

# Methodological Aspects in the Assessment of Treatment Effects in Observational Health Outcomes Studies

Josep Maria Haro,<sup>1</sup> Stathis Kontodimas,<sup>2</sup> Miguel Angel Negrin,<sup>3</sup> Mark Ratcliffe,<sup>2</sup> David Suarez<sup>1</sup> and Frank Windmeijer<sup>4</sup>

- 1 Research and Development Unit, RIRAG Network (FIS G03/061), Sant Joan de Deu-SSM, Fundació Sant Joan de Déu, Sant Boi, Barcelona, Spain
- 2 European Health Outcomes Research, Eli Lilly and Company Limited, Windlesham, UK
- 3 Department of Quantitative Methods, RIRAG Network (FIS G03/061), University of Las Palmas de Gran Canaria, Las Palmas de Gran Canaria, Spain
- 4 Department of Economics, University of Bristol, Bristol, UK

---

## Contents

|  |    |
|--|----|
| Abstract   | 11 |
| 1. The SOHO (Schizophrenia Outpatient Health Outcomes) Study | 13 |
| 2. Methods to Adjust for Selection Bias                      | 14 |
| 2.1 Multivariate Regression Analysis                         | 14 |
| 2.2 Propensity Score Matching                                | 16 |
| 2.3 Instrumental Variables                                   | 16 |
| 3. Repeated Measurements of Outcome                          | 17 |
| 3.1 Inference and Estimation in the Standard Models          | 17 |
| 3.2 Bayesian Hierarchical Model                              | 18 |
| 4. Estimation Results  | 19 |
| 5. Methods to Assess Observer Bias                           | 20 |
| 6. Conclusions   | 23 |

---

## Abstract

Prospective observational studies, which provide information on the effectiveness of interventions in natural settings, may complement results from randomised clinical trials in the evaluation of health technologies. However, observational studies are subject to a number of potential methodological weaknesses, mainly selection and observer bias. This paper reviews and applies various methods to control for selection bias in the estimation of treatment effects and proposes novel ways to assess the presence of observer bias. We also address the issues of estimation and inference in a multilevel setting. We describe and compare the use

of regression methods, propensity score matching, fixed-effects models incorporating investigator characteristics, and a multilevel, hierarchical model using Bayesian estimation techniques in the control of selection bias. We also propose to assess the existence of observer bias in observational studies by comparing patient- and investigator-reported outcomes. To illustrate these methods, we have used data from the SOHO (Schizophrenia Outpatient Health Outcomes) study, a large, prospective, observational study of health outcomes associated with the treatment of schizophrenia.

The methods used to adjust for differences between treatment groups that could cause selection bias yielded comparable results, reinforcing the validity of the findings. Also, the assessment of observer bias did not show that it existed in the SOHO study. Observational studies, when properly conducted and when using adequate statistical methods, can provide valid information on the evaluation of health technologies.

Randomised clinical trials (RCTs) have long been considered the gold standard research methodology for proving the clinical efficacy, safety and quality of medical and pharmaceutical interventions. However, the limitations of RCTs are increasingly receiving attention in the literature. Because RCTs have specific and detailed inclusion criteria and low external validity,<sup>[1]</sup> it has been recommended that observational studies should routinely be conducted post-RCT as a necessary complement to the experimental data.<sup>[2,3]</sup> More specifically, health technology assessment (HTA) bodies and healthcare decision makers across the globe, including the UK's National Institute of Clinical Excellence (NICE), the Australian Pharmaceutical Benefits Advisory Committee (PBAC) and the French Transparency Committee, increasingly require or mandate 'real-life' data generated via observational studies.

A major strength of observational research is that it provides a means to rigorously address issues of real-life effectiveness within the context of a naturalistic setting outside the restrictive experimental environment necessarily generated by clinical trial protocols. In addition, observational studies enable healthcare decision makers to answer a variety of questions on broader treatment effects and humanis-

tic outcomes, disease aetiology and epidemiological issues, thereby giving a depth and breadth of data over and above that provided by the primary endpoints.

However, the strengths of observational research methodology are accompanied by a number of potential methodological weaknesses or biases. In an environment where decision makers will become increasingly reliant on data generated by observational research, it is essential to develop new techniques, and refine existing ones, to assess the validity of the data derived from observational studies. While there is currently much debate in the literature on the relative methodological strengths and weaknesses of RCTs and observational studies,<sup>[4-25]</sup> the objective of this paper is not to review this debate. Rather, we aim to review and apply some of the commonly used methods for quantifying treatment effects and dealing with issues of selection and observer bias within the context of observational research. We also propose new ways to measure observer bias in observational studies. We believe these methods are essential tools for researchers working with observational data and will contribute toward the scientific robustness of measuring treatment effects from observational studies. Specific-

ly, we describe and compare the use of regression methods, propensity score matching, fixed-effects models incorporating investigator characteristics, and a multilevel, hierarchical model using Bayesian estimation techniques. While the first three methods are frequently used statistical techniques for adjusting differences between treatment groups, which would lead to selection bias if not taken into account, the use of Bayesian statistics is relatively new in the area. We also propose new methods of analysis of observer bias, based on the comparison between patient- and investigator-reported outcomes using several statistical techniques. To illustrate the methods presented, we have used data from a large, prospective, naturalistic pan-European observational study of health outcomes associated with the treatment of schizophrenia, the SOHO (Schizophrenia Outpatient Health Outcomes) study.<sup>[26]</sup>

## 1. The SOHO (Schizophrenia Outpatient Health Outcomes) Study

Since the SOHO study is used to illustrate the methods presented in this paper, we will briefly discuss its design. The SOHO study is a 3-year longitudinal, observational study of health outcomes in the outpatient treatment of schizophrenia with antipsychotics, in which 1096 psychiatrists from ten European countries have enrolled 10 972 patients.<sup>[26-29]</sup> Outpatients with schizophrenia, aged at least 18 years who were initiating or changing antipsychotic medication within the normal course of care could be enrolled. Stratified sampling with oversampling was used to provide two patient cohorts of approximately equal size: patients who initiated therapy with or changed to olanzapine; and patients who initiated therapy with or changed to a non-olanzapine antipsychotic. Patients were evaluated at baseline, 3 months, 6 months and then every 6 months up to 36 months post-baseline. The main outcomes assessed in SOHO were clinical severity and health-related quality of life. Clinical severity

was assessed by the investigator using the Clinical Global Impression (CGI) overall severity scale,<sup>[30]</sup> a single-item physician-rated scale with values ranging from 1 (not ill) to 7 (among the most severely ill patients). Quality of life was assessed by the patient using the EuroQol-5 Dimensions (EQ-5D).<sup>[31]</sup> The EQ-5D results allow the calculation of utility scores, which in our case have been calculated using the UK tariffs.<sup>[32]</sup>

The results from the baseline and 3- and 6-month visits of SOHO have been published elsewhere.<sup>[28,33]</sup> Clinical outcomes were reported according to the antipsychotic medication the patient started at the baseline visit. These data are used to illustrate the methods with which we propose to analyse the presence of bias. Here we present the 3-month results comparing olanzapine with other antipsychotics using the outcome measure CGI overall severity scale, using different statistical techniques.

A notable feature of SOHO is that patients can change medication at any point and still remain in the study and be evaluated. Therefore, some of the patients who started a particular antipsychotic at baseline may have changed to another one at the 3-month visit. We have considered that patients who changed treatment at the 3-month visit subsequently provide a new 3-month observation. That is, such patients provide a first 3-month outcome observation by comparison of the baseline and 3-month data and relating this to the medication started at baseline, and also provide a second 3-month observation outcome by comparing the clinical status at the 3- and 6-month visits and relating it to the medication started at the 3-month visit. The analysis takes into account the correlation between the visits by using different statistical techniques as discussed in the following sections.

The analyses presented here include the 6412 patients who started treatment with only one antipsychotic agent at the baseline visit and were evaluated after 6 months follow-up. Those patients pro-

vided 6791 observations. Of them, 3419 were with olanzapine and 3372 with another antipsychotic. The original cohort included 10 205 patients eligible for analysis and the retention rate at 6 months was 88.5%.

## 2. Methods to Adjust for Selection Bias

In a typical prospective study, outcomes associated with exposure to a certain medication are analysed and compared with those in patients not treated with that medication. Treatment group comparisons within observational studies may be influenced by two types of systematic error or bias: selection bias and observer bias. Selection bias occurs when the groups to be compared are not equal in aspects that affect the outcome that is being studied. Since physicians choose treatments on the basis of patient and disease characteristics, such differences between treatment groups occur frequently in observational studies. Selection bias can be overt or hidden. Overt bias occurs when information about the factors that cause or are associated with the bias have been collected. Hidden bias takes place when the required information was not observed or recorded. In RCTs, selection bias is avoided by randomising patients to treatment groups. In observational studies, several methods have been used to control for selection bias, including multivariate regression analysis, propensity score matching and the use of instrumental variables.

### 2.1 Multivariate Regression Analysis

Regression models are the most frequently used method for adjusting for overt bias. They not only adjust for the influence of confounders but also provide direct estimates of the impact of explanatory variables on the outcome of interest. Regression methods usually require some model assumptions and a relatively large sample size when there are a large number of covariates. During the design of observational studies, selection bias (overt and hid-

den) should be controlled for by including all covariates that may be associated with the treatments and/or outcomes. Since there is no way to completely address hidden bias, a small change in the list of covariates may determine whether the study results are reliable.<sup>[34]</sup>

The first issue when creating a multivariate model is to decide which variables to include in the model. Starr and colleagues<sup>[35]</sup> note that “the complementary nature of the relationship between some explanatory variables can lead to compromised inferences if one or the other is innocently omitted from consideration ... [but] the indiscriminate inclusion of any and all variables that might just possibly be important ... [may also compromise inferences].” The variables to be included in the model depend on whether we are trying to adjust for group differences in the comparison of two treatments or we want to know which are the factors associated with an outcome. Since in this paper we are interested in the first case, there is no reason to avoid adjustment for a true covariate and there is little harm in adjusting for factors that were comparable before treatment in ways that are not relevant for the outcomes of interest. This should be limited when the sample size of the study and the categories of the covariate are large enough. However, if there are many covariates, each with some missing data, there may be few subjects with complete data and this may make the analysis more difficult than it would otherwise be. Although sometimes the covariates included are limited to those variables that are statistically significant, there is no reason to believe that the absence of statistical significance implies that the imbalance in the covariate is small enough to be ignored.<sup>[36]</sup> It is advisable to ‘force’ into the model those covariates considered to be the consolidated prognostic factors, even if they do not reach statistical significance in the sample.

As an illustration, we present in section 4 the results of a model where the outcome variable is the

change in reported clinical severity as measured with the CGI overall severity score over a 3-month period using data from the SOHO study (ordinary least squares [OLS] model). In this case, the linear model is specified as (equation 1):

$$\Delta CGI_{it} = x'_{i,t-1} \beta + \gamma D_{it} + \delta S_{it} + u_{it} \quad (\text{Eq. 1})$$

where  $\Delta CGI_{it}$  is the change of CGI from baseline and  $D_{it}$  is the treatment indicator, which is given a value of 1 when a patient is prescribed olanzapine and a value of 0 otherwise. The indicator variable  $S_{it}$  has a value of 1 after a patient has switched medication at 3 months, and for these patients we include two observations. The vector of confounders  $x$  contains information on a number of variables, including age, sex and country of residence of the patients and various co-morbidity measures (table I). For patients who do not switch medication, and for whom there is one observation in the sample,  $x_{i,t-1}$  are the confounders measured at baseline, i.e. the first visit. For those who switch medication,  $x_{i,t-1}$  measures the confounders at baseline for the first observation period. However, it also measures the confounders observed at the second visit for the second observation for this group of patients.

A limitation of this method is that if the treatment groups have different patient characteristics and are not comparable (i.e. lack overlap in covariate values), it will often go unnoticed. In a multiple regression model, it is relatively straightforward to examine whether individual covariate distributions differ among the treatment groups, but difficult to identify overall covariate imbalances. A second limitation is that the results from these analyses rely on the pre-specified functional form of the model (e.g. linearity). For example, when the covariates lack overlap in values, linear models rely on linear extrapolation to obtain treatment estimates, which may not be reliable.

Finally, while the regression method is useful to control for observed covariates, it is not helpful for

attempting to control for unobserved covariates (relevant covariates that have not been measured in the study), except to the extent that they are correlated with the observed covariates. For the model to be valid, we need to control for the relevant observed characteristics and assume that there are no other relevant characteristics which may influence the outcome and are related to the factor we want to analyse (in this case medication).

**Table I.** List of covariates used for adjustment in the models

| Covariate   |
|---|
| Age on presentation   |
| Age squared   |
| Age at first service contact for schizophrenia  |
| Number of suicide attempts 6 months prior to enrolment  |
| Body mass index at previous visit/baseline  |
| Weight (kg) at baseline   |
| Variable indicating whether the patient is using antipsychotics for the first time                      |
| Treatment received 6 months prior to baseline   |
| Receiving an antipsychotic upon presentation to the baseline visit                                      |
| Receiving mood stabilisers upon presentation  |
| Positive symptoms at baseline (CGI)   |
| Negative symptoms at baseline (CGI)   |
| Cognitive symptoms at baseline (CGI)  |
| Depressive symptoms at baseline (CGI)   |
| Overall symptom severity at baseline (CGI)  |
| Presence of extrapyramidal symptoms at baseline   |
| Tardive dyskinesia at baseline  |
| Loss of libido at baseline  |
| Presence of amenorrhoea or other menstrual disturbance at baseline                                      |
| Presence of impotence or sexual dysfunction at baseline   |
| Patient's compliance/adherence to prescribed antipsychotic therapy during the 4 weeks prior to baseline |
| Substance dependency and/or abuse 4 weeks prior to baseline   |
| Alcohol abuse or dependency 4 weeks prior to baseline   |
| Receiving monotherapy or combination treatment at baseline  |
| Psychiatrist indicators   |
| Employment status in the 4 weeks prior to recruitment   |
| Sex   |
| Housing status in the 4 weeks prior to recruitment  |
| Resource use 6 months prior to enrolment  |
| Score of dependent variable at baseline   |

CGI = Clinical Global Impression.

## 2.2 Propensity Score Matching

The propensity score matching approach is a two-stage approach. First, the method summarises the observed covariate information for each subject as a single score, the propensity score, which is the probability of being assigned to a particular treatment conditional on a set of observed covariates.<sup>[37]</sup> Once the propensity score is calculated, subjects are then matched or grouped into sub-classifications based on their score, and comparisons among subjects with a similar propensity score are then generated. The effects of the observed covariates on the outcome are then controlled for using the propensity score stratification. For this to be true, all covariates that affect both the treatment assignment and outcomes must be included in the propensity score model, and all subjects must have some non-zero probability of receiving each treatment. This is referred to as the 'strong ignorability assumption' and it ensures the independence of the treatment assignment and response variable within propensity score subclasses. A second assumption is that treatment assignment depends only on observed covariates. If these assumptions are met, conditional on the true value of the propensity score, the estimate of the treatment effect will be unbiased. In practice, if there is sufficient overlap between the characteristics of the treatment groups, up to 90% of the bias can be removed.<sup>[37,38]</sup>

Propensity score analysis is conceptually similar to multiple regression analysis, although if the two treatment groups are not comparable because there is no overlap in covariate values, the propensity score analysis will detect this at the first stage. Since calculation of the propensity score can use as many covariates as needed, the propensity score methodology may be less sensitive to model misspecification than multiple regression analysis, particularly when quadratic terms are omitted from the model.<sup>[39]</sup> In addition, propensity score matching methods assume no functional form specification for the rela-

tionship between the outcome variable and the observed covariates. The propensity score matching method is completely flexible; however, as with regression methods, a limitation of propensity score matching is that it controls for unobserved covariates, or hidden bias, only to the extent that they are correlated with the observed covariates.

In section 4 we present the results for a propensity score matching analysis to estimate the effect of olanzapine on the change in CGI.

## 2.3 Instrumental Variables

Instrumental variables (IV) estimation uses one or more *instruments* to mimic a randomisation of patients to different likelihoods of receiving alternative treatments. IV are observable factors that influence the choice of treatment but, in contrast to the variables used in the propensity score or multivariate regression, do not directly affect patient outcomes. When valid, the method is analogous to randomisation methods in identifying 'balanced' sources of variation in treatment, so that estimates of treatment effects are not contaminated by selection bias. The IV estimation is well known in econometrics but has been applied infrequently to estimate relationships between medical treatments and health outcomes.<sup>[40-42]</sup>

The IV method assumes that the proposed instruments are not correlated with unobserved differences in characteristics that directly affect outcomes; this absence of correlation is an assumption and its validity can be tested but cannot be proven. Additionally, because the groups being compared differ only in their likelihood of treatment, as opposed to a division into pure treatment and control groups, the method estimates an incremental or 'marginal' effect of treatment only over the range of variation in treatment across the IV groups. Hadley et al.,<sup>[43]</sup> motivated by the potential need to use observational data when making inferences about treatment outcomes when experimental data are not

available, compared two statistical approaches, OLS and IV regression analysis, to estimate the outcomes of three treatments (mastectomy, breast-conserving surgery with radiation therapy, and breast-conserving surgery only) for early stage breast cancer in elderly women. They concluded that no one type of treatment was superior because the OLS and IV point estimates of the differences in survival rates were of similar magnitude, as well as being similar to the estimates of survival difference in RCTs. Moreover, the IV method was challenged in technical terms because of its lack of statistical significance due primarily to the substantial inflation in the values of the standard errors, which is a result typically associated with the use of instrumental variables. These results were possibly due to the fact that the instruments were weak in the sense that they did not explain the endogenous treatment choices well (see, for example, Bound et al.<sup>[44]</sup>).

A potential and major limitation of the IV approach is that it may be difficult to identify an IV among the covariates collected. This is particularly relevant in health services research where the focus is on collecting information relative to potential confounders and not variables related solely to the treatment assignment. For example, Salkever et al.<sup>[42]</sup> applied an IV approach to the analysis of the effects of antipsychotic medication on hospitalisation. They concluded that the IV approach can yield different results from other methods, although it may be difficult to find appropriate instruments in many cases. We explored the existence of instrumental variables in the SOHO study, including variables related to both psychiatrists and patients. None of the psychiatrist variables were related to the medication choice, probably because of the stratified sampling used. On the other hand, patient variables were related to both medication choice and outcome and, thus, are not appropriate instruments.

Related to the IV approach is the sample selection approach of Heckman.<sup>[45]</sup> Instead of specifying

instruments for the endogenous treatment choice and estimating the parameters by two-stage least squares or generalised method of moments, the selection approach makes a parametric assumption about the joint distribution of the unobservables in the model for outcomes and the unobservables in the model for treatment choice. This joint distribution is assumed to be the normal distribution and a so-called 'control function' is added to the model for the outcomes that controls for the conditional expectation of the unobservables given treatment status. Although this model could in principle be estimated without any additional instrumental variables for the treatment choice due to the non-linearity of the control function, identification in that case is achieved purely by the distributional assumption. This means that the parameters are then not non-parametrically identified, and the performance of this estimator can be weak in that case (see, for example, Wooldridge<sup>[46]</sup>). The same issues as discussed in this section for IV also apply when it is difficult to find suitable instruments.

### 3. Repeated Measurements of Outcome

In this section, we discuss the implications for estimation and inference when there is more than one outcome observation for some patients included in SOHO (e.g. for those who change treatment at the 3-month visit). This situation is similar to prospective studies that assess treatment outcomes at various points in time after the baseline assessment. Since outcomes may vary at the different assessment points, the statistical model should include different outcome observations for each patient.

#### 3.1 Inference and Estimation in the Standard Models

Standard statistical methods are not appropriate for the analysis when more than one observation per patient is included. Even though OLS will result in a consistent estimate of the treatment effect under the

assumptions stated in section 2.1, inference based on standard OLS output will be misleading. This is because the estimated standard error will be a biased estimate of the true standard deviation of the treatment effect estimator, as it does not take account of the correlation in unobservables for patients with multiple observations. However, this can easily be remedied by calculating a standard error that is robust for general heteroskedasticity and correlation between repeated observations (see, for example, Lee<sup>[47]</sup>).

For propensity score matching, similar issues arise. The treatment effect will be estimated consistently, but inference has to take account of the correlation structure of the observations. A preferred method to obtain a standard error for the propensity score matching treatment effect estimator is bootstrap resampling. By resampling per patient, the correlation structure is left intact and standard errors thus obtained will be reliable.

For the multivariate regression model, efficiency of an estimator could be improved by taking account of the correlation structure explicitly. In the linear model, for example, efficiency could be improved by estimating random effects models using generalised least squares (GLS) type estimators. An alternative to random effects is a fixed effects procedure that estimates the unobserved patient heterogeneity for repeated observations explicitly by including patient-level dummy indicators. This would be important if unobserved heterogeneity is correlated with, for example, drug prescription, as this may bias the results.

In SOHO, there was an additional level, since besides several observations per patient, a psychiatrist includes several patients. There could possibly be correlation in the unobservables of the patients by psychiatrist. Correcting the simple OLS estimator by using robust measures, as described above, is now not as simple as before. In section 4 we present OLS estimation results for a model where we treat

the psychiatrist effects as fixed effects. This, therefore, allows for possible correlation between the unobserved characteristics of the psychiatrist and the treatment choice.

If there is no correlation between the psychiatrist effects and the other variables in the model, a more efficient way of estimating the parameters is to use the estimation techniques developed for hierarchical, multilevel models. This Bayesian hierarchical model is described and analysed in section 3.2.

### 3.2 Bayesian Hierarchical Model

Spiegelhalter et al.<sup>[48]</sup> and Jones<sup>[49]</sup> were the first to discuss the Bayesian approach in the comparison of health technologies. The most well known advantage of this approach is its dynamic nature, where previously available information is combined with the data to obtain a posterior distribution. This methodology is also promoted as a natural way to manage uncertainty about the parameters of interest, as they are considered as random variables.

Bayesian hierarchical models can be employed for adjusting for covariates when the data are hierarchical. This is frequently the case in economic evaluations, which are often conducted alongside multi-centre and multinational clinical trials, with patients grouped into centres and countries. It is widely recognised that there may be important differences between countries or centres in a range of clinical and economically relevant parameters, such as resource uses or unit costs.<sup>[50]</sup> The effects of centre or country should therefore be incorporated using appropriate analytical methods.

OLS models are the most widely used models in the literature to analyse the centre effects. These models include binary variables referring to hierarchies, such as the psychiatrist or the country, to measure the level effect, and we include these variables in our analysis. However, these models do not properly take into account the hierarchical structure of the data. Furthermore, centre-level or country-

level variables included in an OLS model are considered as if they were measured at a patient level, thus spuriously inflating the amount of information they supply. By contrast, hierarchical or multilevel models (MLMs) can incorporate the hierarchical structure of the data and provide more accurate estimates of patient- and centre-level effects. Several authors have recommended the use of MLMs in health economics to analyse the centre and country effect.<sup>[51-55]</sup>

Bayesian methods can implement MLMs with fewer statistical limitations and with a more natural interpretation than frequentist statistical methods. Bayesian Markov Chain Monte Carlo (MCMC) methods directly estimate the error terms of the hierarchical structure and are easily implemented with the software package WinBUGS.

We have used a Bayesian approach to estimate an MLM that evaluates differences between cohorts in CGI change (equation 2):

$$\Delta CGI_{ipct} = x'_{ipct,t-1} \beta + \gamma D_{ipct} + \delta S_{ipct} + \lambda_c + \varepsilon_{pc} + \eta_{ipc} + v_{ipct} \quad (\text{Eq. 2})$$

where we assume that:

$$v_{ipct} \sim N(0, \sigma_v^2), \eta_{ipc} \sim N(0, \sigma_\eta^2), \varepsilon_{pc} \sim N(0, \sigma_\varepsilon^2), \lambda_c \sim N(0, \sigma_\lambda^2)$$

and the covariances between the various error components are zero. Four levels are considered in this case: visit ( $t$ ), patient ( $i$ ), psychiatrist ( $p$ ) and country ( $c$ ). We include a tendency variable for the second visit assuming a linear growth ( $S_{ipct}$ ). For simplicity, we have assumed a hierarchical structure only in the intercept, which means that the values of  $\gamma$  and  $\delta$  do not vary among levels.

In a Bayesian approach we have to define a prior distribution for the parameters of the model. In the estimation results as presented in section 4 we use a multivariate normal prior distribution for the vector of coefficients,  $\beta$ ,  $\gamma$  and  $\delta$ , and an inverse  $\gamma$  distribution for the variance of the error terms ( $\sigma_v^2$ ,  $\sigma_\eta^2$ ,  $\sigma_\varepsilon^2$ ,  $\sigma_\lambda^2$ ). To facilitate the comparison with classical

models, we assume non-informative prior distributions.

#### 4. Estimation Results

Comparisons between olanzapine and other antipsychotic clinical outcomes as measured with the CGI overall severity score are presented using the statistical techniques described in sections 2.1 and 2.2, which should control for the effects of group differences. If the results are similar using the different methods, this reinforces the robustness of the findings.

To illustrate and compare results from multivariate OLS regression and propensity score matching, we present estimation results for the model of change in CGI as presented in section 2.1. This is for the first 3-month medication period, where two observations are included for patients who switch medication at the 3-month visit.

Table II shows the results of simply taking the differences in effectiveness between olanzapine users and non-olanzapine users. This effect is  $-0.264$ , which would be the estimate of the treatment effect in a RCT setting. In adjusting for confounders, table II next presents estimation results for the olanzapine treatment effect for the linear model estimated by OLS and the matching estimator. For the propensity score matching estimator we specified an Epanechnikov kernel with bandwidth 0.06. The estimated olanzapine treatment effect is similar using these two estimation procedures ( $-0.203$  in the OLS model and  $-0.197$  for the propensity score matching procedure) but is substantially smaller than the simple unadjusted average difference. Analysis of the estimated probabilities shows that there is a large common support, which means that we can readily find controls with a similar estimated probability to be prescribed olanzapine to that of olanzapine users themselves.

Table II also presents the estimation results of the linear model estimated by OLS that incorporates

**Table II.** Comparison of the results of the ordinary least square (OLS) regression model with the propensity score matching and the Bayesian multilevel models (MLMs) model when comparing clinical global impression (CGI) changes for each treatment group during the first 3 months of treatment. (Number of observations = 6791, number of patients = 6412, number of psychiatrists = 952.) Results from the SOHO (Schizophrenia Outpatient Health Outcomes) study<sup>[27,28]</sup>

| Method                                     | Difference | Standard error     | 95% CI         |
|--|------------|--------------------|----------------|
| <b>Simple differences (no controls)</b>    |            |                    |                |
| Olanzapine                                 | -0.264     | 0.024 <sup>a</sup> | -0.331, -0.216 |
| Other antipsychotics                       | 0          |                    |                |
| <b>OLS</b>                                 |            |                    |                |
| Olanzapine                                 | -0.203     | 0.022 <sup>a</sup> | -0.246, -0.161 |
| Other antipsychotics                       | 0          |                    |                |
| <b>Propensity score matching</b>           |            |                    |                |
| Olanzapine                                 | -0.197     | 0.025 <sup>b</sup> | -0.250, -0.124 |
| Other antipsychotics                       | 0          |                    |                |
| <b>OLS with fixed psychiatrist effects</b> |            |                    |                |
| Olanzapine                                 | -0.200     | 0.021 <sup>a</sup> | -0.240, -0.159 |
| Other antipsychotics                       | 0          |                    |                |
| <b>Bayesian MLMs</b>                       |            |                    |                |
| Olanzapine                                 | -0.187     | 0.021              | -0.227, -0.145 |
| Other antipsychotics                       | 0          |                    |                |

a Robust to general heteroskedasticity and within patient correlation.  
b Obtained using bootstrap resampling at the patient level.

psychiatrist indicator variables, thus estimating a fixed effects model as described in section 3.1. The estimated treatment effect (-0.200) is almost identical to the OLS estimate that does not take account of psychiatrist effects, indicating that there is no correlation between psychiatrists' unobservables and the treatment choice. Therefore, we proceeded to estimate the MLM by Bayesian MCMC. The estimated treatment effect (-0.187) was somewhat smaller than that estimated using the other methods (table II).

It is clear from this analysis that it is important to control for confounders in this observational setting; this reduces the estimated treatment effect. The upward bias (in absolute value) of the unadjusted average differences as an estimate of the treatment effect indicates that there is a correlation between treat-

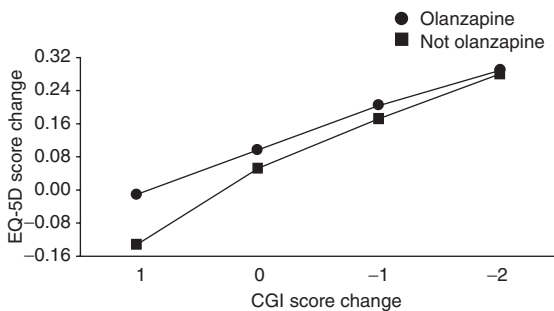
ment choice and patient characteristics. Not taking account of this would spuriously inflate the effectiveness of the treatment.

## 5. Methods to Assess Observer Bias

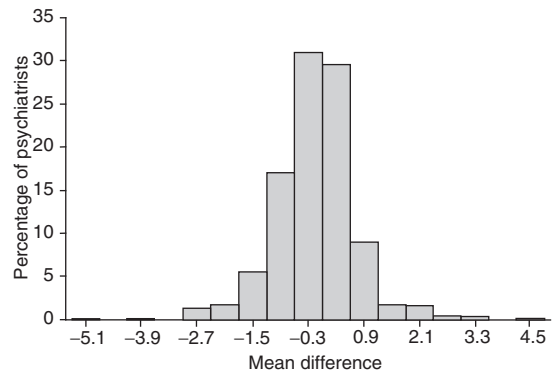
Observer bias occurs when the assessment of the outcome of interest is not the same for the treatment groups that are being compared. In psychiatry this is a common concern in most studies, as outcome measures are often subjectively assessed by evaluators. Observer bias can be especially problematic when the investigators may have an interest in showing the advantage of one treatment over the other, or when the research is funded by a sponsor with a vested interest in a particular outcome. In RCTs, observer bias is minimised by blinding the observer or investigator to the treatment the patient is receiving. However, blinding in observational studies is usually not feasible for logistic and cost reasons. When blinding is not a possibility, a way of avoiding observer bias is to focus on objective outcome measures, such as death or a biological marker. In psychiatry, the most commonly used objective measure of outcome is hospitalisation.<sup>[42,56]</sup> However, hospitalisation may be an applicable outcome only in studies of patients with severe disease, who are often readmitted. The SOHO study is a study of outpatients, with a low number of patients admitted to hospital. More importantly, the use of hospitalisation as an outcome measure may be influenced by many factors that vary enormously between country and region, such as the availability and organisation of services.<sup>[57]</sup> Therefore, outcome measures used are often subjective and it is important to address the issue of observer bias.

One way of assessing the extent to which observer bias may be present is to compare the effects of treatment as assessed by the clinicians with another measure that is not assessed by them. In SOHO, we can measure the presence and extent of assessment bias for each of the treatment groups by comparing

the clinical evaluation by the psychiatrists based on the total CGI score with the assessment made by the patients on their perceived quality of life, as measured by EQ-5D. Since patients are unaware of the hypothesis of the study and have no overt or hidden interest in any differential benefit between the treatments, we can assume that any reporting bias, if existent, would not differ among the treatment alternatives. If observer bias is present, we should find a discrepancy between the difference in EQ-5D and CGI scores by treatment. If there is a bias towards one of the treatments, we would expect that a reported improved CGI score will be accompanied by a smaller change in EQ-5D than for the other treatments. A worsening of the CGI score may then be accompanied by a larger decrease in the utility score. In SOHO, one would expect that observer bias, if present, would benefit the drug olanzapine, as it is manufactured by the study's sponsor. The graph of the average change in EQ-5D in utility score per treatment group per change in total CGI (figure 1) shows that there is no evidence of this bias. For example, when the change in CGI is -2 (an improvement), the average change in the utility score is larger for olanzapine than for the other drugs. Also, when the change in CGI is +1 (a deterioration), the change in utility score is less negative than for the other treatments. This means that a



**Fig. 1.** Comparison of the change in Clinical Global Impression (CGI) score, as assessed by the psychiatrist, with the EuroQol-5 Dimensions (EQ-5D) utility score, as assessed by the patient, for the two treatment cohorts. Results from the SOHO (Schizophrenia Outpatient Health Outcomes) study.<sup>[27,28]</sup>

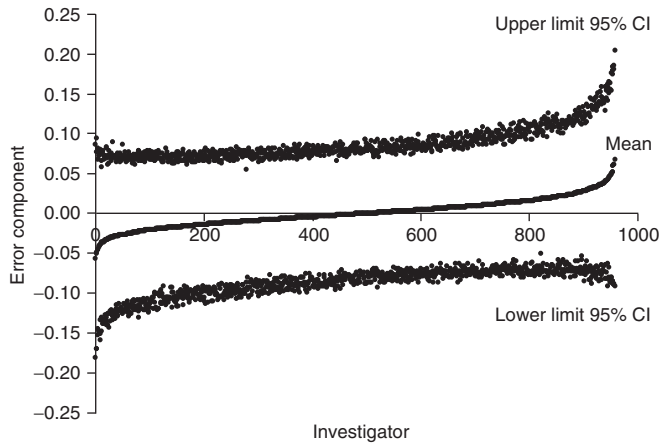


**Fig. 2.** Distribution of mean differences in Clinical Global Impression between treatment groups, by psychiatrist. Results from the SOHO (Schizophrenia Outpatient Health Outcomes) study.<sup>[27]</sup>

positive change in CGI score as assessed by the clinician is accompanied by a smaller self-reported decline in EQ-5D for olanzapine patients than for non-olanzapine patients, whereas a better, negative change in CGI score is accompanied by a larger self-reported increase in EQ-5D for olanzapine patients.

These descriptive results were confirmed with a multiple regression analysis that looked at the relationship of changes in reported EQ-5D with changes in reported CGI scores and showed that the comparison group (no olanzapine) had smaller mean changes in utility score at similar CGI changes than the olanzapine group (estimate -0.034, 95% CI -0.05, -0.02), an indication that there seems to be no evidence of observer bias in favour of olanzapine.

For differences in outcomes between treatment groups, we can also examine whether these are distributed homogeneously between investigators or whether they are accounted for by a limited number of investigators, who would be driving an observer bias. Using data from the SOHO study, we have plotted the distribution of the mean difference in outcomes between the treatment groups, measured as the change in CGI ratings from baseline to 3 months, for each investigator (figure 2). This descriptive analysis shows that the differences between treatment cohorts are approximately normally distributed among investigators, which reinforces



**Fig. 3.** Error component for each investigator. The error components are the discrepancies between the mean incremental effectiveness and the incremental effectiveness for each investigator. Investigators are ordered by the size of error component. Mean and 95% credibility interval (CI; confidence interval) are shown. Results from the SOHO (Schizophrenia Outpatient Health Outcomes) study.<sup>[27]</sup>

the hypothesis that there are no investigator-related systematic group differences. If the assessments of some investigators were dominating the resulting difference between the two treatment groups, we would expect to find a bimodal or skewed distribution.

Finally, Bayesian hierarchical methods can also be used to create a model to estimate the effect of the investigator on the results of the evaluation. The Bayesian method discussed in section 3.2 included a random term for the psychiatrist in the intercept. That term reflects the fact that patients with the same characteristics, and receiving the same treatment, can achieve different levels of effectiveness depending on the psychiatrist who evaluates them. In this previous model, the coefficient  $\gamma$  reflected the incremental effectiveness of the treatment versus the control and was considered fixed for all the psychiatrists. We can study the existence of observer bias by including a random term ( $\varepsilon_{2,pc}$ ) at investigator level in this treatment effect (equation 3):

$$\gamma_p = \gamma + \varepsilon_{2,pc} \tag{Eq. 3}$$

where

$$(\varepsilon_{pc}, \varepsilon_{2,pc}) \sim N\left(0, \Sigma_\varepsilon = \begin{pmatrix} \sigma_\varepsilon^2 & \sigma_{\varepsilon,\varepsilon_2} \\ \sigma_{\varepsilon,\varepsilon_2} & \sigma_{\varepsilon_2}^2 \end{pmatrix}\right)$$

We can estimate the discrepancy between the mean incremental effectiveness  $\gamma$  and the incremental effectiveness for each investigator, after adjusting for covariates, through the term  $\varepsilon_{2,pc}$ . Observer bias occurs when the assessment of the effectiveness for the psychiatrist is not the same for the treatment groups that are being compared. In the presence of observer bias, it is unlikely that the bias will be the same for all the psychiatrists. The absence of significant discrepancies between investigators would support the idea that there is no observer bias. It is important to note that differences in the assessment of the outcome between investigators would not confirm the existence of observer bias because they can be due to some characteristics of the investigator that are not considered in the model. Figure 3 shows the mean and 95% credibility interval (confidence interval) of the estimated discrepancy from the mean

effect for each investigator ( $\hat{\varepsilon}_{2,pc}, p = 1, \dots, 935$ ). The random terms for each psychiatrist are ordered from the lowest score to the highest score. The overlap of the lowest and the highest credibility interval sup-

ports the hypothesis of no differences between investigators, confirming the absence of observer bias.

## 6. Conclusions

The richness and diversity of observational data for the analysis of treatment effects pose a number of design and analytical challenges. First, those who appear in a treatment group in an observational study are likely to be systematically different from those who are in the other, comparison group. We have discussed a number of methods to estimate treatment effects in the presence of selection bias. Our study, using data from SOHO, demonstrates that not controlling for selection bias might spuriously inflate estimated treatment effects. We recommend that extra care is taken at the design phase of an observational study to ensure that the most relevant information is collected during the study in order to be able to control for selection bias. The absence of evaluation of relevant predictors may lead to the impossibility to control for important selection biases.

Secondly, the clustering of multiple observations within subjects over time must be taken into account when each patient has multiple observations. Standard statistical techniques in this case will lead to wrong inferences and perhaps even inconsistent estimates. We have discussed a number of techniques that will allow a researcher to take into account multiple observations of patients over time as well as multiple observations per psychiatrist. In the event that there is no correlation between psychiatrist/patient effects and the other variables in the model, in particular the treatment choice, we present a way of estimating treatment effect using techniques developed for hierarchical, multilevel models.

Assessment of outcomes may be subject to observer bias. We propose descriptive as well as analytical techniques that can be used to assess the presence of observer bias. Use of some of the tech-

niques proposed in this paper, such as the incorporation of an additional measure of outcome completed by the patients, will help assess the presence of observer bias.

This paper has presented a number of techniques to strengthen the armament of the researcher-scientist to be able to perform high quality and scientifically robust analyses using observational data, which will complement the information generated from randomised experiments. Using some of these strategies will allow the production of valid results from observational studies.

## Acknowledgements

Frank Windmeijer received monetary compensation from Eli Lilly and Company for econometric advice. Josep Maria Haro received monetary compensation from Eli Lilly and Company for his participation in the SOHO Advisory Board. David Suarez is providing statistical consultancy work for Lilly. Stathis Kontodimas and Mark Ratcliffe are Eli Lilly and Company employees.

## References

1. Hofer A, Hummer M, Huber R, et al. Selection bias in clinical trials with antipsychotics. *J Clin Psychopharmacol* 2000; 20: 699-702
2. Wells KB. Treatment research at the crossroads: the scientific interface of clinical trials and effectiveness research. *Am J Psychiatry* 1999; 156: 5-10
3. Rubin DB. Estimating causal effects from large data sets using propensity scores. *Ann Intern Med* 1997; 127 (8 Pt 2): 757-63
4. Concato J, Shah N, Horwitz RI. Randomised, controlled trials, observational studies, and the hierarchy of research designs. *N Engl J Med* 2000 Jun 22; 342 (25): 1887-92
5. Benson K, Hartz AJ. A comparison of observational studies and randomised, controlled trials. *N Engl J Med* 2000 Jun 22; 342 (25): 1878-86
6. Black N. Why we need observational studies to evaluate the effectiveness of healthcare. *BMJ* 1996 May 11; 312 (7040): 1215-8
7. Black N. What observational studies can offer decision makers. *Horm Res.* 1999; 51 Suppl. 1: 44-9
8. Britton A, McPherson K, McKee M, et al. Choosing between randomised and non-randomised studies: a systematic review. *Health Technol Assess* 1998; 2 (13): 1-iv, 1-124
9. Pocock SJ, Elbourne DR. Randomised trials or observational tribulations? *N Engl J Med* 2000 Jun 22; 342 (25): 1907-9
10. Friedman HS. Observational studies and randomised trials [letter]. *N Engl J Med* 2000 Oct 19; 343 (16): 1195-6

11. Kunz R, Khan KS, Neumayer HH. Observational studies and randomised trials [letter]. *N Engl J Med* 2000 Oct 19; 343 (16): 1194-5
12. Sacks HS. Observational studies and randomised trials [letter]. *N Engl J Med* 2000 Oct 19; 343 (16): 1195
13. Liu PY, Anderson G, Crowley JJ. Observational studies and randomised trials [letter]. *N Engl J Med* 2000 Oct 19; 343 (16): 1195
14. Smith RP, Meier P. Observational studies and randomised trials [letter]. *N Engl J Med* 2000 Oct 19; 343 (16): 1196
15. Hlatky MA, Califf RM, Harrell Jr FE, et al. Comparison of predictions based on observational data with the results of randomised controlled clinical trials of coronary artery bypass surgery. *J Am Coll Cardiol* 1988; 11: 237-45
16. Kunz R, Oxman AD. The unpredictability paradox: review of empirical comparisons of randomised and non-randomised clinical trials. *BMJ* 1998; 317: 1185-90
17. Wolfe RA. Observational studies are just as effective as randomised clinical trials. *Blood Purif* 2000; 18 (4): 323-6
18. Greene T. Are observational studies 'just as effective' as randomised clinical trials? *Blood Purif* 2000; 18 (4): 317-22
19. Vineis P. Proof in observational medicine. *J Epidemiol Commun Health* 1997; 51: 9-13
20. Holmberg L, Baum M, Adami HO. On the scientific inference from clinical trials. *J Eval Clin Pract* 1999 May; 5 (2): 157-62
21. Wilson GT. The clinical utility of randomised controlled trials. *Int J Eat Disord* 1998 Jul; 24 (1): 13-29
22. Ellenberg JH. Selection bias in observational and experimental studies. *Stat Med* 1994 Mar 15-Apr 15; 13 (5-7): 557-67
23. Kemler MA, de Vet HC. Does randomisation introduce bias in unblinded trials? *Epidemiology* 2000 Mar; 11 (2): 228
24. Rabeneck L, Viscoli CM, Horwitz RI. Problems in the conduct and analysis of randomised clinical trials. Are we getting the right answers to the wrong questions? *Arch Intern Med* 1992 Mar; 152 (3): 507-12
25. Caplan LR. Is the promise of randomised control trials ("evidence-based medicine") overstated? *Curr Neurol Neurosci Rep* 2002; 2: 1-8
26. Haro JM, Edgell ET, Jones PB, et al. The European Schizophrenia Outpatient Health Outcome (SOHO) study: rationale, methods and recruitment. *Acta Psychiatr Scand* 2003; 107: 222-32
27. Haro JM, Edgell ET, Frewer P, et al. The European Schizophrenia Outpatient Health Outcomes (SOHO) study: baseline findings across country and treatment. *Acta Psychiatr Scand Suppl* 2003; 416: 7-15
28. Haro JM, Edgell ET, Novick D, et al. Effectiveness of antipsychotic treatment for schizophrenia: 6-month results of the pan-European Schizophrenia Outpatient Health Outcomes (SOHO) study. *Acta Psychiatr Scand* 2005; 111: 220-31
29. Haro JM, Kamath SA, Ochoa S, et al. The clinical global impression-schizophrenia (CGI-SCH) scale: a simple instrument to measure the diversity of symptoms present in schizophrenia. *Acta Psychiatr Scand Suppl* 2003; 416: 16-23
30. Guy W. Clinical global impression. In: Guy W, editor. *ECDEU assessment manual for psychopharmacology*, revised. Rockville (MD): National Institute of Mental Health, 1976: 217-22
31. Williams A. EuroQol: a new facility for the measurement of health-related quality of life. *Health Policy* 1990; 16: 199-208
32. Kind P, Hardman G, Macran S. UK population norms for EQ-5D. York Centre for Health Economics Discussion Paper, 172. York: Centre for Health Economics, University of York, 1999
33. Lambert M, Haro JM, Novick D, et al. Olanzapine vs. other antipsychotics in actual out-patient settings: six months tolerability results from the European Schizophrenia Out-patient Health Outcomes study. *Acta Psychiatr Scand* 2005; 111: 232-43
34. Rosenbaum PR. *Observational studies*. 2nd ed. New York: Springer-Verlag, 2002
35. Starr TB, Dalcorso RD, Levine RJ. Fertility of workers: a comparison of logistic regression and indirect standardization. *Am J Epidemiol* 1986; 123: 490-8
36. Altman DG. Comparability of randomised groups. *Statistician* 1985; 34: 125-36
37. Rubin DB, Thomas N. Matching using estimated propensity scores: relating theory to practice. *Biometrics* 1996; 52: 249-64
38. Cochran WG. The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics* 1968; 24: 295-313
39. Drake C. Effects of misspecifications of the propensity score on estimators of treatment effects. *Biometrics* 1993; 49: 1231-6
40. Hylan TR, Crown WH, Meneades L, et al. Tricyclic antidepressant and selective serotonin reuptake inhibitors antidepressant selection and health care costs in the naturalistic setting: a multivariate analysis. *J Affect Disord* 1998; 47: 71-9
41. Lu M. The productivity of mental health care: an instrumental variable approach. *J Ment Health Policy Econ* 1999; 2: 59-71
42. Salkever DS, Slade EP, Karakus M, et al. Estimation of antipsychotic effects of hospitalisation risk in a naturalistic study with selection of unobservables. *J Nerv Ment Dis* 2004; 192: 119-28
43. Hadley J, Polsky D, Mandelblatt JS, et al. An exploratory instrumental variable analysis of the outcomes of localized breast cancer treatments in a Medicare population. *Health Econ* 2003; 12: 171-86
44. Bound J, Jaeger DA, Baker RM. Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. *J Am Stat Assoc* 1995; 90: 443-50
45. Heckman JJ. Sample specification bias as a specification error. *Econometrica* 1979; 47: 153-61
46. Wooldridge JM. *Econometric analysis of cross section and panel data*. Cambridge (MA): MIT Press, 2002
47. Lee M-J. *Panel data econometrics: method-of-moments and limited dependent variables*. San Diego (CA): Academic Press, 2002
48. Spiegelhalter DJ, Feedman LS, Parmar MKB. Bayesian approaches to randomized trials (with discussion). *J R Stat Soc Ser A Stat Soc* 1994; 157: 357-416
49. Jones DA. A Bayesian approach to the economic evaluation of health care technologies. In: Spiker B, editor. *Quality of life and pharmacoconomics in clinical trials*. 2nd ed. Philadelphia (PA): Lippincott-Raven, 1996: 1189-96

- 
50. Sculpher MJ, Pang FS, Manca A, et al. Generalisability in economic evaluation studies in health care: a review and case-studies. *Health Technol Assess* 2004; 8: 1-206
  51. Rice N, Leyland A. Multilevel models: applications to health data. *J Health Serv Res Policy* 1996; 1: 154-64
  52. Leyland AH, Goldstein H. *Multilevel modelling of health statistics*. Chichester: Willey, 2001
  53. Goldstein H, Browne W, Rasbash J. Multilevel modelling of medical data. *Stat Med* 2002; 21: 3291-315
  54. Manca A, Rice N, Sculpher MJ, et al. Assessing generalisability by location in trial-based cost-effectiveness analysis: the use of multilevel models [published erratum appears in *Health Econ* 2005; 14 (5): 486]. *Health Econ* 2005; 14: 47-85
  55. Grieve R, Nixon R, Thompson SG, et al. Using multilevel models for assessing the variability of multinational resource use and cost data. *Health Econ* 2005; 14: 185-96
  56. Chien CF, Steinwachs DM, Lehman A, et al. Provider continuity and outcomes of care for persons with schizophrenia. *Ment Health Serv Res* 2000; 2: 201-11
  57. Haro JM, Salvador-Carulla L, Cabases J, et al. Utilisation of mental health services and costs of patients with schizophrenia in three areas of Spain. *Br J Psychiatry* 1998; 173: 334-40
- 

Correspondence and offprints: *Josep Maria Haro*, Sant Joan de Déu-SSM, Fundació Sant Joan de Déu, Dr. Antoni Pujades, 42, 08830 – Sant Boi de Llobregat (Barcelona), Spain.

E-mail: [jmharo@fsjd.org](mailto:jmharo@fsjd.org)