Journal of Health Economics 30 (2011) 1280–1292

Contents lists available at ScienceDirect

Journal of Health Economics

journal homepage: www.elsevier.com/locate/econbase

Inverse probability weighted estimation of social tariffs: An illustration using the SF-6D value sets☆

Ildefonso Méndez*, Jose M. Abellán Perpiñán, Fernando I. Sánchez Martínez, Jorge E. Martínez Pérez

University of Murcia, Spain

ARTICLE INFO

Article history:
Received 2 April 2010
Received in revised form 18 July 2011
Accepted 26 July 2011
Available online 26 August 2011

JEL classification:
I10

Keywords:
Inverse probability weighting
Propensity score
Preference-based health measure
SF-6D

ABSTRACT

This paper presents a novel approach to model health state valuations using inverse probability weighting techniques. Our approach makes no assumption on the distribution of health state values, accommodates covariates in a flexible way, eschews parametric assumptions on the relationship between the outcome and the covariates, allows for an undetermined amount of heterogeneity in the estimates and it formally tests and corrects for sample selection biases. The proposed model is semi-parametrically estimated and it is illustrated with health state valuation data collected for Spain using the SF-6D descriptive system. Estimation results indicate that the standard regression model underestimates the utility loss that the Spanish general population assigns to departures from full health, particularly so for severe departures.

1. Introduction

Preference-based measures of health status are increasingly being used to evaluate the outcomes of health care interventions and to inform resource allocation decisions. A number of health state descriptive systems have been designed for the characterization of health states and estimation methods have been applied for calculating a preference-based single index value for every state defined within some of these systems. The HUI3 (Feeny et al., 2002), the EQ-5D (EuroQol Group, 1990) or the SF-6D (Brazier et al., 1998) are examples of these systems, and all of them describe health states by defining a number of dimensions or attributes (e.g. pain, physical functioning, ability for self-care, etc.) each admitting different levels of severity or impairment. Since most descriptive systems define many more health states than it is feasible to elicit direct valuations for in an empirical study, choices have to be made about how best to estimate values for all states from direct observations on a subset of those states.

The standard approach to model health state values uses a set of dummy indicator variables describing health states in terms of their level of severity in different dimensions of health to explain the individual valuations obtained. Under the assumption of normally distributed errors, a regression of health state values on the set of dummy variables identifies the valuation effect of departures from full health. The estimates are then used to predict the value associated to the health states not directly valued.

The main advantadge of the standard approach is its simplicity. However, it has some limitations that are likely to undermine its benefits. First, the normality assumption is not likely to hold in practice given the skewed, truncated, non-continuous and hierarchical nature of health state valuation data (Brazier et al., 2002) and, thus, the estimates are likely to be biased. Second, the standard approach does not provide any guidance on how to improve the specification of the regression model by accounting for interactions between the severity indicators or for personal characteristics. On the one hand, the linear regression model is not the appropriate framework for meaningfully incorporating the large number of interactions between the severity indicators that can be defined in any health state descriptive system (Brazier et al., 2002). On the other hand, the traditional way of accommodating personal characteristics to the standard model by introducing them additively contravenes both the goal of estimating one preference-based tariff for the whole community (Dolan, 1997) and the theoretical requirement of the intercept being equal to unity (Brazier et al., 2002, 2004).

As a result, most articles that use the standard approach do not control for personal characteristics nor for severity interactions,

* Corresponding author at: Departamento de Economía Aplicada, Facultad de Economía y Empresa, 30100, Espinardo, Murcia, Spain. Tel.: +34 868883732; fax: +34 868883745.

0167-6296/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.jhealeco.2011.07.013

making misspecification of the regression model more likely. This, in turn, leads to biased estimates. Moreover, these articles implicitly assume that the valuation impact of a departure from full health is the same for respondents with different personal characteristics and is independent of the severity of departures in other dimensions of health. This restrictive homogeneity assumption has been rejected by the evidence in Dolan and Roberts (2002) and Kharroubi et al. (2007a), among others. They find that some respondents' characteristics impact on the value they give to health states and that this effect varies with the severity of the health state at examination.

The debate on the valuation effect of personal characteristics is related to that on whose values should count when evaluating health state intervention outcomes and informing resource allocation decisions. The common recommendation of using the preferences of the whole population (Gold et al., 1996; NICE, 2004) calls for obtaining population valid estimates, that is, to adjust for the distribution of the covariates in the population. The standard approach tries to fulfil this requirement by defining samples that are representative for the population of interest with regard to the sex and age interval distributions. The representativeness of the estimates is then analyzed by comparing the descriptive statistics of a large set of covariates in the sample to those in the population of interest. Lastly, corrective weights intended to adjust for the age and sex interval population distributions are introduced if relevant discrepancies are observed between the sample and population descriptive statistics.

This way of proceeding does not provide the user with the appropriate tools for testing and correcting for sample selection biases. Non-response issues and the drop of respondents providing inconsistent responses results in sometimes relevant discrepancies between the "representative" sample design and the estimation sample. Moreover, there are many personal characteristics that affect health state values whose sample distribution is not necessarily that in the population even if age intervals and sex are equally distributed in both instances. For example, Dolan and Roberts (2002) find that marital status and the respondents' ability to cope with usual activities (i.e. one of the dimensions of their own health state) affect health state valuations. Additionally, Kharroubi et al. (2007a) find that the individual's employment status, educational level and own physical and social functioning have a significant effect on health state values.

The comparison of the univariate descriptive statistics of the covariates in the estimation sample to those in the population of interest is not a formal test of sample selection biases and, thus, it might lead to wrong conclusions. In particular, it raises doubts as to in how many covariates we have to find a significant difference of a given magnitude between the sample and population means for the estimates not to be valid at the population level. Moreover, finding no significant difference between the sample and the population means of a continuous variable is not necessarily very informative about the presence of relevant discrepancies in other moments of the distributions. Furthermore, multivariate distributions can differ significantly even if univariate descriptive statistics in the estimation sample are close to those in the population of interest.

Finally, the corrective weights used to ensure representativeness of the regression estimates suffer from the curse of dimensionality problem, that is, its feasibility lowers as the number of personal characteristics where relevant discrepancies are observed between the sample and population descriptive statistics increases. This problem is circumvented in practice by restricting the set of personal characteristics used to construct the weights to the respondent's sex and age group. However, this way of proceeding does not remove sample selection biases in other personal dimensions and, thus, it is not likely to produce representative estimates.

This paper presents a new approach to estimating preference-based measures of health status based on inverse probability weighting (IPW) techniques. The IPW approach makes no assumption on the distribution of health state valuations, allows for an undetermined amount of heterogeneity in the estimates, accommodates covariates in a flexible way, formally tests and corrects for sample selection biases and uses the distribution of personal characteristics in the population of interest to guarantee the representativeness of the estimates. The estimators that we propose are semi-parametrically estimated and their large sample properties are derived. Additionally, and as opposed to the nonparametric Bayesian approach in Kharroubi et al. (2007a,b), our approach provides the user with a simple table of estimated coefficients that defines the estimated preference function, which results in relevant efficiency and transparency gains. We illustrate our approach with the SF-6D descriptive system. Notwithstanding, the IPW approach could be equally applied to other systems (e.g. the EQ-5D) provided that some requirements regarding the selection of states which are directly valued in the sample are met, as it will be further discussed.

The paper has four more sections. Section 2 presents the proposed approach and compares its properties to those of the standard parametric one. Section 3 describes the SF-6D descriptive system and the data used to illustrate the proposed estimators. Section 4 presents and discusses the estimation results and, finally, Section 5 concludes.

2. Modelling

The standard model of health state valuations can be written as

$$Y_{ij} = \alpha + \beta' Z_j + \varepsilon_{ij} \quad (1)$$

where Y_{ij} is the utility that individual i assigns to health state j, Z_j is a vector of dummy indicator variables Z_{kw} that equal one if health state j reaches level of severity w in dimension k and zero otherwise, for $w = 2, 3, \ldots, W_k$ and $k = 1, 2, \ldots, K$, α is the intercept and ε_{ij} is a zero mean error term.[1] Model (1) is the "main effects" model, as opposed to other specifications that also control for level effects, interactions between the elements of Z_j or personal characteristics. The model is estimated using the Ordinary Least Squares (*OLS*) or the Random Effects (*RE*) estimators, that is, assuming that ε is normally distributed. Most researchers use the *RE* estimator since it takes account that the same individual values several health states, increasing the efficiency of the estimates relative to the *OLS* estimator.

The main advantage of the standard approach is that it can be easily implemented in any statistical package. However, it also has some relevant limitations that are likely to undermine its benefits. First, the normality assumption is not likely to hold with health state valuation data (Brazier et al., 2002) and, thus, the regression estimates are likely to be biased.[2] Second, the linear regression model is severely limited in the way it controls for interaction effects and personal characteristics. This model does not provide

[1] For the SF-6D descriptive system $K = 6$ and W_k ranges from 4 to 6. Equivalently, $K = 5$ and $W_k = W = 3$ for the EQ-5D.

[2] Dolan et al. (1996) find evidence that the distribution of health state values obtained using the time trade-off method was non-normal for each health state. Johnson et al. (1998) find departures from normality when estimating US-based population weights using the EQ-5D questionnaire. Diagnostic tests in Brazier et al. (2002) reveal non-normal residuals in the estimation of a preference-based measure of health for the UK general population using the SF-36. Many other studies, like Tsuchiya et al. (2002) and Lamers et al. (2006), simply provide no formal test of the underlying distributional assumption.

the appropriate framework for meaningfully incorporating the large number of interactions that can be defined in any health state descriptive system (Brazier et al., 2002). Additionally, the introduction of personal characteristics additively contravenes both the goal of estimating one preference-based tariff for the whole community (Dolan, 1997) and the requirement of the intercept being equal to unity (Brazier et al., 2002).[3]

As a result of the second limitation, the great majority of articles using the standard approach do not control for interaction terms nor for personal characteristics. These articles implicitly assume that the valuation impact of a departure from full health is the same no matter the severity of the deviation in the remaining dimensions of the health state under evaluation nor the respondents' characteristics. Such an homogeneity assumption has been rejected by the evidence in Dolan and Roberts (2002) and Kharroubi et al. (2007a), among others, who find that some respondents' characteristics impact on the value they give to health states and that this effect varies with the severity of the health state at examination. Moreover, by reducing the set of feasible specifications the second limitation increases the risk of misspecification of the regression model and, thus, the probability of obtaining biased estimates since regression models rely heavily on extrapolation when differences in the covariate distributions for compared respondents are large.

The identification strategy is presented for β_{kw}, the coefficient associated with Z_{kw} that measures the average health state valuation impact of moving from level of severity 1 to level of severity w in dimension k. Let X_i be the vector of characteristics of individual i that potentially affect his valuations. The sample is restricted to respondents valuing levels of severity 1 and w in dimension k and the individual and health state subscripts i and j are dropped out to simplify the notation. The coefficient of interest for the sample with personal characteristics x that value health states with level of severity w' in dimension k' for $k' \neq k$ and $w' = 2, 3, \ldots, W_{k'}$ is[4]

$$\beta_{kw}(x, z_{k'w'}) = E[Y|Z_{kw} = 1, X = x, Z_{k'w'} = z_{k'w'}] - E[Y|Z_{kw} = 0, X = x, Z_{k'w'} = z_{k'w'}]$$

where $z_{k'w'} = \{0, 1\}$. Equivalently, the valuation effect for an individual randomly drawn from the estimation sample is

$$\beta_{kw} = E[\beta_{kw}(h)] = E[Y|Z_{kw} = 1] - E[Y|Z_{kw} = 0] \quad (2)$$

where H comprises X and the set of dummy variables $Z_{k'w'}$ for $k' \neq k$ and $w' = 2, 3, \ldots, W_{k'}$ and expectations are defined over the distribution of H in the estimation sample. This way of writing β_{kw} makes it clear that we are not imposing the homogeneous valuation impact assumption inherent to most applications of the standard approach. In fact, we allow for the utility loss of moving from level of severity 1 to level w in dimension k to vary with any of the elements in H.

The estimator that we develop for β_{kw} can be better interpreted in the context of the treatment effects literature. This literature provides answers to questions concerning the efficacy of a particular programme or policy initiative. In this setting β_{kw} is the average valuation effect of a binary treatment that consists in valuing health states where dimension k reaches level of severity w instead of level 1. The causal interpretation of β_{kw} follows from the assumption that unobserved individual characteristics do not affect health state valuations or their overall average impact is zero.[5]

Among the broad list of available treatment effect estimators, we opt for the so-called Inverse Probability Weighting estimators for three reasons.[6] First, they are easy to implement and provide consistent and in some cases asymptotically efficient estimates of the parameter of interest under fairly standard regularity conditions. Second, they exhibit the best overall finite sample performance among the broad class of treatment effect estimators analyzed in Busso et al. (2009). This is particularly relevant in the current context since estimation samples are of modest size in most empirical applications. Finally, weighting estimators can be used to assess the effect of changes in the distribution of X on the outcome of interest (DiNardo et al., 1996) and, thus, they allow estimation of preference functions for the population of interest from non-representative samples.

Some additional notation is needed at this point. Let $p_{kw}(h) = P(Z_{kw} = 1|H = h)$ be the conditional probability of receiving treatment given H. This variable is the *propensity score* in the treatment effects literature. The research value of the propensity score rests on its power to solve the dimensionality problem. Adjusting for between-groups differences on a high dimensional vector of covariates can be either difficult or impossible, as is the case when using the standard corrective weights. Rosenbaum and Rubin (1983) show that the propensity score captures all of the variance on the covariates relevant for adjusting between-group comparisons, that is, treated and control units with the same value of the propensity score have the same distribution of the elements in H.

Additionally, the following overlap assumption on the joint distribution of treatments and explanatory variables is necessary for the estimation problem to be well defined: $0 < P(Z_{kw} = 1|H) < 1$. This common support condition requires that for a given value of H there is some fraction of the estimation sample in the treatment and control groups to be compared. That is, a necessary condition for the effect of Z_{kw} to be identified is that no other element of H predicts treatment status perfectly. An implication of this condition is that some respondents valuing levels of severity 1 and w in dimension k will not contribute to the estimation of β_{kw}. In particular, units with propensity scores close to zero or one will be particularly influential in the estimation of β_{kw}, making the estimation imprecise. That is the case for respondents whose distributions of the elements of H substantially differ from those for respondents in the other treatment group.

The common support condition is not commonly invoked when estimating a linear regression model because the regression model uses its functional form to work off the common support in the distribution of the elements of H when estimating β_{kw}. However, that can be highly misleading given the previously discussed specification problems inherent to the standard approach. The common support condition ensures that identification does not rest on functional form assumptions. This condition has relevant implications on the selection of health states valued in the sample for β_{kw} to be identified. In particular, it states that there is no level of severity w' in dimension k', for $w' = 1, 2, 3, \ldots, W_{k'}$ and $k' \neq k$, valued only by respondents of a given treatment status. Otherwise, the effect of interest cannot be separately identified from that for $Z_{k'w'}$ unless we rely on extrapolation, like the regression model does. As previously discussed, extrapolation results in biased estimates

[3] There are strong theoretical reasons for restricting the intercept to unity since it captures the utility associated with full health, which equals one on the conventional full health-death scale used to estimate QALYs.

[4] Existence of expectations is assumed throughout.

[5] The assumption is known as *selection on observables* (Barnow et al., 1981) or *strong ignorable treatment assignment* (Rosenbaum and Rubin, 1983) and is also implicit in the standard regression model.

[6] Imbens (2004) provides an overview of the estimators used in the treatment effects literature under the selection on observables assumption.

if differences between the covariate distributions of respondents of different treatment status are relevant and the parametric relationship between the outcome and the regressors is not properly specified.[7]

The expectations in (2) can be written as

$$E[Y|Z_{kw}=t]=\int Yf(Y|H)g(H|Z_{kw}=t)dh,\ \text{for}\ t=\{0,1\} \tag{3}$$

According to the latter expression, each expectation in (2) is calculated using the distribution of H in the corresponding treatment group. However, for β_{kw} to be identified we need the same distribution of H in the two expectations. In particular, we use the distribution of H in the sample of respondents valuing levels of severity 1 or w in dimension k that satisfy the common support condition. Formally, let $g(H)$ and $g(H|Z_{kw}=t)$ be the joint density of H in the estimation sample and in the collective of respondents with treatment status t, respectively, and observe that by definition

$$g(H)=\frac{g(H|Z_{kw}=t)P(Z_{kw}=t)}{P(Z_{kw}=t|H)},\quad \text{for}\ t=\{0,1\}$$

That is, the distribution of H in the collective of respondents with treatment status t can be changed for the distribution in the estimation sample $g(H)$ by simply introducing the appropriate weighting function λ_t in (3)

$$E[Y|Z_{kw}=t]=\lambda_t\underbrace{\frac{P(Z_{kw}=t)}{P(Z_{kw}=t|H)}}Yf(Y|H)g(H|Z_{kw}=t)dh$$

$$=\int Yf(Y|H)g(H)dh,\quad \text{for}\ t=\{0,1\}$$

The effect of interest can now be written as

$$\beta_{kw}=E\left[\frac{Z_{kw}Y}{p_{kw}}\right]-E\left[\frac{(1-Z_{kw})Y}{1-p_{kw}}\right] \tag{4}$$

This suggests the following estimator of β_{kw} which we name as the *IPW*1 estimator

$$\widehat{\beta}_{kw,\ IPW1}=n^{-1}\sum\nolimits_{i=1}^{n}\frac{Z_{kwi}Y_{ij}}{\widehat{p}_{kwi}}-n^{-1}\sum\nolimits_{i=1}^{n}\frac{(1-Z_{kwi})Y_{ij}}{1-\widehat{p}_{kwi}} \tag{5}$$

This equation suggests a simple two-step method to estimate β_{kw}. First, estimate the propensity score using a binary discrete choice model like the logit or probit models. Second, plug the fitted values into the sample analog of (5). The *IPW*1 estimator identifies the effect of interest if the estimation sample is representative for the population of interest, that is, if there are no sample selection biases. However, since we cannot be sure *a priori* that the sample distribution of the elements of X is that in the population, we improve on the latter estimator by accounting for the probability that an individual randomly drawn from the population of interest is in the estimation sample.[8] We do so by rewriting expression (3) so that the two expectations are averaged over the distribution of X in the population of interest. Obviously, the feasibility of this approach rests on whether we have an external representative sample that contains information on X. Conditioned on the availability of the external representative sample, the effect of interest is now written as[9]

$$\beta_{kw}=E\left[\frac{D_sZ_{kw}Y}{p_{kw}p_s}\right]-E\left[\frac{D_s(1-Z_{kw})Y}{(1-p_{kw})p_s}\right]$$

where the estimation sample comprises that used for the *IPW*1 estimate of β_{kw}, D_s is a binary indicator variable that equals one if the individual is in the estimation sample and zero if he is in the external representative sample and $p_s(x)=P(D_s=1|X=x)$ is the conditional (on X) probability of being in the estimation sample. The set of indicator variables $Z_{k'w'}$ for $k'\neq k$ and $w'=1,2,3,\ldots,W_{k'}$ is not included in $p_s(x)$ since representativeness is analyzed with regard to the distribution of personal characteristics. The internal sample analog of expression (6) is the *IPW*2 estimator of β_{kw}

$$\widehat{\beta}_{kw,\ IPW2}=\overline{P}_sn^{-1}\sum\nolimits_{i=1}^{n}\frac{Z_{kwi}Y_{ij}}{\widehat{p}_{kwi}\widehat{p}_{si}}-\overline{P}_sn^{-1}\sum\nolimits_{i=1}^{n}\frac{(1-Z_{kwi})Y_{ij}}{(1-\widehat{p}_{kwi})\widehat{p}_{si}} \tag{6}$$

where $\overline{P}_s=P(D_s=1)$ is the proportion of individuals in the estimation sample. As before, the *IPW*2 estimate of β_{kw} is obtained in two steps. First, estimate discrete choice models for the two propensity scores, compute the fitted values for the estimation sample and ensure the common support condition in the two propensities. Second, plug the fitted values into the sample analog of (7). Under this scheme, a two-step weighted average of the outcome variable recovers β_{kw} in the population of interest. In a first step, the distribution of health state values for respondents of a given treatment status is weighted-down (up) for those values of the elements of H that are (under) over-represented among respondents with that treatment status. In a second step, the distribution of health state values for sample respondents is weighted-down (up) for those values of the elements of X that are (over) under-represented among individuals in the external representative sample.

This estimator can also be interpreted in the related framework of imputation for missing data.[10] To appreciate this, we follow Rubin (1974) and define β_{kw} in terms of potential outcomes. Let Y_t be the valuation that individual i would have given had he received treatment status t. We only observe the realized outcome $Y=D_sZ_{kw}Y_1+D_s(1-Z_{kw})Y_0$ but want to know about the effect of the treatment for an individual randomly drawn from the population of interest (β_{kw}). In this setting, β_{kw} is the difference between the population averages of Y_1 and Y_0, which we label μ_1 and μ_0. We only observe Y_1 for treated individuals in the estimation sample and the probability of a "complete case" i is $p=p_s(x)\times p_{kw}(h)$. As Lunceford and Davidian (2004) point out, weighting by the inverse of the product of propensity scores allows observation i to count for himself and $(p^{-1}-1)$ other "missing" subjects with like covariates h in estimating μ_1.

The propensity score $p_s(x)$ adjusts for the distribution of personal characteristics in the population of interest. Notice that this propensity score is a generalization of the traditional corrective sample weights used to ensure the representativeness

[7] Some studies that use the standard approach report evidence of misspecification of the regression model. A non-exhaustive list includes Dolan (1997) and Johnson et al. (1998). Brazier et al. (2002) express their surprise with the result of no specification problems according to the Ramsey RESET test given the skewness of their *RE* estimation residuals. Many other studies simply provide no formal test of misspecification of the regression model.

[8] On the one hand, deviations from the original sample design due to, for example, nonresponse issues or to the exclusion of respondents providing inconsistent responses might result in non-representative samples. On the other hand, it is difficult or even impossible to define representative samples for the population of interest with regard to the whole set of covariates that have been found to be correlated with health state valuations (Dolan and Roberts, 2002; Kharroubi et al., 2007a).

[9] The availability of such a sample is not likely to be a problem for most countries. For example, the Census and the European Community Household Panel provide us with the distribution of many sociodemographic, employment and health related individual and household characteristics of the Spanish population. In particular, we use data from the European Community Household Panel for Spain.

[10] Each one of the two terms in (7) approximates the average outcome for units of a given treatment status using a weighted sample mean estimator of Horvitz–Thompson type. Horvitz and Thompson (1952) introduced this type of estimator to analyze samples drawn without replacement with unequal selection probabilities from finite universes.

of the regression model estimates in the standard approach.[11] In particular, the corrective weights used in the standard framework provide a nonparametric estimate of the propensity score $p_s(x)$ when X only includes dummy indicator variables. The propensity score allows us to overcome the dimensionality problem in the construction of sample weights and, thus, to account for sample selection biases in as many discrete and continuously measured personal characteristics as necessary.

As discussed in Imbens (2004), the estimator in (7) is not necessarily an attractive estimator for β_{kw} since the weights for observations of a given treatment status t do not add up to unity. Indeed, these weights add up to 1 conditioned on treatment status t in expectation terms, but because the variance of the sum is positive the corresponding sample analog is likely to deviate from one. Thus, we normalize the weights to unity and obtain the following estimator:

$$\widehat{\beta}_{kw,IPW2} = \left(\sum_{i=1}^{n} \frac{Z_{kwi}}{\widehat{p}_{kwi}\widehat{p}_{si}}\right)^{-1} \sum_{i=1}^{n} \frac{Z_{kwi}Y_{ij}}{\widehat{p}_{kwi}\widehat{p}_{si}} - \left(\sum_{i=1}^{n} \frac{(1-Z_{kwi})}{(1-\widehat{p}_{kwi})\widehat{p}_{si}}\right)^{-1} \sum_{i=1}^{n} \frac{(1-Z_{kwi})Y_{ij}}{(1-\widehat{p}_{kwi})\widehat{p}_{si}} \quad (7)$$

The consistency and large sample properties of the *IPW*1 and *IPW*2 estimators are derived in Appendix A using the theory of *M*-estimation.[12] The *IPW*2 estimator can also be non-parametrically estimated by simply producing non-parametric estimates of the propensity scores and plugging the fitted values into (8).[13] However, the number of observations required to attain an acceptable precision for this type of non-parametric estimator increases rapidly with the dimension of X. Moreover, a non-parametric estimate conditioned on particular values of X version of these estimators may be difficult to interpret if the dimension of X is larger than two. Furthermore, the net gains of moving from the standard approach to an alternative one decrease as the implementation of the proposed estimator becomes more challenging. Thus, we focus on semi-parametric approximations to *IPW*2 where the propensity scores are parametrically estimated using standard discrete choice models like the logit or probit models.

The *IPW*2 estimator is member of a class of semi-parametric consistent estimators developed in Robins et al. (1994) for general missing data problems. Robins et al. (1994) show that the estimator within the class having the smallest large-sample variance combines regression on the explanatory variables and propensity score weighting. Contrary to the parametric standard model, the regression model in the semi-parametric efficient estimator is incorporated only as a way of gaining efficiency over the *IPW*2 estimator, that will still be consistent. The asymptotically efficient estimator is doubly robust in the sense that it provides consistent estimates of β_{kw} if either the propensity score or the regression model are correctly specified. Anyway, the double robust estimator cannot be implemented in our context because respondents with treatment status t do not value any possible combination of level of severity w or 1 in dimension k with the levels of severity that can be defined in the remaining dimensions of health. That is, we cannot regress health state values on H within each subsample with treatment status t. Indeed, we can just regress health state values on X and some elements of Z for respondents with treatment status t, where that subset of elements of Z is likely to vary with treatment status t and also with the estimation subsamples used to identify each element of β. Anyway, as shown in Busso et al. (2009), the small sample properties of the double robust estimator are close to those for weighting estimator like the *IPW*2 estimator, with the former estimator being slightly more variable and more biased than the one developed in this paper.

Finally, the requirement of the intercept being equal to unity is satisfied by using the transformed outcome variable $Y^* = Y - 1$ instead of the original one.

3. Data and measurement issues

The data comes from a survey performed in the Spanish region of Murcia over a period of two months in 2007. The sample ($n = 1020$) was designed using the age interval and sex distributions of the Spanish general population. The goal of the survey was to obtain direct valuations for a selection of health states described according to the SF-6D classification system. Later on such valuations would be modeled using our approach in order to predict values for the 18,000 health states that the SF-6D system can define. In what follows we briefly describe the SF-6D instrument, the selection of health states valued, and the valuation survey.

3.1. The SF-6D

The SF-6D is a preference-based measure of health that attaches utility scores to a set of health states by using an algorithm based on preferences of the general population. A total of 18,000 health states are defined by means of a classification system composed of six dimensions: physical functioning (*PF*), role limitations (*RL*), social functioning (*SF*), pain (*PAIN*), mental health (*MH*), and vitality (*VIT*). Each dimension has between four to six levels of severity and every SF-6D health state is defined by selecting one level from each dimension.

Brazier et al. (2002) using a variant of the standard gamble (*SG*) method, elicited preferences for a selection of 249 health states from a sample ($n = 611$) of the UK general population. Next, *OLS* and *RE* models were estimated to predict all 18,000 SF-6D health states. The model recommended by the authors for use in cost-utility analysis was an *OLS* model using mean health state values. Brazier et al. (2004) improved the previous model by removing non-significant estimates and aggregating those coefficients which were inconsistent between them. They referred to such a model as the "parsimonious consistent model".

In contrast to previous algorithms, which relied on parametric models, Kharroubi et al. (2007b) – using the same UK data set as Brazier and colleagues – estimated a set of non-parametric (Bayesian) utility scores for the SF-6D. A drawback of this non-parametric model is that it cannot be defined by a simple table of coefficients as in parametric models.

New SF-6D algorithms have been estimated by using the standard regression approach for other countries apart from the UK (Lam et al., 2008; Brazier et al., 2009; Ferrerira et al., 2010; Abellán-Perpiñán et al., 2011). The Spanish SF-6D value set derived in Abellán-Perpiñán et al. (2011) uses the same database as in this paper. The novelty of that value set is that has a minimum (a floor) which is significantly lower than those estimated in previous studies, making the range of SF-6D values more similar to the EQ-5D scores range.

[11] Tsuchiya et al. (2002) and Kharroubi et al. (2007a) introduce corrective weights to reflect the non-representative age and sex distribution of their respondents in the standard and the nonparametric Bayesian approaches, respectively.

[12] A STATA code that implements the *IPW*1 and *IPW*2 estimators is available from the authors upon request.

[13] Craig and Busschbach (2009, 2011) develop non-parametric approaches to health valuation. However, these estimators do not control for personal characteristics.

3.2. Selection of health states and their valuation

Because of the descriptive richness of the SF-6D system, it is impossible to value all possible permutations of each dimension. Hence, a subset of health states has to be identified in order to estimate additive or multiplicative specifications. A total of 78 health states were selected. Forty-nine were chosen by using the same orthogonal design employed by Brazier et al. (2002) in order to identify the minimum sample of health states required to estimate an additive function. Twenty-nine additional states were included in order to account for more complex specifications.

A lottery equivalence method (McCord and de Neufville, 1986) was chosen for the valuation of the health states. Such methods compare two risky prospects or lotteries in such a way that the potential overvaluation of the sure outcome in comparison to a risky prospect (the well-known certainty effect reported by Kahneman and Tversky, 1979) leading to biased SG measurements (very high utilities reflecting extreme risk aversion attitude) is presumably avoided or, at least, minimized.

Specifically our method asks the respondents to state the probability p^* that made them indifferent between prospect (*FH*, p, *Death*) and prospect (*FH*, 0.5, h), where *FH* stands for full health and h stands for the health state to be valued. Abellán-Perpiñán et al. (2011), after discussing other possible reasons, concluded that the wider range of their SF-6D value set was mainly due to the usage of this (probability) lottery equivalent method. Both theoretical and empirical arguments supporting this finding are provided in their paper.

The procedure used to search for indifference was based on a multiple sequence of choices, in such a way that the interval of values from which the indifference probability would be finally selected became narrower as the respondent made a new choice. Initially, p was fixed as 0.5 to know if the respondent considered that the health state was better or worse than death. If for that initial probability the respondent preferred the first lottery (*FH*, p, *Death*), then the health state h was regarded as worse than death, so the indifference probability should be lower than 0.5. On the contrary, if the respondent preferred the second lottery (*FH*, 0.5, h) then h was regarded as better than death and p^* should be higher than 0.5.

Utility scores were calculated under expected utility from the indifference values stated by the respondents as $U(h) = 2p^* - 1$, assuming the conventions of $U(FH) = 1$ and $U(Death) = 0$. Utility scores calculated in such a way range from −1 to 1.

3.3. The valuation survey

The total sample was divided into 17 subsamples ($n = 60$ each) retaining representativeness with respect to age and sex, in such a way that each of the 17 groups of respondents valued a different subset of five health states. This between-subject design allowed us to obtain a higher number of valuations per health state than Brazier et al. (2002), in whose study each health state was only valued an average of 15 times.

The survey consisted of a computer assisted questionnaire. All the interviews were run on notebook computers. Responses were collected in personal interview sessions. Average time per interview was about 20 min.

Before valuating the health states, a brief explanation of the SF-6D classification system was presented to the respondents who were then asked to rate the five SF-6D health states (anonymously labeled as *V*, *W*, *X*, *Y*, *Z*) by means of a visual analogue scale. The main section of the interview consisted in the valuation of those five health states with the lottery equivalence method previously described. In the final part of the questionnaire information about both health status and socioeconomic characteristics (sex, age, studies, income level, etc.) was collected. Three instruments were used to ask the respondents how healthy they felt: the EQ-5D self-report questionnaire, the SF-36 questionnaire and a visual scale similar to that presented previously for the valuation of the hypothetical states. Table 1 provides descriptive statistics of the sociodemographic variables used in the analysis.

Table 1
Characteristics of sample respondents and Spanish population.

	Sample	Population[a]
Female	50.0	52.05
Age	43.60 (16.64)	46.97 (19.04)
MarStat1	59.84	63.59
MarStat2	6.53	11.25
Mid-educ	34.54	17.56
High-educ	31.02	20.55
Children (presence)	48.80	25.64
Children (number)	1.82 (0.66)	1.41 (0.63)
Income2	28.31	17.39
Income3	29.82	21.76
Income4	18.98	11.47
Smoke2	16.57	9.19
Smoke3	8.63	13.52
Smoke4	1.71	4.74
Own2	10.44	22.25
Own3	1.20	10.72
N	4980	11,515

Notes: The table reports percentages for discrete variables and means and standard errors (in brackets) for continuous variables.

[a] The statistics are calculated using the Spanish sample of the European Community Household Panel for the year 2001.

4. Estimation results

We first analyze the results of implementing the standard approach and provide some evidence on the valuation effect of the respondents' characteristics. Then, we compare the parametric and the semi-parametric estimates of β.

4.1. The standard approach

In Table 2 we present *OLS* and *RE* estimates of β coming from the "main effects" model (columns 1 and 2) and from an expanded regression model that additively incorporates the respondents' characteristics that were collected in the survey (columns 3 and 4). In all cases, the Ramsey RESET and Jarque–Bera tests reject the null hypothesis that the model is correctly specified and that the estimation residuals are normally distributed, respectively.

Let us first comment on the estimates of β. As is commonly found in most applications of the standard approach, the *OLS* and *RE* estimates are quite close in magnitude to each other. No clear direction of change in the magnitude of the estimated β is observed when taking into account that a respondent values several health states, that is, when moving from the *OLS* to the *RE* estimates. In fact, the only qualitative differences between the *OLS* and the *RE* estimates concentrate on mild departures from full health in the "physical functioning" and "pain" dimensions. The coefficient associated with these variables is only found to be significantly different from zero in the *RE* estimates.

There are no inconsistencies in the estimated β and both the *OLS* and *RE* estimates indicate that being limited in the kind of work or other activities as a result of physical health (*RL2*) has no significant effect on health state valuations. An inconsistency occurs if the coefficient estimated for Z_{kw} is not strictly higher than that for

Table 2
Standard approach estimates.

	Main effects model		Expanded model	
	OLS	*RE*	*OLS*	*RE*
c	1.000	1.000	1.000	1.000
PF2	−0.015	−0.025**	−0.016	−0.025**
PF3	−0.034***	−0.056***	−0.031***	−0.054***
PF4	−0.090***	−0.120***	−0.088***	−0.118***
PF5	−0.111***	−0.107***	−0.103***	−0.106***
PF6	−0.338***	−0.335***	−0.332***	−0.333***
RL2	−0.014	0.007	−0.014	0.005
RL3	−0.038***	−0.045***	−0.041***	−0.046***
RL4	−0.070***	−0.088***	−0.078***	−0.091***
SF2	−0.037***	−0.071***	−0.036***	−0.070***
SF3	−0.060***	−0.078***	−0.063***	−0.079***
SF4	−0.203***	−0.194***	−0.203***	−0.194***
SF5	−0.208***	−0.239***	−0.210***	−0.240***
PAIN2	−0.018	−0.044***	−0.016	−0.043***
PAIN3	−0.034***	−0.047***	−0.033***	−0.048***
PAIN4	−0.198***	−0.172***	−0.202***	−0.174***
PAIN5	−0.202***	−0.230***	−0.208***	−0.232***
PAIN6	−0.318***	−0.343***	−0.318***	−0.342***
MH2	−0.066***	−0.026***	−0.064***	−0.025***
MH3	−0.078***	−0.050***	−0.080***	−0.050***
MH4	−0.096***	−0.072***	−0.096***	−0.073***
MH5	−0.224***	−0.196***	−0.226***	−0.197***
VIT2	−0.058***	−0.043***	−0.055***	−0.042***
VIT3	−0.121***	−0.093***	−0.120***	−0.094***
VIT4	−0.157***	−0.158***	−0.154***	−0.155***
VIT5	−0.199***	−0.181***	−0.197***	−0.180***
Female			0.015**	0.015
Age			−0.003***	−0.003**
Age squared			0.003***	0.003**
MarStat1			0.059***	0.063***
MarStat2			−0.014	−0.018
Mid-educ			0.003	−0.002
High-educ			−0.020**	−0.022
Children[a]			−0.010**	−0.010*
Income2			0.021**	0.025*
Income3			0.036***	0.039**
Income4			0.055***	0.051***
Smoke2			−0.001	−0.011
Smoke3			0.026**	0.030
Smoke4			−0.047*	−0.054
Own2			0.007	0.011
Own3			0.038	0.041
Adj. R^2	0.850	0.855	0.856	0.861
Ramsey's reset[b]	0.000	0.000	0.000	0.000
Jarque–Bera[b]	0.000	0.000	0.000	0.000
N	4990	4990	4980	4980

[a] Number of children in the household.
[b] We report *p*-values for the null hypothesis that the model has no omitted variables (Ramsey's Reset) and that the errors are normally distributed (Jarque-Bera).
* Significance at the 10% level.
** Significance at the 5% level.
*** Significance at the 1% level.

$Z_{kw'}$, for $w' > w$. However, there are two exceptions to this consistency rule in the SF-6D. Firstly, levels 5 and 6 of the "physical functioning" dimension ("your health limits you a little/a lot in bathing and dressing") do not necessarily imply a poorer condition than that of levels 3 or 4 ("your health limits you a little/a lot in moderate activities"). In a similar way, level 3 of the "role limitations" dimension ("you accomplish less than you would like as a result of emotional problems") does not reveal a worse health condition than that described in level 2 ("you are limited in the kind of work or other activities as a result of emotional problems").

Estimates in columns 3 and 4 show that health state values are correlated with some of the respondents' characteristics even after adjusting for differences in the severity of the health states being valued. In particular, valuations are significantly affected by the respondent's age and marital status and by other household level characteristics like household income and the number of children at home.[14] According to *RE* estimates, health state valuations are primarily affected by the respondent's age and marital status and by the level of income he enjoys at home.

The estimated non-linear effect of age implies that valuations increase slowly from the age of 18 to about the age of 47, fall slowly up to about 70 and then fall sharply in later years. This means that a 20 year old individual gives about the same value than an otherwise equivalent 70 year old individual. This non-linear association between the age of the respondent and health state values was also found in Dolan and Roberts (2002) and Kharroubi et al. (2007a) for the United Kingdom Time Trade-Off and Standard Gamble valuations of the EQ-5D and SF-6D, respectively. In particular, Dolan and

[14] The significant *OLS* estimates obtained for the sex and educational level of the respondent are not confirmed by the *RE* estimates.

Roberts (2002) also find that the age that maximizes valuations is about 45 years. In contrast, the corresponding age in Kharroubi et al. (2007a) is between 60 and 65 years.

The estimates in Table 2 indicate that there is a positive, monotonic and quantitatively relevant correlation between household income and health state valuations. The values of respondents whose household income is between 2000 and 3000 euros per month are, on average, 0.040 higher than those of respondents whose household income is below 1500 euros per month. That difference amounts to 0.053 if we compare the latter group to those whose total household income is above 3000 euros per month.

The positive association between household income and health state values can be interpreted in the light of the results in Lubetkin et al. (2005). They find a positive and relevant association between personal income and health-related quality of life in a large sample of the United States general population using the EQ-5D. That is, *ceteris paribus* and on average terms, high-income people enjoy better health than low-income people and, thus, we hypothesize that they are more likely to assign a low chance to the event of a bad health outcome when it is presented to them. Moreover, even if respondents judge the likelihood of the valued health states independently of their disposable income, the negative consequences of the realization of a bad health outcome are likely to be very different for low- and high-income individuals. The positive coefficient estimated for household income in Table 2 is compatible with the hypothesis that respondents value health states according to the utility losses that they expect should that health state be realized.

The estimates in Dolan and Roberts (2002) confirm the presence of systematic differences in the valuations of married and single respondents. However, while we find that the valuations of married or cohabiting people are, on average, 0.067 higher than the valuations of single people, they find that the average valuation of the latter collective is 0.006 higher than that of the former one.

Regarding children, Kharroubi et al. (2007a) find no significant association between the presence of children aged under 16 years in the household and the respondents' valuations. We obtain the same result when we control for whether there is a child aged under 12 years in the household or not.[15] However, when we allow for the presence and number of children in the household we obtain a negative and significant association between the number of children at home and the respondent's valuations.

Although existing studies disagree on the sign and magnitude of the effect of some personal characteristics, they provide robust evidence on the relevance of accounting for personal characteristics when estimating preference-based value functions.[16]

4.2. The IPW approach

The first two columns of Table 3 present *IPW*1 and *IPW*2 estimates of β calculated using the set of personal characteristics in Table 2. The external sample necessary to obtain *IPW*2 estimates comes from the Spanish sample of the European Community Household Panel for the year 2001, the latest available year. To facilitate the comparison, the *RE* main effects model estimates in Table 2 are displayed in column 3. We take these numbers as being representative of the standard approach estimates since, as shown in Table 2, they are numerically equivalent to the *OLS* estimates and to those coming from more complex specifications that also control for personal characteristics. The semi-parametric estimates in columns 4 and 5 are calculated by restricting the elements of X to the respondents' sex and age groups and, finally, in the last column we use corrective weights to adjust the *RE* main effects model estimates to the sex and age group distributions of the Spanish adult population, as is frequently done to ensure the representativeness of the standard model estimates.

There are relevant differences between the *IPW*2 and *RE* estimates of β in columns 2 and 3. While just one out of the 25 regression estimates is non-significantly different from zero, the semi-parametric estimates indicate five non-significant estimates. According to these estimates, being slightly limited in vigorous activities (*PF*2), being limited in social activities most of the time (*SF*3), having pain that interferes with normal work a little (*PAIN*3) and feeling tense or downhearted most of the time (*MH*4) has no significant effect. The same holds for being limited in the kind of work or other activities as a result of physical health (*RL*2) according to both the standard and the semi-parametric estimates. Conditioning on the estimated coefficients being significantly different from zero in both approaches, the semi-parametric estimates are higher in absolute value in 16 out of 20 coefficients and the difference between the *IPW*2 and the *RE* estimates is of much larger magnitude in those cases. On average, while the *IPW*2 estimates are 61% higher than the *RE* ones for the 16 coefficients for which the *IPW*2 estimates are larger in absolute value, the *RE* estimates are just 12% higher than the corresponding *IPW*2 estimates for the remaining 4 coefficients. That is, the IPW approach provides lower valuation impacts of departures from full health than the standard approach does.

One of the reasons why the standard and the IPW estimates differ are the different estimation samples used by the two approaches. In particular, while regression models use their functional form to extrapolate and overcome lack of overlap in the covariate distributions between treatment groups, the IPW approach uses the common support condition to obtain estimates that are not sensitive to the choice of specification. As previously discussed, the common support condition implies dropping units with extreme values of the propensity score. In practice, instead of using ad hoc methods for trimming the sample we follow Crump et al. (2009) and discard observations with estimated propensity score outside an interval $[\alpha, 1-\alpha]$, where the optimal cut-off value α is determined by the marginal distribution of the propensity score. This results in relevant precision gains. In particular, we calculate optimal cut-off values for each of the propensity scores involved in each semi-parametric estimate. In most cases the optimal value is close to 0.1.[17] The main cost of the approach developed in Crump et al. (2009) is that potentially some external validity is lost by focusing on a subset of the original sample. This cost is minimized in our case since the *IPW*2 estimator changes the sample distribution of X to that in the population of interest and, thus, it removes sample selection biases based on personal characteristics in X. Moreover, we have analyzed the stability of the estimates of β for different values of α. The estimates, available upon request to the authors, are stable and increase their precision as α gets closer to its optimal value.[18]

Next, differences between the *IPW*1 and *IPW*2 estimates in columns 1 and 2 indicate that the distribution of personal

[15] These estimates are available upon request to the authors.

[16] It is beyond the scope of this paper to explain these discrepancies. They might totally or partially be the result of differences in the elicitation methods, specifications and estimation methods used or they might simply reflect cross-country differences in the distribution of personal characteristics or in the effect of those characteristics.

[17] Crump et al. (2009) find that most of the precision gains are captured by using a rule of thumb to discard observations with the estimated propensity score outside the range [0.1, 0.9].

[18] We reach to the same conclusion for the multiple specifications of the propensity score that have been used in order to improve its balancing power.

Table 3
Standard and semi-parametric estimates of β.

	*IPW*1	*IPW*2	*RE*	*IPW*1[a]	*IPW*2[a]	*RE*[b]
c	1.000	1.000	1.000	1.000	1.000	1.000
PF2	−0.030	−0.011	−0.025**	−0.026	−0.028	−0.026***
PF3	−0.050**	−0.062*	−0.056***	−0.060**	−0.063***	−0.057***
PF4	−0.138***	−0.162***	−0.120***	−0.156***	−0.158***	−0.121***
PF5	−0.145***	−0.149***	−0.107***	−0.148***	−0.152***	−0.110***
PF6	−0.392***	−0.451***	−0.335***	−0.396***	−0.398***	−0.334***
RL2	−0.022	−0.027	0.007	−0.027	−0.028	0.009
RL3	−0.029**	−0.044**	−0.045***	−0.041*	−0.040*	−0.043***
RL4	−0.125***	−0.134***	−0.088***	−0.117***	−0.116***	−0.088***
SF2	−0.043*	−0.051*	−0.071***	−0.033*	−0.035*	−0.074***
SF3	−0.022	0.007	−0.078***	−0.030	−0.030	−0.080***
SF4	−0.162***	−0.157***	−0.194***	−0.162***	−0.165***	−0.197***
SF5	−0.221***	−0.219***	−0.239***	−0.233***	−0.232***	−0.238***
PAIN2	−0.054**	−0.074**	−0.044***	−0.059**	−0.062**	−0.045***
PAIN3	−0.058**	−0.029	−0.047***	−0.081***	−0.078***	−0.042***
PAIN4	−0.194***	−0.209***	−0.172***	−0.193***	−0.194***	−0.172***
PAIN5	−0.247***	−0.274***	−0.230***	−0.244***	−0.244***	−0.228***
PAIN6	−0.320***	−0.326***	−0.343***	−0.300***	−0.303***	−0.343***
MH2	−0.073***	−0.063**	−0.026***	−0.102***	−0.103***	−0.024***
MH3	−0.128***	−0.136***	−0.050***	−0.171***	−0.172***	−0.051***
MH4	−0.056*	−0.019	−0.072***	−0.062*	−0.064**	−0.072***
MH5	−0.169***	−0.176***	−0.196***	−0.171***	−0.170***	−0.195***
VIT2	−0.077***	−0.090***	−0.043***	−0.088***	−0.089***	−0.043***
VIT3	−0.162***	−0.165***	−0.093***	−0.189***	−0.185***	−0.089***
VIT4	−0.219***	−0.220***	−0.158***	−0.223***	−0.220***	−0.160***
VIT5	−0.232***	−0.247***	−0.181***	−0.230***	−0.234***	−0.178***

[a] The elements of X are restricted to the respondents' sex and age group.
[b] We use corrective weights to adjust to the sex and age groups distribution in the Spanish sample of the ECHP for 2001.
* Significance at the 10% level.
** Significance at the 5% level.
*** Significance at the 1% level.

characteristics in the sample significantly differs from that in the Spanish adult population. These differences are of lower magnitude than those found when comparing the *RE* to the *IPW*2 estimates. In most cases the *IPW*2 estimate exceed the corresponding *IPW*1 one. On average, the estimate of β_{kw} increases by 13% in absolute value when using the distributions of personal characteristics in the Spanish adult population instead of those in the estimation sample.

The finding that adjusting to the population distribution of the covariates results in relevant variations in the magnitude of the semi-parametric estimates contrasts with the evidence in Kharroubi et al. (2007a) and Dolan and Roberts (2002). These articles find that the standard model estimates are almost invariant to the inclusion of corrective weights that adjust to the age interval and sex distributions in the population. In fact, we reach the same conclusion when comparing the unweighted *RE* estimates in column 3 to those in column 4 where we use corrective weights defined over the respondents' sex and age groups. Interestingly, the semi-parametric estimates also lead to the same result once the elements of X are restricted to the respondents' sex and age groups. As shown in columns 4 and 5, the *IPW*1 and *IPW*2 estimates of β_{kw} obtained using the restricted set of personal characteristics are almost identical for any k and any w. Moreover, the restricted semi-parametric estimates are close in magnitude to the *IPW*1 estimates in column 1 but they differ substantially from the *IPW*2 estimates in column 2, where we adjust for the population distribution in the whole list of personal characteristics in Table 2. This suggests that adjusting to the population distributions of a reduced set of discretely measured personal characteristics is not enough to guarantee the population validity of the estimates. The propensity score seems far more effective in removing sample selection biases.

The estimation of the propensity score $p_s(x)$ allows us to formally test for sample selection biases, that is, to identify the characteristics whose sample distribution differs from that in the Spanish adult population. A significant coefficient in the estimation of the discrete choice model for the propensity score indicates that the distribution of the corresponding characteristic is not balanced between the population and the sample. As an illustrative example, in Table 4 we present the results of estimating a logit model for the propensity score $p_s(x)$ in the estimation of the coefficient associated to dimension "personal functioning" in its second level of severity (*PF*2). We find that the sample distribution of any characteristic but the respondents' sex significantly differs from that in the Spanish adult population. This finding that there are relevant compositional differences between the sample and the population

Table 4
Logit estimation of propensity score $p_s(x)$ for the coefficient associated to level of severity 2 in dimension "personal functioning".

Variable	Coefficient
Constant	−2.708***
Female	−0.038
Age	0.052***
Age sq.	−0.013
Marstat1	−1.626***
Marstat2	−1.620***
Mid-educ	0.870***
High-educ	0.375***
Children [a]	1.074***
Income2	1.235***
Income3	0.952***
Income4	1.131***
Smoke1	0.461***
Smoke2	−0.884***
Smoke3	−1.629***
Own2	−0.804***
Own3	−1.950***
Psedo R^2	0.236
N	13,509

[a] Number of children in the household.
*** Significance at the 1% level.

Table 5
Standard and semi-parametric consistent estimates.

	*IPW*2	*RE*
c	1.000	1.000
PF2	−0.011	−0.025**
PF3	−0.062*	−0.056***
PF4	−0.162***	−0.120***
PF5	−0.149***	−0.107***
PF6	−0.451***	−0.335***
RL2	−0.027	0.007
RL3	−0.044**	−0.045***
RL4	−0.134***	−0.088***
SF2	−0.051*	−0.071***
SF3		−0.078***
SF4		−0.194***
SF34[a]	−0.090***	
SF5	−0.219***	−0.239***
PAIN2	−0.074**	−0.044***
PAIN3		−0.047***
PAIN4		−0.172***
PAIN34[b]	−0.100***	
PAIN5	−0.274***	−0.230***
PAIN6	−0.326***	−0.343***
MH2	−0.063**	−0.026***
MH3	−0.136***	−0.050***
MH4		−0.072***
MH5		−0.196***
MH45[c]	−0.138***	
VIT2	−0.090***	−0.043***
VIT3	−0.165***	−0.093***
VIT4	−0.220***	−0.158***
VIT5	−0.247***	−0.181***

[a] Dummy indicator variable that equals one if the "social functioning" dimension reaches levels of severity 3 or 4, and zero otherwise.
[b] Dummy indicator variable that equals one if the "pain" dimension reaches levels of severity 3 or 4, and zero otherwise.
[c] Dummy indicator variable that equals one if the "mental health" dimension reaches levels of severity 4 or 5, and zero otherwise.
* Significance at the 10% level.
** Significance at the 5% level.
*** Significance at the 1% level.

of interest is common to the estimation of $p_s(x)$ for any of the *IPW*2 estimates in Table 3.[19]

Following Brazier and Roberts (2004), we estimate a *IPW*2 parsimonious consistent model by aggregating levels of a given dimension when inconsistencies are found, that is, when the coefficient estimated for Z_{kw} is not higher than that estimated for $Z_{kw'}$, for $w' > w$. Bearing in mind the previously discussed exceptions to this rule, we find three inconsistencies. The estimate for *PAIN*2 is larger in absolute value than that for *PAIN*3, the coefficient for *MH*4 is smaller in absolute value than that associated to *MH*3 and, finally, the estimate for *SF*3 is not significantly different from zero, while those for *SF*2 and *SF*4 are lower than zero. As opposed to the standard model, the IPW approach does not require the estimation of the full vector β once an inconsistency is detected. The *RE* model estimates are included for comparability purposes, since no inconsistencies are found in these estimates.

The consistent estimates presented in Table 5 allow us to estimate values for the 18,000 health states defined by the SF-6D classification system. The resulting *RE* and *IPW*2 estimated tariffs are summarized in Table 6 and their densities are depicted in Fig. 1. As expected given the preceding discussion, the semi-parametrically estimated values tend to be significantly lower than those predicted using the standard regression model. These discrepancies are observed in any of the distributional

[19] These estimates and also those for the first step estimate of the propensity score $p_{kw}(x)$ are available upon request to the authors.

Table 6
RE and *IPW*2 tariffs. Summary statistics of the values predicted for the 18,000 health states defined by the SF-6D.

	*IPW*2	*RE*
Mean	0.354	0.440
S.D.	0.249	0.212
Minimum	−0.553	−0.382
Percentiles		
10	0.021	0.156
25	0.192	0.299
50	0.368	0.452
75	0.532	0.594
90	0.667	0.708
Negative values (%)	8.82	2.53

moments and they tend to be higher the lower is the predicted utility of a health state, that is, the higher is its severity. The difference between the *z*th percentile of the *RE* and *IPW*2 distributions of values lowers from 0.135 for $z = 10$ to 0.041 for $z = 90$. In relative terms, the 10th and 90th percentiles of the *IPW*2 distribution are 86.5 and 5.8% higher than the corresponding percentiles of the *RE* distribution. A similar picture emerges when looking at the proportion of predicted negative values. While only 2.5% of the values predicted on the basis of the standard approach estimates are lower then zero, the corresponding number for the semi-parametric values is 3.5 times higher. In particular, almost 9% of the semi-parametric values are strictly lower than zero. The densities of the predicted tariffs in Fig. 1 confirm that the standard approach underestimates the utility loss that the Spanish adult population gives to deviations from full health, particularly so when dealing with severe deviations.

The IPW and the standard models cannot be directly compared in terms of their predictive ability since they use different estimation samples. While the regression model uses its functional form to work off the common support when estimating β_{kw}, the IPW model restricts the sample to respondents valuing levels of severity 1 and w in dimension k whose estimated propensity score is not close to zero or one. To get a feeling about the predictive ability of the models, we restrict the sample to respondents not excluded in the semi-parametric estimation of none of the coefficients used to predict the value of a particular health state. We find that the predictive performance of the models is similar with a root mean squared error of 0.139 for the *RE* model, 0.167 for the *IPW*2 consistent model and 0.137 for the *IPW*1 model associated to the *IPW*2 consistent model. The *IPW*1 model is included for comparability

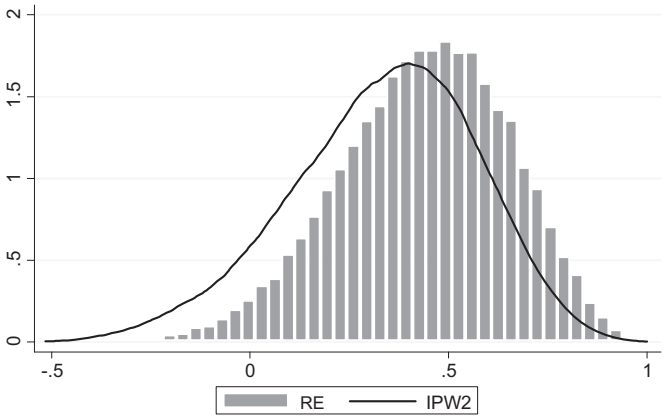

Fig. 1. A comparison of the *RE* and *IPW*2 tariffs' predicted values. Note: The graph presents the densities of the values predicted for the 18,000 health states defined by the SF-6D using the *RE* and *IPW*2 consistent estimates in Table 5.

purposes, since the *IPW*2 model imposes the distribution of characteristics in an external sample and, thus, it is expected to perform poorer in terms of within sample predictive ability. The associated *IPW*1 model performs slightly better than the *RE* model.

5. Conclusions

This paper presents a novel approach to model health state valuations using inverse probability weighting (IPW) techniques with important advantages over the standard regression model. Our approach makes no assumption on the distribution of health state values, accommodates covariates in a flexible way, eschews parametric assumptions on the relationship between the outcome and the regressors, allows for the valuation impact of departures from full health to be heterogeneous in personal characteristics and in the severity of departures in other dimensions of health. Additionally, unlike the standard approach, our approach produces population valid estimates even if the estimation sample is not representative for that population with regard to many discrete and continuously measured variables. The proposed estimators are semi-parametrically estimated and we also derive its large sample properties.

The standard model estimates are likely to be very sensitive to differences in the covariate distributions of respondents valuing different health states since it applies regression models that use extrapolation to deal with limited support in the distribution of the covariates. In contrast, our approach calls for selecting the health states for which direct valuations are obtained for identification not to rest on extrapolation. That will be the case in the main effects model if there is common support in the level of severity of the remaining dimensions of health between respondents valuing health states with a departure from full health in a particular dimension and those valuing health states with full health in that dimension.

We illustrate our approach with the SF-6D descriptive system. The results indicate that the standard and the semi-parametric estimates differ to a great extent, with the utility loss of a departure from full health being higher in most cases when estimated using our approach. In fact, when the estimated coefficients are used to predict utilities for the 18,000 health states defined in the SF-6D we find that the standard approach systematically underestimates the valuation impact of departures from full health and that the magnitude of the underestimation increases with the severity of the health state at examination.

Moreover, we also find evidence that the standard approach fails in its goal of producing population valid estimates by introducing corrective weights that adjust to the sex and age intervals distribution in the population of interest. We consider continuous generalizations of the corrective weights that easily overcome their limitations and allow us to test and correct for sample selection biases.

The IPW approach is easier to implement and interpret than the nonparametric Bayesian approach in Kharroubi et al. (2007b). In particular, and contrary to the nonparametric estimator, our approach provides the user with a table of estimated coefficients that defines the estimated preference function, which results in efficiency and transparency gains.

Finally, the IPW approach is suitable for application to any of the existing multi-attribute descriptive systems. In particular, it would be of interest to semi-parametrically estimate the EQ-5D value sets, since this is the most widely used instrument for the description and valuation of health states. With reference to this, it should be stressed the importance of the common support condition in the selection of the health states valued in the sample. If we look at the studies which have derived the EQ-5D tariffs for the United Kingdom, Spain, the Netherlands and Japan, we conclude that this overlap requirement in the levels of severity of the dimensions of health is met in six out of the 10 estimated coefficients in the case of the UK (Dolan, 1997) and Spain (Badia et al., 2001) estimations, whereas the condition is satisfied in just one out of 10 estimated coefficients in the Dutch (Lamers et al., 2006) and Japanese (Tsuchiya et al., 2002) tariffs. Consequently, the validity of the Dutch and Japanese estimates rests on whether the corresponding regression models were correctly specified or not. That is, the Dutch and Japanese tariffs are less likely to be robust to the misspecification of the regression model than those in Dolan (1997) and Badia et al. (2001).

Acknowledgement

José María Abellán Perpiñán, Jorge Eduardo Martínez Pérez and Fernando Ignacio Sánchez Martínez also acknowledge financial support from Ministerio de Ciencia e Innnovación grant ECO2010-22041-C02-02. This research was supported by the Health and Consumption Department of the Autonomous Community of Murcia.

Appendix A. Asymptotic properties

We derive the asymptotic properties of $\widehat{\beta}_{kw,IPW2}$ and present those of $\widehat{\beta}_{kw,IPW1}$ as a particular case. The subscript *IPW*2 is dropped out to reduce the notation. The properties of $\widehat{\beta}_{kw}$ are derived by viewing it as an *M*-estimator, that is, as the solution to a set of estimating equations.[20] In particular, $\widehat{\beta}_{kw}$ is one element of the vector $\widehat{\theta}$ that solves the vector equation

$$\sum_{i=1}^{n} \psi(W_i, \widehat{\theta}) = 0$$

where $W_i = [Y_i, Z_{k'w'}, X_i]$, for $k' \neq k$ and $w' = 2, 3, \ldots, W_{k'}$ and $\theta = [\delta, \gamma, \beta_{kw}]$. The vector equation ψ has three equations and can be written as

$$\sum_{i=1}^{n} \psi_1(W_i, \theta) = \sum_{i=1}^{n} \frac{D_{si} - p_{si}(X_i, \delta)}{p_{si}(X_i, \delta)[1 - p_{si}(X_i, \delta)]} \frac{\partial p_{si}(X_i, \delta)}{\partial \delta} = 0$$

$$\sum_{i=1}^{n} \psi_2(W_i, \theta) = \sum_{i=1}^{n} \frac{Z_{kwi} - p_{kwi}(H_i, \gamma)}{p_{kwi}(H_i, \gamma)[1 - p_{kwi}(H_i, \gamma)]} \frac{\partial p_{kwi}(H_i, \gamma)}{\partial \gamma} = 0$$

$$\sum_{i=1}^{n} \psi_3(W_i, \theta) = \sum_{i=1}^{n} \left\{ A(1) \frac{Z_{kwi} Y_i}{\widehat{p}_{kwi} \widehat{p}_{si}} - A(0) \frac{(1 - Z_{kwi}) Y_i}{(1 - \widehat{p}_{kwi}) \widehat{p}_{si}} - \beta_{kw} \right\} = 0$$

where $p_{si} = p_{si}(X_i, \delta)$, $p_{kwi} = p_{kwi}(H_i, \gamma)$, $\widehat{p}_{kwi} = (H_i, \widehat{\gamma})$, $\widehat{p}_{si} = (X_i, \widehat{\delta})$, $A(t) = N(\sum_{i=1}^{n} (Z_{kwi}^{t}(1 - Z_{kwi})^{1-t})/(\widehat{p}_{kwi}^{t}(1 - \widehat{p}_{kwi})^{1-t} \widehat{p}_{si}))^{-1}$ for $t = \{0, 1\}$ and N is the total number of individuals in the estimation sample. The solutions to equations $\psi_1(W_i, \theta)$ and $\psi_2(W_i, \theta)$ are the maximum likelihood estimates of δ and γ, the coefficients of the binary response models used to estimate the propensity scores p_s and p_{kw}, respectively. We estimate the propensity scores using the logistic regression model, where $p(Q, \varphi) = \{1 + \exp(-Q^T \varphi)\}^{-1}$. The solution to equation $\psi_3(W_i, \theta)$ is the coefficient of interest.

By standard results on M-estimation, under the true parameter value θ

$$\sqrt{n}(\widehat{\theta} - \theta) \longrightarrow N(0, A(\theta)^{-1} B(\theta) \{A(\theta)^{-1}\}^T)$$

[20] Stefanski and Boos (2002) provide an excellent review of the theory of M-estimation. Additionally, Lunceford and Davidian (2004) derive the asymptotic properties of the *IPW*1 estimator.

where

$$A(\theta) = E[-\tilde{\psi}(W,\theta)]$$

with $\tilde{\psi}(W,\theta) = \partial\psi(W,\theta)/\partial\theta^T$ and

$$B(\theta) = E[\psi(W,\theta)\psi(W,\theta)^T]$$

To estimate the asymptotic variance use

$$\widehat{A} = \frac{1}{n}\sum\nolimits_{i=1}^{n} -\frac{\partial\psi(W_i,\widehat{\theta})}{\partial\theta^T}$$
$$\widehat{B} = \frac{1}{n}\sum\nolimits_{i=1}^{n} \psi(W_i,\widehat{\theta})\psi(W_i,\widehat{\theta})^T$$

where the derivative of ψ can be calculated as

$$\frac{\partial\psi(W,\theta)}{\partial\theta^T} = \begin{pmatrix} \frac{\partial\psi_1(W,\theta)}{\partial\delta^T} & \frac{\partial\psi_1(W,\theta)}{\partial\gamma^T} & \frac{\partial\psi_1(W,\theta)}{\partial\beta_{kw}^T} \\ \frac{\partial\psi_2(W,\theta)}{\partial\delta^T} & \frac{\partial\psi_2(W,\theta)}{\partial\gamma^T} & \frac{\partial\psi_2(W,\theta)}{\partial\beta_{kw}^T} \\ \frac{\partial\psi_3(W,\theta)}{\partial\delta^T} & \frac{\partial\psi_3(W,\theta)}{\partial\gamma^T} & \frac{\partial\psi_3(W,\theta)}{\partial\beta_{kw}^T} \end{pmatrix}$$

where

$$A_{11i} = \frac{\partial\psi_1(W,\theta)}{\partial\delta^T} = -\frac{1}{p_{si}(1-p_{si})}P_\delta P_\delta^T$$
$$A_{12i} = \frac{\partial\psi_1(W,\theta)}{\partial\gamma^T} = A_{13i} = \frac{\partial\psi_1(W,\theta)}{\partial\beta_{kw}^T} = 0$$
$$A_{21i} = \frac{\partial\psi_2(W,\theta)}{\partial\delta^T} = A_{23i} = \frac{\partial\psi_2(W,\theta)}{\partial\beta_{kw}^T} = 0$$
$$A_{22i} = \frac{\partial\psi_2(W,\theta)}{\partial\gamma^T} = -\frac{1}{p_{kwi}(1-p_{kwi})}P_\gamma P_\gamma^T$$
$$A_{31i} = \frac{\partial\psi_3(W,\theta)}{\partial\delta^T} = -\left[\frac{D_{si}D_{kwi}Y_i^*}{p_{kwi}{p_{si}}^2} - \frac{D_{si}D_{kwi}Y_i^*}{(1-p_{kwi}){p_{si}}^2}\right]P_\delta$$
$$A_{32i} = \frac{\partial\psi_3(W,\theta)}{\partial\gamma^T} = -\left[\frac{D_{si}D_{kwi}Y_i^*}{p_{kwi}^2 p_{si}} + \frac{D_{si}D_{kwi}Y_i^*}{(1-p_{kwi})^2 p_{si}}\right]P_\gamma$$
$$A_{33i} = \frac{\partial\psi_3(W,\theta)}{\partial\beta_{kw}^T} = -1$$

where $P_\delta = \partial/\partial\,\delta\{p_{si}\}$, $P_\gamma = \partial/\partial\,\gamma\{p_{kwi}\}$ and $Y_i^* = D_{kwi}A(1)Y_i + (1 - D_{kwi})A(0)Y_i$. Equivalently, the elements of B are calculated as

$$B_{11i} = \psi_1(W_i,\theta)\psi_1(W_i,\theta)^T = \frac{1}{p_{si}(1-p_{si})}P_\delta P_\delta^T$$
$$B_{12i} = \psi_1(W_i,\theta)\psi_2(W_i,\theta)^T = 0$$
$$B_{13i} = \psi_1(W_i,\theta)\psi_3(W_i,\theta)^T = \left[\frac{D_{si}D_{kwi}Y_i^*}{p_{kwi}{p_{si}}^2} - \frac{D_{si}D_{kwi}Y_i^*}{(1-p_{kwi}){p_{si}}^2}\right]P_\delta$$
$$B_{21i} = \psi_2(W_i,\theta)\psi_1(W_i,\theta)^T = 0$$
$$B_{22i} = \psi_2(W_i,\theta)\psi_2(W_i,\theta)^T = \frac{1}{p_{kwi}(1-p_{kwi})}P_\gamma P_\gamma^T$$
$$B_{23i} = \psi_2(W_i,\theta)\psi_3(W_i,\theta)^T = \left[\frac{D_{si}D_{kwi}Y_i^*}{p_{kwi}^2 p_{si}} + \frac{D_{si}(1-D_{kwi})Y_i^*}{(1-p_{kwi})^2 p_{si}}\right]P_\gamma$$
$$B_{31i} = \psi_3(W_i,\theta)\psi_1(W_i,\theta)^T = B_{13i}^T$$
$$B_{32i} = \psi_3(W_i,\theta)\psi_2(W_i,\theta)^T = B_{23i}^T$$
$$B_{33i} = \psi_3(W_i,\theta)\psi_3(W_i,\theta)^T = \left(\frac{D_{si}D_{kwi}Y_i^*}{p_{kwi}p_{si}} - \frac{D_{si}(1-D_{kwi})Y_i^*}{(1-p_{kwi})p_{si}} - \beta_{kw}\right)^2$$

Finally, it can be shown that the large-sample variance of β_{kw} is

$$V(\beta_{kw}) = A_{33}^{-1}(B_{33} - B_{23}^T B_{22}^{-1} B_{23} - B_{13}^T B_{11}^{-1} B_{13})(A_{33}^{-1})^T$$

The expression of the large-sample variance of β_{kw} in the case where γ and δ are known is $A_{33}^{-1}B_{33}(A_{33}^{-1})^T$. The additional two terms in the parenthesis are the adjustment in the large-sample variance of the effect of interest coming from the first step estimation of the two propensity scores. Interestingly, it results that estimation of the propensity scores leads to smaller large-sample variance for these *IPW*estimators than using the true values. That is, as Lunceford and Davidian (2004) point out, even if the functional form of the propensity score is known exactly, it is beneficial from an efficiency viewpoint to estimate it. Hirano et al. (2003) explain this result in the context of the Generalized Method of Moments and the Empirical Likelihood estimators.

The expression for the variance of the *IPW*1 estimator includes only the first two terms in the parenthesis in the latter expression, where B_{23} and B_{33} are now calculated as the sample average of the following expressions evaluated at the estimated value of the elements of θ

$$B_{23i} = \left[\frac{D_{kwi}Y_i^*}{p_{kwi}^2} + \frac{(1-D_{kwi})Y_i^*}{(1-p_{kwi})^2}\right]P_\gamma$$
$$B_{33i} = \left(\frac{D_{kwi}Y_i^*}{p_{kwi}} - \frac{(1-D_{kwi})Y_i^*}{(1-p_{kwi})} - \beta_{kw,IPW1}\right)^2$$

Appendix B. Variable definitions and sources

The variables for the Spanish population are constructed using the Spanish sample of the European Community Household Panel (ECHP) for the year 2001, the latest available year, provided by Eurostat. In the empirical analysis we control for the sex and age (in years) of the respondent, whether he is married or cohabiting (*MarStat*1) or separated, divorced or widow (*MarStat*2) and whether the respondent has attained a secondary level of education (Mid-educ) or an university degree (High-educ). We also classify respondents according to whether their monthly total household income is below 1500 euros, between 1500 and 2000 euros (*Income*2), between 2000 and 3000 euros (*Income*3) or above 3000 euros (*Income*4). Regarding their smoking behaviour, we distinguish between non-smokers and respondents who actually smoke less than 10 cigarettes per day (*Smoke*2), between 10 and 20 cigarettes (*Smoke*3) and more than 20 cigarettes per day (*Smoke*4). Additionally, we construct two dummy variables that indicate if the respondent thinks that his general health is fair (*Own*2) or bad/very bad (*Own*3). The remaining categories in the answer to the question of how is your health in general are good and very good. We control for the number of children in the household. However, while the variable constructed using ECHP data refers to children under the age of 16 years, the corresponding variable from the collected data refers to children aged under the age of 12.

References

Abellán-Perpiñán, J.M., Sánchez-Martínez, F.I., Martínez-Pérez, J.E., Méndez, I., 2011. Lowering the 'floor' of the SF-6D scoring algorithm using a lottery equivalent method. Health Economics, doi:10.1002/hec.1792.

Badia, X., Roset, M., Herdman, M., Kind, P., 2001. A comparison of United Kingdom and Spanish general population time trade-off values for EQ-5D health states. Medical Decision Making 21, 7–16.

Barnow, B., Cain, C., Goldberger, A., 1981. Selection on observables. Evaluation Studies Review Annual 5, 43–59.

Brazier, J.E., Roberts, J., 2004. The estimation of a preference-based measure of health from the SF-12. Medical Care 42 (9), 851–859.

Brazier, J.E., Roberts, J., Deverill, M., 2002. The estimation of a preference-based measure of health from the SF-36. Journal of Health Economics 21, 271–292.

Brazier, J.E., Roberts, J., Tsuchiya, A., Busschbach, J., 2004. A comparison of the EQ-5D and SF-6D across sever patients groups. Health Economics 13, 873–884.
Brazier, J.E., Usherwood, T., Harper, R., Thomas, K., 1998. Deriving a preference-based single index from the UK SF-36 health survey. Journal of Clinical Epidemiology 51 (11), 1115–1128.
Brazier, J.E., Fukuhara, S., Roberts, J., Kharroubi, S., Yamamoto, Y., Ikeda, S., Doherty, J., Kurokawa, K., 2009. Estimating a preference-based index form the Japanese SF-36. Journal of Clinical Epidemiology 62, 1323–1331.
Busso, M., DiNardo, J., McCrary, J., 2009. New Evidence on the Finite Sample Properties of Propensity Score Matching and Reweighting Estimators. IZA Discussion Paper 3998.
Craig, B.M., Busschbach, J.J., 2009. The episodic random utility model unifies time trade-off and discrete choice approaches in health state valuation. Population Health Metrics 7, 16–25.
Craig, B.M., Busschbach, J.J., 2011. Toward a more universal approach in health valuation. Health Economics 20, 864–875.
Crump, R.K., Hotz, V.J., Imbens, G.W., Mitnik, O.A., 2009. Dealing with limited overlap in estimation of average treatment effects. Biometrika 96, 187–199.
DiNardo, J., Fortin, N.M., Lemieux, T., 1996. Labor market institutions and the distribution of wages, 1973–1992: a semi-parametric approach. Econometrica 64 (5), 1001–1044.
Dolan, P., 1997. Modelling valuations for Euroqol health states. Medical Care 35, 351–363.
Dolan, P., Gudex, C., Kind, P., Williams, A., 1996. The time trade-off method: results from a general population study. Health Economics 5, 141–154.
Dolan, P., Roberts, J., 2002. To what extent can we explain time trade-off values from other information about respondents? Social Science and Medicine 54, 919–929.
EuroQol Group, 1990. EuroQol—a new facility for the measurement of health-related quality of life. Health Policy 16, 199–208.
Feeny, D., Furlong, W., Torrance, G.W., Goldsmith, C.H., Zenglong, Z., Depauw, S., Denton, M., Boyle, M., 2002. Multi-attribute and single-attribute utility function for the Health Utility Index Mark 3 system. Medical Care 40, 113–128.
Ferreira, L.N., Ferreira, P.L., Pereira, L.N., Brazier, J.E., Rowen, D., 2010. A Portuguese Value Set for the SF-6D. Value in Health 13 (5), 624–630.
Gold, M.R., Siegel, J.E., Russell, L.B., Weinstein, M.C., 1996. Cost-Effectiveness in Health and Medicine. Oxford University Press, Oxford.
Hirano, K., Imbens, G.W., Ridder, G., 2003. Efficient estimation of average treatment effects using the estimated propensity score. Econometrica 71 (4), 1161–1189.
Horvitz, D.G., Thompson, D.J., 1952. A generalization of sampling without replacement from a finite universe. Journal of the American Statistical Association 47, 663–685.
Imbens, G.W., 2004. Nonparametric estimation of average treatment effects under exogeneity: a review. The Review of Economics and Statistics 86 (1), 4–29.
Johnson, J.A., Coons, S.J., Ergo, A., Szava-Kovats, G., 1998. Valuation of EuroQOL (EQ-5D) health state in an adult US sample. Pharmaconomics 13 (4), 421–433.
Kahneman, D., Tversky, A., 1979. Prospect theory: An analysis of decision under risk. Econometrica 47, 263–291.
Kharroubi, S., Brazier, J.E., O'Hagan, A., 2007a. Modelling covariates for the SF-6D standard gamble health state preference data using a nonparametric Bayesian method. Social Science and Medicine 64, 1242–1252.
Kharroubi, S., Brazier, J.E., Roberts, J., O'Hagan, A., 2007b. Modelling SF-6D health state preference data using a nonparametric Bayesian method. Journal of Health Economics 26, 597–612.
Lam, C.L., Brazier, J.E., McGhee, S.M., 2008. Valuation of the SF-6D health states is feasible, acceptable, reliable, and valid in a Chinese population. Value in Health 11, 295–303.
Lamers, L.M., McDonnell, J., Stalmeier, F.M., Krabbe, P.F.M., Busschbach, J.J.V., 2006. The Dutch Tariff: results and arguments for an effective design for national EQ-5D valuation studies. Health Economics 15, 1121–1132.
Lubetkin, E.I., Jia, H., Franks, P., Gold, M.R., 2005. Relationship among sociodemographic factors, clinical conditions, and health related quality of life: examining the EQ-5D in the U.S. general population. Quality of Life Research 14, 2187–2196.
Lunceford, J.K., Davidian, M., 2004. Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. Statistics in Medicine 23, 2937–2960.
McCord, M., de Neufville, R., 1986. Lottery equivalents: reduction of the certainty effect problem in utility assessment. Management Science 32 (1), 56–60.
National Institute for Clinical Excellence (NICE), 2004. Guide to the Methods of Technology Appraisal. NICE, London.
Robins, J.M., Rotnitzky, A., Zhao, L.P., 1994. Estimation of regression coefficients when some regressors are not always observed. Journal of the American Statistical Association 89, 846–866.
Rosenbaum, P., Rubin, D.B., 1983. The central role of the propensity score in observational studies for causal effects. Biometrika 70, 41–55.
Rubin, D.R., 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. Journal of Educational Psychologies 66, 688–701.
Stefanski, L.A., Boos, D.D., 2002. The calculus of M-estimation. The American Statistician 56, 29–38.
Tsuchiya, A., Ikeda, S., Ikegami, N., Nishimura, S., Sakai, I., Fukuda, T., Hamashima, C., Hisashige, A., Tamura, M., 2002. Estimating an EQ-5D population value set: the case of Japan. Health Economics 11, 341–353.