Structural Estimation with a Randomized Trial of A Principal Agent Model of Medical Insurance with Moral Hazard

Marcos Vera-Hernández*(YE)
University College London
JEL C35, D82, I1, G22

January 16, 2002

Abstract

Despite the importance of principal-agent models in the development of modern economic theory, there are few estimations of these models. We contribute to fill this gap in a field where moral hazard has traditionally been considered important: the utilization of health care services. This paper presents a model where the individual decides to have treatment or not when she suffers an illness spell. The decision is taken on the basis of comparing benefits and out-of-pocket monetary costs of treatment. In the paper, we recover the estimates of the corresponding principal agent model and obtain an approximation to the optimal contract.

1 Introduction

Contract theory has been extremely important in the development of modern economic theory during the last thirty years. However, the increasing sophistication of the theory

*I would like to thank Michael Creel for his time and useful advices. I have benefited of useful discussions and comments received from Jerome Adda, Pedro Barros, Richard Blundell, Paul Contoyannis, Hide Ichimura, Belén Jerez, Andrew Jones, Roberto Leon, Matilde Machado, Costas Meghir, Nicolás Porteiro, Pedro Rey, Frank Windmeijer, and comments received from participants at York Seminars in Health Econometrics, the 2001 Conference of the International Health Economics Association in York and Ceemap Seminar at UCL. All remainning errors are my responsability. Address for correspondence Universtity College London, Dept.of Economics. Gower Street - London - WC1E 6BT. E-mail:uctpamv@ucl.ac.uk

has not gone hand in hand with empirical validation of the models as Salanié (1997) points out. Chiappori and Salanié (2000) offer us an up to date perspective of the literature that has tried to link econometrics and contract theory. Most of the existing works have used a reduced form approach.[1] Obtaining policy recommendations from reduced form models is usually a difficult task. Firstly, estimates of the raw economic parameters (technology and preferences) are usually necessary to find the policy function. Secondly, the parameters of reduced form models are a function of the raw parameters and agents' constraints, which would in general shift in case a policy change takes place (Lucas, 1976).

The purpose of this paper is to estimate the structural parameters (preferences and technology) of a principal-agent model involving moral hazard. This will allow us to solve for the optimal policy function, i.e. the optimal contract. We concentrate on the problem of health care insurance. Moral hazard arises because health shocks are not contractible, hence contracts are not complete. Consequently, it might be optimal to give incentives to the consumer in order for her not to have expensive treatments for minor health shocks.

Our main contribution is to estimate the parameters of a principal-agent model with moral hazard. This allows to use the principal-agent paradigm when solving for the optimal contract. This presents the following advantages. First, principal-agent models have developed in the last thirty years as a rigorous framework where the moral hazard concept has been widely studied. Second, when following this approach, the analyst has to make a clear distinction between contractible and non contractible variables. The relation between contractible and non-contractible variables provide important information when deriving the optimal contract. In a reimbursement health care setting, it is natural to think that the non contractible variable is a health shock, while the contractible one is the treatment cost. One would expect a strong but stochastic relation between health shocks and treatment costs. This will be important when deriving the optimal contract since a large cost tells much about the health shock. This will also be the basis for a new measure of moral hazard that we will propose: the correlation between the unobservable that influence both severity and treatment costs. Third, the optimal contract is obtained directly from first principles and does not need to do further assumptions on which is the first best level of health care utilization.

Moral hazard in the use of medical services has been one of the most recurrent issues

[1]Some exceptions are Ferrall and Shearer (1999), Margiotta and Miller (2000), Paarsch and Shearer (2000) and Biais *et al.* (1999).

in health economics. Early references about the topic are Arrow (1963), Pauly (1968) and Zeckhauser (1970). The more widespread view is that

"When moral hazard is present, insurance that reduces risk will also cause larger expected losses. In medical care, these losses represent the consumption of units of medical care whose value to the consumer is less than their cost, because the insurance coverage reduce the user price below cost"(Pauly, 1986).

Previous papers have tried to estimate optimal health care insurance contracts.[2] Their methodology is based on optimal taxation rather than asymmetric information theory. Regarding the relation between the insurance problem and optimal taxation, medical insurance might induce a distortion on the consumption of health care services, since it lowers the marginal price of consumption. In this respect, the problem of optimal taxation is similar to optimal health insurance. However as Belsey (1988) points out, there is a crucial difference between them: the insurance problem is against a background of incomplete markets. These previous approaches are based on comparing the welfare loss of a given insurance contract with respect to the situation of no insurance. Consequently they assume that the first best level of health care services correspond to the one where there is no insurance. Ma and Riordan (2001) have shown that this assumption does not verify in presence of income effects. In fact, the implementation of the first best requires the consumer to be responsible for only a fraction of treatment costs, because his marginal valuation of income rises once out of pocket payment is paid out of her income. This corresponds to the early observation by Meza (1983) for which previous approaches might be overestimated welfare losses due to moral hazard, in presence of income effects.

This paper differs from previous approaches in several dimensions. First, in previous approaches, the individual decision was the amount of monetary resources dedicated to health care. Though this is a simplifying assumption, it is undesirable for two reasons. First, since the individual has a demand curve for resources dedicated to health care, then it is assumed that the larger the health care costs, the larger the utility is. However, it seems preferable to disentangle quantity consumed from the cost of producing it, since the individual will derive utility form quantity but not from the cost of production. There is a second reason why it is undesirable to model individual decision as the amount of monetary resources dedicated to health care. This way of modelling consumer behavior contradicts

[2]Feldstein (1973), Feldman and Dowd (1991), Buchanan *et al* (1991), Newhouse (1993) and Manning and Marquis (1996).

one of the basic findings from the RAND Health Insurance Experiment. In fact, Keeler and Rolph (1998) and Newhouse(1993) found that generosity of insurance plans mainly influence the decision whether to seek treatment or not against an illness episode, rather than treatment costs. This is not strange given the asymmetry of information between patient and physician. In our model, individual will decide whether to have treatment or not against an illness spell with some level of severity. The costs of treatment will be given to the individual as a technological relation. Our choice, though more complicated from an econometric point of view, allows to disentangle quantity from costs and mimics better the decision pattern found in health care consumption by Keeler and Rolph (1998) and Newhouse(1993). An important advantage from this approach is that we can exploit the stochastic relation between costs and health shocks when we solve for the optimal contract. If treatment costs tell much about health shocks, then the problem of moral hazard is alleviated, since the insurer can infer the value of the non contractible variables from the contractible ones.[3] This will be our basis to propose a new measure of moral hazard based on the correlation between unobservables that health shocks and the ones that influence cost of treatment. Consequently, this measure is based on the informational content that contractible variables have over non contractible ones.

Our second main difference with previous literature is that we derive the optimal contract from first principles, that is the solution of the principal-agent problem. Consequently, we do not have to do any assumption regarding which is the first best level of health care consumption, as previous approaches have to do.[4] As it is explained above, this is specially important when income effects are present. As we have emphasized before, if income effects are important, previous approaches might have overestimated welfare losses (Meza, 1983). It is important to have a methodology to assess consumer cost-sharing in presence of income effects, since it is not clear yet if they are a prevalent feature of the health care insurance market.[5]

[3]Buchanan *et. al.* (1991) do dissentangle quantity consumed and health care costs, but they disregard the correlation between contractible and non contractible variables and they do not obtain the optimal insurance plan from first principles.

[4]Blomqvist (1997) is an exception, but our model differs from hers in the rest of dimensions we have mentioned before.

[5]Evidence of income effects have been found by Newhouse (1993), Pohlmeier and Ulrich (1995), Manning and Marquis (1996), Gurmu (1997), Holly *et al.* (1998) and Creel and Farrell (2001), while Deb and Trivedi (1997) does not find a statistically significant effect.

Consumer cost sharing is an important policy issue. In the last US presidential elections, an increase in the coverage of pharmaceuticals for the elderly was part of President Bush electoral program. On July 1993 in France, the social security cost sharing for ambulatory care and pharmaceutical goods was reduced in a 5% (Chiappori *et al.* 1998). In an attempt to control rising public health care costs, the Belgian government has raised coinsurance rates several times over the period 1986-1995 with a sharp increase in 1994 (Van de Voorde *et al.* 2000). In the UK, policy change has reduced eligibility for publicly provided treatment, increased copayments for dental, opthalmic services and pharmaceuticals (Propper, 2000) and reduce state financed care at the margins (Propper *et al.*, 2001). In Spain, doctor visits are free but consumers are charged 40% for outpatient drugs charges.

Along this introduction we have used the term *reimbursement* health care insurance. We are studying the case where the consumer gets reimbursed her health care expenses according to the insurance contract, but there is no contractual relation between health care provider and insurer. The contract that we obtain is contingent on a cost function that is partly determined by the incentives that the provider faces. If provider incentives change, this will change our cost function and then the optimal contract. Although it would be desirable, we do not have data available to determine the optimal mix between provider and consumer incentives.[6] However, our model points out how this could be achieved. This is thanks to the separation between quantity demanded and treatment costs. Close to this, we focus on the contract for outpatient non dental and non mental treatments. As Belsey (1988) points out, there is no reason to think that all kinds of health care services should have face the same contract. Dental care is normally subject to higher degrees of patient cost sharing than standard outpatient treatment, while hospitalization expenses usually have much wider insurance coverage. We highlight that previous papers used all types of treatments in order to obtain the insurance contract.

The paper is organized as follows. In Section 2 we describe how the individual chooses to have treatment or not when suffering an illness episode by comparing out of pocket costs of treatment and the health penalty of not obtaining treatment. This follows the tradition health economics research in which the consumer is responsible for the contact decision, but not for the amount of treatment. In Section 3 we describe the data that we use, that comes from the RAND Health Insurance Experiment, that randomly assigned individuals

[6]Ellis and McGuire (1993) argue that demand side cost-sharing might be desirable even in Health Maintance Organizations.

to insurance plans. Consequently, for the estimation, we do not need to assume that observed contracts are optimal and we can condition on the insurance contract without the need of dealing with endogeneity issues. Section 4 describes the econometric strategy followed to estimate the model. Section 5 gives the results of a preliminary analysis and Section 6 discuss the estimates of the structural parameters, and evaluates the suitability of the model. Section 7 sets up the principal-agent problem and explains the numerical technique to solve for the optimal contract. The last section concludes.

2 The demand model

2.1 Individual decision problem

This section is devoted to modeling the individual's decision concerning whether to be treated or not when suffering an illness spell. This model is the basis for the estimation of the parameters of the principal-agent model. In our set-up, the consumer faces a specific insurance contract that will influence her decision. Given that in our dataset, individuals were randomly assigned to insurance contracts, we do not describe how individuals chose among different insurance choices.

Our model is inspired in the baseline model of Ma and Riordan (1997). At the end of this section, we will specify the difference between our model and theirs. We build the model based on the following hypotheses: First, the individual decides whether to be treated or not but she does not decide the cost of treatment. We justify this on the basis of the informational asymmetry between doctor and patient. We will also assume that the doctor chooses treatment costs independently of individual's insurance coverage and income. This corresponds to the situation where the medical guideline that the doctor follows does not take into account individual economic characteristics but gives the most cost-effective treatment. Consequently, we will assume that treatment costs come from a given technological relation.

Second, the individual is rational and compares benefits and costs when she decides. This might be a strong assumption when one is dealing with severe illnesses for which the individual lacks experience and can hardly value the benefits and costs of the treatment. Furthermore, the treatment decision in the case of very severe illnesses might depend on long term effects that we are not ready to model. In the empirical application we will restrict the type of illness spells that we study in order to make our rationality assumption

is more likely to hold.

We assume the individual is endowed with a health capital stock $\overline{s}$, and an income level y. An individual will be ill when she receives a penalty health shock of magnitude $s > 0$. There will be a stochastic relation between s and $\overline{s}$, given by the conditional density function $f_{s|\overline{s}}$. If $s \leq 0$, the individual is not ill and therefore treatment is not demanded. We assume the support of s is $(-\infty, +\infty)$ .With this specification the illness process follows the same determinants as the health penalty process. We will justify this in terms of requirements for identification in the section devoted to the econometric specification.

The individual is under the coverage of an insurance contract that under premium p reimburses the cost of the treatment c, if the individual agrees to pay the quantity $D(c)$, where $D(.)$ is a non-negative function specified in the insurance contract. This function gives the cost sharing agreement between insurer and insured. The health penalty shock and cost of treatment variables will follow a joint density function conditional on the health stock, $\overline{s}$, and cost shifters $\overline{c}$ given by

$$g_{s,c|\overline{s},\overline{c}} = \begin{cases} 0 & \\ f_{s|\overline{s}} & if\ s \leq 0\ and\ c = 0 \\ f_{s,c|\overline{s},\overline{c}} & if\ s > 0\ and\ c > 0. \end{cases} \tag{1}$$

The first part of the density function implies that costs cannot be positive when the individual is not ill ($s \leq 0$), since there is no need for treatment.

The timing of the model is as follows. First the individual receives a random draw of (s, c) from the joint density $g_{s,c|\overline{s},\overline{c}}$. When the individual suffers an illness spell ($s > 0$), she will obtain different utilities depending on her decision regarding treatment. We will assume that the individual knows the health penalty s and the cost c when she decides. We reckon that this might be a strong assumption but it is difficult to figure out an estimable demand model where the consumer knows neither the price nor the good she is buying. In the empirical application, we restrict the type of illness spells that we consider in order to make this assumption more plausible. We also assume that in case the treatment is obtained, there is perfect healing and the initial level of health is recuperated. As before, we think that restricting the types of illness spell that we consider might help to make this assumption more plausible. If treatment at cost c is obtained, the individual will have to pay the quantity $D(c)$. Specifically, the consumer's ex-post utility will be

$$\begin{array}{cc} U(y-p,s), & \text{if ill with health penalty } s \text{ but treatment is not obtained} \\ U(y-p-D(c),0), & \text{if ill and treatment is obtained with } D(c) \text{ as out of pocket payment} \\ U(y-p,0), & \text{if consumer is not ill.} \end{array} \tag{2}$$

We assume that $U(.,.)$ is increasing and concave in the first argument, while decreasing and convex in the second.

In what follows, we will describe the individual decision problem. Given that the set of actions is discrete (to have or not to have treatment), it is of interest to look for the health penalty threshold, that given a cost, leaves the individual indifferent between having treatment or not. For each cost c there is a unique value of $s = \widetilde{s}(c) \geq 0$ such that

$$U(y-p,\widetilde{s}(c)) = U(y-p-D(c),0).$$

Given c, the uniqueness comes from $U(y-p,0) > U(y-p-D(c),0)$ and the fact that the utility function is strictly decreasing in the second argument. Consequently we make the following definition.

Definition 1 *The health penalty threshold is the function $\widetilde{s}(c)$ such that $U(y-p,\widetilde{s}(c)) = U(y-p-D(c),0)$.*

Given a draw of (s,c) from the joint distribution above, the individual will decide to have treatment ($T=1$) or not ($T=0$), according to the following rule:

$$T = \begin{cases} 1 & if\ s > \widetilde{s}(c) \\ 0 & if\ s \leq 0\ or\ 0 < s < \widetilde{s}(c). \end{cases} \tag{3}$$

The intuition is very simple. The individual will decide not to have treatment ($T=0$) either when she is not ill ($s<0$) or when the health penalty shock does not offset the out of pocket monetary cost of the treatment ($0 < s < \widetilde{s}(c)$). Notice that $\widetilde{s}(c)$ depends on the cost sharing function $D(c)$. In particular, if the contract does not specify any out of pocket payment, $D(c) = 0$ for all c, then the individual will decide to have treatment when she is ill independently of the costs of the treatment.

As we said at the beginning of the section, the model is inspired in the baseline model by Ma and Riordan (1997). However, there are some differences. In our model, the illness probability and the health penalty shock are drawn from the same distribution, while in their model, both processes are modelled independently. We need this assumption in order to be able to identify the parameters in estimation. However, in our model, costs are random Ma and Riordan (1997) insurance model is for a specific illness and they assume that costs are fixed and do not vary with the health penalty shock.

2.2 Expected results on observed costs per episode

The structural model above allow us to analyze how the observed costs per episode vary with the copayment rates. To our knowledge, the literature has not discussed this issue using a theoretical demand model before. We will see that observed costs per episodes are decreasing in the generosity of insurance coverage, even if neither the patient nor the doctor decides to spend less in a treatment. The following lemma and proposition clarifies our point:

Lemma 2 *For a given premium, p, and for every cost, c, given $D(c) = kc$, the health penalty threshold function $\widetilde{s}(c)$ is increasing in the copayment, k.*

Proof. This result directly comes from Definition 1. The greater is k, the smaller is the value of $U(y - p - kc, 0)$, and consequently, s must be greater for the equality in Definition 1 to hold. ■

This means that the greater the copayment, the greater must be the health penalty in order to ask for treatment. In the following proposition we will see that this implies that observed costs are expected to be smaller for those individuals with greater copayments.

Proposition 3 *For a given premium, p, observed costs of treatment are decreasing in the copayment rate.*

Proof. Observed costs are given by the following expression:

$$E[c|k, p, T = 1] = \int_0^{+\infty} \int_{\widetilde{s}(c)}^{+\infty} c f_{s,c|\overline{s},\overline{c}}(s, c) \partial c \partial s.$$

Using Leibinz's rule, the derivative the expression with respect to k is:

$$\frac{\partial E[c|k, p, T = 1]}{\partial k} = -\int_0^{+\infty} c f_{s,c|\overline{s},\overline{c}}(\widetilde{s}(c), c) \frac{\partial \widetilde{s}(c)}{\partial k} \partial c < 0,$$

given that $f_{s,c|\overline{s},\overline{c}}(\widetilde{s}(c),c)$ is positive since it is a density function, and the derivative of $\widetilde{s}(c)$ with respect to k is positive according to the previous lemma. ■

The intuition is as follows: given a health penalty, those with a large copayment will ask for treatment only if the treatment is inexpensive enough. Consequently, observed costs are expected to be smaller for those individuals with greater copayments. Notice that this holds in a framework where nobody chooses treatment costs, but they are given by a technological relation. The key assumption is that individuals are able to anticipate treatment costs. We will see in Section 5 that the data verify this relation.

3 The Data

3.1 The experiment

The data that we use come from the RAND Health Insurance Experiment (HIE), a social experiment conducted between 1975 and 1982 in six different cities of USA. Families participating in the experiment were randomly assigned to one of fourteen different fee-for service health insurance plans. More information about the experiment can be obtained in Manning *et al.* (1987) and Newhouse *et al.* (1993). Keeler and Rolph (1988), Manning and Marquis (1996), Marquis and Holmer (1996), Deb and Trivedi (1999) and Guilleski and Mroz (2000) have previously used this data.

We would like to highlight two important characteristics of the dataset. First, insurance plans are exogenous to the individuals. Individuals did not choose the insurance plan they were enrolled, but they were randomly assigned to it. Participation incentives were paid to minimize the risk of attrition bias. Therefore the analyst does not need to face the endogeneity problem of insurance coverage, that is, the possibility that less healthy people, anticipating large medical expenditures, buy more generous insurance coverage.

Second, it gives information on illness episodes. Charges on claims from providers were grouped to create episodes of treatment. The grouping was based primarily on diagnosis, time since last charge for a related diagnosis, and information from the provider on treatment history. For each episode, the dataset contains information on total expenses, beginning and ending date, as well as the type of episode and the principal provider of the service. According to the type of the episode, they were classified as acute, chronic, chronic flare-up,[7] well-care and prenatal/maternity. The classification for providers include

[7] Temporary problem in a usually controlled condition.

hospital inpatient, hospital outpatient, physician, dentist, pharmacy and nonpharmacy supplier. The classification of the provider is hierarchical, that is, outpatient services preceding or following a hospitalization were included in a hospital provider. Drugs and tests were part of the episode in which they were prescribed. This exhaustive information about episodes is important for us, since we are using a behavioral model that assumes that patients know the cost and severity of the treatment. This will be useful when we restrict the type of illness spells that we will use. More information on the way episodes were constructed can be obtained from Keeler and Rolph (1988) and references thereinafter.

Using information on episodes has an important advantage: one can separately study decisions to start episodes of treatment and decisions on the amount of treatment (cost of treatment). The decision to start an episode of treatment is mainly done by the patient, while the doctor, or the doctor jointly with the patient, will decide on the amount of treatment (or cost). Therefore, it is desirable to be able to study these two decisions separately.

The fee-for-service plans of the experiment had different levels of cost sharing that varied over two dimensions: the copayment rate (the percentage of the cost of each insurance claim that the individual pay out of her pocket) and an upper limit on annual out of pocket expenditures called Maximum Dollar Expenditure (MDE). Consequently, the family only paid according to the copayment if the total out of pocket expenditures had not exceed the MDE. The copayment rates were 0, 25, 50 or 95 percent. Depending on the plan, the MDE was 5,10, or 15 percent of the previous year's income, with a maximum of \$1000.[8] In section 3.3, we will comment on the consequences of this MDE.

3.2 Our sample

In this subsection we will make specific the sample we have used in our estimation. We will begin to comment on the types of episodes used. In our structural model, the individual might receive a health penalty shock. This is not compatible with well care episodes nor with prenatal/maternity episodes. Nor are chronic treatments compatible with our model, since they are routine treatments. Therefore, the individual will predict them at the beginning of the contract year. We will not consider inpatient episodes since they

[8] Apart from the ones mentioned above, there was a HMO plan, as well as an individual deductible plan that limited annual out-of-pocket outpatient expenditures to \$150 per person, with a 95% of copayment rate for outpatient expenditures. These ones will not be used in this study.

are probably too severe and the individual could hardly decide on the basis of economic factors. We do not consider either dental episodes, since they were not usually covered by insurance contracts in those years and consumers showed an opportunistic behavior during the experiment. We do not analyze mental care episodes since they have different determinants and, even with generous insurance coverage, relatively little was spent on outpatient mental health care at the time the experiment was conducted.

Consequently, we have only considered acute and chronic-flare-up outpatient episodes. Acute episodes (Newhouse *et al.* 1993) are defined by unforeseen and undeferrable treatment opportunities. From an economic point of view, spending on these episodes will occur only when the patient is temporarily sick. Consequently our model might be appropriate to deal with acute episodes. Chronic flare-up episodes are much as acute episodes, but they are caused by a chronic condition. Since we have not considered inpatient episodes, most of the episodes will be because of relatively minor conditions. It is important to have a relatively homogenous types of illness episodes, since in general the consumption of different health care commodities will imply different insurance contracts associated with them (Belsey 1988).

A subsample of individuals were enrolled in the experiment during five contract years (contracts had a duration of 365 days but not natural years, so they were called contract years), but most of them were enrolled only three contract years. We will use observations on the third contract year, so we are sure that individuals were familiar with the conditions of the insurance contract offered by the experiment.[9]

Our model does not allow for multiple spells. As we will see, the econometric implementation needs to account for sample selection in costs since we do not observe the treatment costs of those that were ill but did not seek treatment. Consequently, the model is already quite complicated in order to account for multiple spells. Therefore, our dependent variable will be a dichothomous variable that takes value 1 if the individual started to be treated by a new episode during the first month of the contract year.[10] As Table 4.1 indicates few people started more than one episode in the first month. As Guilleski

[9]In the section devoted to preliminary analysis, we explain why we chose the third year instead of the second one.

[10]More specifically, during the second day of the contract-year and the following 30 days. Medical expenses in the new contract year that were generated by illness episodes that started in the preceeding contract year were recorded as starting the first day of the new contract year. We want to exclude them, since they are not decisions taken in the current contract year.

(1998), in case of starting more than one episode in the month, we take the first one. Therefore, we only have one observation per person.

Moreover, we only consider people older than 17, in order to consider episodes which are decided by the individual and not by their parents.[11] We do not consider either people that were self-employed. They were very few in the sample and might differ much in their opportunity costs of having a treatment. Since we do not have information about this opportunity cost, we prefer not to include them in the analysis. Finally we deleted observations with missing values in relevant variables. Table 4.3 gives the description of the variables used, as well as the descriptive statistics of the observations used in the estimation. All the monetary variables are in 1973 dollars.[12] We used monthly medical consumer price index to deflate treatment costs. Individuals participating in the experiment did not pay any insurance premium.

3.3 The maximum dollar expenditure and the copayment rate

In our estimation we will condition on the individual's copayment rate. Therefore, it is important that the copayment rate we condition on is the same one that the individual uses to decide on demand for treatment. As Keeler *et al.* (1977), Ellis (1986), Keeler and Rolph (1988) and Newhouse *et al.* (1993) have noted, the existence of a cap on out of pocket payments may make the effective marginal price and the nominal price differ (the copayment rate we condition on). If individuals were able to anticipate exceeding the cap with certain probability, then the effective marginal price will be smaller than the nominal one.

We have reasons to think that in our case this is a minor problem. First, according to Newhouse *et al.* (1993) "Few participants [in the HIE experiment] proved able to anticipate exceeding the MDE, which allowed us to ignore this factor and to obtain a much more tractable estimation problem". In addition, they show that the size of the remaining MDE was important in the decision to initiate hospital episodes, but not to initiate other episodes type. Consistent with the same idea, Keeler and Rolph (1988) showed that people participating in the experiment adopted a mixture of myopic and inflexible behavior. That is, if expenditures did not exceed the MDE, then they responded to current copayments. Once out of pocket expenditures exceeded the MDE, people did not instantly adapt to

[11] Eligible individuals for the experiment were younger than 61.

[12] We have deflacted because we have several years, as noted in the beginning of this section

the zero price of medical care, but they took some time to do so.

Our analysis is based on the first month and for episodes that took as much one month. Given the references above that apply to the same data that we use, families should be far from either exceeding the MDE or anticipating to exceed the MDE. Therefore, we feel confident on conditioning on the current copayment rate.

4 Econometric specification

4.1 Functional form assumptions

In this section, we will give specific functional forms to the demand model described above in order to carry out the estimation. We will start by the density function. The health penalty random variable, s, follows a normal distribution with mean $\overline{s} = x_s\beta_s$. That is,

$$s = x_s\beta_s + \varepsilon_s,$$

where x_s is a vector of covariates, β_s a conformable vector of parameters and ε_s is an unobservable component that follows a normal distribution with zero mean and variance σ_s^2. Consequently,

$$f_{s|\overline{s}} = \frac{1}{\sigma_s} * \phi(\frac{s - x_s\beta_s}{\sigma_s}),$$

where $\phi(.)$ is the standardized normal density function.

The cost per episode will be given by

$$\ln C = \alpha \ln(s+1) + x_c\beta_c + \varepsilon_c,$$

where x_c is a vector of covariates, β_c a conformable vector of parameters and ε_c is an unobservable component. Notice that the individual is ill only when $s > 0$, so the function above is well defined in the appropriate range. Inside the logarithm in the left hand side, we add one to the health penalty. This ensures that the cost increases with α for any $s > 0$. The functional form above is convenient since the predicted cost is always positive for any value of the parameters. All the previous studies cited above that estimated cost equations also used a logarithm transformation.

It will be assumed that $(\varepsilon_s, \varepsilon_c)$ follows a bivariate normal distribution with null means and variance covariance matrix given by:

$$\Sigma = \begin{bmatrix} \sigma_s^2 & \rho\sigma_s\sigma_c \\ \rho\sigma_s\sigma_c & \sigma_c^2 \end{bmatrix}.$$

By means of a change of variable we can obtain the joint density of (s, c), when both variables take positive values:

$$f_{s,c|x_s,x_c} = \frac{1}{c\sigma_s\sigma_c} * b(\frac{s - x_s\beta_s}{\sigma_s}, \frac{\ln c - \alpha \ln(s+1) - x_c\beta_c}{\sigma_c}; \rho), \tag{4}$$

where $b(.,.;\rho)$ is the standardized bivariate normal with correlation coefficient equal to ρ.

We will specify two types of utility functions, previously proposed by Ma and Riordan (1997). When the individual is ill with health penalty s, but does not have treatment, the utility function will be either,[13]

$$\begin{aligned} U_I(y-p,s) &= U(y-p) - s, \text{ or} \\ U_{NI}(y-p,s) &= U(y-p-s). \end{aligned}$$

In case of having treatment the utilities will be

$$\begin{aligned} U_I(y-p-k*c,0) &= U(y-p-k*c), \text{ or} \\ U_{NI}(y-p-k*c,0) &= U(y-p-k*c) \text{ accordingly.} \end{aligned}$$

Using U_I, given a cost, the individual will have an episode of illness treated when the health penalty is larger than the health penalty threshold previously defined:

$$s > \widetilde{s}_I(c) = U(y-p) - U(y-p-k*c), \tag{5}$$

where k is the copayment rate. Notice that income, y directly influences $\widetilde{s_I}(c)$, therefore income will influence the decision to have treatment. In this case, we will say that there are income effects. Notice that in this case, income and health are not directly comparable, since the comparison is done by means of the marginal utility of income. Since the existence or not of income effects is a controversial issue in the literature, it is desirable to consider a specification where there are no income effects. The second utility function is useful for this purpose. For the second utility function the individual will have treatment when:

$$s > \widetilde{s}_{NI}(c) = k * c. \tag{6}$$

In this case, health and money are directly comparable and there are no income effects, that is, income does not influence individual demand for treatment.

[13] As explained below, the subscript I will refer to income effects, while NI refers to No-Income effects.

We still must give a functional form for $U(.)$. The exponential utility function is very convenient for our purposes. Ferrall and Shearer (1999) comment that from a computational standpoint, the exponential utility is perhaps the only feasible functional form when solving the principal-agent problem.[14] Support to the exponential utility is given by Manning and Marquis (1996) and Marquis and Holmer (1996). Also using data from the RAND HIE experiment, they both argue in favor of the constant absolute risk aversion hypothesis. Therefore, based on both the complexities that arise using different specifications and the support found in favor of the constant absolute risk aversion hypothesis, we will use

$$U(z) = -\exp(-\theta z),$$

where θ stands for the constant absolute risk aversion coefficient. Marquis and Holmer (1996) found that θ did not vary with income or other demographics, so we will not parametrize it as a function of individual characteristics.

4.2 The likelihood function

Estimation will be done by maximum likelihood. Here we will give its formulation. Whether the individual had treatment or not gives us different information. For individuals with treatment, we can observe the cost of the treatment, however this is not the case for individuals without treatment. This is a sample selection issue that the model must accommodate.

The likelihood contribution for individual i if she had treatment ($T = 1$) at a cost $c > 0$, will be

$$L_{1,i} = \int_{\widetilde{s}_j(c_i)}^{+\infty} \frac{1}{c_i \sigma_s \sigma_c} * b\left(\frac{s - x_{si}\beta_s}{\sigma_s}, \frac{\ln c_i - \alpha \ln(s+1) - x_{ci}\beta_c}{\sigma_c}; \rho\right) \partial s, \qquad j = I, NI. \tag{7}$$

The integral above is the probability of having treatment at cost c_i. This is the area of the density function (4) in the region at which it is optimal to have treatment. As given by (5) or (6), this occurs when the health penalty is large enough. We emphasize that, in general, the integration limit depends on the utility function used.

[14] Others authors that have used the exponential utility function are Towsend (1994), Mace (1991), Haubrich (1994) and Marigotta and Miller (2000).

The likelihood contribution for individual i if she did not have treatment ($T = 0$) will be

$$L_{0,i} = \Phi(-\frac{x_{si}\beta_s}{\sigma_s}) + \int_0^{+\infty} \int_0^{\widetilde{s}_j(c)} \frac{1}{c\sigma_s\sigma_c} * b(\frac{s - x_{si}\beta_s}{\sigma_s}, \frac{\ln c - \alpha \ln(s+1) - x_{ci}\beta_c}{\sigma_c}; \rho)\partial s \partial c, \qquad j = I, NI. \tag{8}$$

where $\Phi(.)$ denotes the cumulative distribution function of the standardized normal. This first term is the probability of not being ill, while the second is the probability of being ill but with a health penalty not large enough to offset the monetary out of pocket costs. The cost has to be integrated out, since it is not observed for those that did not have treatment.

Individuals with a copayment rate of zero, $(k = 0)$, do not pay anything for the treatment. In this case, $\widetilde{s}_j(c) = 0$ for all c. In this case, since the individual does not pay anything, the second term of (8) vanishes and the integral in (7) has a lower limit of zero. That is, the probability of having treatment is the probability of being ill. In this case, even if we use the utility function that exhibits income effects, income will not influence the demand decision. The model that does not exhibit income effects is robust to the choice of the functional form $U(.)$ since it does not enter in $\widetilde{s}_{NI}(c)$.

The log-likelihood function is given by

$$\ln L = \sum_{i=1}^{N} 1[T_i = 1] * \ln(L_{1,i}) + \sum_{i=1}^{N} 1[T_i = 0] * \ln(L_{0,i}).$$

Computation of the log-likelihood function requires numerical integration. The integration over costs in done using Hermite quadrature which is specially convenient to deal with integrands related to the normal density function (Judd 1998). To compute this integral, one needs to write the likelihood in terms of $\exp(-\xi^2)$ where ξ is the variable to integrate. We achieve this by doing the following change of variable:

$$\xi = \frac{\ln(c) - x_c\beta_c}{\sigma_c}.$$

As a result, we obtain integration limits that go from minus infinity to plus infinity, which is an additional requirement to apply Hermite quadrature. The integral over s is done by Legendre quadrature. This integration routine requires fixing the limits of integration to finite quantities. Therefore the plus infinity in (7) was replaced by the mean of s plus eight times its standard deviation. Since s is a normal random variable, this is

practically the same as integrating up to infinity. All the integration routines used are Gaussian quadrature techniques that outperform the alternative Newton-Cotes formula (Judd 1998). We used 16 points of quadrature in the Hermite integration and 24 in the Legendre.[15]

4.3 Identification

First of all, we would like to comment on the restriction imposed in the theoretical model that both the illness process and the health penalty are generated by the same random variable s. In order to consider a more general formulation where illness and health penalty were generated by different processes we would need to observe a variable indicating that the individual was ill but decided not to have treatment. Since we only observe to have treatment or not, we have to constrain to a unique process that determines both illness and health penalty.

In order to consider more issues related to identification we will express the model in the following way:

$$\Pr(T=1|c) = \Pr(s > \widetilde{s}_j(c)), \quad j = I, NI \tag{9}$$

$$\ln c = \alpha \ln(s+1) + x_c\beta_c + \varepsilon_c, \ if \quad s > 0, \text{ where}$$

$$s = x_s\beta_s + \varepsilon_s.$$

$$U_j(y-p, \widetilde{s}_j(c)) = U_j(y-p-kc, 0). \tag{10}$$

Therefore it is a simultaneous equation model in which the dependent right hand side variables are not censored and therefore we do not require any restriction for coherency of the system. In order to obtain non-parametric identification, it is necessary to have at least one continuous restriction in both x_s, and x_c.

Another issue of concern is identification of the parameters of the discrete choice equation. For the model that does not exhibit income effects we have that equation (9) is

$$\Pr(T=1|c) = \Pr(u_s > \frac{kc - x_s\beta}{\sigma_s}),$$

where $\varepsilon_s = \sigma_s u_s$, k is the copayment rate and therefore it is not a parameter to estimate, but a data we condition on. Consequently, if there are enough observations with $k > 0$

[15]The Hermite routine needs less points since it is designed to deal with normal related integrands as ours. The Legendre is a more general purpose routine so we have chosen more points for the Legendre than for the Hermite.

then, σ_s is identified because the product kc changes among individuals. However, if there are not enough elements in the sample for which $k \neq 0$, this might result in poor identification. For the model that exhibits income effects, equation (9) is

$$\Pr(T = 1|c) = \Pr(u_s > \frac{\exp(-\theta y) * (-1 + \exp(\theta kc)) - x_s\beta_s}{\sigma_s}). \tag{11}$$

In this case, the product kc is affected by the parameter θ and therefore the previous argument does not hold. Notice that since θ enters non linearly in the fraction, this might allow us to identify the parameters of the model. However, identification that relies on non linear restrictions can give poor results. This is the case in our model and we have decided to constrain one of the parameters. We use the value of θ estimated by Marquis and Holmer (1996) that used also data from the RAND HIE. However, we will not constrain θ directly, rather we will constrain for a value of σ_s that gives as a corresponding point estimate for θ equals to the estimated value by Marquis and Holmer (1996). With this strategy, all the observations in the sample will contribute to the identification of the parameters. If on the contrary, we constrained the value of θ directly, the observations with $k = 0$ will not contribute to the identification of the parameter because in that case the first term of the numerator of (11) vanishes.

5 Preliminary analysis

In this section we will analyze Tables 3 to 6 that show the results of a preliminary analysis of the data. Table 3 shows how frequency of episodes treated varies with copayment rates. Those that enjoy free care seek care more often than those that face cost-sharing contracts. Differences among copayment plans are not so clear. This is not strange since Newhouse *et al.* (1993) found that the largest decrease in use of outpatient services occurs between the free and the 0.25 copayment rate.[16] From Table 3, it is clear that the copayment group with less observations are those with 50% copayment. Table 4 shows the estimates of a standard probit model for TREAT as a dependent variable. The results on the copayment rates give us basically the same information that Table 3. The coefficients of the dummy variable for copayment groups are negative and statistically significant, indicating that copayment influences the probability of having treatment in the expected way. Apart

[16] Furthermore, in table 4.8 in Newhouse *et al.* (1993) the coefficients over the 0.5 and 0.95 copayment rates were very similar (-0.46 vs -0.49) in a Negative Binomial regression of number of acute episodes.

from copayments, other significant variables are FEM and GHEA. As expected, women are more likely to have episodes treated then men. People with poor health conditions also have episodes treated with greater frequency. AGE and DISEA have expected signs but they are not statistically significant. In fact, GHEA is an index of general health that might measure health capital better than other more crude proxies as age or the number of chronic diseases. On top of this, we are only considering people older than 17 and younger than 62, what might be contribute to not to find a significant effect of age.

Table 5 shows how the average of episode treatment cost for those treated varies with copayment rates. As our theoretical demand model predicted in section (2.2), it seems clear that the higher the copayment rate, the lower is the average observed cost. According to our model, the higher the copayment rate, the more inexpensive the treatment has to be in order to ask for treatment, given an illness episode of the same severity. Therefore, one expects to observe this pattern in the costs per episode treated.

Table 6 shows a OLS regression for $\ln(c)$ over covariates and dummies for copayments, for those people that had a episode treated in the first month of the third year. Age have a positive and significant effect on the logarithm of costs per episodes. Since we have 248 observations on this regression it is hard to find significant effects, given the variability that medical costs usually have. APPT, AVSP and AVMD are the individual's satisfaction with the length of time for a medical appointment, as well as satisfaction with medical specialists and family doctors respectively. We interpret them as capacity indicators: a higher satisfaction implies a higher capacity and then smaller costs. The signs are consistent with this interpretation and APPT is statistically different from zero at 14%. The effect of copayments in the logarithm of observed costs per episodes have the expected sign, but not significantly different from zero except for the one with 50% copayment.[17] Keeler and Rolph (1988) and Newhouse (1993) obtain the same conclusion using data from all the sample period, and not only the first month as we do here.

In this section we have accomplished two objectives. First, we have assessed which covariates are statistically significant for the two dependent variables of interest: treatment decision and cost per episode. Second, we have checked that though we restrict ourselves to the first month of the third year of the experiment, we obtain the same conclusions that previous analyst have obtained using the whole sample period: copayments mainly

[17] As it is expected from the random assignment of copayments, the conclusions are the same if we do not use any covariates in the regression except the copayment dummies.

influence the decision to have medical episodes treated while copayments very weakly influence costs per episodes Keeler and Rolph (1988) and Newhouse (1993).[18]

6 Structural estimation results

6.1 Model results

In this section we will comment on the results obtained in the estimation of the model developed in section (4). Firstly, we comment on the exclusions restrictions made in X_s, X_c which are necessary to non parametrically identify the model. The variables APPT, AVMD and AVSP are satisfaction variables related to the capacity of doctors in the area where the participants live. Assuming that capacity translates to cost of treatment, in the sense that a larger capacity implies smaller costs because of more severe competition in the market, then these variables should not influence health, conditional on costs. Hence, we do not include them in X_s. APPT, AVMD and AVSP are good instruments since its inclusion in the OLS regression of costs presented in the previous section doubled the value of the R^2 statistic. From X_c we exclude SOC, which is a social contact indicator, since it did not turn out significant in the OLS regression presented in the previous section and its t-statistic in the probit regression is not very low (-1.46).[19] On top of this, we do not have any a priori reason to think why an index of social contacts might influences costs of treatment, nor we have found evidence of this in the literature.

We estimated the structural models presented in section (4), both the model with income effects and the model without income effects. As we mentioned above, the scale of the model with income effect is not identified, unless a parameter is fixed. We decided to fix the risk aversion coefficient to $\theta = 0.00309$, which is the value estimated by Marquis and Holmer (1996) for the same dataset that we are using.[20] We found α to be negative

[18]When we repeated the exercise for the first month of the second year, we found significant effects of copayments on costs. This made us think that the first month of the second year was not a good representative of the information contained in the experiment, and the we decided to concentrate on the third one.

[19]The fact that a variable is not significant in the OLS regression for costs presented in the previous section does not directly implies that it was not in the cost function of our structural model. In the structural model we have a cost function unconditional on the event of treatment, while the OLS regresion of the previous section was conditional on that. Still we think that the OLS results are a useful guide to specify the structural model.

[20]The value of 0.00309 is expressed in 1973 dollars, as income and costs are.

and not statistically significant from zero (t-values:-0.761 and -0.632). We checked the robustness of our results to functional form, but the negative sign persisted. Given that it is quite implausible to think that costs decreases with health severity, we think that the negative sign is a indication that the size of the true effect must be quite small and consequently we decided to estimate both models with α restricted to zero. We highlight that this does not mean that health severity does not influence costs. Costs and health severity are still related through their correlation between unobservables: ε_s and ε_c.

Tables 7 to 9 show the estimation results of the structural models. Table 7 reports the estimates of the severity equation. The two models measure severity in a very different way: the model with income effect measures severity in utility units of consumption, while the model without income effect measure it in monetary terms. That is the reason why the size of the coefficients differ very much. As in the probit regression presented in the previous section, the statistically significant variables are FEMALE and GHEA. Other variables as AGE and DISEA have the expected signs, but they are not statistically significant from zero at 95% confidence level.

Table 8 shows the results for the cost equation. In line with the OLS regression of the previous section, we do not have significant variables, except the constant term and AVSP in the model for the no income effect model. We would like to highlight that even the OLS regression showed this lack of significance, hence it is not due to our structural model but to the number of observations with positive costs (248). Whether this is a major problem or not depends on the use of these estimates. We do not plan to assess how changes in the covariates influence our policy function (optimal contract), but to obtain treatment cost for a representative individual with covariate values the sample average. Hence, we do not see it as a major inconvenience.

Table 9 presents the value of ρ, the correlation coefficient between the unobservables: ε_s and ε_c. We highlight that in both models it is positive, quite high and statistically different from zero. Consequently, our model predicts that costs increase positively with severity in a stochastic fashion.[21] As we will explain in depth below, this high correlation means that the insurer can learn much about the severity by observing the cost.

[21]This is not due to restricting alpha to be zero. In the unrestricted model, the correlation coefficient took a value of 0.78 and 0.85 in the income and no income effect model respectively.

6.2 Model evaluation

The purpose of this section is twofold: to compare the relative performance of both models, as well as to analyze the suitability of the structural models.

We have computed the value of Andrews' (1988) goodness of fit test for both models. This is a conditional moment test for the difference between the average estimated probability of TREAT=1 and the frequency of ones in the sample. We obtained a P-value of 0.70 for the model with income effects and a P-value of 0.87 for the model without income effects, consequently none of the structural models is rejected at usual confidence levels by this test.

In Tables 10 and 11, we show the results of montecarlo simulations based on 5000 replications of the model using our sample. In the second column of both tables 10 and 11 we report the frequency of treatment and average observed costs by copayment rates that we find in the data. Under the columns labelled "Mean" we report the point predicted value that we obtain using the simulations. The P2.5 and P97.5 columns refer to the 2.5 and 97.5 percentiles of the series of simulated values. They cannot be interpreted as confidence intervals since they just pick up the uncertainty of the error terms, but not the one related to the parameter estimates.[22] Therefore, it is likely that the confidence intervals are wider.[23] In table 10, we do not observe any large difference between the mean frequencies of treatment and its real value. The larger discrepancy is for 50% copayment group, however this is group with smaller sample size. Overall, the model with no income effects fits better the frequencies of treatment as expected from the selection test previously presented. Table 11 give the equivalent results for the other dependent variable: observed costs. For the 0% and 25% copayment group, the mean predictions of the model are very close the to the real value. The prediction for the 50% copayment group presents the largest discrepancy, though this is the group with smaller sample size, thus the uncertainty around the average real value is also larger. The point prediction for the group with 95% copayment is a bit closer to the real value for the model with income effects, that for the model without income effects. However, the real value is within the uncertainty bands that we have built for the model with no income effects. Overall, it seems that both models

[22] In order to provide confidence intervals, one would have to bootstrap, which in our case is computationally very costly.

[23] We present these percentiles in order to have a first measure of uncertainty around the mean value of the prediction when the model paramaters are considered fixed.

are able to provide predictions reasonably close to real values in both dependent variables of the models.

7 Solving the principal agent problem

Once one has estimated the parameters of the principal-agent problem, one can solve for the optimal contracts and therefore estimate them. This is possible thanks to the fact that we have estimated the parameters of the theoretical model in the previous section. In this section, we will define the optimal contracts: the first best, the second best and optimal copayment. We would like to highlight that, for the estimation, we did not need to assume that contracts were optimal. This was possible thanks to the experimental design of the data we used.

It is important to describe the timing of events as well as the contractibility assumptions. At $t = 1$, the contract is offered to the individual. The contract might specify the premium p, the cost sharing function $D(.)$, and, if possible, the region of the (s, c) space at which treatment will be covered by the insurance scheme. The cost sharing function $D(.)$ will specify how much the individual will pay out of pocket for receiving the medical treatment. She accepts or rejects. If she accepts, at $t = 2$, she will obtain a draw of (s, c) from the joint distribution $g_{s,c|\overline{s},\overline{c}}$. If she is ill, $s > 0$, conditional on the contract, the individual will decide to have treatment or not. At this stage, the individual knows the realization of the two random variables. In the First Best, the two random variables are contractible. However in the second best case, the only contractible variable is the cost, and the health penalty shock will be individual's private information. This informational asymmetry is the source of moral hazard.[24]

In the first best problem, since both costs and health penalty shock are contractible, the insurer does not need to assume that the individual will behave optimally. That is, the contract offered to the individual will be a complete contingent plan on both s and c, that maximizes her expected utility conditional on the zero profit restriction for the insurance company, which comes from the assumption on competition.[25] The first best contract

[24]We say this is a moral hazard problem since it is an informational asymmetry after the signature of the contract (Laffont-Tirole 1993, Macho and Pérez 1993). Since it is more an informational advantage than an action, other authors would call it hidden information.

[25]In fact, the problem could easily be re-specified to allow for a positive level of expected profits. By solving the problem for different level of profits one would obtain the utility-profit frontier.

would determine a premium, a cost sharing function that depends on both s and c, as well as in which combinations of (s, c) this cost sharing function applies. On one hand, since we have two contractible variables, the first best is difficult to characterize since one must find the region of (s, c) for which the insurer will cover treatment costs.[26] On the other hand, though the first best solution is usually an interesting benchmark from the theoretical point of view, it does not provide any policy function. Hence we will concentrate on solving the second best, that is, the optimal contract under no contractibility of severity.

7.1 Optimal contract

The second best problem assumes that the insurer can just contract on the costs but not on the health penalty shock. Moreover, since there is noise between these two random variables (ε_s and ε_c are not perfectly correlated), the insurer cannot perfectly recuperate the health penalty from the observed cost. Consequently, in the second best problem no argument of the contract can depend on the health penalty shock. Therefore, the second best contract will be

$$c_{sb} = \{p, D(c)\},$$

where p is the premium and $D(c)$ the cost sharing function. The cost sharing function will give how much the individual will pay out of pocket for receiving medical treatment which costs c. At the second best, the insurer cannot design a region of (s, c) for which treatment is covered, since s is not contractible. On the contrary, the insurer will take into account that the insured will behave optimally according to his optimal decision rule (3). Therefore, if the treatment episode is c, the individual will demand treatment when $s > \widetilde{s}(c)$ where $\widetilde{s}(c)$ is implicitly defined by $U(y - p, \widetilde{s}(c)) = U(y - p - D(c), 0)$.

[26] The second best will be easier to characterize becuase the relation between severity and costs are given by the incentive compatibility constraint which is known in our problem.

The second best problem is defined by

$$\underset{\{P,D(c)\}}{Max} \Pr(s<0|\overline{s})*U(y-p,0)+\int_0^{+\infty}\int_0^{\widetilde{s}(c)} U(y-p,s)g_{s,c|\overline{s},\overline{c}}(s,c)\partial s\partial c \tag{12}$$

$$+\int_0^{+\infty}\int_{\widetilde{s}(c)}^{+\infty} U(y-p-D(c),0)g_{s,c|\overline{s}}(s,c)\partial s\partial c.$$

$$st: p=\int_0^{+\infty}\int_{\widetilde{s}(c)}^{+\infty}(c-D(c))g_{s,c|\overline{s},\overline{c}}(s,c)\partial s\partial c. \tag{13}$$

$$U(y-P,\widetilde{s}(c))=U(y-p-D(c),0). \tag{14}$$

Notice that the problem has two constraints. The first one ensures that the insurance contract is optimally fair, that is the premium is equal to the expected cost. The function that appears on the integration limit is a constraint given by the optimal individual's behavior.

7.2 Approximate solution to optimal contract

One of the advantages of structural estimation is the possibility to solve for policy measures. This is possible because one recuperates the raw parameters of the theoretical model. In this case, we are interested in obtaining the optimal contract, or an approximation to it. This is done by finding the premium p and cost-sharing function $D(c)$ that solves the principal-agent problem above.

Principal-Agent problems often lack of closed form solution. See, for instance Haubrich (1994) and Judd (1998) to find examples of problems solved by numerical methods. Our problem is quite complicated and hence we will also solve it using numerical methods. In order to solve the problem, we need to solve for the premium p, and the function $D(c)$. The way we will approach the problem is to parametrize $D(c)$ in a flexible way and then find the value of the parameters that solve the maximization problem. We have chosen the following parametrization

$$D(c|a_0,a_1,a_2,a_3)=L(a_0+a_1*c+a_2*c^2+a_3*c^3)*c \tag{15}$$

where $L(.)$ stands for the cumulative distribution function of the Logistic. This parametrization has the advantage that restricts the value of $D(c)$ to be between 0 and c, since a cumulative distribution function always takes a value between 0 and 1. The argument of

L is a polynomial what allows the function $D(c)$ to follow a flexible profile on c.[27]

We have used Simulated Annealing as maximization routine. This is a very robust random search algorithm that is able to escape from local optima. It also easily allows to restrict the value of the variables to maximize within a certain bounds, which is very useful to avoid overflows (Goffe et al. 1994). [28] This maximization algorithm does not need to take derivatives and it only needs to evaluate the function. The coefficients of the polynomial are the inputs to the maximization routine. At each combination of polynomial coefficients, we numerically solve the premium constraint and then evaluate the objective function. The solutions to the problem with income effects are found in table 12 and some values given by the cost-sharing function in table 13.

As it is clear from table 12, the percentage that the consumer pays out of pocket decreases from 81% when medical costs are \$1, to 46% when medical treatments are \$125. The out of pocket function, $D(c)$, shows a concave pattern that gives more coverage to the individual in case of higher costs. Given the high correlation between severity and costs unobservables (0.735), the optimal contract gives a higher coverage when illness are more likely to be severe. This concave profile is frequently observed in health insurance contracts: they do not provide coverage at low costs, for higher costs they cover a fixed percentage and when costs exceeds and upper bound, the insurance company cover all the costs (Cutler and Zeckhauser, 2000). We must point out that the cost sharing numbers should include other monetary and non-monetary costs of having treatment, apart from medical costs, since we have not considered them in the estimation.

The premium that corresponds to the approximation found to the optimal contract is \$1.27. In order to give a correct interpretation to this parameter, one should have in mind that it is in 1973 dollars and that the insurance contracts refers to acute conditions that do not need hospitalization and only for those illnesses that start in a given month. We also estimated more simple contracts that yield a smaller level of utility. The estimated value of the optimal deductible corresponds to \$42, while the estimated optimal copayment corresponds to 63%. From a comparison of the optimal copayment and the cost sharing values obtained in table 12 for the optimal contract, it is clear that non linear contracts

[27] One could incorporate further terms in the polynomial. We chosed a third degree polynomial since the results were very similar to the ones given by a second degree polynomial.

[28] The optimal values found are in the interior of the specified bounds and then the optimal values do not hit the constraint.

give more generous coverage at high values of the costs.

Even if extensive previous work has been done using the RAND Health Insurance Data, it is difficult to compare our results with those in previous work. Firstly, our model deals with illness episodes that starts in a month, rather that all the illness episodes of a year. Secondly, we have not included chronic expenditures as they are not random, nor episodes that have required hospitalization. As Belsey(1988) points out, different health care services will require optimally different cost sharing rules, as they have different elasticities. Hence, we have chosen to follow this recommendation and focus only on acute episodes that has not required any hospitalization.

7.3 Moral hazard measure

In this subsection we will give a new measure of the extent of moral hazard. The health economics literature has traditionally relied on the elasticity of utilisation with respect to insurance coverage. As we pointed out in the introduction, the suitability of this measure relies on the absence of income effects. Moreover, it is not directly linked to the extent to which informational asymmetries are important. The measure we will propose here is completely related to informational asymmetries and it does not hinge on the existence of income effects. The availability of this measure is a product of our modelling choice: to do a clear distinction between contractible and non contractible variables.

Given that moral hazard occurs because severity s is not contractible, there is moral hazard unless we can recuperate s from the observed costs c. If the correlation between ε_s and ε_c was one then there would be a one to one non relation between s and c. This would allow to recuperate s from c, given both principal and agent know the parameters of the model. Consequently, it will be the same to contract on both s and c that only on c and there would not be moral hazard. This form the basis for our measure of moral hazard: ρ. Moral hazard will be more prevalent the closer ρ will be to zero, since then there is no any stochastic relation between contractible and non contractible variables. In our case, we estimated ρ to be 0.735. Though there is lack of experience with this measure to know in what extent 0.73 is high or low, it seems moderately close to one and hence the extent of moral hazard seems reduced.

8 Conclusions

Unlike previous empirical research, we have used the Principal-Agent paradigm to estimate the optimal insurance contract for reimbursement health care insurance. Consequently, the optimal contract is derived from first principles and it is robust to the presence of income effects. The empirical implementation allows to estimate a new measure of moral hazard based on the correlation between unobservables influencing contractible and non contractible variables. We have also disentangled the treatment decision from the cost of treatment, following previous research that attributed to the consumer the decision whether to seek treatment or not, while the cost decision is mostly left to the doctor.

The data that we use come from the RAND Health Insurance Experiment, a social experiment conducted between 1975 and 1982 in six different cities of USA. Families participating in the experiment were randomly assigned to insurance plans, what allows to consider the insurance status as exogenous. Estimation of the structural parameters was carried out by maximum likelihood with numerical integration to accommodate that costs of treatment are only observed for those that decided to seek medical treatment.

Our theoretical model predicts the empirical pattern than observed costs are decreasing with the copayment. This is due to the fact that people with high copayment will only seek treatment in case it is inexpensive enough. We estimate a high correlation between unobservables influencing severity and those influencing costs, what reduces the extent of moral hazard. The optimal out of pocket function shows a concave profile on costs, providing a larger coverage for episodes with larger costs. Given the high correlation between severity and costs, this means that the optimal insurance contract provides a more generous coverage in case of severe illness episodes.

Due to lack of data, we have not considered the optimal mix of consumer and provider incentives (Ellis and McGuire, 1993). Consequently, we have look for the optimal insurance contract restricted to the set of those that only gives incentive to the consumer. Finally, we would like to point out that this paper has followed the tradition of a sovereign consumer that is able to evaluate correctly health status, and in particular for our model, severity. More research is needed to evaluate this point and its consequences. We expect that this paper might also be useful for those incorporating non monetary costs in consumer behaviour and its consequences over insurance coverage.

9 Tables

Table 1 *Distribution of number of episodes that started in the first month*

Number of episodes	Number of people
0	1617
1	217
2	28
3	3

Table 2. Description of variables

Variable	**Mean**	**S.D.**	**Description**
			Endogenous
Treat	0.13	0.33	=1 if treated by an episode that started in the first month, 0 on the contrary
Costs	29.23	44.36	episode treatment cost. 1973 dollars
			Exogenous
Copay	0.31	0.37	copayment rate: 0, 0.25, 0.5, 0.95
Inc	469.2	399.9	montly per-capita family income at enrollement. 1973 dollars
Fem	0.55	0.49	=1 if female
Ghea	0.70	0.14	general health index divided by 100. Higher values indicates better health
Soc	0.70	0.24	index of social contacts divided by 100 Higher values indicates more contacts
Educ	1.21	0.29	Number of years of education divided by 10
Disea	1.21	0.88	Index for number of diseases divided by 10 Higher values indicates more diseases
Age	3.64	1.16	age divided by 10
Appt	0.44	0.29	Satisfaction with length of wait for medical appointments Higher values indicate greater satisfaction. Divided by 100
Avmd	0.39	0.24	Satisfaction with availability of family doctors Higher values indicate greater satisfaction. Divided by 100
Avsp	0.55	0.26	Satisfaction with availability of medical specialists Higher values indicate greater satisfaction. Divided by 100

Table 3. Distribution of episodes treated

	Observ.	Percentage of episodes treated
Copay. 0	832	0.17
Copay. 25	472	0.13
Copay. 50	146	0.08
Copay. 95	415	0.09

Table 4. Probit estimates for having treatment

	Estimate and Se.
Constant	-0.35 (0.30)
Soc	-0.22 (0.15)
Female	0.21 (0.08)
Ghea	-0.67 (0.28)
Disea	0.06 (0.04)
Age	-0.06 (0.03)
Educ	0.03 (0.13)
Copay. 25	-0.19 (0.09)
Copay. 50	-0.42 (0.16)
Copay. 95	-0.39 (0.10)

Table 5 *Average cost of treatment by copayments rates*

Copay. 0	31.88
Copay. 25	28.64
Copay. 50	13.96
Copay. 95	25.07

Table 6 *OLS estimates of LnC*

	Estimate and Se.
Constant	3.38 (0.24)
Disea	-0.08 (0.08)
Ghea	-0.33 (0.50)
Age	0.14 (0.06)
Female	-0.22 (0.14)
Soc	0.35 (0.28)
Educ	-0.16 (0.24)
Appt	-0.43 (0.23)
Avsp	-0.44 (0.25)
Avmd	-0.47 (0.28)
Copay. 25	-0.05 (0.16)
Copay. 50	-0.56 (0.31)
Copay. 95	-0.08 (0.19)

Table 7. Estimates of Severity equation

	With Income		No Income	
	Coefficient	Se	Coefficient	Se
Constant	-0.052	0.047	-29.23	22.46
Ghea	-0.119	0.055	-60.23	26.36
Soc	-0.034	0.023	-16.2	10.82
Female	0.032	0.015	16.39	6.9
Age	-0.009	0.006	-3.53	2.81
Disea	0.007	0.008	3.3	3.98

Table 8 Estimates of Cost equation

	Income Effect		No Income Effect	
	Coefficient	Se	Coefficient	Se
Constant	2.555	0.573	2.347	0.593
Education	-0.135	0.212	-0.078	0.215
Disea	-0.093	0.098	-0.081	0.101
Ghea	-1.105	0.638	-1.088	0.659
Age	0.102	0.069	0.117	0.074
Appt	-0.454	0.249	-0.479	0.253
Avmd	-0.36	0.289	-0.321	0.298
Avsp	-0.463	0.246	-0.487	0.246

Table 9. Estimates of the cholesky decomposition and other information of interest.

	Income Effect		No Income Effect	
	Coefficient	Se	Coefficient	Se
a	0.153	0.052	78.204	20.238
b	0.989	0.246	1.059	0.264
c	-0.911	0.063	0.899	0.072
ρ^*	0.735	0.09	0.76	0.09
Log-lik	-1754.39		-1752.192	
P-Andrews	0.7		0.87	

The value of ρ is obtained from the estimates of b and c.

Its standard error is computed using the delta method.

Table 10. Predictions of frequency of treatment

Copay	Real	Income effects			No Income effects		
		P2.5	Mean	P97.5	P2.5	Mean	P97.5
Copay 0	0.17	0.120	0.143	0.168	0.126	0.150	0.174
Copay 25	0.13	0.103	0.134	0.165	0.105	0.135	0.165
Copay 50	0.08	0.075	0.122	0.178	0.068	0.116	0.171
Copay 95	0.09	0.084	0.114	0.147	0.074	0.101	0.130

Table 11 Predictions of observed costs

Copay	Real	Income effects			No Income effects		
		P2.5	Mean	P97.5	P2.5	Mean	P97.5
Copay 0	31.8	23.7	31.2	41.7	24.6	32.4	43.6
Copay 25	28.6	19.5	27.7	39.5	19.7	27.8	39.1
Copay 50	13.9	14.3	26.8	48.0	13.5	24.8	41.7
Copay 95	25.0	16.4	23.4	33.2	14.3	19.7	26.6

Table 12 *Solution to optimal contract with income effects*

a_0	1.48084
a_1	-0.0249
a_2	1.11e-4
a_3	-1.19e-7
p	1.2744
U	-0.2363

Table 13 Cost sharing values given by optimal contract

c	$D(c)$	$D(c)/c$
1	0.81	0.81
10	7.75	0.77
20	10.72	0.73
30	20.88	0.69
50	31.67	0.62
70	39.13	0.56
90	46.07	0.51
125	58.26	0.46

References

[1] Andrews, D.W.K. (1988), Chi-Square Diagnostic Tests for Econometric Models. Introduction and Applications, Journal of Econometrics, 37, 135-156.

[2] Arrow, K. (1963), Uncertainty and the Welfare Economics of Medical Care. *The American Economic Review,* 53, 941-973.

[3] Biais, B., Bisière, C. and Décamps, J.P. (1999), A Structural Econometric Investigation of the Agency Theory of Financial Structure. GREMAQ, IDEI. Mimeo.

[4] Buchanan, J. L., Keeler, E., Rolph, J. and Holmer, M. (1991), Simulating Health Expenditures under Alternative Insurance Plans. *Management Science,* 37, 1067-1090.

[5] Cameron, A.C., Trivedi, P.K., Milne, F, and Pigott, J. (1988), A Microeconometric Model of the Demand for Health Care and Health Insurance in Australia. *Review of Economic Studies,* 55, 85-106.

[6] Coulson, N.E., Terza, J., Neslulan, C.A. and Stuart, C.B. (1995), Estimating the Moral Hazard Effect of Supplemental Medical Insurance in the Demand for Prescription Drugs by the Elderly, *American Economic Review Papers and Proceedings*, 85, 122-126.

[7] Chiappori, P.A., Durand, F., and Geoffard., P-Y. (1998). Moral Hazard and the Demand for Physician Services: First Lessons from a French Natural Experiment. *European Economic Review,* 42, 499-511.

[8] Chiappori, P.A., Salanié, B. (2000), Testing Contract Theory: a Survey of Some Recent Work. Invited lecture to the World Congress of the Econometric Society.

[9] Cutler, D. and Zeckhauser, R. (2000), The Anatomy of Health Insurance, in Handbook of Health Economics, vol. 1A, Culyer, A. and Newhouse, J. (eds). North Holland.

[10] Deb, P. and Trivedi, P. (1999), The Structure of Demand for Health Care: Latent Class versus Two Part Models. Indiana University Purdue Univiersity. Mimeo.

[11] Feldman, R. and Dowd, B. (1991), A New Estimate of the Welfare Loss of Excess Health Insurance. *The American Economic Review, 81,1, 297-302.*

[12] Feldstein, M. (1973), The Welfare Loss of Excess Health Insurance. *Journal of Political Economy.* 81, 251-80.

[13] Ferrall , C. and Shearer, B. (1999), Incentives and Transaction Costs within the Firm: Estimating an Agency Model Using Payroll Records. *Review of Economic Studies, 66, 309-338.*

[14] Goffe, W.L., Ferrier, G.D. and Rogers, J. (1994), Global Optimization of Statistical Functions with Simulated Annealing, *Journal of Econometrics*, 60, 65-99.

[15] Haubrich J. (1994), Risk Aversion, Performance Pay and the Principal-Agent Problem. *Journal of Political Economy.102, 258-276.*

[16] Holly, A., Gardiol, L., Domenighetti, G. and Bisig, B. (1998), An Econometric Model of Health Care Utilization and Health Insurance in Switzerland, *European Economic Review*, 42, 513-522.

[17] Judd, K. (1998), *Numerical methods in economics.* The MIT Press, Cambridge, Massachusetts.

[18] Keane, M. and Wolpin, K. (1997), Introduction to the JBES Special Issue on Structural Estimation in Applied Microeconomics. *Journal of Business and Economic Statistics, 15,2, 111-114.*

[19] Keeler, E., Newhouse, J., ans Phelps, C. (1977), Deductibles and the Demand for Medical Care Services: the Theory of a Consumer Facing a Variable Price Schedule under Uncertainty. *Econometrica.* 45, 641-656.

[20] Keeler, E. and Rolph, J. (1988), The Demand for Episodes of Treatment in the Health Insurance Experiment. *Journal of Health Economics.* 7, 337-67.

[21] Ellis, R.P (1986), Rational Behavior in the Presence of Coverage Ceilings and Deductibles, *RAND Journal of Economics*, 158-175.

[22] Ellis, R.P. and McGuire, T. G. (1993), Supply-Side and Demand-Side Cost Sharing in Health Care. *Journal of Economic Perspectives, 7,4, 135-151.*

[23] Guilleskie, D. and Mroz, T (2000), Estimating the Effects of Covariates on Health Expenditures. NBER working paper w7942.

[24] Gurmu (1997) Medicaid, JAE.

[25] Laffont, J. and Tirole, J. (1993), *A theory of incentives in procurement and regulation.* The MIT Press. Cambridge, Massachusetts.

[26] Lucas, R. (1976), *Econometric policy evaluation: a critique.* Carnegie-Rochester conference series on public policy, 1, 19-46.

[27] Ma, A. C. and Riordan, M. H. (2001), Health Insurance, Moral Hazard and Managed Care. Forthcoming in Journal of Economics and Managment Strategy.

[28] Macho, I. and Pérez, D. (1994), *Introducción a la economía de la información.* Ariel, Barcelona.

[29] Manning, W.G., Newhouse, J. P. *et al.* (1987), Health Insurance and the Demand for Medical Care: Evidence from a Randomized Experiment. *The American Economic Review,* 77**,** 251-277.

[30] Manning, W. G. and Marquis, M. S. (1996), Health Insurance: the Tradeoff Between Risk Pooling and Moral Hazard. *Journal of Health Economics*, 15, 609-639.

[31] Manning, W. G. and Marquis, M. S. (2001), Health Insurance: Tradeoffs Revisited. *Journal of Health Economics,* 20, 289-293.

[32] Mace (1991), Full Insurance in the Presence of Aggregate Uncertainty. *Journal of Political Economy*, 99, 928-956.

[33] Marigotta and Miller (2000), Managerial Compensation and the Cost of Moral Hazard. *International Economic Review*, 41, 3, 669-719.

[34] Marquis, M.S. and Holmer, M.R. (1996), Alternative Models of Choice under Uncertainty and the Demand for Health Insurance. *Review of Economic and Statistics,* 78(3), 421-427.

[35] Meza, D (1983), Health Insurance and the Demand for Medical Care. *Journal of Health Economics, 2, 47-54.*

[36] Newhouse, J.P.et.al. (1993), *Free for all? Lessons from the Rand Health Insurance Experiment.* Harvard University Press, Cambridge, Massachusetts.

[37] Paarsch, H. and Shearer, B. (2000), Piece Rates, Fixed Wages, and Incentive Effects: Statistical Evidence from Payroll Records. *International Economic Review,* 41,1, 59-92.

[38] Pauly, M. (1968), The Economics of Moral Hazard. A Comment. *The American Economic Review,* 58*, 531-537.*

[39] Pauly, M. (1986), Taxation, Health Insurance and Market Failure in the Medical Economy. *Journal of Economic Literature,* 24*, 629-675.*

[40] Pohlmeier, W. and Ulrich, V. (1995), An Econometric Model of the Two-Part Decisionmaking Process in the Demand for Health Care, *The Journal of Human Resources,* 30,2, 339-361.

[41] Propper, C. (2000), The demand for private health care in the UK, *Journal of Health Economics,* 19, 855-876.

[42] Propper, C., Rees, H. and Green, K. (2001), The Demand for Private Medical Insurance in the UK: A Cohort Analysis, *The Economic Journal,* 111, C180-C200.

[43] Salanié, B. (1997). *The Economics of Contracts: a primer.* MIT Press, Cambridge, Massachusetts.

[44] Street, A. Jones, A. and Furuta, A (1999), Cost-Sharing and Pharmaceutical Utilization and Expenditure in Russia, *Journal of Health Economics*, 18, 459-472.

[45] Towsend (1994), Risk and Insurance in Village India. *Econometrica*, 62, 3, 539-591.

[46] Van de Voorde, C., Doorslaer., E. V., Shokkaert., E. (2000), Do Doctors Induce Demand to Offset Patient's response to Cost Sharing? Evidence from a Natural Experiment in Belgium. Ninth European Workshop of Econometrics and Health Economics. Tinbergen Institute. Amsterdam.

[47] Zeckhauser, R. (1970), Medical Insurance: a Case Study of the Tradeoff Between Risk Spreading and Appropiate Incentives. *Journal of Economic Theory*, 2, 10-26.