fedea
Fundación de
Estudios de
Economía Aplicada

The Influence of BMI, Obesity and Overweight on Medical Costs: A Panel Data Approach

by
Toni Mora*
Joan Gil**
Antoni Sicras-Mainar**

Documento de Trabajo 2012-08

SERIE: Economía de la Salud y Hábitos de Vida
CÁTEDRA Fedea-la Caixa

October 2012

* Universitat Internacional de Catalunya and IEB (UB).
** CAEPS and University of Barcelona.
*** Badalona Serveis Assistencials (BSA).

ISSN: 1696-750X

Jorge Juan, 46
28001 Madrid -España
Tel.: +34 914 359 020
Fax: +34 915 779 575
infpub@fedea.es

N.I.F. G-78044393

THE INFLUENCE OF BMI, OBESITY AND OVERWEIGHT ON MEDICAL COSTS: A PANEL DATA APPROACH*

TONI MORA[a], JOAN GIL[b] & ANTONI SICRAS-MAINAR[c]

[a] *Universitat Internacional de Catalunya and IEB (UB), Barcelona, Spain*
[b] *CAEPS and University of Barcelona (UB), Barcelona, Spain*
[c] *Badalona Serveis Assistencials (BSA), Badalona, Barcelona, Spain*

Abstract

This paper estimates the impact of the BMI, obesity and overweight on direct medical costs. We apply panel data econometrics and use a two-part model with a longitudinal dataset of medical and administrative records of patients in primary and secondary healthcare centres in Spain followed up over seven consecutive years (2004-2010). Our findings show a positive and statistically significant impact of the BMI, obesity and overweight on annual medical costs after accounting for data restrictions, different subsamples of individuals and various econometric approaches.

JEL Classification:
Keywords: BMI and Obesity; Healthcare costs; Panel data; Two-part models.

* The authors wish to thank Badalona Serveis Assistencials (BSA) for providing us with the core dataset to carry out this research. We also acknowledge the computational resources provided by the Centre for Scientific and Academic Services of Catalonia (CESCA) and we are indebted to the Catalan Health Department and IDESCAT (the Catalan Statistics Office) for giving us access to the population census data. Toni Mora and Joan Gil gratefully acknowledge financial support from the Generalitat of Catalonia's grant programmes 2009-SGR-102 and 2009-SGR-359, respectively.

1. Introduction

Obesity is a complex, multifactorial, chronic disease involving genetic, perinatal, and environmental components. Its prevalence in Europe in the last two decades has tripled and 150 million adults and 15 million children and adolescents in the region are today estimated to be obese (Berghöfer et al., 2008). After the United Kingdom, Spain is the EU country to have recorded the highest increases in its standardised rate of obesity over this period (OECD, 2012) and ranks high in terms of overweight and obesity levels on the continent. The latest data from the European Health Survey (2009) report that 38% (16%) of Spanish adults are overweight (obese) (cf. OECD, 2012).

The epidemic is a major public health concern since obesity is a key risk factor for a range of chronic conditions (including, hypertension, diabetes, cholesterol, heart disease, stroke, gallbladder disease, biliary calculus, narcolepsy, osteoarthritis, asthma, apnoea, dyslipidaemia, gout and certain cancers) that tend to reduce the quality of life and ultimately result in death (Alberti et al., 2009; López-Suárez et al., 2008). Additionally, a significant number of obese patients tend to suffer mental disorders and social rejection leading to a loss of self-esteem, a particularly sensitive issue in the case of children (Gariepy et al., 2010). Given its prevalence and association with multiple chronic illnesses, obesity tends to increase healthcare resource utilisation and costs substantially.

The connection between obesity and the cost of healthcare in the health economics literature lies rooted in Grossman's model (1972) so that obesity impacts both the demand for health and healthcare services through the depreciation of the stock of health. Empirical evidence indicates that the obese tend to reduce the demand for health while increasing the demand for healthcare resources, thus impacting healthcare budgets.

The aim of the paper is to estimate the impact of the BMI, obesity and overweight on direct medical costs (i.e., diagnosis and treatment) by applying a two-part model (2PM). More specifically, the paper contributes to the literature in two main respects. First, we use panel data econometrics to estimate medical costs for a longitudinal dataset based on medical and administrative records of around 100,000 patients followed up over seven consecutive years (2004-2010). This is, as far as we know, the first application exploring the impact of body weight on healthcare costs using longitudinal information and its corresponding methods. Likewise, we exploit administrative data that contain objective health, weight and height (and consequently the BMI) measurements. Hence, the problems associated with self-reported data

are not an issue here. Second, we report findings for the impact of body weight on healthcare costs in a European country whose healthcare centres operate under a typical national healthcare system and strict cost-containment policies were implemented during the period of analysis. Thus, we expect a lower impact on direct medical costs compared to, for instance, the impact reported for the US, based basically on a private healthcare system.

The paper is organised as follows: Section 2 presents the related literature; Section 3 describes the empirical strategy; Section 4 describes the data; Section 5 presents the results, Section 6 discusses the main policy implications of the findings and Section 7 concludes.

2. Related Literature

A sizeable body of literature quantifies the magnitude of healthcare expenditure associated with the obesity epidemic. Barrett et al. (2011) distinguish two different lines of research on the subject. Thus, one set of studies concerns itself with the estimation of annual direct costs of obesity at an aggregate level. Most of them follow an "etiologic fraction" approach and consider the most frequent obesity-related diseases (Wolf and Colditz, 1998; Colditz, 1999; Sander and Bergemann, 2003; Vazquez-Sanchez and Alemany, 2002; Müller-Riemenschneider et al., 2008), while others make estimates relying on representative sample data (Finkelstein et al., 2004; Arterburn et al., 2005). These studies report that the proportion of national health care expenditure attributable to obesity ranges from 5.3 to 7% for the US and from 0.7 to 2.6% in other countries. In Spain, the share is reported to reach 7% of total health care expenditure.[1] A second set of studies takes a lifetime perspective and employs medical records in order to estimate the impact of BMI categories on resource utilisation and direct costs. Most are based on US data (Quesenberry et al., 1998; Thompson et al., 2001; Raebel et al., 2004; Finkelstein et al., 2005) and very few on data from other countries (Borg et al., 2005; Kakamura et al., 2007; van Baal et al., 2008).

The study we report here is conducted in line with this second set of studies. But while we employ microdata and take a longitudinal perspective, the methods adopted differ significantly. We specifically apply panel data methods which have been widely recognised in the literature on the estimation and prediction of healthcare expenditure using cross-section data. Namely, our paper is methodologically similar to those of Cawley and Meyerhoefer

[1] Among studies of this type, a number estimate medical costs and obesity based on survey data (Sturm, 2002; Andreyeva et al., 2004; Von Lengerke et al., 2006).

(2012) and Wolfenstetter (2012), although their estimations of the medical costs of obesity and overweight rely on cross-section data.[2]

3. Empirical Method

There is a plethora of investigations in the field of health economics exploring the advantages and drawbacks of the empirical methods proposed to analyse the use of healthcare services and their associated medical costs.[3] The (cross-section) datasets used for analysing such healthcare outcomes typically contain a large proportion of zero observations (non-users) as well as a long right-hand tail of individuals who make a heavy use of healthcare services and who incur high costs (skewness). Given these characteristics, OLS estimation is biased and inefficient. A good alternative for analysing these outcomes and dealing with such data problems is the well-known "hurdle" or "two-part model", which assumes that the censoring mechanism and the outcome may be modelled using two separate processes or parts (Manning et al., 1981; Duan et al., 1983; Duan et al., 1984). For instance, in explaining individual annual hospital expenses, the first part determines the probability of hospitalization, while the second part explains associated hospital expenditures conditional on being hospitalised.[4]

The traditional candidates for modelling the first part in this literature are binary regression models (i.e., probit and logit). However, much controversy exists regarding the estimation of the dependent variable in the second part. On the one hand, researchers have proposed the log transformation of costs (also the square root) before OLS estimation in order to accommodate or reduce skewness.[5] As nobody is interested in log model results *per se* (e.g., log dollars) such estimates must subsequently be retransformed to the original scale, but these retransformations can be problematic due to the impact of, for instance, heteroskedasticity (Manning, 1998). On the other hand, generalised linear models (GLMs) have recently been proposed as an alternative approach when there are unknown forms of heteroskedasticity (Mullahy, 1998; Manning and Mullahy, 2001; Buntin and Zaslavsky 2004;

[2] This is the first paper to estimate the (causal) impact of obesity on medical costs using the MEPS 2000-2005 data and applying the aforementioned methods in health econometrics.
[3] See Jones (2010) for a review of these econometric methods and their comparative performance.
[4] These two distinct processes can be understood from the perspective of a principal-agent model where the decision to contact a physician is made by the patient but the frequency of visits or continuation of treatment is decided by the doctor.
[5] Estimates based on logged models are actually often much more precise and robust than direct analyses of the unlogged original dependent variable (Manning, 1998). They may also reduce (but not eliminate) heteroskedasticity.

Manning et al. 2005). These models specify a distribution function (e.g., gamma, Poisson, or Gaussian) that reflects the relationship between the variance and the raw-scale mean functions and a link function that relates the conditional mean of medical costs to the covariates. Interestingly, GLM estimates are performed on the raw medical cost scale, so there is no need for retransformation. A further advantage is that this approach allows for heteroskedasticity through the choice of the distribution function.

3.1 Two-part model strategy

In line with previous studies, this paper estimates direct medical costs by means of a 2PM taking into account the panel structure of the data. Interestingly, in our dataset medical costs are zero for 16% of the sample and positive medical costs are highly skewed to the right. Thus, the first part of the 2PM models the probability of incurring a positive cost ($y_i > 0$) using a random-effects (RE) logit or probit binary model of the type,

$$E(y \mid x_{it}\beta, \alpha_i) = \Pr(y_i > 0 \mid x_{it}\beta, \alpha_i) = F(\alpha_i + x_{it}\beta) \qquad (1)$$

where the non-linear function $F(\cdot)$ is the logistic or the standard normal cumulative distribution function, x_{it} are the regressors and α_i is the unobserved time-invariant and individual-specific effect that is normally distributed, $\alpha_i \sim N(0, \sigma_\alpha^2)$. Then, the second part of the 2PM uses linear panel data methods to predict the mean direct medical costs conditional on positive costs. Notice that these two parts are assumed to be independent and are estimated separately. Specifically, two specifications are analysed here:

i) First, a RE generalised least squares (GLS) regression of log medical costs (log y) on a set of controls,

$$E(\log y \,|\, y_i > 0, \alpha_i, x_{it}) = x_{it}'\delta + (\alpha_i + \varepsilon_{it}) \qquad (2)$$

where x_{it} are the regressors, $\alpha_i \sim N(0, \sigma_\alpha^2)$ and $\varepsilon_{it} \sim N(0, \sigma_\varepsilon^2)$ is the idiosyncratic error term. Given that the combined error is $u_{it} = \alpha_i + \varepsilon_{it}$ [with $Var(u_{it}) = \sigma_\alpha^2 + \sigma_\varepsilon^2 = \sigma_u^2$ and $Cov(u_{it}, u_{is}) = \sigma_\alpha^2$, $s \neq t$], it follows that the RE model permits serial correlation over time: $\rho_u = Corr(u_{it}, u_{is}) =$

$\sigma_\alpha^2 / (\sigma_\alpha^2 + \sigma_\varepsilon^2)$ for all s≠t. In this model, the individual-specific effect is assumed to be uncorrelated with the explanatory variables.

If the (combined) residuals from the log medical costs in equation (3) are lognormal and homoscedastic, then the retransformation to raw scale medical costs using the exponentiation function is not a serious problem. The problems become more evident when we deviate from these circumstances. If the error terms of the logged or transformed model are not normally distributed, but are homoscedastic, the usual alternative for the retransformation has been to rely on Duan's (1983) smearing or retransformation factor, as applied in several RAND Health Insurance Experiment studies (e.g., Duan et al., 1983, 1984; Manning et al. 1987). In this case the expected value of medical costs at levels conditional on positive costs is,

$$E\left(y \mid y_i > 0, \alpha_i, x_{it}\right) = e^{\left(\hat{\alpha} + x_{it}^{'}\hat{\delta}\right)}\hat{D} \qquad (3)$$

where $\hat{\alpha}$ and $\hat{\delta}$ are consistent parameter estimates of equation (2) and $\hat{D}$ is the smearing factor, that is, the average of the exponentiated OLS residuals of the logged dependent variable ($\hat{D} = N^{-1}\sum_{i=1}^{N} e^{(\hat{u})}$) where $\hat{u} = \log y - \hat{\alpha} - x_{it}^{'}\hat{\delta}$.[6] As the typical value for the smearing factor lies between 1.5 and 4.0 in healthcare costs applications, ignoring the retransformation can result in a substantial underestimation of mean medical costs.

However, according to Manning (1998) and Mullahy (1998) this strategy is problematic when transformed errors have a heteroskedastic distribution with a variance that depends on the regressors in a non-trivial manner (i.e., $Var(u \mid x) = \sigma_u^2 h(x)$, where *h(x)* is some function of the covariates *x* that determines the heteroskedasticity). Both authors point out that OLS estimates of $E\left(y \mid y_i > 0, \alpha_i x_{it}\right)$ that ignore the possible dependence of the retransformation factor on the regressors and which, therefore, use the (homoscedastic) smearing factor instead are likely to yield biased estimates of key parameters of interest including marginal effects or elasticities.

Given the presence of heteroskedasticity (detected by means of the Breusch-Pagan or White tests), if it is produced by several covariates, some of which are continuous (i.e.,

[6] When errors are lognormally distributed and homoskedastic, u ~ $N(0,\sigma_u^2)$, then Equation (3) becomes $E\left(y \mid y_i > 0, \alpha_i x_{it}\right) = e^{\left(\hat{\alpha} + x_{it}^{'}\hat{\delta} + 0.5\sigma_u^2\right)}$.

complex heteroskedasticity), one alternative is to assume a parametric structure for the heteroskedastic error term. Here, in line with Mullahy (1998), we assume the exponential conditional mean (ECM) specification accounting for the panel structure of the data: $\sigma_u^2 h(x) = e^{(\alpha + x\gamma)}$ which ensures the positivity of the variance function. Therefore, the heteroskedasticity adjusted retransformation of the expected response of medical costs on the explanatory variables is,

$$E\left(y \mid y_i > 0, \alpha_i, x_{it}\right) = e^{\left(\hat{\alpha} + x_{it}^{'}\hat{\delta} + 0.5e^{(\alpha + x^{'}\hat{\gamma})}\right)} \tag{4}$$

where $\hat{\gamma}$ are the estimated coefficients for the logarithmic regression $\log(u^2) = \alpha + \chi_0 + \gamma_1 x_1 + \gamma_2 x_2 + ... + \gamma_k x_k + e$ and their significance indicates the main variables contributing to the heteroskedasticity. Note that equation (4) rests on the assumption of the lognormality of residuals.

As long as the purpose is to recover the estimation of the conditional expected direct medical costs in levels for the entire sample under a 2PM setting, we can write,

$$E\left(y \mid \alpha_i, x_{it}\right) = F(\alpha_i + x_{it}^{'}\hat{\beta}) e^{\left(\hat{\alpha} + x_{it}^{'}\hat{\delta} + 0.5e^{(\alpha + x_{it}^{'}\hat{\gamma})}\right)} \tag{5}$$

where F(·) is the logistic or standard normal distribution. Notice that equation (5) adopts the heteroskedasticity adjusted retransformation of the second part of the 2PM.

ii) Second, a GLM panel regression of (positive) direct medical costs on a set of controls,

$$E(y \left| y_i > 0, \alpha_i, x_{it}\right.) = \mu_i = f(\alpha_i + x_{it}^{'}\delta) \tag{6}$$

where the link function f(·), the first component of the GLM, relates the conditional mean of costs directly to the covariates. The second component is a distribution function that specifies the relationship between the variance and the conditional mean. This is often specified as a power function: $Var(y \mid y > 0, \alpha, x) = E(y \mid y > 0, \alpha, x)^{\upsilon} = u^{\upsilon}$. In order to determine which specific link (e.g., logarithm, square root or linear function) and distribution functions (e.g., gamma, Poisson or Gaussian) best fits the data, we calculated Pregibon's link test and the

Park (1966) test, respectively. However, the most frequently used GLM specifications in healthcare cost studies are the log link function and the Gamma distribution (see, for example, Manning and Mullahy, 2001; Manning et al., 2005). In this case, the expected value of medical costs for the entire sample is computed as,

$$E(y|\alpha_i, x_{it}) = F(\alpha_i + x_{it}^{'}\hat{\beta}) f(\hat{\alpha} + x^{'}\hat{\delta}) \quad (7)$$

where F(·) is again the logistic or standard normal cumulative distribution function.

In selecting these two competing approaches to analyse the impact of the BMI (or obesity categories) on mean medical costs, we are aware of their respective advantages and drawbacks. For instance, general linear modelling is recommended, as opposed to log estimation with retransformation, when complex heteroskedasticity is present and residuals are not lognormally distributed. However, Manning and Mullahy (2001) point out that GLM estimation suffers a substantial loss in precision in the face of heavy-tailed, log scale residuals or when the variance function is misspecified (see, also, Buntin and Zaslavsky, 2004; Baser, 2007). A general finding that seems to emerge from the literature that compares the performance of these two models for positive expenditure (among other methods) in terms of consistency and precision (Manning and Mullahy, 2001; Buntin and Zaslavsky, 2004; Manning et al., 2005; Baser, 2007; Hill and Miller, 2010) is that no one method dominates the other and there are important trade-offs in terms of precision and bias, mainly when different subgroups of population or types of medical costs are analysed (Hill and Miller, 2010; Jones, 2010). Notwithstanding, Mihaylova et al's (2011) literature review confirms that 2PM models perform better.

Finally, given the difficulties of finding adequate exclusion restrictions in the data, the usual procedure when estimating 2PM models is to assume the same type of regressors in both parts of the equations. Fortunately, our data provide information about the patients' relatives, so that we can construct the binary indicator of living with relatives (value 1) or alone (value 0). This indicator is then used as an exclusion restriction since we assume that living with relatives influences the decision to seek care and, hence, the incurring of positive healthcare costs (first equation), but it is irrelevant when estimating the amount of medical costs incurred (second equation).

3.2 Marginal and incremental effects in two-part models

The derivation of marginal effects (MEs) and incremental effects (IEs) in non-linear models is not as straightforward as it is in linear regression models (see Hertz, 2010). In this paper, we are interested in estimating both the ME of the BMI regressor, x_k, and the IE of the obesity regressor, x_d, on direct medical costs (measured in levels) in a two-part framework, using the above specifications in the second part.

When we estimate mean medical costs using the heteroskedasticity adjusted retransformation model, to calculate the ME of the BMI we take the partial derivative of equation (5) with respect to x_k holding the remaining covariates constant,

$$\frac{\delta E(y \mid \alpha, x)}{\delta x_k} = \left(\frac{\delta F(\alpha + x'\beta)}{\delta x_k} e^{\left(\alpha + x'\delta + 0.5e^{(x'\gamma)}\right)} \right) + \left(\frac{\delta \left(e^{\left(\alpha + x'\delta + 0.5e^{(x'\gamma)}\right)} \right)}{\delta x_k} F(\alpha + x'\beta) \right) \tag{8}$$

Now if we assume that F(·) is the cumulative logistic distribution, $\Lambda(\alpha + x'\beta) = \frac{e^{(\alpha + x'\beta)}}{1 + e^{(\alpha + x'\beta)}}$, then the ME becomes:

$$\begin{aligned} \frac{\delta E(y \mid \alpha, x)}{\delta x_k} = & \left(\beta_k \Lambda\left(\alpha + x'\beta\right)\left[1 - \Lambda\left(\alpha + x'\beta\right)\right] e^{\left(\alpha + x'\delta + 0.5e^{(\alpha + x'\gamma)}\right)} \right) \\ & + \left(\Lambda\left(\alpha + x'\beta\right)\left(e^{\left(\alpha + x'\delta + 0.5e^{(\alpha + x'\gamma)}\right)} \right)\left[\delta_k + 0.5\gamma_k e^{(\alpha + x'\gamma)} \right] \right) \end{aligned} \tag{9}$$

where the first term in equation (9) is the ME of the probability of positive medical costs with respect to the BMI and the second term measures the ME of the heteroskedasticity adjusted conditional medical costs on positive values with respect to the same regressor.

Now if we calculate the ME using the GLM specification of the second part of the 2PM model and assume the standard normal cdf for the first part $\Phi(\alpha + x'\beta) = \int_{-\infty}^{\alpha + x'\beta} \phi(z)dz$, then the partial derivative of equation (7) is,

$$\frac{\delta E(y \mid \alpha, x)}{\delta x_k} = \left(\beta_k \phi\left(\alpha + x'\beta\right) f(\alpha + x'\delta) \right) + \left(\Phi\left(\alpha + x'\beta\right) f'(\alpha + x'\delta) \right) \tag{10}$$

3.3 Econometric issues

Some of the econometric challenges posed by our panel data were adequately addressed in the estimations. First, a patient's weight and height are not always measured when visiting their doctor, which means that for a subset of individuals their BMI may present a missing value in time *t*. To overcome this problem, we restricted the sample to those individuals who had at least one weight and height measurement. Based on this information we were able to infer the individuals' BMI for the period 2004-2010. Second, since not having weight and height measurement information may induce sample selection bias, we followed Wooldridge's (2005, page 581) proposal to accommodate this impact. In other words, we ran a robust probit estimation of not having covariate measurements for each period *t* and then saved the inverse Mill's ratios. These were later added to the two-part model equations.

Third, a further issue to be addressed is that of estimating the models with fixed effects (FEs) or random effects (REs) in a panel data context. Although FEs should control for unobserved heterogeneity at the individual level, we preferred the REs option. This decision was driven by the infeasibility of estimating the same FEs in the two parts of the 2PM. To the best of our knowledge, no standard procedure can perform this. Therefore, we used REs panel estimation, which relies on the normality of the errors and the fact that errors are uncorrelated with the observed covariates (x_{it}). Fourth, to allow for the possibility that the observed BMI may be correlated with the time-invariant and individual-specific effect (α_i), we parameterised this association.[7] However, here we followed the Mundlak (1978) procedure, which uses within-individual means of the BMI rather than separate values for each year. As a consequence, the original set of regressors is augmented with the global BMI mean. Fifth, to further control for heterogeneity we considered the impact of the previous year's BMI on our regressions. Notice that although some endogenous effects may still be present, such as a health status shock (e.g., accident or a job loss) that would have a marked impact on medical spending (on traumatology or psychiatric services), we assumed that no other effects at the individual level could be controlled for.

Sixth, we also specified a dynamic panel regression model by including the medical costs incurred in the previous year as an additional regressor to capture state dependence. To deal with the initial conditions problem, we followed Albouy *et al's* (2010) proposal which modifies Wooldridge's (2005) approach. In fact, these authors proposed using the generalised

[7] In line with Chamberlain (1980), one option could be to assume that $\alpha_i = \alpha' BMI_i + u_i \sim idd\ N(0, \sigma^2)$ where $BMI_i = (BMI_{i1},..,BMI_{iT})$ are the values of the BMI for every year of the panel, and $\alpha = (\alpha_1,....., \alpha_T)$.

residual of a simple model in cross-section at the initial date but taking into account the two-part model framework. The latter can be considered the best available estimation of the over or under propensity to consume at the initial date. Seventh, a further sample selection issue of concern occurs if during the analysed period individuals drop out from the panel because of immigration, incapacity, death, etc. We found that around 3% of our total observations suffered attrition as a consequence of death. Here, the strategy adopted involved simply including a dummy on the occurrence of death rather than including an additional probability of individuals' dropping out from the panel. Eighth, to control for non-linearity, we alternatively modelled the impact of the BMI categories (e.g., overweight and obesity compared to normal weight) on both equations of the two-part model. Finally, the marginal effects were computed manually as a consequence of having transformed data and were conveniently bootstrapped.[8]

4. Data and variables

Observational and longitudinal data are drawn from the administrative and medical records of patients followed up over seven consecutive years in six primary care centres (Apenins-Montigalà, Morera-Pomar, Montgat-Tiana, Nova Lloreda, Progrés-Raval and Marti i Julià) and two reference hospitals (Hospital Municipal de Badalona and Hospital Universitari Germans Trias i Pujol), serving more than 110,000 inhabitants in the north-eastern sector of Barcelona. This population is mostly urban, of lower-middle socioeconomic status from a predominantly industrial area. Our sample includes patients aged 16+ who had at least one contact with the healthcare system between 1 January 2004 and 31 December 2010, and who were assigned to one of the aforementioned healthcare centres during this period. The study also considers those who died during the period analysed. However, we exclude subjects that were transferred or who moved to other centres and patients from other areas or regions.

This dataset incorporates a rich set of information about the individual patients' use of healthcare resources (including, number of visits to the GP; specialist and emergency care; number of hospitalizations and bed days; laboratory, radiology and other diagnostic tests; and consumption of medicines), their clinical measurements of height and weight, and each patient's chronic conditions and other diagnosed diseases (according to the ICPC-2), any functional limitations, their date of admission and discharge, type of healthcare professional(s) contacted and the motive of their visit. Moreover, the dataset includes details

[8] We thank Partha Deb for providing with the Stata codes to perform these calculations.

of each patient's age, gender, employment status (active/retired), place of birth and habitual residence.

Owing to a unique identifier, the data from the administrative and medical records can be merged with the Population Census allowing us to incorporate new variables for each patient (e.g., education or marital status) not available in the original sample.

4.1 Data on Healthcare Costs

In addition to its longitudinal nature, the dataset provides a wide array of information on healthcare costs. This includes the specific characteristics of the primary and hospital healthcare centres considered and also the extent of development of their information systems. In addition to these internal sources, costs were also calculated (where necessary) using data taken from invoices for intermediate products issued by a number of different providers and from the prices fixed by the Catalan Health Service.

The computation of healthcare costs follows a two-stage procedure: first, incurred expenditures (financial accounting) are converted into costs (analytical accounting), which are then allocated and classified accordingly.[9] Depending on the volume of activity, we consider two types of costs: fixed or semi-fixed costs and variable costs. The former include personnel (wages and salaries, indemnifications and social security contributions paid by the health centre), consumption of goods (intermediate products, health material and instruments), expenditures related to external services (cleaning and laundry), structure (building repair and conservation, clothes, and office material) and management of healthcare centres, according to the Spanish General Accounting Plan for Healthcare Centres. The latter include costs related to diagnostic and therapeutic tests and pharmaceutical consumption.[10]

Our unit of measurement is the cost per treated patient during the period in which the subject was observed and all the direct cost concepts imputed for the set of diagnosed episodes. Table 1 presents our estimates of the resulting unitary cost rates for the years 2004 and 2010. As such, the total medical costs per patient in each period are calculated as the sum of fixed and semi-fixed costs (i.e., average cost per medical visit multiplied by the number of medical visits) and variable costs (i.e., average cost per test requested multiplied by the

[9] Expenditures not directly related to care (e.g. financial spending, losses due to fixed assets, etc.) were excluded from the analysis.

[10] For instance we considered: (i) laboratory tests (haematology, biochemistry, serology and microbiology), (ii) conventional radiology (plain film requests, contrast radiology, ultrasound scans, mammograms and radiographs), (iii) complementary tests (endoscopy, electromyography, spirometry, CT, densitometry, perimetry, stress testing, echocardiography, etc.); iv) pharmaceutical prescriptions (acute, chronic or on demand).

number of tests + retail price per package at the time of prescription multiplied by the number of prescriptions). Note that in this study we do not account for the computation of 'out-of-pocket payments' paid by the patient or family, as they are not registered in the database. Healthcare costs figures were converted to 2010 Euros using the Consumer Price Index (CPI).

[Insert Table 1 around here]

4.2 Other variables

The body mass index (BMI) of each patient, our continuous variable of interest, was calculated as weight (in kilograms) divided by the square of height (in metres) using clinical or measured information, thus avoiding the traditional problems found with self-reported data. Notice that in our sample not all patients were measured when they visited the physician; however, others were measured on more than one occasion. We also computed the impact of obesity and overweight on medical costs by using the WHO classification that distinguishes between normal-weight ($18 \leq BMI \leq 24.9$ kg/m^2), overweight ($25 \leq BMI \leq 29.9$ kg/m^2) and obesity (BMI of ≥ 30 kg/m^2).[11]

To identify the impact of the BMI (or, alternatively, of obesity and overweight) on medical costs we included a wide range of covariates. First, we controlled by the patients' demographic characteristics, including age and gender, and also by immigrant status, since there is evidence that the immigrant population presents a different pattern of use and access to healthcare services. Note that non-linear age effects were considered after running the modified Hosmer-Lemeshow test. We also added a set of dummies to control for their employment status (active/retired), whether the individual was the main beneficiary of the public health insurance, and whether Catalan was their usual language of communication. Two groups of indicators were employed with respect to the individuals' health conditions that affected medical costs. On the one hand, we included the Charlson comorbidity index for each patient and the individual case-mix index obtained from the 'Adjusted Clinical Groups' (ACG), a patient classification system for iso-consumption of resources.[12] On the other hand

[11] Although the BMI is the most widely used measure of obesity, it is not free of problems. For instance, the BMI does not take into consideration body composition (adiposity vs. lean weight) or body fat distribution. This means it may fail to predict obesity among very muscular individuals and the elderly.

[12] A task force consisting of five professionals (a document administrator, two clinicians and two technical consultants) was set up to convert the ICPC-2 episodes to the International Classification of Diseases (ICD-9-CM). The criteria used varied depending on whether the relationship between the codes is null (one to none), univocal (one to one) or multiple (one to many). The operational algorithm of the Grouper ACG ® Case-Mix

we considered the number of medical episodes suffered by each patient during the period analysed as a proxy for the individual's health status. Merging these data with the Population Census allowed us to control medical costs by the patients' educational level and marital status.

We have an initial balanced panel dataset containing 706,473 observations for the whole period 2004-2010. However, when we restrict the sample to patients presenting at least one weight and height measurement, the final sample is reduced to 452,108 observations, that is, 64% of the original.

5. Results

5.1 Summary statistics

Descriptive statistics for the main set of variables used in the empirical exercise are presented in Tables 2-4. Table 2 shows that the unconditional mean annual medical costs per patient for the period is 755.11€ (in 2010 Euros), which is considerably higher than the unconditional median of 306.92€ (less than half that of the mean cost in our final sample). The skewness statistic (5.91 compared to 0 for symmetric data) and the kurtosis coefficient (82.97 compared to 3 for normal data) indicate that the distribution of costs in levels is highly skewed to the right. As expected, the logarithmic transformation reduces the range of variation of costs, narrowing the degree of skewness: the mean medical cost (5.02€) approximates to that of the median (5.73€) and the skewness (kurtosis) statistic falls to -0.97 (2.85).[13]

[Table 2 around here]

Direct medical costs are zero for 16.4% of the sample (74,144 obs.), a non-negligible portion of zeros, while the number of observations with positive medical costs is 377,964. As Table 3 shows the mean positive annual costs per patient reaches 903.09€. This figure is significantly

System consists of a series of consecutive steps to obtain the 106 mutually exclusive ACG groups, one for each patient. The application of ACG provides the resource utilization bands (RUB) so that each patient, depending on his/her overall morbidity, is grouped into one of five mutually exclusive categories (1: healthy users or very low morbidity; 2: low morbidity; 3: moderate morbidity; 4: high morbidity; and 5: very high morbidity).

[13] A comparison with the initial sample, Table 2 shows that medical costs have increased. This indicates that patients without any weight or height measurements, after having visited their physician, enjoyed a better health status and incurred lower costs.

higher for women (949.40€) than it is for men (845.96€). As expected, medical costs increase with patients' age, with a higher Charlson comorbidity index and with terminal illness.

[Table 3 around here]

Finally, Table 4 summarises the mean and standard deviation values of the variables of interest and of the controls. In our sample, the mean BMI in the period of study (2004-2010) is 26.70, corresponding to a prevalence of obesity (overweight) of 23% (36%). As expected, the mean measured BMI is slightly higher among men (26.75) than it is among women (26.67), with the prevalence of obesity being higher among women (25% vs. 21%) and overweight among men (42% vs. 31%). Notice that women represent 54% of the sample and that they are slightly older than men (48.86 vs. 47.52 years of age). The mean Charlson comorbidity index is similar for both genders although the mean number of episodes is higher among women (2.28 vs. 1.73). As for labour status, around 67% of the sample is active and the percentage of individuals who have to be dropped from the sample due to death is higher among men (3% vs. 2%).

[Table 4 around here]

5.2 BMI and direct medical costs

In Tables 5-9 we present the results of our RE panel data estimations of direct medical costs using a 2PM. These tables show the bootstrapped estimates of the MEs (IEs) of the patients' measured BMI (obesity and overweight) on medical costs using different econometric specifications. Accompanying these estimates, we also report measures of goodness of fit and of the predictive performance for each model (i.e., the auxiliary R^2, the root mean square error – RMSE, and the mean absolute prediction error - MAPE). Note that these estimations account for a wide list of controls (see Section 4.2), health district dummies and time dummy variables. In addition, as discussed previously, each model incorporates the inverse Mill's ratio of not having weight and height measurements, the global mean BMI (i.e., the Mundlak correction procedure), one-year lagged measured BMI, a dummy for the occurrence of death and a dichotomous exclusion restriction. The number of bootstrap replications is set at 200.

The first set of results (Table 5) presents the estimation of the ME of (measured) BMI on direct costs in levels following equation (9). It should be noted that the first part predicts

the probability of any medical costs being incurred assuming a panel data logit model, while the second part, in the case of positive costs, specifies a heteroskedasticity adjusted retransformed panel OLS estimation on log costs. The Shapiro-Wilk normality test of residuals rejects the null hypothesis that the log residuals are normally distributed (W=18.13, p-value=0.000). We find evidence of heteroskedasticity when regressing the squared residuals of log costs on the set of covariates (chi-squared=$1.18*10^6$, p-value=0.000). A variant of the Park test suggests that several covariates contribute to this heteroskedasticity, which justifies the adjustment of the retransformed log costs. According to the first specification in Table 5, we find that one additional unit of the BMI results in an increase of 4.712€ in annual medical costs per patient. A dynamic version of the model is also investigated in which the (log) medical costs incurred in the previous year and a one-year lagged cost indicator are included in the model (the second specification in Table 5). Interestingly, while we report a statistically significant lower marginal impact (ME of 2.775€), the auxiliary R^2 (from a regression of actual log costs on the predicted values) increases markedly up to 40.5%, suggesting an improved goodness of fit, while the RSME and MAPE errors, which measure the precision of the predictions, are significantly reduced.

[Table 5 around here]

However, as discussed above, a significant drawback of the log OLS approach is that the retransformation of the estimates back to the original scale requires knowledge of the degree and form of heteroskedasticity. As pointed out by the empirical literature (cf. Hill and Miller, 2010) such regression models, tend to perform poorly in terms of their bias and predictive accuracy, making the GLM more attractive for the second part of the two-part model. This alternative approach is additionally favoured by the fact that the Kurtosis index of log residuals from a panel OLS regression of direct medical costs has an average value of 2.9 in the data. Although this is slightly lower than the normal distribution (3), we believe that GLMs should be reasonably efficient with this degree of skewness (Manning and Mullahy 2001).[14]

Thus, in Table 6 we estimate the ME of (measured) BMI on annual direct medical costs according to equation (10). Notice that the first part specifies a panel data probit model to estimate positive medical costs while the second part uses GLM panel data regression.

[14] Cawley and Meyerhoefer (2012) follow the same strategy when estimating their models.

According to the first specification, based on Gamma GLM with log link (widely used in the literature on health care costs),[15] we find that one additional unit of BMI results in an increase of 7.458€ in annual medical costs per patient, a significantly higher impact than that estimated in Table 5. Notice that with our data the GLM model performs much better than the OLS log costs estimation as long as the RMSE and MAPE (auxiliary R^2) measures decrease (increase) substantially. Interestingly, the dynamic specification shows a lower marginal impact (5.459€) on annual medical costs caused by a one-unit rise in the BMI, but a relatively better performance is achieved here compared to that recorded with the non-dynamic specification.

[Table 6 around here]

5.3 Obesity, overweight and medical costs

In addition to the impact of the BMI, we also investigated the effect of obesity and overweight categories on healthcare costs. Table 7 reports the bootstrapped estimated incremental effect (IE) of obesity and overweight (since they are both dummy variables) on direct medical costs using the same approach as in Table 6, namely, a GLM procedure for the second part based on a Gamma distribution and the log link function.[16] Notice, however, that here we excluded the Mundlak correction procedure and the one-year lagged BMI regressor, when the rest of the econometric issues posed by the data set (Section 3.3) were accounted for. Generally our results show a highly significant and positive estimated IE of obesity and overweight on medical costs. Under the first "static" specification we find that a one unit increase in the prevalence of obesity raises direct medical costs by 51.868€ per patient and year. As expected the impact of the overweight status on such costs is notably lower (16.559€). Notwithstanding this, according to the dynamic specification, the IE of both obesity and overweight on costs is much stronger (77.737€ and 41.040€, respectively). Again, the accuracy and goodness of fit achieved with this estimation is greater.

[Table 7 around here]

[15] The Pregibon link test gives an estimated value of $-0.591*10^{-5}$ (p-value=0.000) which is practically 0, suggesting the logarithm as the link function. The Park (1966) test gives a coefficient $\upsilon = 1.79$ (p-value=0.000) which is consistent with a gamma-class distribution.

[16] Notice that the equation used to compute the IEs or discrete changes differs slightly from equation (10).

5.4 Robustness checks

To assess how sensitive the above estimations are with respect to the impact of the BMI on medical costs, several robustness checks have been performed (see Table 8). Notice that the reference estimation is the last specification from Table 6 based on a Gamma GLM with log link and using a dynamic approach (i.e., an ME of 5.459€). We begin the sensitivity analysis by dividing the sample by sex, given the evidence of a marked differentiated pattern in the utilization of healthcare resources by gender in most western countries. This set of new estimates, however, includes the same controls as those accounted for in the previous tables. Interestingly, the first two rows of Table 8 show a marked differential impact of gender on healthcare costs. While we find a stronger and statistically significant ME of the BMI on direct medical costs per patient for males (11.021€), this effect is much weaker for females (2.859€). Although not shown here, if we restrict the sample to patients aged 20-64 our estimations report a relatively similar effect of the BMI on medical costs compared to the reference case. So, although elderly patients consume the highest share of medical resources, as highlighted in Table 3, the BMI tends to peak at a much younger age.

Finally, the last row of Table 8 verifies how sensitive the impact of BMI is when key covariates affecting medical costs (i.e., patients' medical conditions) are dropped from the model. Under these conditions, our dynamic version predicts a significant and slightly higher ME of the BMI on costs (7.995€ vs. 5.459€) since part of the variation in medical costs attributable to such health conditions are now captured by the individuals' body mass.

[Table 8 around here]

5.5 Instrumenting BMI by means of biological information

In a final step we followed Cawley and Meyerhoefer's (2012) proposal, and one that is widely used in the literature, and instrumented the individuals' BMI with the BMI of a biological relative (i.e., children's information).[17] Although i) our weight and height data are clinically measured and, as such, the BMI does not suffer any misreporting, ii) we control for specific chronic diseases and iii) we use longitudinal information to control for unobserved heterogeneity. Even so we opted to use their approach in our estimations given that some kind

[17] Given that we linked our dataset to census information we were able to obtain household and parental identifiers.

of heterogeneity might still be present. Moreover, because various primary care programs (principally, the Healthy Child Program) specifically targeted children, we have considerably more information on children's BMI to construct the instrument than was the case in Cawley and Meyerhoefer's (2012) study. We considered non-linearities in the instrument (quadratic and cubic terms).

Table 9 reports the new IV results.[18] This table contains two sections: section A presents the ME of the BMI on direct medical costs, and section B does the same for the IE of obesity and overweight. For comparative purposes the first row of each section shows the ME (IE) of BMI (obesity, overweight) using the same sample size as that used under the IV estimation, which of course is greatly reduced. The second rows report our IV estimations.

In line with Cawley and Meyerhoefer (2012), our findings indicate that the IV estimates of the impact of BMI or obesity and overweight on direct costs are notably higher than those without instrumenting. Thus, the ME of the BMI is 39% greater than that without instrumenting (10.003€ vs. 7.201€). However, more marked increases were observed for the non-linear estimations for the ME of the BMI. Thus, being obese (overweight) increases direct medical costs by 96.155€ (78.814€) per patient and year, which is 84% (291%) higher than in the non-instrumented case.[19]

[Table 9 around here]

6. Conclusion

This study has examined the impact of BMI, obesity and overweight on direct medical costs. We have applied panel data econometrics and used a two-part model with a longitudinal dataset of medical records of patients followed up over seven consecutive years (2004-2010). This is the first application in the literature of this methodology based on longitudinal information and BMI measurements as opposed to self-reported data.

One of the consequences of obesity is the higher health care costs borne by the entire society (i.e., negative externality) through higher insurance premiums or taxes to cover the extra funding. Hence, understanding the link between body mass or obesity and medical costs should be then crucial to achieve a more sustainable growth of health expending; especially at

[18] The sample is considerably reduced as we only take into account individuals with children.
[19] Note that these results provide an estimate of the Local Average Treatment Effect (LATE) of one additional BMI unit on medical costs for a sample of individuals with children.

a time of increased pressure to cut successively public budgets. But it should also serve as a way to stimulate the allocation of more resources into prevention actions to tackle the development of the epidemic.

Our estimations indicate that a one unit increase in individual BMI increases direct medical costs by between 5 and 10€ per patient and year. Similarly, obesity (overweight) increases direct medical costs by between 50 and 96€ (17 and 79€) per patient and year. This means that if half the analysed population (i.e., individuals using the healthcare centres at least once during the study period) experienced a one unit increase in their BMI, annual direct costs would increase by between 250,000 and 500,000€. Similarly, if half the Spanish population experienced the same BMI increase, then the annual rise in direct healthcare costs would represent around 0.025% of GDP (256 million €). These magnitudes are similar to the recent budget cuts suffered by the Spanish healthcare system.

As expected, the impact of bodyweight on healthcare costs for our sample of primary and secondary health centres is lower than that reported by Cawley and Meyerhoefer (2012) as the Spanish healthcare system provides universal coverage and its services are free at the point of delivery. Furthermore, during the period of analysis, strict cost-containment policies were in operation.

References

Alberti, KG, Eckel RH, Grundy SM, Zimmet PZ, Cleeman JI, Donato KA, Fruchart JC, James WP, Loria CM, Smith SC Jr., 2009. Harmonizing the metabolic syndrome: a joint interim statement of the International Diabetes Federation Task Force on Epidemiology and Prevention; National Heart, Lung, and Blood Institute; American Heart Association; World Heart Federation; International Atherosclerosis Society; and International Association for the Study of Obesity. Circulation; 120: 1640-1645.

Albouy, V., Davezies, L., Debrand, T., 2010. Health expenditure models: a comparison using panel data. Economic Modelling, 27, 791-803.

Andreyeva T., Sturm R., Ringel JS., 2004. Moderate and severe obesity have large differences in health care costs. Obes Res 12: 1936-1943.

Aranceta Bartrina J., Serra Majem Ll., Foz Sala, Moreno Esteban B., 2005. Grupo Colaborativo SEEDO. Prevalencia de la obesidad en España. Med. Clin. (Barc.) 125: 460-466.

Arterburn D.E., Maciejewski M.L., Tsevat J., 2005. Impact of morbid obesity on medical expenditures in adults. International J Obes 29: 334-339.

Barrett AM, Colosia AD, Boye KS, Oyelowo O. Burden of obesity: 10-year review of the literature on costs in nine countries. ISPOR 13th Annual International Meeting, May 2008, Toronto, Ontario, Canada.

Baser O., 2007. Modeling transformed health care cost with unknown heteroskedasticity. App. Econ. Res. Bull. 1: 1-6.

Basu A., Rathouz P., 2005. Estimating marginal and incremental effects on health outcomes using flexible link and variance function models. Biostatistics 6(1): 93-109.

Berghöfer A., Pischon T., Reinhold T., Apovian C.M., Sharma A.M., Willich S.N., 2008. Obesity prevalence from a European perspective: a systematic review. BMC Public Health. 2008; 8: 200.

Borg S., Persson U., Odegaard K., Berglund G., Nilsson J.A., Nilsson P.M., 2005. Obesity, survival, and hospital costs-findings from a screening project in Sweden. Value Health: 562-71.

Buntin M.B., Zaslavsky A.M., 2004. Too much ado about two-part models and transformation? Comparing methods of modelling Medicare expenditures. Journ. Health Econ., 23: 525-542.

Cawley J., Meyerhoefer C., 2012. The medical care costs of obesity: an instrumental variables approach. Journ. Health Econ., 31: 219-230.

Chamberlain G., 1980. Analysis of covariance with qualitative data. Rev. Econ. Stu. 47: 225-238.

Colditz G.A., 1999. Economic costs of obesity and inactivity. Med. Sci. Sports Exerc. 31 (11 Suppl): S663-S667.

Duan N., 1983. Smearing estimate: a nonparametric retransformation method. J. Amer. Statist. Assoc. 78: 605-610.

Duan N., Manning, W.G., Morris C.N., Newhouse, J.P., 1983. A comparison of alternative models for the demand for medical care. J. Bus. Econ. Stat. 1(2): 115-126.

Duan N., Manning, W.G., Morris C.N., Newhouse, J.P., 1984. Choosing between the sample-selection model and the multi-part model. J. Bus. Econ. Stat. 2(3): 283-289.

Finkelstein E.A., Fiebelkorn I.C., Wang G., 2004. State level estimates of annual medical expenditures attributable to obesity. Obes. Res. 12: 18-24.

Finkelstein E.A., Fiebelkorn I.C., Wang G., 2005. The costs of obesity among full-time employees. Am. J. Health Promot. 20: 45-51.

Gariepy G., Nitka D., Schmitz N., 2010. The association between obesity and anxiety disorders in the population: a systematic review and meta-analysis. Int. J. Obes. (Lond). 34: 407-419.

Grossman M., 1972. On the concept of health capital and the demand for health. Journ Pol. Eco. 80: 223-255.

Hertz T., 2010. Heteroskedasticity-robust elasticities in logarithmic and two-part models. Applied Economics Letters 17: 225-228.

Hill S., Miller G., 2010. Health expenditure estimation and function form: applications of the Generalised Gamma and Extended Estimating Equations models. Health Econ., 19: 608-627.

Jones A. M., Rice N., Bago d'Uva M.T. Balia S., 2007. Applied Health Economics, (Routledge Advanced Texts in Economics and Finance), Routledge, UK.

Jones A. M., 2010. Models for Health Care. HEDG Working Paper 10/01.

López Suárez A., Elvira González J., Beltrán Robles M., Alwakil M., Saucedo J.M., Bascuñana Quirell A., Barón Ramos M.A., Fernández Palacín F., 2008. Prevalence of obesity, diabetes, hypertension, hypercholesterolemia and metabolic syndrome in over 50-year-olds in Sanlúcar de Barrameda, Spain. Rev. Esp. Cardiol. 61: 1150-1158.

Manning, WG., Morris, CN, Newhouse, JP., 1981. A two-part model of the demand for medical care: preliminary results from the Health Insurance Study. In: van der Gaag, J., Perlman, M. (Eds.), Health, Economics, and Health Economics. North Holland, Amsterdam, pp. 103-123.

Manning W.G., 1998. The logged dependent variable, heteroscedasticity and the retransformation problem. Journ. Health Econ. 17: 283-295.

Manning W.G., Mullahy J., 2001. Estimating log models: to transform or not to transform? Journ. Health Econ. 20: 461-494.

Manning W.G., Basu A., Mullahy J., 2005. Generalised modelling approaches to risk adjustment of skewed outcomes data. Journ. Health Econ. 24: 465-488.

Mihaylova, M., Briggs, A., O'Hagan, A., Thompson, SG., 2011. Review of statistical methods for analysing healthcare resources and costs. Health Econ., 20: 897-916. doi:10.1002/hec.1653

Mullahy, J., 1998. Much ado about two: reconsidering retransformation and the two-part model in health econometrics. Journ. Health Econ. 17: 247-281.

Müller-Riemenschneider F., Reinhold T., Berghöfer A., Willich SN., 2008, Health-economic burden of obesity in Europe. Eur J. Epidemiol. 23: 499-509.

Mundlak Y., 1978, On the pooling of time series and cross-section data. Econometrica. 46: 69-85.

Nakamura K., Okamura T., Kanda H., Hayakawa T., Okayama A., Ueshima H., 2007, Health Promotion Research Committee of the Shiga National Health Insurance Organizations. Medical costs of obese Japanese: a 10-year follow-up study of National Health Insurance in Shiga, Japan. Eur. J. Public Health. 17(5): 424-429.

OECD, 2012. Obesity updates 2012.

Quesenberry C.P Jr., Caan B., Jacobson A., 1998, Obesity, health services use and health care costs among members of a health maintenance organization. Arch. Intern. Med. 158: 466-472.

Raebel M.A., Malone D.C., Conner D.A., Xu S., Porter J.A., Lanty F.A., 2004, Health services use and health care costs of obese and non-obese individuals. Arch. Intern. Med. 164: 2135-2140.

Sander B., Bergemann R., 2003, Economic burden of obesity and its complications in Germany. Eur. J. Health Econ. 4: 248-253.

Sturm R., 2002, The effects of obesity smoking, and drinking on medical problems and costs. Health Aff (Millwood) 21: 245-253.

Thompson D., Brown J.B., Nichols G.A., Elmer P.J., Oster, G., 2001, Body mass index and future healthcare costs: a retrospective cohort study. Obes. Res. 9: 210-218.

van Baal P.H.M., Polder J.J., de Wit G.A., Hoogenveen R.T., Feenstra T.L. et al., 2008, Lifetime medical costs of obesity: prevention no cure for increasing health expenditure. PloS Med 5(2), e29, (DOI http://dx.doi.org/10.1371/journal.pmed.0050029).

Vázquez-Sánchez R., López Alemany J.M., 2002, Los costes de la obesidad alcanzan el 7% del gasto sanitario. Rev. Esp. Econ. Salud, Sept-Oct 1(3).

Von Lengerke T., Reitmeier P., John J., 2006, Direct medical costs of (severe) obesity: a bottom-up assessment of over vs. normal weight adults in the KORA-study region (Augsburg, Germany). Gesundheitswesen 68: 110-115.

Wolf A.M., Colditz G.A., 1998, Current estimates of the economics costs of obesity in the United States. Obes. Res. 6: 97-106.

Wolfenstetter, SB., 2012. Future direct and indirect costs of obesity and the influence of gaining weight: Results from the MONICA/KORA cohort studies, 1995-2005. Economics and Human Biology 10: 127-138.

Wooldridge J.M., 2005, Simple solutions to the initial conditions problem in dynamic, non-linear panel data models with unobserved heterogeneity. J. Appl. Econometrics, 20: 39-54.

Table 1. Unit cost estimates per patient in 2004 and 2010

Healthcare resources	Unit costs (€) 2004	Unit costs (€) 2010
Medical visits:		
Visits to Primary Medical Care	16.09	24.37
Visits to Emergency Care	79.49*	123.48
Hospitalization (per day)	217.03*	337.13
Visits to Specialist Care	71.30*	110.76
Complementary tests:		
Laboratory tests	18.33	22.64
Conventional radiology	14.64	18.79
Diagnostic/therapeutic tests	21.37	37.76
Pharmaceutical prescriptions	PVP	PVP

Note: Figures for years 2004-2010 are estimated from linear interpolation based on observed data in 2003 and 2009. Figures for the year 2010 are derived using the same growth rates. (*) These figures were estimated using the growth rate experienced by primary care visits during the period 2003-2009. PVP is retail price.
Source: BSA analytical accounts.

Table 2. Mean Annual Direct Medical Costs per Patient 2004-2010 (Euros year 2010)

	Initial Sample		Final Sample	
	Costs (in Euros)	**Log Costs**	**Costs (in Euros)**	**Log Costs**
Unconditional Mean	544.04	4.01	755.11	5.02
Unconditional Median	139.93	4.95	306.92	5.73
Standard Deviation	1,138.78	2.92	1,309.96	2.55
Skewness	6.70	-0.36	5.91	-0.97
Kurtosis	103.72	1.62	82.97	2.85
N (Number of obs.)	706,473	706,473	452,108	452,108

Table 3. Mean Positive Annual Direct Medical Costs per Patient 2004-2010 (Euros year 2010)

	Final Sample with Positive Costs		
	Both Genders	**Male**	**Female**
Full sample	903.09 (1,382.42)	845.96 (1,378.48)	949.40 (1,383.88)
		By subgroups of the population:	
Ages 16-24	335.29 (425.99)	325.67 (418.85)	344.10 (432.24)
Ages 24-40	390.40 (607.38)	380.78 (664.52)	398.32 (555.83)
Ages 40-54	624.72 (852.38)	574.61 (855.90)	664.21 (847.53)
Ages 54-65	1,049.15 (1,246.88)	974.56 (1,212.95)	1,113.64 (1,271.99)
Ages + 65	1,911.87 (2,097.58)	1,862.60 (2,167.37)	1,947.54 (2,044.84)
Active (labour status)	493.28 (678.66)	467.65 (673.02)	515.50 (682.74)
Charlson index (>0)	1,777.23 (2,057.78)	1,693.65 (1,992.99)	1,863.36 (2,119.18)
Immigrant status	411.74 (698.34)	383.81 (764.77)	435.35 (635.88)
Deceased individuals	3,302.33 (4,727.91)	3,411.68 (5,066.23)	3,173.23 (4,292.89)
N (Number of obs.)	377,964	169,199	208,765

Table 4. Descriptive statistics of control variables. Period 2004-2010

	Final Sample		
	Both Genders	**Male**	**Female**
BMI	26.70 (5.18)	26.75 (4.54)	26.67 (5.67)
Obesity	0.23 (0.42)	0.21 (0.41)	0.25 (0.43)
Overweight	0.36 (0.48)	0.42 (0.49)	0.31 (0.46)
Age	48.24 (19.23)	47.52 (18.84)	48.86 (19.54)
Female	0.54 (0.50)		
Immigrant status	0.05 (0.22)	0.05 (0.23)	0.05 (0.22)
Active (labour status)	0.67 (0.47)	0.70 (0.46)	0.65 (0.48)
Charlson comorb. index	0.07 (0.35)	0.07 (0.37)	0.06 (0.32)
Average number episodes	2.02 (2.05)	1.73 (1.84)	2.28 (2.18)
Deceased individuals	0.03 (0.17)	0.03 (0.18)	0.02 (0.15)
N (Number of obs.)	452,108	209,637	242,471

Note: Figures are mean values between 2004-2010. Standard deviations are reported in parentheses.

Table 5. Bootstrapped Marginal Effects of Measured BMI on Annual Direct Medical Costs (in Euros year 2010): OLS log costs panel data estimation

Two-Part Model	ME of BMI	RMSE	MAPE	Auxiliary R^2
OLS on Log(y) + Heteroskedasticity-adjusted Retransformed Model (N=318,276)	4.712 (1.10)***	416,295	737.90	0.216
OLS on Log(y) + Heteroskedasticity-adjusted Retransformed Model + Lagged Costs + Lagged Costbin (N=258,900)	2.775 (1.18)**	344,489	675.89	0.405

Notes: Auxiliary R^2 denotes the R-squared from a regression of actual costs on the predicted values; RMSE denotes the root mean squared error; MAPE is the mean absolute prediction error. Estimations account for an extensive list of covariates, health district dummies and time dummy variables. In addition, all regressions contain one-year lagged measured BMI, the Mundlak correction and a dichotomous exclusion restriction for the first part. N sample units refers to the second part.

Table 6. Bootstrapped Marginal Effects of Measured BMI on Annual Direct Medical Costs (in Euros year 2010): GLM panel data estimation

Two-Part Model	ME of BMI	RMSE	MAPE	Auxiliary R^2
GLM- Log link + Gamma distr. (N=318,276)	7.458 (1.47)***	296,512	525.14	0.516
GLM- Log link + Gamma dist. + Lagged Costs & Lagged Costbin (N=258,900)	5.459 (1.50)***	258,719	508.58	0.556

Notes: Auxiliary R^2 denotes the R-squared from a regression of actual costs on the predicted values; RMSE denotes the root mean squared error; MAPE is the mean absolute prediction error. Estimations account for an extensive list of covariates, health district dummies and time dummy variables. In addition, all regressions contain one-year lagged measured BMI, the Mundlak correction and a dichotomous exclusion restriction for the first part. N sample units refers to the second part.

Table 7. Bootstrapped Incremental Effects of Obesity and Overweight on Annual Direct Medical Costs (in Euros year 2010): GLM panel data estimation

Two-Part Model	IE Obesity	IE Overweight	RMSE	MAPE	Auxiliary R^2
GLM- Log link + Gamma dist. (N=373,058)	51.868 (3.06)***	16.559 (2.33)***	318,853	442.60	0.514
GLM- Log link + Gamma dist. + Lagged Costs & Lagged Costbin (N=258,900)	77.737 (3.88)***	41.040 (5.42)***	258,813	508.76	0.556

Notes: Auxiliary R^2 denotes the R-squared from a regression of actual costs on the predicted values; RMSE denotes the root mean squared error; MAPE is the mean absolute prediction error. Estimations account for an extensive list of covariates, health district dummies and time dummy variables. Regression contains a dichotomous exclusion restriction for the first part. N sample units refers to the second part.

Table 8. Robustness Analysis: GLM panel data estimation with Log link + Gamma distr. + Lagged Costs & Costbin

Two-Part Model	ME of BMI	RMSE	MAPE	Auxiliary R^2
Male sample (N= 111,862)	11.021 (2.75)***	168,867	505.17	0.544
Female sample (N=147,038)	2.859 (1.14)**	195,295	509.35	0.569
Without health controls (N=259,775)	7.995 (1.36)***	257,807	503.56	0.625

Notes: Auxiliary R^2 denotes the R-squared from a regression of actual costs on the predicted values; RMSE denotes the root mean squared error; MAPE is the mean absolute prediction error. Estimations account for an extensive list of covariates, health district dummies and time dummy variables. In addition, all regressions contain one-year lagged measured BMI, the Mundlak correction and a dichotomous exclusion restriction in the first part. N sample units refers to the second part.

Table 9. IV estimates: GLM panel data estimation with Log link + Gamma distr. + Lagged Costs & Costbin

Section (A)

Two-Part Model	ME of BMI	RMSE	MAPE	Auxiliary R^2
Non IV estimation (N=140,137)	7.201 (1.44)***	164,780	441.16	0.510
IV estimation (N= 140,137)	10.003 (1.60)***	164,899	441.49	0.511

Section (B)

Two-Part Model	IE Obesity	IE Overweight	RMSE	MAPE	Auxiliary R^2
Non IV estimation (N=139,703)	52.170 (4.18)***	20.152 (2.89)***	164,848	441.34	0.510
IV estimation (N=139,703)	96.155 (6.53)***	78.814 (5.08)***	164,321	439.85	0.508

Notes: Auxiliary R^2 denotes the R-squared from a regression of actual costs on the predicted values; RMSE denotes the root mean squared error; MAPE is the mean absolute prediction error. Estimations account for an extensive list of covariates, health district dummies and time dummy variables. Regressions contain one-year lagged measured BMI, the Mundlak correction and a dichotomous exclusion restriction in the first part. N sample units refers to the second part.

ÚLTIMOS DOCUMENTOS DE TRABAJO

2012-08: "The Influence of BMI, Obesity and Overweight on Medical Costs: A Panel Data Approach **Toni Mora, Joan Gil y Antoni Sicras-Mainar**.

2012-07: "Strategic behavior in regressions: an experimental", **Javier Perote, Juan Perote-Peña y Marc Vorsatz**.

2012-06: "Access pricing, infrastructure investment and intermodal competition", **Ginés de Rus y M. Pilar Socorro**.

2012-05: "Trade-offs between environmental regulation and market competition: airlines, emission trading systems and entry deterrence", **Cristina Barbot, Ofelia Betancor, M. Pilar Socorro y M. Fernanda Viecens.**

2012-04: "Labor Income and the Design of Default Portfolios in Mandatory Pension Systems: An Application to Chile", **A. Sánchez Martín, S. Jiménez Martín, D. Robalino y F. Todeschini**.

2012-03: "Spain 2011 Pension Reform", **J. Ignacio Conde-Ruiz y Clara I. Gonzalez**.

2012-02: "Study Time and Scholarly Achievement in PISA", **Zöe Kuehn y Pedro Landeras**.

2012-01: "Reforming an Insider-Outsider Labor Market: The Spanish Experience", **Samuel Bentolila, Juan J. Dolado y Juan F. Jimeno**.

2011-13: "Infrastructure investment and incentives with supranational funding", **Ginés de Rus y M. Pilar Socorro**.

2011-12: "The BCA of HSR. Should the Government Invest in High Speed Rail Infrastructure?", **Ginés de Rus**.

2011-11: "La rentabilidad privada y fiscal de la educación en España y sus regiones", **Angel de la Fuente y Juan Francisco Jimeno**.

2011-10: "Tradable Immigration Quotas", **Jesús Fernández-Huertas Moraga y Hillel Rapoport**.

2011-09: "The Effects of Employment Uncertainty and Wealth Shocks on the Labor Supply and Claiming Behavior of Older American Workers", **Hugo Benítez-Silva, J. Ignacio García-Pérez y Sergi Jiménez-Martín**.

2011-08: "The Effect of Public Sector Employment on Women's Labour Martket Outcomes", **Brindusa Anghel, Sara de la Rica y Juan J. Dolado**.

2011-07: "The peer group effect and the optimality properties of head and income taxes", **Francisco Martínez-Mora**.

2011-06: "Public Preferences for Climate Change Policies: Evidence from Spain", **Michael Hanemann, Xavier Labandeira y María L. Loureiro.**

2011-05: "A Matter of Weight? Hours of Work of Married Men and Women and Their Relative Physical Attractiveness", **Sonia Oreffice y Climent Quintana-Domeque**.

2011-04: "Multilateral Resistance to Migration", **Simone Bertoli y Jesús Fernández-Huertas Moraga**.

2011-03: "On the Utility Representation of Asymmetric Single-Peaked Preferences", **Francisco Martínez Mora y M. Socorro Puy**.

2011-02: "Strategic Behaviour of Exporting and Importing Countries of a Non-Renewable Natural Resource: Taxation and Capturing Rents", **Emilio Cerdá y Xiral López-Otero**.

2011-01: "Politicians' Luck of the Draw: Evidence from the Spanish Christmas Lottery", **Manuel F. Bagues y Berta Esteve-Volart**.

2010-31: "The Effect of Family Background on Student Effort", **Pedro Landeras**.

2010-29: "Random–Walk–Based Segregation Measures", **Coralio Ballester y Marc Vorsatz**.

2010-28: "Incentives, resources and the organization of the school system", **Facundo Albornoz, Samuel Berlinski y Antonio Cabrales**.

2010-27: "Retirement incentives, individual heterogeneity and labour transitions of employed and unemployed workers", **J. Ignacio García Pérez, Sergi Jimenez-Martín y Alfonso R. Sánchez-Martín**.

2010-26: "Social Security and the job search behavior of workers approaching retirement", **J. Ignacio García Pérez y Alfonso R. Sánchez Martín**.

2010-25: "A double sample selection model for unmet needs, formal care and informal caregiving hours of dependent people in Spain", **Sergi Jiménez-Martín y Cristina Vilaplana Prieto**.

2010-24: "Health, disability and pathways into retirement in Spain", **Pilar García-Gómez, Sergi Jiménez-Martín y Judit Vall Castelló**.

2010-23: Do we agree? Measuring the cohesiveness of preferences", **Jorge Alcalde-Unzu y Marc Vorsatz**.

2010-22: "The Weight of the Crisis: Evidence From Newborns in Argentina", **Carlos Bozzoli y Climent Quintana-Domeneque**.

2010-21: "Exclusive Content and the Next Generation Networks"·, **Juan José Ganuza and María Fernanda Viecens**.

2010-20: "The Determinants of Success in Primary Education in Spain", **Brindusa Anghel y Antonio Cabrales**.

2010-19: "Explaining the fall of the skill wage premium in Spain", **Florentino Felgueroso, Manuel Hidalgo y Sergi Jiménez-Martín**.

2010-18: "Some Students are Bigger than Others, Some Students' Peers are Bigger than Other Students' Peers", **Toni Mora y Joan Gil**.