Laboratoire d'Economie de Dauphine

WP n°2/2017

Document de travail

Sélection on moral hazard in Supplementary Health Insurance

Mathilde Péron
Brigitte Dormont

Pôle Laboratoire d'Economie et de Gestion des Organisations de Santé (LEGOS)
Place du Maréchal de Lattre de Tassigny 75775 Paris Cedex 16
Tél (33) 01 44 05 44 46 Fax (33) 01 44 05 40 67
Site : www.legos.daupine.fr

Selection on moral hazard in Supplementary Health Insurance [*]

Mathilde PÉRON[†], Brigitte DORMONT[‡]

August 2017

1 Introduction

It is critical for insurers to evaluate the possible effect of health insurance on care consumption when they design their contracts and set their prices. However, when insurance is voluntary, the estimated relationship between health insurance coverage and healthcare consumption is influenced by endogeneous selection: individual characteristics, such as health status, age, gender, income, supply side constraints or preferences are likely to explain both individuals' consumption of healthcare and demand for health insurance. Einav et al. (2013) distinguish two sources of endogeneous selection: classical adverse selection and selection on moral hazard. Classical adverse selection is linked to individual heterogeneity as regards demand for healthcare. Basically, some individuals consume more healthcare than others and are also more likely to buy insurance in order to reduce the financial risk associated with their healthcare expenditures. Selection on moral hazard appears when there is individual heterogeneity as regards the behavioral response to health insurance. In this case, some individuals might be more prone to buy insurance because they expect an increase in their healthcare consumption due to better coverage.

Empirical contributions that aim to estimate the causal effect of insurance on healthcare use acknowledge that there is heterogeneity in the demand for healthcare and control for classical adverse selection (Cameron et al. 1988, Coulson et al. 1995, Holly et al. 1998, Vera-Hernández 1999, Schellhorn 2001,

[*]We gratefully acknowledge financial supports from the Health Chair - a joint initiative by PSL, Université Paris Dauphine, ENSAE, MGEN and ISTYA under the aegis of the Fondation du Risque (FDR).

[†]Corresponding author, University of York. E-mail: mathilde.peron@york.ac.uk

[‡]PSL, Université Paris Dauphine

Buchmueller and Couffinhal 2004, Jones et al. 2006). In this literature, the response to health insurance is often assumed to be homogeneous across individuals and moral hazard is estimated through a single parameter associated with the price elasticity of demand for healthcare. In this framework, results based on randomization such as the RAND Health Insurance Experiment (Manning et al. 1987, Newhouse 1993), or quasi-natural experiments (Chiappori et al. 1998) are usually considered as a gold standard. Of course, randomization is an elegant solution to eliminate selection bias from the estimation of the impact of insurance on care use. But this approach is not necessarily of interest when insurance is voluntary. Because these analyses remove the endogenous choice component from the equation, they are not able to estimate a potential selection on moral hazard and predict the impact of a voluntary insurance on healthcare consumption. The question of selection on moral hazard has been addressed empirically by Einav et al. (2013). They use individual-level panel data from an American firm where employees can choose among different level of coverage. They find heterogeneity on moral hazard together with selection on moral hazard: individuals who buy more comprehensive coverage exhibit greater moral hazard.

Assuming that individuals select themselves in connection with their expected response to insurance can be particularly relevant, especially when one wants to predict the effect of copayments and deductibles on healthcare expenditures. Suppose that an insurer wants to supply an additional contract with better coverage. If he relies on average estimates of the price elasticity of demand[1], he will underestimate the increase in costs due to moral hazard. Indeed, contracts with more comprehensive coverage will attract individuals whose healthcare consumption would increase more strongly. On the contrary, if the insurer wants to introduce copayments to limit medical spending, he will overestimate the effect of such a decision: higher copayments will firstly attract individuals who are less sensitive to healthcare prices. Of course, these concerns are relevant only if the insurance under review is voluntary and not mandatory. Actually, this situation deserves attention because it is often encountered: it concerns all the cases where individuals can buy supplementary health insurance. However, the empirical literature addresses issues that are relevant mostly in the case of mandatory health insurance.

In this paper we investigate the relationships between the demand for healthcare, the decision to take out health insurance and the behavioral response to better coverage with a structural model that specifies individual heterogeneity in demand for healthcare and response to insurance (i.e. moral hazard). We set the analysis in the French context where individuals can voluntarily take out supplementary health

[1] That would be estimated, for instance, by a random assignment procedure like in the Rand experiment.

insurance (SHI) which covers medical goods and services with higher quality than the basic healthcare basket covered by mandatory national health insurance (NHI). We especially focus on the demand for specialist who balance bill their patients, i.e. charge them more than the regulated fee set by NHI. We estimate the causal effect of voluntary SHI on the demand of specialist consultations with balance billing, taking into account both classical adverse selection and selection on moral hazard. The econometric analysis is performed on a French database of 58,519 individuals observed in 2012.

In France, the NHI offers universal, yet partial, coverage. Individuals can take out SHI to enhance their coverage and limit out-of-pocket expenditures, either voluntary in the individual market or through their employer. For ambulatory care, the NHI sets a regulated price and reimburses only a fraction of it to patients (70% of the regulated fee for specialist consultations). On top of NHI copayments, patients may also have to pay balance billing. Indeed, patients have the choice to visit two types of specialists: 'sector 1' (S1) specialists are mandated to charge the NHI regulated fee whereas 'sector 2' (S2) specialists are allowed to balance bill, i.e. charge a fee that exceeds the regulated price, which is the basis for NHI reimbursement. S1 and S2 specialists are supposed to provide the same medical service. However, because S2 is restricted to physicians who have been practicing in a qualifying hospital setting, S2 consultations can be associated by patients with a higher level of quality. Because they charge higher fees, waiting lists are also likely to be shorter for S2 specialists. Almost 95% of the French population is covered by a SHI contract, which covers at least the 30% NHI copayment. Still, there are important differences between SHI contracts in terms of balance billing coverage: in polls, only 48.5% of SHI policyholders state that they are well covered against balance billing (Célant et al. 2014).

In the specific context of demand for balance billing coverage we can expect both forms of selection, e.g. classical adverse selection and selection on moral hazard. Indeed, in Dormont and Péron (2016) we gave evidence of individual heterogeneity in balance billing consumption related to demand for more comprehensive SHI coverage. Our estimates were based on a French panel data set of 43,111 individuals observed in 2010 and 2012. In 2010, the whole sample was covered by the same SHI contract, with no coverage against balance billing. We were able to observe the same individuals in 2012 after 3,819 of them had switched to other SHI contracts that cover balance billing. Using individual fixed effects and instrument variables we were able to deal with the non-exogeneity of the decision to switch insurer and estimate the change in balance billing consumption between 2010 and 2012 due to a better coverage. Our estimates show that those who ask for better coverage consume, *ceteris paribus*, more balance billing than the rest of the sample, even when they are not covered for balance billing. This would reveal classical adverse selection in the demand for balance billing coverage. Heterogeneity in the response to better coverage can be linked to unobservable individual heterogeneity, and to observable characteristics. First, the response to a better balance billing coverage is likely to be influenced by unobservable individual characteristics. Indeed, the demand for S2 visits relies strongly on perceived quality of care. Preferences and beliefs, which are unobserved, are likely to be heterogeneous: they can explain both heterogeneous response to a better coverage and decision to take out SHI resulting in selection on moral hazard. Second, heterogeneity in moral hazard might as well be influenced by observable characteristics such as gender, age, income or living area. In Dormont and Péron (2016) we found evidence of moral hazard only for individuals living in areas where there are few specialists who do not balance bill their patients (S1 specialists)[2]. Turning to a possible impact of income, we can refer to Nyman's contribution to the debate on moral hazard (Nyman 1999, 2003). Traditional models of health insurance (Friedman and Savage 1948, Pauly 1968) see moral hazard as a pure price effect: because better insurance coverage reduces the price faced by patients and assuming the negative price-elasticity of healthcare demand, patients with insurance coverage should increase their healthcare consumption. However, Nyman considers that better coverage also creates an income effect which releases the budget constraint and gives patients access to care that they could not afford without insurance. Within this framework, low income individuals should react more to an improvement in coverage than rich individuals.

In the econometric literature, selection on moral hazard is more generally known as selection on returns or essential heterogeneity. Assuming that there is individual heterogeneity in treatment effects, essential heterogeneity arises when individuals decide to take the treatment in relation with their expected response to the treatment. Heckman and Vytlacil (2007) show that in the presence of essential heterogeneity, instrumental variable (IV) methods, which are frequently used to control for endogeneous selection, do not estimate an average treatment effect (ATE), nor a treatment effect on treated. Indeed, IV methods only estimate a local average treatment effect (LATE), specific to individuals who would react to the shock induced by the instrument. In the presence of essential heterogeneity, this local effect cannot be extended to the average population. Another consequence is that different instruments are likely to give different estimates of the treatment effect because they rely on compliers with different reactions to the treatment. Beyond the objective to estimate unbiased causal effects, we can question the relevance of estimating an

[2]This is because the effect of insurance on the relative price of S1 and S2 consultations depends on the search and waiting time costs associated with a S1 consultation, which are strongly influenced by S1 availability in each area

ATE in a context where individuals can decide to participate or not in the treatment. Indeed, in this case, we pay more attention to the treatment effect of those who are more likely to take the treatment rather than to the average effect on the whole population. Marginal treatment effects (MTE) estimators have been developed to capture the impact of a treatment likely to vary within a population in correlation with observed and unobserved characteristics, in a setting where individuals select themselves into treatment. First defined by Bjorklund and Moffitt (1987), MTE have been comprehensively described by Heckman and Vytlacil (2001) and Heckman et al. (2006). Empirically, MTE have been used to capture returns in education (Carneiro et al. 2011), breast cancer treatment effects (Basu et al. 2007) or the effect of family size on children's outcome (Brinch et al. 2012). Recently, Kowalski (2015) uses MTE in an experimental framework to assess the external validity of the Oregon health insurance experiment.

MTE are the appropriate tools when one focuses on the effect of voluntary health insurance on balance billing consumption. First, essential heterogeneity is only a concern if individuals can decide to take the treatment and if unobservable characteristics can influence their outcome. In our setting, individuals can choose their level of balance billing coverage while their preferences for higher quality of care, which are unobservable to the econometrician, are likely to influence their balance billing consumption. Second, MTE rely on a structural approach that links the output (the demand for balance billing), the decision to take the treatment (take out SHI) and the treatment effect (moral hazard). This unified framework identifies complex relationships between demand for higher quality of care and comprehensive SHI. It allows to identify different motives of the demand for balance billing coverage, either to cover expected expenditures or to increase balance billing consumption. Third, MTE fully take into account individual heterogeneity in the response to treatment, due to both observable and unobservable characteristics. The structural approach further associates the heterogeneous treatment effect to different mechanisms related to income, supply side constraints or preferences. We are indeed able to give some 'content' to moral hazard, especially in terms of access to S2 specialists, and go beyond the homogeneous price effect usually reported in the literature.

In this paper, we estimate the marginal treatment effect of SHI coverage on balance billing consumption. We take into account observed and unobserved individual heterogeneity in the demand for S2 consultations and in moral hazard. We also control for classical adverse selection and selection on moral hazard. Our empirical analysis is built on a structural model that links (i) the demand for balance billing, (ii) the decision to take out more comprehensive SHI and (iii) the behavioral response to better coverage. Thanks to this unified framework we are able to give insights on the determinants of the demand for higher quality of care and the role of health insurance in terms of access to care, especially for low income individuals.

Our database stems from administrative data provided by a French insurer, the Mutuelle Générale de l'Education Nationale (MGEN). We use cross-sectional data which provide for 58,519 individuals information on healthcare claims and reimbursements by the NHI and SHI in 2012. We are able to observe two groups of individuals: MGEN-SHI subscribers and better-SHI subscribers. The former are not covered for balance billing. The latter were previously covered by the same MGEN-SHI contract but decided in 2011 to switch towards another SHI insurer: in 2012 they benefit from balance billing coverage. The better-SHI subscribers are used as a treatment group to estimate the heterogeneous effect of SHI coverage on balance billing consumption and test for the existence of classical adverse selection and selection on moral hazard.

We find evidence of individual heterogeneity in the response to better coverage and of selection on moral hazard. Individuals with unobserved characteristics that make them more likely to take out better SHI are also those who exhibit stronger moral hazard, i.e. a larger increase in balance billing per consultation. We also find that individuals' income is a strong determinant of balance billing consumption and influence the behavioral response to better coverage. Without coverage, the poor consume less balance billing than the rich but increase their consumption more sharply once covered for balance billing. They are also more likely to subscribe to comprehensive coverage.

The fact that unobserved characteristics influence both the decision to take out SHI and the magnitude of moral hazard is firstly a concern for insurers. Indeed, when providing comprehensive balance billing coverage, insurers have to take into account that their contract is likely to attract individuals who are more sensitive to healthcare prices and respond more sharply than average to better coverage. In a context where SHI is voluntary, the inflationary impact of SHI coverage might be worsened by selection on moral hazard. Our policy conclusions as regards the role of income are of different nature. We argue that the negative effect of income on the demand for S2 consultations coupled with its positive effect on moral hazard reveals that insurance plays an important role in terms of access to care.

This paper is organized as follows. Section 2 presents the MTE method. In section 3 we present our data and empirical strategy. The empirical specification is developed in section 4. Results are presented in section 5. Section 6 concludes.

2 Method: Marginal Treatment Effects

Allowing for heterogeneity in treatment effects potentially yields essential heterogeneity. This term means that the assignment to treatment, or the choice to be treated, is correlated with the treatment impact. In our case, some people would choose to take out better supplementary insurance because they know their healthcare consumption will respond positively to better coverage. As stated by Heckman et al. (2006), when treatment effects are likely to be heterogenous, it is reasonable to allow for a correlation between the choice for treatment and the treatment impact.

Consider the two potential outcomes $Y_{i,1} = \alpha_1 + U_{i,1}$ and $Y_{i,0} = \alpha_0 + U_{i,0}$ which are observed if the individual is respectively treated ($D_i = 1$) or not treated ($D_i = 0$). The observed outcome is:

$$\begin{aligned} Y_i &= D_i Y_{i,1} + (1 - D_i) Y_{i,0} \\ &= \alpha_0 + ((\alpha_1 - \alpha_0) + (U_{i,1} - U_{i,0})) D_i + U_{i,0} \end{aligned}$$

Here the treatment impact varies across individuals. One has: $Y_i = \alpha_0 + \tau_i D_i + U_{i,0}$ with $\tau_i = Y_{i,1} - Y_{i,0} = (\alpha_1 - \alpha_0) + (U_{i,1} - U_{i,0})$.

To estimate this model one has to deal with two possible selection problems: (i) a correlation between D_i and $U_{i,0}$, which is due to a selection on the level of the outcome without treatment; (ii) a correlation between D_i and τ_i, i.e. a selection on the expected impact of the treatment (essential heterogeneity). In case of essential heterogeneity, the use of instrumental variables is not straightforward. Firstly, the IV method does not provide a consistent estimation of the mean treatment effect $\overline{\tau}$. [3] Secondly, if there is selection on the gains from treatment, the IV estimate must be interpreted as a local average treatment effect "which is only informative about the average causal effect of an instrument-induced shift in D" (Brinch et al. 2012). As shown by Heckman et al. (2006), the solution is to estimate marginal treatment effects (MTE). MTE are computed from a model that explicitly specifies the decision to be treated, and gives the treatment impact for someone who is at the margin, i.e. who is indifferent between being treated or not. Moreover, MTE produce a function that is invariant to the choice of instruments.

[3] One has: $\tau_i = (\alpha_1 - \alpha_0) + (U_{i,1} - U_{i,0}) = \overline{\tau} + \eta_i$

From $Y_i = \alpha_0 + \tau_i D_i + U_{i,0}$, one has: $Y_i = \alpha_0 + \overline{\tau} D_i + (U_{i,0} + \eta_i D_i)$

To provide a consistent estimate of $\overline{\tau}$, the IV Z must be uncorrelated with $U_{i,0} + \eta_i D_i$. In the case of essential heterogeneity this condition is not satisfied, even if Z is not correlated with $U_{i,0}$ and η_i. Indeed, $E(\eta_i D_i \,|Z_i\,) = E(\eta_i \,|D_i = 1, Z_i\,) \Pr(D_i = 1\,|Z_i\,)$, and the first term of the right-hand side is different from 0 if the decision to treat is correlated with the individual gain of the treatment.

2.1 The Generalized Roy model

To introduce MTE, Heckman et al. (2006) consider the Generalized Roy model, which is a switching regression model that allows a structural approach to policy evaluation.[4] For the sake of simplicity, the subscripts i are omitted hereafter. The model specifies the two potential outcomes (Y_0, Y_1) and the decision to participate in the treatment ($D = (0, 1)$). The choice of receiving treatment is modeled as a function of observables Z and unobservables V, and linked to the observed outcome Y through a latent variable D^*. In addition to the previous model, we now assume that the outcomes depend on exogenous regressors X. Hence, the treatment has an impact on unobserved heterogeneity (from U_0 to U_1) and on the effect of covariates X (from β_0 to β_1):

$$
\begin{aligned}
Y &= DY_1 + (1 - D)Y_0 && (1)\\
Y_1 &= X\beta_1 + U_1 && (2)\\
Y_0 &= X\beta_0 + U_0 && (3)\\
D^* &= Z\gamma - V && (4)\\
D &= \begin{cases} 1 & \text{if } D^* > 0 \\ 0 & \text{if } D^* \leq 0 \end{cases} && (5)
\end{aligned}
$$

In the general case, it is assumed that U_0, U_1 and V are independent of Z, conditional on X. In addition, the probability of treatment is a non-trivial function of Z, conditional on X : $Pr(D|X = x, Z = z) \neq Pr(D|X = x)$ (Basu et al. 2007).

The propensity score $P(Z)$ is the probability of receiving treatment conditional on Z:

$$P(Z) \equiv Pr(D = 1|Z = z) = Pr(V < Z\gamma|Z = z) = F_V(Z\gamma)$$

where F_V is the cumulative distribution function of V, hence a monotonic and absolutely continuous function.

An individual chooses to be treated if the latent variable D^* is positive:

$$D = 1 \Leftrightarrow D^* > 0 \Leftrightarrow Z\gamma > V \Leftrightarrow F_V(Z\gamma) > F_V(V) \Leftrightarrow P(Z) > F_V(V)$$

[4] Our description of the statistical framework follows closely that found in Heckman et al. (2006) and Brave et al. (2014).

Defining $U_D = F_V(V)$, the condition to be treated is that the propensity score is greater than U_D : $P(Z) > U_D$.

Without a loss of generality we can assume that U_D is a uniformly distributed random variable between 0 and 1. In this case the p^{th} quantile of U_D is p and different values of U_D correspond to different quantiles of V.

The propensity score must be interpreted as the incentive to choose the treatment, for given covariates Z. As for U_D, it can be seen as the individual idiosyncratic disutility of taking the treatment. Conditionally on her characteristics z, which provide a propensity score p, an individual will ultimately take the treatment if her disutility u_D is lower than p (and be indifferent if $u_D = p$). For the econometrician, variables Z are observed and realizations u_D are not observed. Given that values of U_D are quantiles of V, it is possible to compare $P(Z)$ and U_D on the same interval $[0, 1]$ on the horizontal axis (Figure 1):

2.2 Marginal Treatment Effects

In our framework, decision to participate in the treatment and treatment impact vary across individuals. MTE capture the treatment effect $(Y_1 - Y_0)$ for the 'marginal individual' who is indifferent between being treated or not, conditional on her observed characteristics $X = x$. By definition, the marginal individual has a propensity score equal to her disutility of taking the treatment: $U_D = p$.

$$MTE \equiv E(Y_1 - Y_0 | X = x, U_D = p) \tag{6}$$

Heckman et al. (2006) show how MTE can be identified by taking the derivative of $E(Y|X = x, Z = z)$ with respect to $P(Z)$. First, note that

$$E(Y|X = x, Z = z) = E\{Y|X = x, P(Z) = p\} \tag{7}$$

Following Heckman and Vytlacil (2001), the observed outcome can be written as:

$$\begin{aligned} E\{Y|X = x, P(Z) = p\} &= E(Y_0|X = x) + E(Y_1 - Y_0|X = x, D = 1)p & (8) \\ &= E(Y_0|X = x) + \int_0^p E(Y_1 - Y_0|X = x, U_D = u_D)du_D & (9) \end{aligned}$$

As a consequence,

$$\frac{\partial E\{Y|X=x,P(Z)=p\}}{\partial p}=E(Y_1-Y_0|X=x,U_D=p) \tag{10}$$

Expression (10) shows how the derivative of $E(Y|X=x,Z=z)$ identifies marginal treatment effect, i.e the expected treatment effect conditional on X and U_D[5]. As noted by Heckman et al. (2006), "a high value of $P(Z)=p$ identifies MTE at a value of $U_D=u_D$ that is high - that is associated with nonparticipation". Indeed, that individuals with a high propensity score are indifferent between being treated or not implies that they have a very high idiosyncratic disutility of taking the treatment u_D. Therefore, MTE with high p values identify returns for individuals who are less likely to take the treatment. Conversely, MTE with low values of p identify returns for individuals prone to take the treatment.

2.3 Estimation

Ideally, a continuous instrument with sufficient variation conditional within all $X=x$ would allow for a fully non-parametric estimation of the MTE, conditional on X, and would produce a separate MTE curve for each value of X. However, as noted by Cornelissen et al. (2016), such an instrument is rarely available and further assumptions are often required to identify MTE. A first assumption is to condition the outcome on X in a parametric linear way and model potential outcomes as $Y_1=x\beta_1+U_1$ and $Y_0=x\beta_0+U_0$ and the selection equation as $D_i^*=Z\gamma-V$.

A second assumption is to assume that the shape of the MTE curve is independent of X. Only the intercept of the MTE curve is allowed to vary with X. This is implied by the full independence assumption $(X,Z)\perp\!\!\!\perp(U_0,U_1,V_1)$, which is stronger than the conditional independence assumption $Z\perp\!\!\!\perp(U_0,U_1,V_1)|X$ necessary for fully non-parametric approach[6].

The linear separability and the full independence assumptions imply that the MTE is additively separable into an observed and unobserved component. Combining (8) with the linear expressions of Y_1 and Y_0, one obtains:

$$E\{Y|X=x,P(Z)=p\}=x\beta_0+x(\beta_1-\beta_0)p+K(p)\ , \tag{11}$$

[5]The ATE, by contrast, is the average treatment effect, conditional on X. Note that the ATE can be constructed as a weighted average of MTE by integrating over U_D (Heckman and Vytlacil 2001, Heckman et al. 2006), providing that the support of U_D covers $[0,1]$: $ATE\equiv E(Y_1-Y_0|X=x)$.

[6]Alternatively, Brinch et al. (2012) assume additive separability between the observed and unobserved component in the expected outcomes conditional on $U_D=u_D$. This allows a non-parametric estimation of the MTE with a binary instrument, as used in Kowalski (2015)

$$\text{with } K(p) = E\{U_0|P(Z)=p\} + E\{U_1 - U_0|P(Z)=p\}p \tag{12}$$

$K(p)$ is a nonlinear function of the propensity score. It serves here as a control function, as defined by Heckman and Robb (1985). It takes into account the fact that the difference between the outcome and the specification on the right-hand side is a function of p. Hence, a regression applied on (11) consistently estimates parameters (β_0, β_1). Note that the fact that $K(p)$ does not depend on X reflects the assumption that the slope of the MTE curve does not depend on observable characteristics.

As stated above, the MTE are computed as the partial derivative of the conditional expectation of Y with respect to $P(Z)$:

$$\frac{\partial E\{Y|X=x, P(Z)=p\}}{\partial p} = x(\beta_1 - \beta_0) + \frac{\partial K(p)}{\partial p} \tag{13}$$

The estimation of the outcome equations requires a first stage estimation of the propensity score for each individual, $\hat{P}(z) = Pr(Z\gamma > V|Z=z) = p$. The propensity score can be fitted by a probit or logit model[7].

Writing the control function $K(p)$ as a polynomial in p , equation (11) becomes:

$$E\{Y|X=x, P(Z)=p\} = x\beta_0 + \{x(\beta_1 - \beta_0)\}p + \sum_{i=1}^{\vartheta} \phi_i p^i \tag{14}$$

A parametric estimation of the MTE can be obtained from:

$$MTE\{X=x, P(Z)=p\} = x(\beta_1 - \beta_0) + \sum_{i=1}^{\vartheta} i\phi_i p^{i-1}, \tag{15}$$

using the estimations of $\widehat{\beta_1 - \beta_0}$ and $\widehat{\phi_i}$ obtained from the linear regression implied by (14).
Alternatively one can adopt a semi-parametric approach by running a local polynomial regression (Fan and Gijbels 1996) on

$$\widetilde{y} = y - x\widehat{\beta}_0 - \{x(\widehat{\beta_1 - \beta_0})\}p.$$

The semi-parametric estimator can only be estimated on the common support of the propensity score. Precisely, the common support assumption requires that there exist positive frequencies of $\hat{P}(z)$ for individuals that receive $(D=1)$ and do not receive $(D=0)$ the treatment. It is worth noting that,

[7]It is preferable not to consider a linear probability model because it does not allows to constrain the range of $\hat{P}(z)$ to be $(0,1)$, see Brave et al. (2014).

although a parametric estimator of MTE can be estimated on the whole range $[0,1]$, its precision also crucially depends on the common support (Brave et al. 2014). Therefore, our interpretation of the results will be limited to the common support.

3 Data and empirical strategy

We use a data set from a French supplementary insurer: *Mutuelle Générale de l'Education Nationale* (MGEN), which is a not-for-profit insurer who provides mandatory basic health insurance for teachers and Ministry of education's employees. MGEN also supplies supplementary health insurance in the form of a unique[8] contract (MGEN-SHI) which offers a minimal supplementary coverage: it covers only copayments and not balance billing. People can subscribe to this MGEN-SHI on a voluntary basis, or take out another SHI. For historical reasons, MGEN manages both basic (NHI) and supplementary insurance (MGEN-SHI). Our data stemmed from administrative MGEN data: they provide, for each policyholder, detailed information about her medical bills and reimbursements for basic health insurance and for supplementary insurance when the individual is a MGEN-SHI subscriber.

In France, ambulatory care is mostly provided by self-employed physicians paid on a fee-for-service basis. Since 1980, physicians can choose between two contractual arrangements with the regulator. If they join "Sector 1", physicians are not permitted to balance bill. They agree to charge their patients the reference fee (23€ in 2012 for a routine visit), and get fiscal deductions in return. If they join "Sector 2", they are allowed to set their own fees. Access to Sector 2 being strongly limited for GPs since 1990, most of them belong to Sector 1: they are 87% in 2012. Hence the issue of balance billing concerns mostly specialists. Balance billing adds 35% to the annual earnings of Sector 2 specialists. The average proportion of specialists operating in Sector 2 amounts to 42% in 2012. This proportion varies dramatically across specialties: for instance, the proportion of specialists operating in Sector 2 is 19% for cardiologists, 73% for surgeons and 53% for ophthalmologists.

Actually, we do not observe the coverage of balance billing for people who subscribed to another SHI than MGEN-SHI. However, MGEN used to send a questionnaire to people who switched to another SHI. This allows us to know, for people who have terminated a MGEN-SHI contract, if they have subscribed to another SHI. For this reason, we selected, for year 2012, a sample of subscribers of MGEN-SHI and

[8]This is true for our observational period. From 2016 on, MGEN started to supply a choice between different contracts for SHI.

of subscribers of another SHI, who were in 2010 subscribers of MGEN-SHI and have terminated their contract in 2011. In this case, we know that their new coverage will be at least equal and probably better than before, because MGEN-SHI coverage on balance billing is zero. We name this new contract 'better-SHI'.

Because in France balance billing concerns mostly specialists, our analysis focuses on the impact of coverage of balance billing on the use of specialists. We leave the differences in differences approach used in Dormont and Péron (2016) to specify, on a cross section of individuals observed in 2012, a Roy model for the issue at stake. It is a switching regression model that explains together the decision to take out coverage for balance billing (better-SHI), and the demand for consultations with balance billing when the individual is – or is not – covered for balance billing. As stated above, such a specification enables us to estimate the impact of better coverage on the use of balance billing in case of essential heterogeneity. For that purpose, we use an instrument which explains the decision to take out better coverage and which is not directly related to balance billing consumption.

Our original sample was composed of 91,629 subscribers of MGEN-SHI and 8,249 subscribers of better-SHI. We excluded individuals who live outside continental France as well as the top 1% of care users in 2012. Because we focus on specialist consultations, we only keep individuals who have at least one visit to a specialist in 2012, with or without balance billing. Our final sample includes 58,519 individuals: 53,456 subscribers of MGEN-SHI and 5,063 subscribers of better-SHI, observed in 2012, who have visited a specialist at least once in 2012.

Our empirical strategy requires the use of an instrument to explain the decision to terminate MGEN-SHI contract in order to take out a better-SHI. A valid instrument must be correlated to the decision to quit MGEN-SHI and be uncorrelated to the consumption of balance billing (in the Roy model, we assume that $U_0, \quad U_1$ and V are independent of Z and X). The decision to retire in 2011 for people younger than 55 years-old, that we used in Dormont and Péron (2016), is a reliable instrument. The age threshold refers to a specific right for teachers and civil-servant who raised three or more children to retire before 55. This right has been revoked in January 2012, creating an important incentive for individuals meeting the criteria to retire in 2011. Indeed, MGEN-SHI premiums raise from 2.97% of wages before retirement to 3.56% of pensions after. We argue that this retirement policy change creates an exogeneous shock that gives individuals incentives to terminate their MGEN-SHI contract for a better-SHI, but has no reason to drive their balance billing consumption. We do not observe eligibility (the number of children raised is not available) so we use the decision to retire before 55. In our sample, 368 individuals decided to retire in 2011 and half of them quit MGEN-SHI the same year. We included retirement before 55 as a covariate in a simple log-linear model that explained balance billing consumption in 2010, when all individuals had the MHEN-SHI coverage; the coefficient was non significantly different from zero. Therefore, we decided to rely on the 'early retirees' instrument to explain the decision to subscribe to better-SHI[9].

Our data provide, for each individual in 2012 the number of visits to a specialist Q, including the number of visits to S2 specialists who charge balance billing, $Q2$, as well as the total amount of balance billing, BB. We focus on four variables of interest: the number of specialist consultations, Q (with $Q \geq 1$), the proportion of S2 consultations, $Q2/Q$, the average balance billing per consultation, BB/Q and the average balance billing per S2 consultation $BB/Q2$ (computed only for individuals who have at least one S2 consultation in 2012). We are able to distinguish three dimensions in the demand for specialist consultations: quantity of specialist consultations, quality in terms of choice between S1 and S2 specialists and finally the average price per consultation to a S2 specialist, which might be linked to quality.

Given that our data do not provide the fee for each consultation, we compute for each individual an annual average of balance billing per consultation. However, we are able to control for the individual's needs regarding medical specialties. This is important because, as shown in Dormont and Péron (2016), the availability of S1 and S2 specialists varies dramatically from a specialty to another in France. Gynecologists, ophthalmologists, surgeons and ENT specialists[10] charge balance billing in a larger proportion than their colleagues. As a matter of fact, patients' choice to visit a S2 is likely to be far more constrained when they need to visit one of these specialties. We therefore use a dummy variable called 'expensive physicians' ($ExpPhy$) which equals 1 when the individual visited one of these specialists at least once in 2012.

Our information on individual characteristics include gender, age, income and health status, all measured in 2012. The impact of age is modelled with three age groups: 20-40, 40-60 and over 60 years old. Our income variable is based on individuals' wage or pension used by MGEN to compute MGEN-SHI

[9]Note that the condition of independence between the instrument and balance billing consumption is more demanding with cross-sectional data than it was with panel data in Dormont and Péron (2016), where the specification of individual fixed effects makes it possible to deal with time-invariant sources of non exogeneity. In this framework, the need of excluded instruments was only dictated by possible unobservable health or information shocks that would have explained both the switch of SHI and a 'change' in balance billing consumption. Here we need an instrument that is not correlated with the 'level' of balance billing consumption. As explained above, this condition is fulfilled for 'retirement before the age of 55'. But it is not the case for the fact of 'moving out to another *département*'. We cannot use this variable as an instrument for our cross-section analysis.

[10]Ear, Nose and Throat specialists

premiums. The dummy CD, which equals 1 if individuals have a chronic disease, is used as an indicator of health status. Access to S1 or S2 specialists is not only a question of price (balance billing or not), but also a question of geographical access (transportation costs) or waiting time. To measure the respective availability of S1 or S2 specialists, we use the 'specialist : population ratios' (SPR) provided by national statistics in 2012. The SPR is the number of specialists either in S1 ($SPR1$) or in S2 ($SPR2$) per 100,000 inhabitants in each *département*.

3.1 Basic features of the data

Table 1 displays the characteristics of the 58,519 individuals of our final sample: there is a high proportion of women (72.5%), the average age is close to 58 years, the average income amounts to €2,500 and 22% have a chronic disease. In comparison, the average wage is in France equal to €2,157[11] and 19.5%[12] of people have a chronic disease. These characteristics derive from the fact that (i) MGEN covers teachers and civil servant who have a certain education level and are mostly women; (ii) we have restricted our sample to those who visited a specialist at least once in 2012.

Compared to MGEN-SHI holders, better-SHI holders are on average 12 years younger, count more women (82% vs 72.5%) and less individuals with chronic disease (9.4% vs 22%). To sum up, those who decided to quit MGEN-SHI are on average younger and healthier. This is a common result in the literature on switching behavior: in the USA (Buchmueller and Feldstein 1997, Strombom et al. 2002), Switzerland (Dormont et al. 2009) or in the Netherlands (Duijmelinck and van de Ven 2016), switchers are invariably younger and also tend to be healthier. We discuss the motivations to subscribe to better-SHI further in the paper.

Table 2 displays statistics about the use of specialist visits and balance billing in 2010 and 2012 for MGEN-SHI holders and future better-SHI holders (who are covered by MGEN-SHI in 2010 and better-SHI in 2012)[13]. Of course, in 2012, better-SHI holders are likely to have a better coverage for balance billing than MGEN-SHI holders. Whereas the total number of visits (Q) is not significantly different between MGEN-SHI and better-SHI holders, the latter consume significantly more balance billing, both in quantity ($Q2 = 1.7$ for better-SHI holders vs 1.3 for MGEN-SHI holders) and price ($BB/Q2 = 26.1$ vs 24.2). Consequently, better-SHI holders' mean consumption of balance billing, (BB), amounts to €46.9 in 2012, which is 42.6% higher than for MGEN-SHI holders.

[11] Average net mensual wage in 2012; source: INSEE

[12] source: ESPS survey

[13] This comparison is not possible for all the 58,519 individuals observed in 2012 since only 43,612 of them used at least a specialist visit in 2010.

These differences might reflect adverse selection, as well as moral hazard and, if there is heterogeneity in moral hazard, possible selection on moral hazard. Actually, our data design enables us to observe the use of balance billing by better-SHI subscribers in 2010, before they take out better coverage. In 2010, all individuals in our sample, including future better-SHI, are all MGEN-SHI holders, hence not covered for balance billing. Table 2 shows that in 2010 the future better-SHI holders, who will quit MGEN-SHI the next year, consumed more balance billing than those meant to stay under MGEN-SHI contract. This reveals classical adverse selection: those who ask for better coverage consume more balance billing than others.

4 Empirical specification

The aim of this paper is to estimate the effect of health insurance on the consumption of balance billing when moral hazard is heterogeneous. Assuming that moral hazard may be related to the decision to choose a better coverage for balance billing, we estimate MTE to capture heterogeneity in response to health insurance and to test for essential heterogeneity. Also, our estimation strategy enables us to evaluate the effect of observable characteristics, such as income, on the consumption of balance billing, on the demand for better SHI coverage and on moral hazard.

4.1 Model and estimation

Following the generalized Roy model presented in section 2.1, we specify a choice equation explaining the individual's decision to take out another SHI to enjoy better coverage (better-SHI) than the one provided by MGEN-SHI. The estimation of this choice equation enables us to understand coverage choices' determinants and provides the propensity scores that are used to identify MTE.

The choice is specified through the binary variable D, which is equal to 1 if the individual chooses to take out better-SHI in 2011. In 2012, people covered by better-SHI benefit from balance billing coverage whilst MGEN-SHI enrollees (those who stayed) do not. The decision depends on the sign of a continuous latent variable D^*:

$$
\begin{aligned}
D^* &= x\gamma_1 + \gamma_2 EarlyRetiree - V = Z\gamma - V && (16)\\
D &= \begin{cases} 1 & \text{if } D^* > 0 \\ 0 & \text{if } D^* \leq 0 \end{cases} && (17)
\end{aligned}
$$

$EarlyRetiree$ is our excluded instrument: the decision to retire before 55 years old is correlated with the decision to subscribe to better-SHI, but not with the consumption of balance billing. x is a vector of covariates which includes individuals' gender, age, income and whether they suffer from a chronic disease. It includes also local availability of specialists of sector 1 (S1, not allowed to charge balance billing) and 2 (S2, allowed to charge balance billing) and for the individual's needs as regards medical specialty (the proportion of S2 specialists is particularly high for ophthalmologists, gynaecologists and ENT). V is an unobservable random variable corresponding to the individual idiosyncratic disutility of choosing better-SHI (linked with unobservable individual characteristics such as disutility of administrative switching costs, belief that sector 2 doctors provide better quality of care, and risk aversion, i.e. utility of coverage for given risk level).

$P(Z)$ is the propensity score, i.e. the probability of choosing better-SHI conditional on Z. As explained in section 2, it is useful to define $U_D = F_V(V)$, where F_V is the cumulative function of V. U_D is a random variable uniformly distributed between 0 and 1 and values of U_D correspond to quantiles of V. For a given level of Z, individuals who have a large U_D are less likely to take out better-SHI.

$$D = 1 \Leftrightarrow Z\gamma > V \Leftrightarrow F_V(Z\gamma) > F_V(V) \Leftrightarrow P(Z) > U_D$$

We rely on the parametric and semi-parametric approaches presented in section 2.3 to estimate MTE. We estimate the propensity score $\hat{P}(z) = p$ for each individual with a Probit model[14]. We then determine the common support, i.e. the values of $\hat{P}(z) = p$ for which we have positive frequencies of individuals who decided to take out better-SHI ($D = 1$) and of individuals who remained MGEN-SHI enrollees ($D = 0$).

Then we perform OLS on equation (14), assuming that the function $K(p)$ is a polynomial of degree 3:

$$y = x\beta_0 + \{x(\beta_1 - \beta_0)\}p + \phi_1 p + \phi_2 p^2 + \phi_3 p^3 \tag{18}$$

[14]The results are robust to the use of a Logit model.

y is the log-transformation of one of our four variable of interest: Q the number of specialists consultations, $Q2/Q$ the proportion of S2 consultations in the total of visits to a specialist, BB/Q the average amount of balance billing per visit, $BB/Q2$ the average amount of balance billing per S2 visit. As for the choice equation, x is a vector of covariates which includes individuals' gender, age, income, chronic disease, local availability of S1 and S2 specialists, and individual's needs regarding ophthalmologists, gynaecologists and ENT. Subscript 1 (respectively, 0) refers to better-SHI enrollees (respectively, to $MGEN$-SHI enrollees). Better-SHI enrollees benefit from balance billing coverage, but this is not the case for MGEN-SHI enrollees. According to the Roy model, when an individual chooses to switch from MGEN-SHI to better-SHI, his or her behavior switches from $Y_0 = X\beta_0 + U_0$ to $Y_1 = X\beta_1 + U_1$.

The parametric estimator of MTE is computed for given values x as

$$MTE\{x, p\} = x(\beta_1 - \beta_0) + \phi_1 + \phi_2 p + \phi_3 p^2 \tag{19}$$

In our setting, MTE capture the effect of having better balance billing coverage for the individual 'at the margin', who is indifferent between subscribing to better-SHI or remaining enrolled in MGEN-SHI ($U_D = p$).

We also compute a semi-parametric estimator of MTE by running a local polynomial regression of $\widetilde{y}$ on p with:

$$\widetilde{y} = y - x\widehat{\beta}_0 - \{x(\widehat{\beta_1 - \beta_0})\}p. \tag{20}$$

Note that the semi-parametric approach differs only in the estimation of the unobserved component $K(p)$. To run the estimations, we use the Stata command ***margte*** (Brave et al. 2014) with a polynomial of degree 3 to estimate the parameters of the MTE. We use an epanechikov kernel function in the nonparametric estimation. Standard errors are computed using bootstrap (50 reps). Parametric and semi-parametric MTE are computed at mean values of x as in equations (21) and (22):

$$MTE\{\overline{x}, p\} = \overline{x}(\beta_1 - \beta_0) + \phi_1 + \phi_2 p + \phi_3 p^2 \tag{21}$$

$$MTE\{\overline{x}, p\} = \overline{x}(\beta_1 - \beta_0) + \frac{\widehat{\partial K(p)}}{\partial p} \tag{22}$$

4.2 Interpretation of the estimates

Our empirical specification allows for a detailed analysis as regards the impacts of observable characteristics:

- β_0 captures the impacts of individual characteristics on the demand for S2 consultations without balance billing coverage;
- γ captures their effect in the decision to switch;
- In addition, we estimate the change $(\beta_1 - \beta_0)$ in the impact of regressors which is due to better coverage.

Note that in our model the fact that the impacts of regressors can be modified by better coverage is a source of heterogeneity in moral hazard that comes in addition to the heterogeneity linked to unobserved characteristics. Suppose that $(\beta_1 - \beta_0) < 0$ for income. This would mean that low-income people react more strongly to insurance.

In what follows, we first examine the estimates obtained for β_0, γ and $(\beta_1 - \beta_0)$. Then we compare their signs to identify the situations of *classical adverse selection* (relationship between β_0 and γ) and the situations of *selection on moral hazard* (relationship between $(\beta_1 - \beta_0)$ and γ).

As regards essential heterogeneity, Heckman et al. (2006) propose a simple test to explore the assumption of a variable treatment effect due to unobservable characteristics. The joint significance of the polynomial coefficients ϕ_1, ϕ_2, ϕ_3 in equation (18) reveals the presence of essential heterogeneity. Indeed, the signs of ϕ_2 and ϕ_3 determine the slope of the curve that characterizes the relationship between the treatment effect and the value of U_D. Precisely, $\phi_1 = \phi_2 = \phi_3 = 0$ would mean that the treatment effect does not vary with unobservable characteristics, i.e. there is no evidence of essential heterogeneity. On the contrary, depending on the values of ϕ_2 and ϕ_3 , one can find that individuals with a low (or high) disutility to switch benefit more (or less) from better balance billing coverage.

Because the common support is not defined for all values of U_D between 0 and 1, we are not able to compute an ATE with the semi-parametric approach. Note that although parametric MTE are estimated on $[0, 1]$, their precision strongly decreases for $U_D > 0.35$ which makes the value of MTE difficult to interpret for higher values of U_D. So, in any case, we restrict our analysis of MTE on the values corresponding to the common support.

5 Results

Results are presented in Tables 3 to 7. Table 3 displays the effects of observable individual characteristics on the demand for better-SHI. Table 4 displays the effect of observed characteristics on consumption for balance billing without coverage and on moral hazard. Table 5 summarizes the influence of observed characteristics and gives evidence of adverse selection and selection on moral hazard. Tables 6 and 7 show evidence of heterogeneity on moral hazard. Figures 3 and 4 display respectively parametric and semi-parametric MTE over U_D evaluated at mean values of x with 95% confidence intervals computed from a non-parametric bootstrap.

5.1 Influence of observed characteristics: consumption of balance billing without coverage

The determinants of the amount of balance billing paid by patient who do not benefit from insurance coverage are captured by the coefficients β_0 (Table 4). Income, medical needs and availability of S1 and S2 specialists appear as the main determinants. The average amount of balance billing per consultation significantly increases with income: a 10% increase in income drives up BB/Q by 5.3%. Individuals aged of 60 years old and more, those who suffer from a chronic disease or visit gynaecologists, ophthalmologists or ENT specialists consume also more balance billing than others. The availability of S1 and S2 specialists has also a very strong impact on the amount of balance billing paid by patients. BB/Q is 18% higher for patients living in *départements* where the number of S1 specialists is low and 56% higher for those who lived in *départements* where S2 specialists are numerous.[15]

5.2 Influence of observed characteristics: demand for better coverage

The effects of observed individual characteristics on the probability of subscribing to better-SHI are captured through the coefficients γ in the first step of the estimation (Table 3). We find that in our sample, young and healthy (with no chronic disease, CD=0) individuals are more likely to quit MGEN-SHI. Low income individuals are more likely to take out better-SHI than high income. Individuals who live in *départements* where there are few S1 specialists or a lot of S2 specialists are also more likely to take out better-SHI.

[15] For the sake of interpretation, we use three categories for SPR1: low SPR1 includes the first third of *départements* in terms of SPR1 ($SPR1 \in [20, 41[$), medium SPR1 the second third ($SPR1 \in [41, 52[$), high SPR1 the last third ($SPR1 \in [52, 56]$). We proceed with the same method for SPR2 but only present two categories: low SPR2 includes the first third of *départements* in terms of SPR2 ($SPR2 \in [2, 15[$; medium and high SPR2 includes the second and last third ($SPR2 \in [15, 29]$).

5.3 Influence of observable characteristics: shift in the impact of better coverage

We find that better coverage induces significant changes ($\beta_1 - \beta_0$) in the impacts of regressors, resulting in heterogeneous moral hazard linked to observed characteristics (Table 4): the reaction to better coverage appears to be significantly heterogenous between different levels of income, age, genders, availibility of S1 specialist. More precisely, the effect of insurance on balance billing consumption is consistently and significantly decreasing with income: the poor react more to insurance than the rich. They increase more strongly their proportion of S2 visits and consult more expensive S2 specialists. Women react also more to a better coverage as concerns their number of consultations. The increase in quantity of consultations, Q, is 89% higher for women than for men. However, because the effect on the ratio $Q2/Q$ is also 19% lower for women, it seems that the quantity effect is mainly due to an increase in S1 visits. Compared to 40-60 years old, individuals over 60 react more to balance billing coverage. Finally, consistently with our results in Dormont and Péron (2016), moral hazard on BB/Q is 156% higher (+€21.11) in *départements* with low SPR1 and 178% higher (+€25.36) in *départements* with high SPR2.

5.4 Influence of observed characteristics: relationships between balance billing consumption, demand for insurance and reaction to better coverage

Classical adverse selection means that patients with a higher balance billing consumption without coverage are more likely to take out better coverage: it can be captured through the relationship between γ and β_0. Selection on moral hazard means that patients with a stronger reaction to balance billing coverage are more likely to take out better coverage: it can be captured through the relationship between γ *and* ($\beta_1 - \beta_0$). Table 5 summarizes our findings for different explanatory variables: it shows that classical adverse selection and moral hazard do not always go in the same direction.

Selection on moral hazard appears clearly as regards income. Indeed, the impact of income on the decision to take out better coverage is negative ($\gamma < 0$), positive for the use of balance billing BB/Q with no coverage for it ($\beta_0 > 0$), and its influence on balance billing decreases with better coverage ($\beta_1 - \beta_0 < 0$). We can deduce from this that low income individuals present a relatively low classical adverse selection but react strongly to health insurance and are more likely to switch. This findings that low income people react more to an improvement in coverage seems to us particularly interesting. Assuming that all individuals have the same marginal rate of substitution between medical services and consumption of other goods, such a result can be seen as an empirical evidence of Nyman's interpretation of moral hazard (Nyman

1999, 2003). Poor people would react more to coverage than others because better coverage not only changes the relative price of consultations with balance billing, but also releases their budget constraint.

Table 5 shows also that individuals living in *départements* with few S1 specialists show both classical adverse selection and selection on moral hazard, which explains their high motivation to switch. On the contrary, old individuals who also consume a lot of balance billing and would react strongly to health insurance are less likely to switch. The switching costs are probably too high considering that, for individuals over 60, MGEN premiums are on average lower than the competition which generally uses age-based premiums.

5.5 Heterogeneity in moral hazard depending on unobserved characteristics

Is moral hazard heterogeneous depending on unobserved characteristics? Is it related to the decision to quit MGEN-SHI? A simple test of joint significance on the terms of the propensity score polynomial shows that we have to reject the hypothesis of a homogenous treatment effect (Table 6). Furthermore, the signs of p^2 and p^3 give us the form of the MTE function depending on U_D. Table 7 compares IV estimates, as well as semi-parametric estimates of ATE and MTE for different values of p. Figure 4 plots the semi-parametric MTE depending on U_D with 95% confidence intervals and all covariates at their mean value. Because the common support is relatively restricted (Figure 2), roughly for p included in $[0.02, 0.35]$, we cannot interpret the MTE results for $U_D > 0.35$. Similarly, Figure 3 plots parametric MTE depending on U_D with 95% confidence intervals and all covariates at their mean value. Results are very close to semi-parametric estimates.

The MTE of better health insurance on $Q2/Q$, BB/Q and $BB/Q2$ is decreasing in U_D. This shows selection on moral hazard: individuals who are more likely to take out better coverage have a stronger reaction to health insurance because of unobserved characteristics. We find the contrary for the MTE of better health insurance on Q: it is increasing in U_D: those who are the less prone to take out better coverage show moral hazard in the number of specialist consultations only (of any sector, 1 or 2).

To interpret this result, we need to go back to the model specified in equation (16). U_D corresponds to quantiles of V. For a given propensity score, the decision to take out better SHI depends on the value of V ($Z\gamma > V$). The lower V, the higher the probability of choosing better-SHI. V can be linked with unobserved individual characteristics such as disutility (V_1) of administrative switching costs, belief (V_2) that sector 2 doctors provide better quality of care, or risk aversion (V_3). Assuming for simplicity that risk aversion is homogenous across individuals, the decision depends on $V_1 - V_2$: better-SHI subscription is restrained by the disutility of switching costs (V_1) but encouraged by faith in better quality (V_2). Following this interpretation, individuals who are the more prone to switch for better SHI are those with the stronger faith in the quality of care provided in sector 2[16].

Our findings give empirical support for such a story: the highest impact of better coverage on balance billing consumption (BB/Q) is observed for the first switchers. For the first decile of U_D (i.e. of V), they increase their balance billing per consultation by €111.9 (Table 7). Then MTE decrease for higher values of U_D and become non significant for values between 0.2 and 0.3 (Figure 4). Similar results are found for $log(Q2/Q)$, $log(BB/Q)$, and $log(BB/Q2)$, which are all variables measuring the use of sector 2 consultations.

The reverse is found for $log(Q)$, i.e. the number of specialist consultations (either in sector 1 or 2). For this variable, MTE are increasing with U_D as concerns the parametric estimation and are increasing with U_D but generally non significant in the semi parametric estimation. In any case, they are not significant for low values of U_D. These individuals do not believe that sector 2 specialists provide better quality of care (or do not value this quality). Hence the disutility of administrative costs delays their decision to take out better-SHI. Also, the improvement in coverage has no impact on their use of sector 2 specialists. If any significant impact, it is only on the number of consultations without distinction between sectors.

Obviously, this interpretation is based on a story on the 'content' of the unobserved components of the decision to subscribe to better-SHI. Nevertheless, the contrast between the decreasing profiles of MTE regarding balance billing use ($Q2/Q$, BB/Q and $BB/Q2$) and the increasing or flat profile of MTE regarding the use of specialist consultations provide a strong support to our econometric approach. In any case, our results are coherent with the expected effect of heterogeneous beliefs in the quality of sector 2 specialists.

6 Conclusion

When insurance is voluntary, some individuals may buy insurance because they expect an increase in their consumption due to better coverage. Defined as 'selection on moral hazard' by Einav et al. (2013),

[16]In our specification, Z is by definition uncorrelated with V, U_1 and U_0, while V can be correlated with the unobserved components, U_1 and U_0, in the demand for balance billing or for consultations. While there is only one V driving the decision to switch, U_1 and U_0 are different for each of our four variables of interest Q, $Q2/Q$, BB/Q and $BB/Q2$.

this phenomenon is likely to play a preponderant role in a context of supplementary health insurance, where subscription is voluntary.

In this paper we investigate the relationships between healthcare use, decision to take out supplementary health insurance and response to better coverage. We use a model that specifies individual heterogeneity in demand for healthcare and in moral hazard. We focus on the demand for specialists who balance bill their patients, i. e. charge them more than the regulated fee set by NHI. Indeed, the demand for specialists who balance bill relies on preferences and beliefs in quality of care. Individuals are likely to be heterogeneous in their preferences and beliefs, while these unobserved characteristics both drive demand for care and decision to take out SHI, resulting in selection on moral hazard.

In the econometric literature, selection on moral hazard is generally known as 'essential heterogeneity'. Marginal treatment effects estimators have been developed to capture the impact of a treatment likely to vary across individuals. We use MTE to estimate the causal effect of SHI coverage on balance billing consumption on a French database of 58,519 individuals observed in 2012.

We find evidence of individual heterogeneity in the response to better coverage and of selection on moral hazard. Individuals with unobserved characteristics that make them more likely to subscribe to comprehensive SHI are also those who exhibit stronger moral hazard, i. e. a larger increase in balance billing per consultation. As concerns the influence of observed characteristics, we also find that individuals' income is a determinant of balance billing consumption and influences the behavioral response to better coverage. Without coverage, the poor consume less balance billing than the rich but increase their consumption more sharply once covered. They are also more likely to take out comprehensive coverage.

In a context where SHI is voluntary, the inflationary impact of SHI coverage on balance billing might be worsened by selection on moral hazard. Our policy conclusions as regards the role of income are of different nature. The negative effect of income on the demand for balance billing consultations coupled with its positive effect on moral hazard provides evidence that insurance plays an important role in terms of access to care for low-income individuals.

Tables and Figures

Figure 1: Treatment choice for given propensity score $P(Z)$ and values of disutility U_D

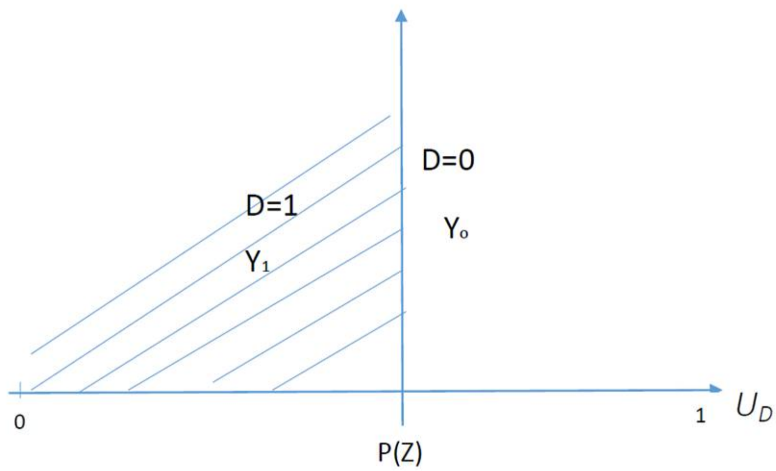


Table 1: Number of MGEN-SHI and better-SHI holders and individual characteristics in 2012 for individuals with at least one visit to a specialist ($Q \geq 1$)

	N	Women %	Age mean (sd)	Income mean (sd)	Chronic Disease %
MGEN-SHI holders	53,456	72.5	57.7 (15.2)	2,499 (764)	22
better-SHI holders	5,063	82***	45.2*** (13.3)	2,406*** (712)	9.4***

*** Significantly different from MGEN-SHI holders, p<0.01

MGEN sample: 58,519 individuals with at least one specialist consultation in 2012

Table 2: Number of specialist visits and amount of balance billing in € in 2010 and 2012 for individuals with at least one visit to a specialist ($Q \geq 1$) in 2010 and 2012

	Q mean (sd)	Q2 mean (sd)	Q2/Q mean (sd)	BB mean (sd)	BB/Q mean (sd)	BB/Q2 mean (sd)
In 2010						
MGEN-SHI	3.6 (4.6)	1.5 (2.8)	44% (0.43)	35.1 (79.5)	10.4 (12.6)	22.8 (11.6)
Future better-SHI	3.7 (4.6)	1.8*** (3.0)	52%*** (0.43)	47.5*** (88.0)	13.2*** (13.6)	24.7*** (11.6)
In 2012						
MGEN-SHI	3.3 (3.4)	1.3 (2.1)	43% (0.43)	32.9 (68.1)	10.5 (13.4)	24.2 (11.7)
Better-SHI	3.3 (2.3)	1.7*** (2.4)	51%*** (0.43)	46.9*** (81.5)	13.7*** (14.3)	26.1*** (12.2)

*** Significantly different from MGEN-SHI holders, p<0.01

MGEN sample: 43,612 individuals with at least one specialist consultation in 2010 and 2012

$BB/Q2$: subsample of 26,557 individuals with at least one S2 specialist consultation in 2010 and 2012

Table 3: Effect of covariates and excluded instruments on the probability of taking out better coverage (PROBIT)

$\Pr(QUIT = 1)$	coef.
γ	
Women	0.08***
Log(income)	-0.14***
20-40	0.60***
40-60	ref.
60+	-0.29***
CD	-0.20***
Exp. Phy	0.09***
High SPR1	ref.
Med SPR1	0.01
Low SPR1	0.21***
Low SPR2	ref.
Med & High SPR2	0.24***
Excluded instrument	
Early retirees	1.36***
N	58,519

Figure 2: Common support

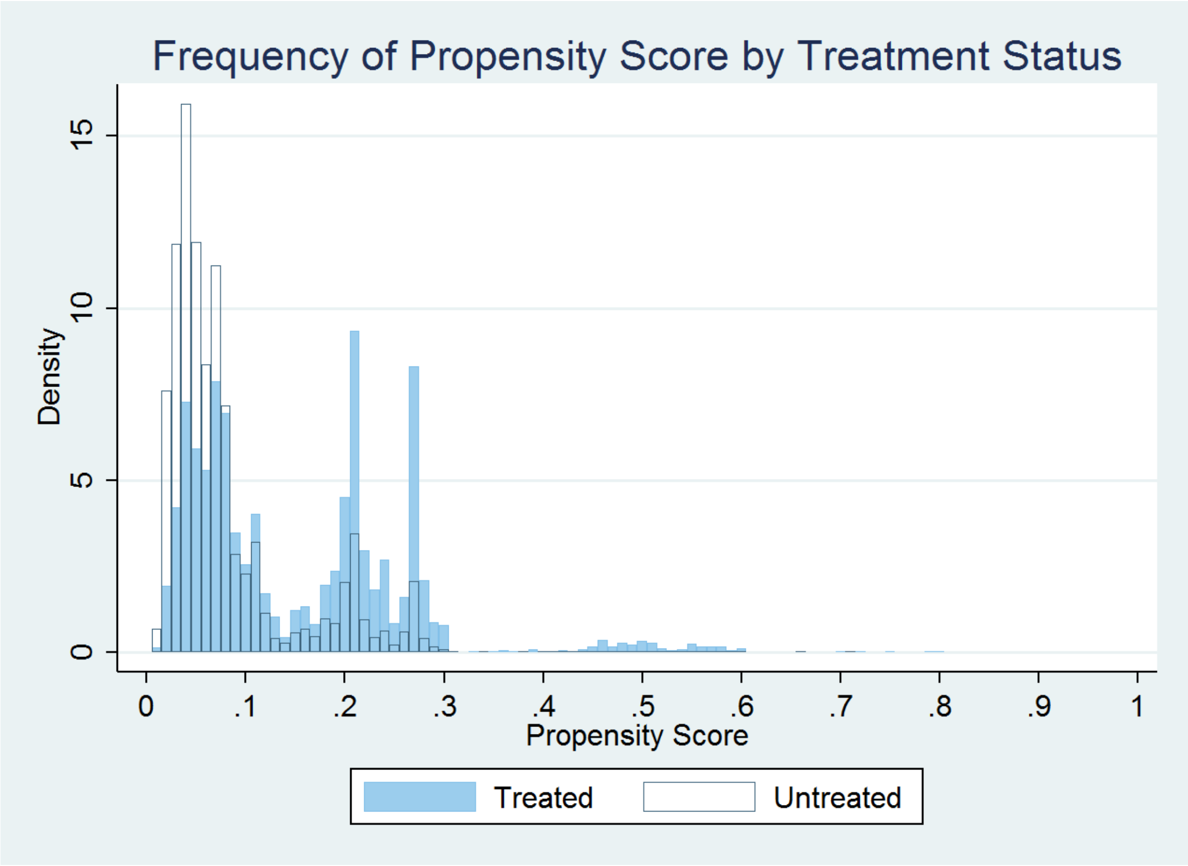


Table 4: Effect of covariates on the consumption of balance billing and on moral hazard

	log(Q)	log(Q2/Q)	log(BB/Q)	BB/Q	log(BB/Q2)
α_0	0.70***	-0.68***	-4.40***	-52.63***	0.26
β_0					
Women	0.10***	-0.01	0.00	0.13	0.05**
Log(income)	0.01	0.09***	0.53***	6.11***	0.26***
20-40	0.25**	-0.00	-0.11	-3.74**	-0.24*
40-60	ref.	ref	ref.	ref.	ref.
60+	-0.21***	0.02*	0.17***	2.45***	0.16***
CD	0.15***	0.01	0.17***	2.68***	0.13***
Exp. Phy	0.16***	0.14***	0.81***	4.22***	0.16***
High SPR1	ref.	ref	ref.	ref.	ref.
Med SPR1	-0.13***	-0.01	-0.11***	-1.74***	-0.13***
Low SPR1	-0.03	0.05***	0.18***	0.22	-0.02
Low SPR2	ref.	ref	ref.	ref.	ref.
Med & High SPR2	0.12***	0.12***	0.56***	2.73***	0.13***
$(\beta_1 - \beta_0)$					
Women	0.89***	-0.19***	-0.68***	-8.48***	-0.28*
Log(income)	-0.08	-0.19***	-1.27***	-15.26***	-0.74***
20-40	0.16	-0.15	-0.75	-7.59	0.11
40-60	ref.	ref	ref.	ref.	ref.
60+	3.01***	0.69***	4.35***	48.34***	0.47
CD	1.19***	-0.12	-0.49	-8.23***	-0.09
Exp. Phy	-0.02	-0.01	0.15	8.48***	0.43**
High SPR1	ref.	ref	ref.	ref.	ref.
Med SPR1	0.25**	-0.03	-0.01	0.77	0.17
Low SPR1	0.07	0.22***	1.56***	21.11***	0.69***
Low SPR2	ref.	ref	ref.	ref.	ref.
Med & High SPR2	0.39*	0.27***	1.78***	25.36***	0.76***
N	58,519	58,519	58,519	58,519	33,332

Table 5: Obervables: summary of relationships between probability of switching, demand for S2 specialists without coverage and moral hazard - average balance billing per consultation (BB/Q)

	Switch γ	Demand for BB β_0	Moral hazard $(\beta_1 - \beta_0)$
Women	+	NS	-
Income	-	+	-
60+	-	+	+
CD	-	+	NS
Exp. phy.	+	+	NS
Low SPR1	+	+	+
High SPR2	+	+	+

Table 6: Polynomial coefficients and joint test of significance

	log(Q)	log(Q2/Q)	log(BB/Q)	BB/Q	log(BB/Q2)
p	-6.02***	2.63***	17.41***	241.66***	10.80***
p^2	16.20***	-3.74**	-27.20***	-433.62***	-16.93***
p^3	-14.17***	3.05**	21.89***	325.62***	12.12**
chi-square statistic	14.50	12.22	25.79	56.76	19.50
p-value	0.002	0.007	0.000	0.000	0.000

Table 7: Capturing Moral hazard and the effect of unobserved characteristics: OLS, IV, empirical ATE and semi-parametric MTE

	log(Q)	log(Q2/Q)	log(BB/Q)	BB/Q	log(BB/Q2)
OLS	0.02**	0.04***	0.21***	2.03***	0.07***
IV	-0.03	0.04	0.19	1.29	-0.06
Empirical ATE	0.72	0.56***	3.21***	30.22*	0.63
MTE p=0.1	**-0.97**	**0.98****	**7.04****	**111.91****	**3.69****
lower bound	-2.56	0.34	3.76	81.89	1.13
upper bound	0.62	1.62	10.33	141.94	6.24
p=0.2	**1.15****	**0.43****	**2.23****	**23.40****	**0.48**
lower bound	0.05	0.09	0.49	1.78	-0.54
upper bound	2.24	0.76	3.96	45.01	1.50
p=0.3	**2.58***	**0.43**	**0.15**	**25.99**	**0.40**
lower bound	-0.04	-0.11	-0.10	-37.08	-1.97
upper bound	5.21	0.98	0.39	89.05	2.77

Empirical ATE: computed by STATA program ‘margte’ on the common support only

Figure 3: Parametric MTE - $log(Q)$, $log(Q2/Q)$, $log(BB/Q)$, BB/Q, $log(BB/Q2)$

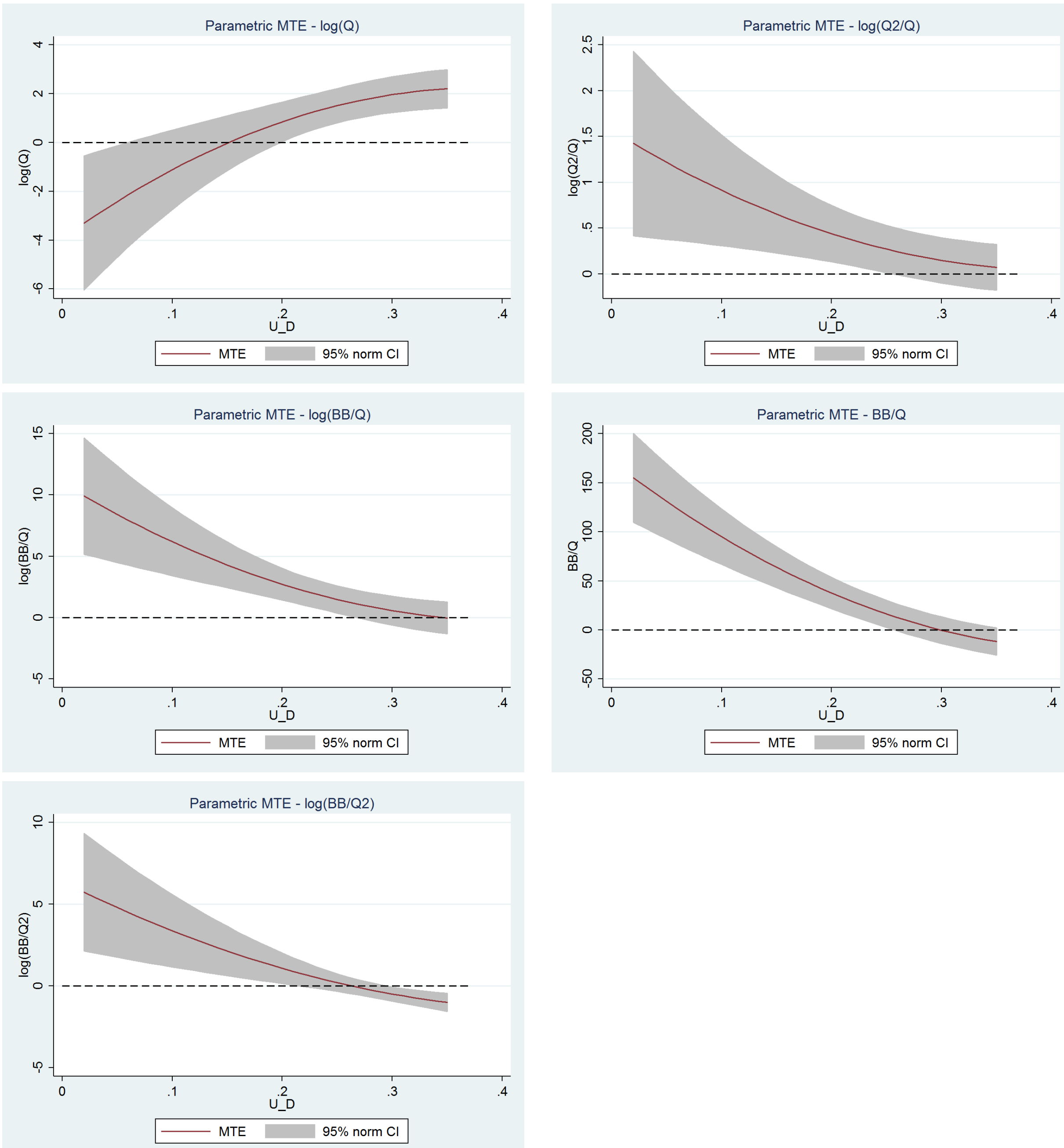


Figure 4: Semi-parametric MTE - $log(Q)$, $log(Q2/Q)$, $log(BB/Q)$, BB/Q, $log(BB/Q2)$

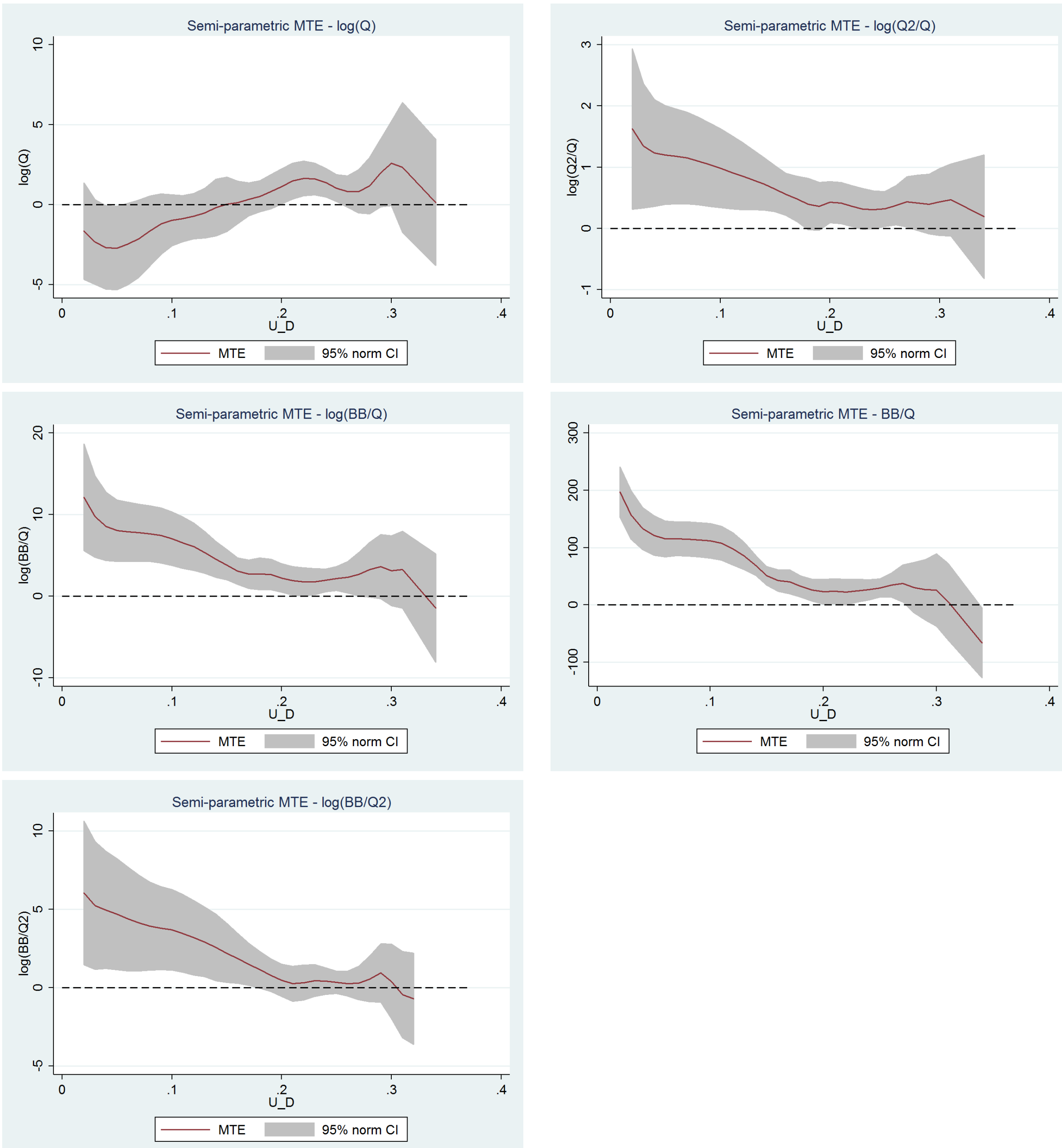
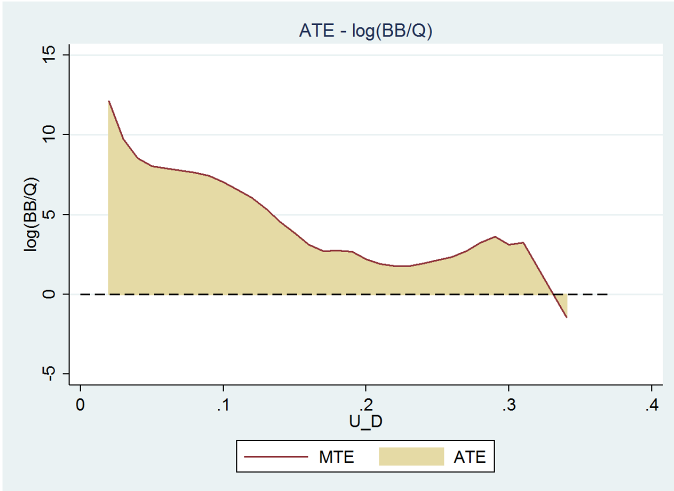


Figure 5: Empirical ATE on $log(BB/Q)$



References

Basu, A., Heckman, J. J., Navarro-Lozano, S. and Urzua, S. (2007), 'Use of instrumental variables in the presence of heterogeneity and self-selection: An application to treatments of breast cancer patients', *Health economics* **16**(11), 1133–1157.

Bjorklund, A. and Moffitt, R. (1987), 'The estimation of wage gains and welfare gains in self-selection models', *The Review of Economics and Statistics* pp. 42–49.

Brave, S., Walstrum, T. et al. (2014), 'Estimating marginal treatment effects using parametric and semi-parametric methods', *Stata Journal* **14**(1), 191–217.

Brinch, C., Mogstad, M. and Wiswall, M. (2012), 'Beyond late with a discrete instrument', *Statistics Norway Discussion Paper* (703), 703.

Buchmueller, T. C. and Couffinhal, A. (2004), 'Private health insurance in france'.

Buchmueller, T. C. and Feldstein, P. J. (1997), 'The effect of price on switching among health plans', *Journal of Health Economics* **16**(2), 231–247.

Cameron, A. C., Trivedi, P. K., Milne, F. and Piggott, J. (1988), 'A microeconometric model of the demand for health care and health insurance in australia', *The Review of economic studies* **55**(1), 85–106.

Carneiro, P., Heckman, J. J. and Vytlacil, E. J. (2011), 'Estimating marginal returns to education', *American Economic Review* **101**, 2754–2781.

Célant, N., Guillaume, S. and Rochereau, T. (2014), 'Enquête sur la santé et la protection sociale 2012', *Les rapports de l'Irdes* (556).

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**(3), 499–511.

Cornelissen, T., Dustmann, C., Raute, A. and Schönberg, U. (2016), 'From late to mte: Alternative methods for the evaluation of policy interventions', *Labour Economics* **41**, 47–60.

Coulson, N. E., Terza, J. V., Neslusan, C. A. and Stuart, B. C. (1995), 'Estimating the moral-hazard effect of supplemental medical insurance in the demand for prescription drugs by the elderly', *The American Economic Review* **85**(2), 122–126.

Dormont, B., Geoffard, P. and Lamiraud, K. (2009), 'The influence of supplementary health insurance on switching behaviour: evidence from swiss data', *Health Economics* **18**(11), 1339–1356.

Dormont, B. and Péron, M. (2016), 'Does health insurance encourage the rise in medical prices? a test on balance billing in france', *Health economics* **25**(9), 1073–1089.

Duijmelinck, D. M. and van de Ven, W. P. (2016), 'Switching rates in health insurance markets decrease with age: empirical evidence and policy implications from the netherlands', *Health Economics, Policy and Law* **11**(02), 141–159.

Einav, L., Finkelstein, A., Ryan, S. P., Schrimpf, P. and Cullen, M. R. (2013), 'Selection on moral hazard in health insurance', *The American Economic Review* **103**(1), 178–219.

Fan, J. and Gijbels, I. (1996), *Local polynomial modelling and its applications: monographs on statistics and applied probability 66*, Vol. 66, CRC Press.

Friedman, M. and Savage, L. J. (1948), 'The utility analysis of choices involving risk', *The journal of political economy* pp. 279–304.

Heckman, J. J. and Robb, R. (1985), 'Alternative methods for evaluating the impact of interventions: An overview', *Journal of econometrics* **30**(1-2), 239–267.

Heckman, J. J., Urzua, S. and Vytlacil, E. (2006), 'Understanding instrumental variables in models with essential heterogeneity', *The Review of Economics and Statistics* **88**(3), 389–432.

Heckman, J. J. and Vytlacil, E. (2001), 'Policy-relevant treatment effects', *The American Economic Review* **91**(2), 107–111.

Heckman, J. J. and Vytlacil, E. J. (2007), 'Econometric evaluation of social programs, part ii: Using the marginal treatment effect to organize alternative econometric estimators to evaluate social programs, and to forecast their effects in new environments', *Handbook of econometrics* **6**, 4875–5143.

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**(3), 513–522.

Jones, A. M., Koolman, X. and Van Doorslaer, E. (2006), 'The impact of having supplementary private health insurance on the use of specialists', *Annales d'Economie et de Statistique* pp. 251–275.

Kowalski, A. E. (2015), 'Marginal treatment effects and the external validity of the oregon health insurance experiment'.

Manning, W. G., Newhouse, J. P., Duan, N., Keeler, E. B. and Leibowitz, A. (1987), 'Health insurance and the demand for medical care: evidence from a randomized experiment', *The American economic review* pp. 251–277.

Newhouse, J. P. (1993), *Free for all?: lessons from the RAND health insurance experiment*, Harvard University Press.

Nyman, J. A. (1999), 'The value of health insurance: the access motive', *Journal of health economics* **18**(2), 141–152.

Nyman, J. A. (2003), *The theory of demand for health insurance*, Stanford University Press.

Pauly, M. V. (1968), 'The economics of moral hazard: comment', *The American Economic Review* pp. 531–537.

Schellhorn, M. (2001), 'The effect of variable health insurance deductibles on the demand for physician visits', *Health Economics* **10**(5), 441–456.

Strombom, B. A., Buchmueller, T. C. and Feldstein, P. J. (2002), 'Switching costs, price sensitivity and health plan choice', *Journal of Health economics* **21**(1), 89–116.

Vera-Hernández, Á. M. (1999), 'Duplicate coverage and demand for health care. the case of catalonia', *Health economics* **8**(7), 579–598.