the month. This resembles the subset of individuals who had some illness during the month before interview. On average these they consume 1 GP visit and approximately 0.4 and 0.3 specialist visits. Overall, this evidence suggests that there is some positive dependence between the three counts.

The next subsection describes in some detail the regressors we include in our empirical model. For an overview of all variables, see **Table 3.**

5.1 FEES AND WAITING TIMES

To specify a model consistently with the theory set out in **Section 3**, we have to measure a proper vector of visit prices and define a measure of waiting times. We also need a price index for all non-food goods, deflating nominal variables by a price deflator for food.

Public specialists are paid according to a schedule of administered fees, while private ones base their fees on competitive pressure from close substitutes in the relevant market. We take the territory of the LHA as the proper relevant market area for medical consultation. In a sense, we define the consumption of physician services outside one's residence LHA as negligible.[11] The average price of each type of visit is known to consumers. We assume that patients believe the alternatives to be priced according to what they observe on average in their local market, and define the vector of observed prices accordingly as: $\boldsymbol{P}=(0, f_{PUB}, f_{PRI})$. We assume individuals foresee the price of a consultation by combining the price signal coming from the LHA area and that available from the broader provincial area. Therefore to obtain a measure of both fees, **FEE PUB** and **FEE PRI**, we consider outlays for the last visit reported by individuals who had at least one visit in the four weeks before the interview[12]. These figures are then averaged across individuals belonging to the same LHA, to obtain the local price signal, and the same province for the broader price signal. Each individual is the assigned, both for public and private specialist visits, an average of the two price signals using population shares as weights[13].

To account for different purchasing power due to regional disparities, we deflated the fees using a measure of household consumption expenditure in the region of

[11] We do not have data on exit rates for specialist care. In hospital admissions for basic medical treatment the exit rate is about 25%.

[12] In this aggregation we had to cope with the way outlays were recorded in the survey. Individuals were asked to indicate in which of seven classes their total outlays for the last visit fell. Therefore individual outlay was calculated as the bracket midpoint.

[13] We excluded individuals belonging to LHAs where fewer than 10 observations for each type of visit outlay were available.

residence. For this purpose we use monthly food expenditures as collected in the Survey on Household Income and Wealth (SHIW 2000) conducted by the Bank of Italy. The outlays were adjusted for family size and aggregated across the 20 regions and 4 size classes of municipality of residence. So fees are expressed as shares of the average monthly food expenditure in the region. On this basis a public specialist consultation costs about 6% of monthly household food expenditure, a private specialist is on average 32%.

Finally, to fulfill the requirements of a coherent incomplete system of demand (Hausman, 1981, Epstein, 1982, LaFrance, 1985) we inserted **CONSRE,** appropriately deflated as above, to represent a proxy for the price of the non-food bundle of goods available in the individual's place of residence. We used the monthly consumption expenditures on non-food goods as collected in SHIW. As above, outlays were adjusted for family size and aggregated across regions and city sizes.

The definition of a proper measure of waiting times to access for the different types of consultation is more straightforward. ISMF provides the information on waiting times for obtaining the last visit, either by a public or a private consultant, as reported by individuals who had at least one visit in the four weeks before the interview. These figures are averaged across individuals at the LHA and province levels as for fees. Each individual in the sample is then assigned the waiting time foreseen on average in his LHA, both for public and private specialist visits (**WAIT PUB** and **WAIT PRI**).

Table 4 gives details on the source of variation we exploit in our econometric exercise to identify demand elasticities.

5.2 QUALITY

Since we estimate demand elasticities by exploiting geographical variability in fees and waiting times, it is of essential to control for other "environmental" effects that might shift demand and undermine identification. A major issue in our case is given by the spatial distribution of the quality of physician consultancy. We consider here the quality of GP consultation as spatially homogeneous, given the low therapeutic and diagnostic value of primary care in Italy. However we cannot exclude that the quality of specialist consultation plays a role in motivating the demand for either type of visit.[14] We proxy quality by a measure of the size of the local pool of physicians and a set of covariates reflecting the general level of quality of public healthcare.

[14] Take for instance the case of low quality in the locally available public specialists for consultation. This might lead to a higher rate of referral to private specialist (demand shift), undermining the proper identification of price effects unless the quality of public specialist consultancy is properly accounted for.

We consider that the quality of the consultation per se depends mainly on the local availability of human capital (physicians). Therefore, we develop a measure of the poverty of the local pool (we call it **P_LPHC**). Private specialist consultations in Italy are mainly supplied by physicians appointed in public hospitals. It is by operating in public hospitals that they acquire skills (by treating many patients) and reputation. Since reputation has a spatial gradient and mobility is costly, physicians tend to offer private consultancy near the appointing hospital. It therefore seems reasonable to use the spatial accessibility to public hospitals in a given area as a proxy for P_LPHC. To achieve this we estimated a linearized version of a gravity model specified as follows:

$$\begin{aligned} T_{ij} &= a * k(P_i) * g(A_j) * h(dist_{ij}) * i(f_{ij} A_j) \\ &= a * \exp(\alpha_{Pi} D\_P_i) * \exp(\alpha_{Aj} D\_A_j) * \exp(\gamma_1 dist_{ij} + \gamma_2 dist_{ij}^2 + \gamma_3 dist_{ij}^3) * (f_{ij} A_j)^{\alpha} \end{aligned}$$

where T_{ij} stands for the number of residents in province i admitted to hospital j, A_j and P_i represent attraction and pull factors respectively referred to the admitting hospital and the province of residence, $dist_{ij}$ is the distance between i and j and f_{ij} is a set of characteristics specific to the flow between i and j. Upon estimation[15], a canonical measure of accessibility of zone i is given by the so-called Hansen measure (1959):

$$P\_LPHC_i = \sum_j \frac{\hat{T}_{ij}}{\hat{T}_{i\bullet}} dist_{ij}$$

This variable is the mean predicted distance the patient must cover to receive hospital admission. On our premises, this captures the poverty of LPHC (see **Table A1** for a ranking of Italian regions by this measure). The higher **P_LPHC**, the lower the overall, public and private, locally available quality of medical consultations (covariation effect). At the same time, the higher **P_LPHC** the lower the quality diversification in the private market (differential effect) implying less incentive to refer to private specialists. When the differential effect prevails, a higher **P_LPHC** will lead to less utilization of private specialists and greater utilization of GPs and public specialists.

The overall quality of public healthcare supply is captured by the public expenditure per-capita (**PUB EXP PCAPITA)** in the LHA, the share of women aged 25 and above that underwent a pap smear test (**PAP SMEAR**) and of those that underwent a mammography (**MAMMOGRAPHY**) in LHA. The motivation for the first regressor

[15] Details on the estimation of the gravity model are available upon request.

is apparent; concerning the last two we posit that promoting preventive care programs signals good quality of physician care. We therefore expect these controls to be correlated positively with use of public supply and negatively with private.

5.3 OTHER COVARIATES

We tried, while drawing on similar specifications in the literature, to keep our specification as parsimonious as possible. In this respect our specification is quite close to Deb and Trivedi (1997) and to Pohlmeier and Ulrich (1995) thus allowing us to make useful comparisons. We included as controls the following variables: **INVALID**, **POOR SRH, GOOD SRH**, **NCHRONIC**, **SMOKER, FEMALE**, **AGE**, **EDUCATION MEDIUM, EDUCATION HIGH, MARRIED** [see **Table 3** for details]. **EXEMPT** identifies those individuals whose health or economic conditions entitle them to free public specialist visits. Since these individuals are insensitive to public fees, our measure of public fees assumes a non-null value only for non-exempt individuals. We also insert a set of variables to capture the local market conditions. **PHYS DENSITY** provides a very rough measure of the availability of doctors in the local area. **CAPOLUOGO** (a dummy for provincial capital**)** and **POPULATION** can capture the travel time and accessibility opportunity cost of going to see a physician.

Since we do not observe physical waiting time, as Coffey (1983) does, we let related aspects, basically the role played by the individual value of time, enter the equations indirectly by conditioning on labor force participation status (Browning and Meghir, 1990) through the variables **EMPLOYED, SELF-EMPLOYED, RETIRED.**

The measure of monthly disposable **INCOME** we include in our model is derived from a matching exercise performed by the Italian National Statistical Institute, as the ISMF survey does not have data on household income. By regression matching each household in the sample was assigned the imputed after-tax income estimated using data from the Survey on Household Income and Wealth conducted by the Bank of Italy.[16] We apply to nominal **INCOME** the same deflator we used for public and private fees, i.e. household monthly food expenditure.

Finally, we account for the possible endogeneity of private health insurance (**INSURED)** in our model of healthcare utilization, an issue that has attracted a good deal of attention since Cameron et al. (1988). To keep endogeneity from undermining the consistency of the estimates, we specify the utilization equations as functions of the latent continuous variable that determines the binary insurance coverage dummy we observe by means of the usual threshold transformation. The latent variable is then

[16] The matching was performed by Proto and Solipaca (2001).

allowed to be simultaneously determined with the count variables. Windmeijer and Santos Silva (1997) show that estimation of the resulting simultaneous equation model through a two-stage procedure leads to consistent estimates of the parameters of the use equations. We therefore use as a regressor for visit counts the linear prediction of the first-stage binary insurance dummy model (**P_INSURED**).[17]

6 RESULTS

6.1 GENERAL COMMENTS ON THE ESTIMATED MODEL

Before discussing the results on cross price and waiting time demand elasticities, some comments on the overall results from the estimation of model (4) will be helpful (**Table 5**). We found that the heterogeneity distribution $G(v)$ can be approximated by a finite distribution with 2 points of support ($K=2$).[18] This result, consistent with other works in the literature, suggests that the population is conveniently represented by two classes of individuals, grouping respectively about 92.5% and 7.5% of the population. Factor loadings θ_{PUB} and θ_{PRI} are positive and precisely estimated, suggesting that unobserved factors leading to lower (higher) consumption of GP visits also decrease (increase) the conditional mean of both types of specialist visits. These findings are consistent with the hypothesis that the heterogeneity factor v captures individual frailty/anxiety, leading to a positive dependence between visit counts. The presence of prices among the regressors of the model is crucial in order to avoid mistaking such positive dependence for complementarity.

In order to gain some heuristic evaluation of the model's goodness of fit we present a comparison of the sample and fitted frequencies of the most relevant observed threesomes of counts (**Table 6**). The fitted frequencies are the sample averages of the individual predicted probabilities evaluated first through the joint trivariate model and then through the model assuming independent counts. The predictions obtained with the estimated joint distribution are fairly close to the actual frequencies, and in all cases they outperform the results obtained under independence.

Table 7 contains the fitted mean number of visits for the two component distributions and the mixture density (last column). For all three types of visit, the fitted means are substantially larger for the second component, which turns out to represent the high-use population, while the first component describes low users. High users

[17] Details on first-stage estimation and significance of the instruments in both the insurance and the use equation are available upon request.
[18] We tried to estimate both models with three points of support but ended up with degenerate approximations.

exhibit an average number of 0.88 GP consultations, more than three times the figure for low users (0.26). The mean number of public specialist consultations displays a great discrepancy between the two subpopulations: the high users' Public visit mean is seven times the low users' (0.6 vs. 0.08). For private specialist consultations, too, the ratio is about the same (0.38 vs. 0.06). Evaluation of the figures for non exempt individuals only leads in general to substantially lower figures, but the pattern still holds.

The regressor coefficients, with few exceptions, are quite precisely estimated and of the "right" sign. It is worth noting that education is a discriminating factor against GP visits: intermediate education levels favor both types of specialist visit, while highly educated individuals tend to patronize only private specialists. The propensity for private insurance positively affects private specialist consultations, while it is statistically irrelevant in the demand for public GP and specialist consultations. Being exempt, or with poor self-assessed health or suffering from a higher number of chronic conditions results in consumption of more of all types of visit: all these variables are proxies of individual health status. Our proxy for quality, the measure of poor access to the local pool of physician human capital (P_LPHC) acts as expected: it does not affect GP visits, increases reliance on public specialists and significantly decreases private specialist consultations. The consumption of GP and public specialist visits increases with age, while private specialist visits do not. Being self-employed or retired picks up the effect of time constraint due to working status: self-employed individuals with a high value of time demand less healthcare, while retirement is associated with a higher number of the three types of visit. Smoking produces health damage that drive up the demand for all types of visits as the individual goes beyond a certain age.

6.2 EVIDENCE ON RATIONING THE DEMAND FOR PUBLIC VISITS

To investigate the main point of inquiry – rationing of healthcare demand – we estimated a full set of (own- and cross-) price and waiting time elasticities from our incomplete system of demand (**Table 8**). We retain the assumption of equilibrium in the local market for public and private consultancy. These demand elasticities therefore represent partial demand response to a small perturbation of the equilibrium. **Table 9** gives the results of an exercise of comparative statics, showing the impact on public specialist demand and individual expenditure of different increases in user-charges.

According to our estimates the demand for public specialist visits is moderately price sensitive. The own-price elasticities we find are in the order of those estimated in the literature (see the survey in Cutler (2002)). Namely, a 10% price increase reduces the average number of visits by 3.1%. That is, user charges for public specialist visits are effective in moderating demand. In our simulation, raising user charges by 25% or 50% reduces the demand for public specialist consultation by 7% or 14%, with a concomitant increase of 15% or 28% in total revenues from patient co-payments. Administrative waiting time plays a less substantial role as a rationing tool. Demand elasticity is about -0.04. The limited impact of waiting time is clearly reflected in the comparative statics.

There is no significant pattern of substitution prevails between public and private specialist consultations. Cross price elasticities are both positive, but very imprecisely estimated. This is even more markedly the case for cross waiting time elasticities.[19] This suggests that in our sample the demands for public and for private specialist consultation are independent of one another. Of course, "economic" independence does not exclude the positive dependence arising through unobserved heterogeneity.

Table 10 gives a welfare analysis of an increase in user charges for public specialist consultation. Inasmuch as substitution effects are negligible, we ignore them and compute the variation in consumer surplus due to price changes. Given that income effects are nil, this measure proves to be correct. We let the user charges on public specialist visits increase by 10%, 25% and 50% and compute the emerging variation in consumer surplus and total expenditure. 50% increase in user charges produces a per-capita consumer welfare loss in the market for public consultation of € 7.30 per year. Meanwhile, expenditure is reduced so much that the net welfare loss amounts to € 3.00 a year. These values are negligible, but they increase if we condition on the unobserved heterogeneity. The rightmost part of **Table 10** shows the welfare calculus for a "high user", for whom the consumer welfare loss comes to € 32.50; net of fiscal compensation, it becomes € 13.50.

7 DISCUSSION

Let us draw out the implications of some empirical results that are relevant to the general debate.

First of all, the demand for public healthcare as proxied by GP and specialist visits does not depend on household income: the estimated coefficients are nil. The demand for private healthcare does increase with income, however, and with the propensity to have private insurance.[20] At the same time, the estimated loading factors imply a

[19] A slight pattern of substitution does exist between GP visits and public specialist visits, due to the effect of administrative waiting time.

[20] Similar findings were obtained by Propper (2000) exploring the demand for private healthcare in the British NHS.

positive dependence between public and private healthcare due to unobserved frailty. That is, other things being equal, unhealthy individuals consume more of both types of service. Note that our estimates do not suggest that the more frail consume relatively more private consultations. This evidence contradicts a basic tenet of the standard argument suggested by the normative literature on the public provision of private goods, namely that it induces the affluent to opt out (see Besley and Coate 1991, Boadway and Marchand 1995). The Italian NHS does not seem to operate any redistribution through such self-selection and opting out. Furthermore there is also no self-selection or opting out by the unhealthy either. Our evidence implies that the Italian NHS provides a basic package that is consumed by the entire population, regardless of income and unobserved health condition, which is supplemented by an imperfect substitute available in the market, chosen more frequently by the richer and the privately insured.

Before commenting on the own- and cross-price and wait elasticities, let us recall that in the framework of Lindsay and Feigenbaum (1984) we have adopted, waiting time reduces the individual valuation of expected health gains from a given treatment and that user charges produce a similar impact if they are paid at the end of the wait. For the marginal consumer, the maximum acceptable wait is assumed to be equal to the expected delay, so that the present value of the rights obtained by being put on a waiting list has to equal the cost. Therefore, in this framework for those getting in the queue, the maximum acceptable wait should be greater to or equal to the expected delay.

Our results on own- and cross-price and wait elasticities suggest that in the public sector the marginal consumer is more sensitive to a price increase than to an increase in waiting time: marginal consumers of public visits are more discouraged by higher fees than by longer waitings. Marginal consumers of private visits (i.e. those obtaining the least health gain from them) are not discouraged at all by fee increases (in our sample) but can be discouraged by an increase in waiting time: the marginal consumer in the private sector is sensitive to waiting time, not to price. At the same time we detect no significant cross-price and cross-waiting effect. This evidence is consistent with consumers expecting little health gain from the public visit. Thus, marginal consumers of public visits discouraged by increased fees or waiting time do not value the lost visit highly enough to switch to the private. To put it another way, marginal consumers of public visits, i.e. those quick to leave the list if user charges are raised, do not value the health gains from quicker access to the private service highly enough to pay the higher fees. The same effect is found also for marginal consumers of private visits discouraged by an increase in waiting time. This suggests that rationing public provision with moderate waiting lists and moderate user charges is effective in curbing demand: such measures make consumption less attractive to marginal consumers who value the expected health gains from public and private visits at less than the full cost of obtaining them. We do not find any evidence that the marginal consumers who are discouraged tend to be richer. This is in contrast with the thesis that higher user charges (Besley 1991, Munro 1992, Balestrino 1995) and longer waiting lists (Bucovetsky 1982, Besley and Coate 1991, Hoel and Saether 2003) for a publicly provided private good can be welfare-improving, insofar as they prompt additional opting-out by the more affluent.

8 CONCLUSIONS

We have investigated the effectiveness of monetary and non-monetary rationing of the demand for physician care in Italy. To account for peculiarities in the Italian NHS we estimate, in the spirit of a seemingly unrelated regression model, a joint model for the number of general practitioner, public specialist and private specialist visits. We account for simultaneity through a latent-class semiparametric approach in which dependence among the three counts stems from a common unobserved heterogeneity component. To identify the parameter of interest we exploit geographical variations in fees and waiting times.

We find that own-price elasticity of the demand for public specialist consultation is about -0.3. Administrative waiting time plays a less important role as a rationing tool. There is apparently no substitution between the demand for public and for private specialist care, so that user charges work as a net deterrent for over-consumption. We compute the consumer welfare change and conclude that increasing the user charges for public visits results in a net welfare loss that is non-negligible for individuals who are high users.

Our evidence does not indicate that the public provision of healthcare induces the richer citizens to opt out, so the Italian NHS apparently does not work any redistribution through self-selection. In the range of values we explore here, the mix of low user charges and moderate waiting time appears to be quite effective in curbing the demand for public healthcare, insofar as it discourages consumption by the individuals who value the expected health gains from public and private visits at less than the full cost of obtaining them. We find no evidence in support of the thesis that the marginal consumers who are discouraged tend to be the wealthier, so user charges and waiting list cannot be claimed to reach redistributive purposes.

REFERENCES

Acton JP. Nonmonetary factors in the demand for medical services: some empirical evidence. Journal of Political Economy 1975; 83(2); 595-614.

Anderson GF, Hussey PS, Frogner BK, Waters HR. Health Spending In The United States and the Rest of the Industrialized World. Health Affairs 2005; 24(4); pg. 903-915.

Balestrino A. User Charges as Redistributive Devices. Journal of Public Economic Theory 1999; 1(4); 511-524.

Besley T, Coate S. Public Provision of Private Goods and the Redistribution of Income. American Economic Review 1991; 81(4); 979–984.

Besley T, Hall J, Preston I. The demand for private health insurance: do waiting lists matter?. Journal of Public Economics 1999; 72(2); 155-181.

Besley T. Welfare improving user charges for publicly provided private goods. Scandinavian Journal of Economics 1991; 93; 495-510.

Boadway R, Marchand M. The Use of Public Expenditures for Redistributive Purposes. Oxford Economic Papers 1995; 47(1), 45–59.

Browning M, Meghir C. The Effects of Male and Female Labor Supply on Commodity Demands. Econometrica 1990; 59(4); 925-951.

Bucovetsky S. On the Use of Distributional Waits. Canadian Journal of Economics 1982; 17(4); 699-717.

Cameron AC, Trivedi PK, Milne F, Piggott J. A microeconometric model of the demand for health care and health insurance in Australia. Review of Economic Studies 1988; 55; 85-106.

Cameron AC, Trivedi PK. Microeconometrics. Methods and Applications, Cambridge University Press: Cambridge; 2005.

Coffey RM. The Effect of Time Price on the Demand for Medical-Care Services. The Journal of Human Resources 1983; 18 (3); 407-424.

Colle AD, Grossman M. Determinants of Pediatric Care Utilization. The Journal of Human Resources 1978; 13; 115-158.

Cullis J, Jones P, Propper C 2000. Waiting Lists and Medical Treatment. In: Newhouse JP, Culyer AJ (Eds), Handbook of Health Economics, vol. 1B. North-Holland: Amsterdam; 2000; p. 1201-1249.

Cutler DM 2002. Public Policy for Health Care. In: Auerbach A, Feldstein M (Eds), Handbook of Public Economics, volume 4. North-Holland: Amsterdam; 2002; p. 2143-2243.

Cutler DM, Gruber J. Does Public Insurance Crowd Out Private Insurance The Quarterly Journal of Economics 1996; 111(2); 391-430.

Cutler DM, Sheiner L. Demographics and medical care spending: standard and non-standard effects. In: Auerbach AJ, Lee RD. (Eds) Demographic change and fiscal policy. Cambridge University Press: Cambridge; 2001; p. 253-91

Cutler DM. The incidence of adverse medical outcomes under prospective payment. Econometrica 1995; 63(1); 29-50.

Deb P, Trivedi P. Demand for medical care by the elderly: a finite mixture approach. Journal of Applied Econometrics 1997; 12; 313-336.

Deb P, Trivedi P. The structure of demand for health care: latent class versus two-part models. Journal of Health Economics 2002; 21; 601-625.

Deb P. A Discrete Random Effects Probit Model with Application to the Demand for Preventive Care. Health Economics 2001; 10(5); 371-83

Docteur E, Oxley H. Health-Care Systems: Lessons From The Reform Experience. OECD Economics Department Working Papers No.374; 2003.

Epstein L. Integrability of Incomplete Systems of Demand Functions. Review of Economic Studies 1982. 49(3); 411-425.

Goldman F, Grossman M. The Demand for Pediatric Care: An Hedonic Approach. The Journal of Political Economy 1978; 86(2); 259-280.

Gravelle, H, Dusheiko, M, Sutton, M. The Demand for Elective Surgery in a Public System: Time and Money Prices in the UK National Health Service. Journal of Health Economics 2002; 21(3); 423-449.

Hamilton B, Bramley Harker R. The impact of NHS reforms on queues and surgical outcomes in England: evidence from hip fracture patients. Economic Journal 1999; 109; 437-462.

Hansen WG. How Accessibility Shapes Land Use. Journal of the American Institute of Planners 1959; 25; 73-76.

Hausman JA. Exact Consumer's Surplus and Deadweight Loss. American Economic Review 1981; 71(4); 662-676.

Heckman J, Singer B. A method for minimizing the impact of distributional assumptions in econometric models for duration data. Econometrica 1984; 52(2); 271-320.

Hoel, M, Saether, EM. Public Health Care with Waiting Time: The Role of Supplementary Private Health Care. Journal of Health Economics 2003; 22(4); 599-616.

Lindsay CM, Feigenbaum B. Rationing by Waiting Lists. American Economic Review 1984; 74(3); 404-417.

Martin, S, Smith, P. Rationing by Waiting Lists: An Empirical Investigation. Journal of Public Economics 1999; 71(1); 141-64.

Mroz, A. Discrete factor approximations in simultaneous equation models: estimating the impact of a dummy endogenous variable on a continuous outcome. Journal of Econometrics 1999; 82; 233-274.

Munro A. Self selection and optimal in-kind transfers. Economic Journal 1992; 102 (414); 1184-1196.

Pohlmeier W, Ulrich V. An econometric model of the two-part decision making process in the demand for health care. Journal of Human Resources 1995; 30; 339-61.

Propper C. The demand for private health care in the UK. Journal of Health Economics 2000; 19(6); 855-876.

Proto G, Solipaca A. L’esperienza del match tra le indagini Multiscopo Salute 1994 e Banca d’Italia 1995, mimeo, Istat: Rome; 2001.

Schoen C, Osborn R, Huynh PT, Doty M, Zapert K, Peugh J, Davis K. Taking the Pulse of Health Care Systems: Experiences of Patients With Health Problems in Six Countries. Health Affairs 2005; 24; 509-525.

Siciliani L, Hurst J. Explaining Waiting-Time Variations for Elective Surgery across OECD Countries. OECD Economic Studies 2004; 0(38); 95-123.

Windemeijer FAG, Santos Silva JMC. Endogeneity in Count Data Models: an Application to Demand for Health Care. Journal of Applied Econometrics 1997; 12; 281-294.

FIGURES AND TABLES

Table 1: Frequency of physician visits by type

N° Visits	GP	Public Specialist	Private specialist	Any type
0	81.0%	92.3%	93.9%	72.4%
1	12.7%	5.7%	4.8%	16.1%
2	4.0%	1.3%	0.9%	6.9%
3	1.2%	0.4%	0.2%	2.3%
4	0.8%	0.2%	0.1%	1.4%
5	0.1%	0.1%	0.0%	0.4%
+5	0.4%	0.2%	0.1%	1.0%
N° Obs.	45601	45601	45601	45601
Positives	8643	3490	2786	12608
Part. Rate	19.0%	7.7%	6.1%	27.6%

Table 2: Mean number of visits. Variance in parenthesis

Visit type	Full sample	Conditional on some “positives”
GP	0.296	1.070
	(0.584)	(1.286)
Public Sp.	0.111	0.400
	(0.231)	(0.722)
Private Sp.	0.080	0.291
	(0.135)	(0.427)
Any type	0.487	0.000
	(1.078)	(0.000)

Table 3: Descriptive statistics

Variable	Description	mean	st.dev.	min	max
Generic visits (GP)	Number of GP office visits in the 4 weeks before the interview	0.296	0.764	0.000	9.000
Public specialist visits	Number of visits by a public specialist in the 4 weeks before the interview	0.111	0.481	0.000	16.000
Private specialist visits	Number of visits by a private specialist in the 4 weeks before the interview	0.080	0.367	0.000	9.000
FEE PUB	Public specialist price as a share of local food expenditure in the LHA	0.062	0.045	0.000	0.166
FEE PRI	Private specialist price as a share of local food expenditure in the LHA	0.324	0.054	0.194	0.531
INCOME	Monthly equalized family disposable income as a share of local food expenditure	5.159	4.156	1.001	160.155
CONSRE	Monthly equalized local non food family expenditure as a share of food expenditure	1.321	0.247	0.840	2.171
WAIT PUB	Waiting time (weeks) for obtaining a public specialist consultation in the LHA	1.846	0.710	0.512	4.075
WAIT PRI	Waiting time (weeks) for obtaining a private specialist consultation in the LHA	0.856	0.395	0.239	3.314
PUB EXP PCAPITA	Public health expenditure per capita in the LHA	0.997	0.225	0.476	1.494
PAP SMEAR	Share of women aged 25 and above that underwent a pap smear test in LHA	0.588	0.136	0.304	0.822
MAMMOGRAPHY	Share of women aged 25 and above that underwent a mammography in LHA	0.361	0.092	0.180	0.555
P_LPHC	Accessibility to the Local Pool of Human Capital (x100 Km) in the LHA	0.565	0.276	0.060	1.798
PHYS DENSITY	Physician density (x1000 inhabitants in residing Provincia)	5.505	1.350	3.394	8.782
POPULATION	Population (x1,000,000) in the residing Comune	0.128	0.398	0.000	2.644
CAPOLUOGO	=1 if the person resides in a Comune that is provincial chieftown	0.268	0.443	0.000	1.000
EXEMPT	=1 if the person is exempt from co-payment	0.293	0.455	0.000	1.000
INVALID	=1 if the person suffers from some invalidity	0.079	0.270	0.000	1.000
NCHRONIC	Number of chronic conditions	1.781	2.162	0.000	27.000
POOR SRH	=1 if self-rated health is poor	0.106	0.308	0.000	1.000
GOOD SRH	=1 if self-rated health is good	0.078	0.269	0.000	1.000
SMOKER	=1 if the person is currently a smoker	0.272	0.445	0.000	1.000
FEMALE	=1 if the person is female	0.252	0.434	0.000	1.000
AGE	Age in years/100	0.558	0.162	0.180	1.050
EDUCATION MEDIUM	=1 if the person holds a secondary school certificate	0.246	0.431	0.000	1.000
EDUCATION HIGH	=1 if the person holds a university degree	0.072	0.259	0.000	1.000
MARRIED	=1 if the person is married	0.674	0.469	0.000	1.000
EMPLOYED	=1 if the person is employed	0.505	0.500	0.000	1.000
SELF-EMPLOYED	=1 if the person is currently self-employed	0.156	0.363	0.000	1.000
RETIRED	=1 if the person is retired	0.377	0.485	0.000	1.000
INSURED	=1 if the person is "covered" by a private health insurance	0.159	0.365	0.000	1.000
P_INSURED	Linear prediction of INSURED latent variable	0.159	0.136	-0.172	0.675

Table 4: Geographical variations in visit counts, nominal and real monetary aggregates (fees and income).

	N-West	N-East	Centre	South	Islands	ITALY
N° Obs	9476	9497	8559	12740	5329	45601
N° LHA	27	26	31	38	13	135
GP visits	0.254	0.278	0.293	0.335	0.313	0.296
Publ.Sp. visits	0.093	0.114	0.111	0.108	0.142	0.111
Priv.Sp. visits	0.075	0.080	0.089	0.082	0.073	0.080
ANY visits	0.422	0.472	0.493	0.524	0.528	0.487
% NO visits	0.742	0.709	0.716	0.724	0.728	0.724
Public fee*	19.651	23.083	21.197	18.147	15.386	19.833
Private fee	80.508	78.863	70.946	70.089	67.108	73.772
Income	1.384	1.425	1.304	0.950	0.903	1.200
Food Exp**	0.256	0.249	0.254	0.198	0.189	0.230
FEE PUB*	0.077	0.093	0.085	0.093	0.082	0.087
FEE PRI	0.316	0.319	0.282	0.358	0.360	0.325
INCOME	5.366	5.716	5.145	4.775	4.735	5.159
WAIT PUB	1.833	2.616	1.769	1.345	1.776	1.826
WAIT PRI	0.842	1.102	0.956	0.701	0.832	0.878
EXEMPT	0.271	0.303	0.289	0.297	0.306	0.292
INSURED	0.190	0.211	0.172	0.108	0.109	0.159

All monetary values are denominated in EU €.

* These aggregates refer to average fees for NON EXEMPT individuals.

**Source: our elaboration on data from the Bank of Italy.

Table 5: Model estimates

	GP		PUBLIC SPECIALIST		PRIVATE SPECIALIST	
	ESTIMATE	STD ERR	ESTIMATE	STD ERR	ESTIMATE	STD ERR
FEE PUB	-0.2955	1.1270	-3.5576 ***	1.4446	1.2759	1.5268
FEE PRI	-0.1447	0.3918	0.2042	0.6066	0.3888	0.6351
INCOME	0.0049	0.0044	0.0019	0.0055	0.0072 *	0.0054
CONSRE	-0.0026	0.1318	-0.0154	0.1768	-0.1825	0.1682
WAIT PUB	0.0655 *	0.0406	-0.0573 *	0.0386	-0.0058	0.0509
WAIT PRI	-0.0120	0.0479	0.0056	0.0812	-0.1561 **	0.0699
PUB EXP PCAPITA	0.0851	0.1282	0.2462 *	0.1821	0.1600	0.2002
PAP SMEAR	-0.2978	0.3634	-0.2793	0.6959	-0.7604	0.6409
MAMMOGRAPHY	1.2767 ***	0.4347	0.2662	0.7283	0.8735 *	0.6799
P_LPHC	-0.0551	0.1127	0.2008	0.1723	-0.2498 *	0.1704
PHYS DENSITY	0.0177	0.0167	0.0694 ***	0.0285	0.0018	0.0259
POPULATION	0.0016	0.0662	-0.0283	0.0774	-0.0740	0.0787
CAPOLUOGO	-0.1466 ***	0.0535	0.0934 *	0.0602	-0.0396	0.0836
EXEMPT	0.2905 ***	0.1040	0.3572 ***	0.1305	0.2256 *	0.1396
INVALID	-0.0005	0.0381	0.2687 ***	0.0522	0.1188 *	0.0755
NCHRONIC	0.1349 ***	0.0053	0.1434 ***	0.0087	0.1720 ***	0.0104
POOR SRH	0.5303 ***	0.0337	0.7409 ***	0.0606	0.7150 ***	0.0675
GOOD SRH	-0.7994 ***	0.0882	-0.9310 ***	0.1362	-0.6340 ***	0.1273
SMOKER	0.5280	0.4745	1.0475 *	0.6679	0.9243 *	0.6362
FEMALE	0.1620 ***	0.0378	0.0834 *	0.0628	0.4856 ***	0.0700
AGE	1.2653 *	0.9070	4.0960 ***	1.1333	0.8031	1.0753
AGE SQUARED	-0.0791	0.7241	-3.7780 ***	0.9318	-0.6986	0.8746
AGE*SMOKER	-2.8224 *	1.7317	-4.5958 **	2.5205	-3.8400 *	2.4137
AGE SQUARED*SMOKER	2.7146 **	1.5099	4.1239 **	2.3044	3.2330 *	2.1664
EDUCATION MEDIUM	-0.0842 **	0.0448	0.1033 *	0.0733	0.1735 ***	0.0644
EDUCATION HIGH	-0.3005 ***	0.0678	0.0906	0.1119	0.1723 *	0.1085
MARRIED	0.0421	0.0367	0.0331	0.0530	0.1123 **	0.0612
P_INSURED	-0.2926	0.3318	-0.3644	0.5333	1.8183 ***	0.4947
EMPLOYED	-0.0442	0.0559	-0.0075	0.0851	0.1159	0.0992
SELF-EMPLOYED	-0.1878 ***	0.0584	-0.1110	0.1044	-0.1778 **	0.0864
RETIRED	0.0680 **	0.0384	0.1358 **	0.0659	0.2707 ***	0.0780
CONSTANT	-3.0288 ***	0.4217	-4.6036 ***	0.9248	-4.2214 ***	0.6743
ALPHA	1.2759 ***	0.0780	2.6780 ***	0.2297	3.1943 ***	0.3164
θ_{PUB}	1.6702	0.1227				
θ_{PRI}	1.5559	0.1184				
N_2	1.2122	0.0791				
Π_2	0.0749	0.0109				
# OBSERVATIONS	45601					
# PARAMETERS	160					
LOGLIKELIHOOD	-54444.7					

***, **, * denote significance at 1%, 5% and 10% levels respectively.

Each equation contains 19 dummies for regional fixed effects.

Numerical maximization of the loglikelihood has been performed with the NLPQN call function in SAS, using the Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm. We checked invariance of the results to different sets of starting points. Upon convergence, robust standard errors of the parameter estimates, accounting for clustering at the LHA level, are computed using the sandwich formula.

Table 6: Model goodness of fit

Event	Actual	Joint	Independent
0,0,0	72.4%	71.9%	71.2%
1,0,0	9.9%	11.0%	11.1%
0,1,0	3.4%	3.8%	4.0%
0,0,1	2.9%	3.2%	3.4%
1,1,0	1.9%	0.8%	0.8%
1,0,1	1.5%	0.6%	0.6%
0,1,1	3.5%	0.2%	0.2%
1,1,1	0.1%	0.1%	0.0%

Table 7: Observed and predicted mean number of visits

	Predicted			**Observed**	
	First Component	Second Component	ALL	ALL	Conditional on some positives
			Full sample		
GP	0.262	0.880	0.308	0.296	1.070
Public Sp.	0.080	0.608	0.120	0.111	0.400
Private Sp.	0.060	0.397	0.085	0.080	0.291
Any type	0.402	1.885	0.513	0.487	1.760
			NOT Exempt		
GP	0.162	0.546	0.191	0.193	0.939
Public Sp.	0.044	0.337	0.066	0.068	0.333
Private Sp.	0.048	0.320	0.069	0.069	0.333
Any type	0.255	1.202	0.326	0.330	1.604

Table 8: Price and waiting time elasticities across percentiles.

	GP visits	Public Specialist visits	Private Specialist visits
FEE PUB	-0.026	-0.307***	0.110
	(0.097)	(0.125)	(0.132)
FEE PRI	-0.046	0.065	0.124
	(0.125)	(0.193)	(0.202)
WAIT PUB	0.050*	-0.044*	-0.004
	(0.031)	(0.030)	(0.039)
WAIT PRI	-0.009	0.004	-0.119**
	(0.037)	(0.062)	(0.053)

***, **, * denote significance at 1%, 5% and 10% levels respectively. Elasticities and their robust standard errors are evaluated at the median level of each variables according to (6). For public visits median values refer to non exempt individuals.

Table 9: Simulation effects of public fees increase and waiting time increase on public specialist demand and overall expenditure*

	WAIT PUB at initial value		**WAIT PUB** +10%		**WAIT PUB** -10%	
FEE PUB	Number	expenditure	number	expenditure	number	expenditure
+10%	-3.0%	6.5%	-4.0%	5.3%	-2.0%	7.7%
+25%	-7.4%	15.3%	-8.3%	14.0%	-6.4%	16.5%
+50%	-14.2%	27.6%	-15.1%	26.2%	-13.2%	29.0%

* The values are obtained by predicting demand for each NOT EXEMPT individual and then aggregating over this sub-sample.

Table 10: Welfare effects of public fees increase*

	ALL			LOW USERS			HIGH USERS		
FEE PUB	Consumer surplus	Expendit.	Net variation	Consumer surplus	Expendit.	Net variation	Consumer surplus	Expendit.	Net variation
+10%	-1.6	1.0	-0.5	-1.0	0.7	-0.4	-7.9	5.2	-2.7
+25%	-3.8	2.4	-1.4	-2.5	1.6	-0.9	-19.3	12.3	-7.0
+50%	-7.3	4.4	-2.9	-4.9	2.9	-2.0	-37.1	22.2	-14.9

* The values are obtained by predicting demand for each NOT EXEMPT individual and then aggregating over this sub-sample. Values are in per-capita EU€ and expanded to a year base.

APPENDIX

Table A1: Regions ranking according to P_LPHC indicator

REGIONE	P_LPHC*100
EMILIA-ROMAGNA	43.7
LAZIO	44.4
MARCHE	46.1
VENETO	46.3
LOMBARDIA	47.8
CAMPANIA	47.9
UMBRIA	49.7
LIGURIA	49.9
ABRUZZO	50.7
TOSCANA	53.4
PUGLIA	56.9
FRIULI VG	58.0
PIEMONTE	59.4
TRENTINO AA	64.5
MOLISE	68.2
SARDEGNA	74.0
CALABRIA	94.8
SICILIA	100.1
BASILICATA	101.3
VALLE D'AOSTA	113.8