Statistical and econometric methods for the evaluation of cost-effectiveness using observational data

Dimitrios Rovithis[1], Stavros Petrou[1,2], Borislava Mihaylova[2]

1. National Perinatal Epidemiology Unit, University of Oxford, Old Road Campus, Headington, Oxford OX3 7LF | E-mail: dimitrios.rovithis@npeu.ox.ac.uk
2. Health Economics Research Centre, University of Oxford

Presented at the Health Economists' Study Group Meeting

London School of Economics and Political Science, January 2010

Work in progress – not to be cited without authors' permission

December 2009

Abstract

Aims: The use of observational clinical and economic data for health economic evaluations can potentially produce biased results. A number of different analytical methods that deal with this problem have been proposed in the biomedical and economics fields. The aim of this paper is twofold: first, to identify the analytical methods currently employed to adjust for confounding in treatment effects in the health economic evaluation literature; second, to highlight the available approaches in both disciplines that can be used to evaluate cost-effectiveness of health care interventions using observational data.

Methods: A search of online bibliographic databases complemented with review of the bibliography of selected studies. Relevant studies were appraised using a pre-designed review template.

Results: Despite the availability of different analytical methods both in the health economics and the biomedical literature, preliminary results of the review revealed that so far, their use for evaluating cost-effectiveness has been limited, with the majority of identified economic evaluations focusing only on observed sources of bias.

Conclusions: In the absence of comparative work, the interpretation of the results obtained from most studies identified by this review can be deemed controversial. The use of some available econometric methods for evaluating cost-effectiveness has yet to be studied and show potential for future research.

Keywords: selection bias, confounding, cost-effectiveness, observational data

JEL classification: I10, I19, C01

Introduction

Many health economic evaluations are based to some extent on observational clinical and/or economic evidence. Because this evidence originates from non-randomised comparison, health economic studies using such evidence might be biased (Jones 2009). Confounding, arising from selection bias in particular, constitutes a major threat to the internal validity of a study (Hennekens, Buring, and Mayrent 1987) and unless its presence can be minimised, the estimated effects do not necessarily imply a cause and effect relationship (Hoppe et al. 2009).

The purpose of this paper is twofold: First, to present a brief overview of the methods used in the biomedical and economic literature to adjust for confounding in treatment effects when using observational evidence. Second, to complement and extend recent reviews published elsewhere, such as Heckman (2008), Blundell and Costa Dias (2008), Jones (2009), Imbens and Wooldridge (2009), and Jones and Rice (2009), by identifying which of these methods are currently employed for confounding adjustment in the health economic evaluation literature and elucidating issues surrounding their use.

The evaluation problem

Over the years, a number of methods from statistics and econometrics that deal with confounding arising from selection bias have been employed, depending on whether the source of bias is observed or not (Imbens and Wooldridge 2009). The key idea behind these methods is the construction of the counterfactual outcome when the evaluation problem is the measurement of a treatment effect, broady defined as the effect of a health care intervention, policy or programme on an outcome variable (Blundell and Costa Dias 2008). Construction of the counterfactual, that is how the outcome differs under different situations, is seen as the fundamental problem that these methods seek to address, since when evaluating the impact of an intervention each patient can either be in the treatment group or the control group, but never in both (Jones and Rice 2009). Central to the evaluation problem is also the decision relating to estimating a single average effect or looking into the heterogeneity in patients' responses to the technology or policy in question. Evaluation methods usually measure average treatment effects (ATE) across the study population (Imbens 2004; Jones 2009).

FIGURE 1 HERE

In experimental studies, assignment to treatment is typically random, independent of covariates and potential outcomes (Collins and MacMahon 2001). In contrast, in observational studies treatment assignment is based on non-random selection which can lead to selection bias (MacMahon and Collins 2001). Figure 1 illustrates selection in observational studies; non-random selection can be dependent on observable (selection on observables) and fully or partially unobservable (selection on unobservables) covariates (Jones and Rice 2009). The result of selection bias is

that the causual relationship between treatment and potential outcome cannot be directly observed as it is confounded by known and/or unknown factors (Imbens and Wooldridge 2009). Different evaluation methods can provide us with some information as to what would have happened to the treated patients had they not been treated (and vice versa), even in the absence of randomisation.

Overview of evaluation methods

The use of the appropriate evaluation method will ultimately depend on a number of factors including the research question, the nature and overall quality of the available data, and the way that patients are assigned to treatment (Blundell and Costa Dias 2008). The focus here is on analytical methods commonly used in the analysis of effects of treatments in cost-effectiveness analysis using observational data.

Selection on observables

Regression analysis

Regression analysis evaluates the relationships between two or more variables by quantifying the level of change in a dependent variable (the outcome) resulting from a given level of change in an independent variable (the predictor). An advantage of this form of analysis is that it allows assessment of outcomes (eg. costs and effects) and treatment effects, while at the same time it takes into consideration the influence of potential confounders (Jones 2009). In regression models each coefficient represents an estimate of effect of the predictor in question on the outcome under the statistical model used. The choice of the appropriate model usually depends on the available data, as well as the outcome under evaluation (Imbens 2004).

Matching

The aim of matching is to synthetically balance confounding characteristics in the absence of randomisation (Heckman, Ichimura, and Todd 1998). Matching in its various manifestations achieves this by creating treatment groups which share the same characteristics, first by identifying observable confounding factors, and then pairing patients in the treatment groups according to these confounders. Patients can be paired on a range of confounding variables including personal characteristics or environmental factors.

Stratification

Stratification identifies confounding factors, but in contrast with matching, patients are placed in groups (strata) according to the same level of the confounder (Kirkwood and Sterne 2003). The analysis is then carried out in each stratum within which the confounding variable remains constant and summarised across the stratas.

Propensity scores

Propensity score analysis combines the strengths of different analytical methods into a two-step approach. First, the propensity score is calculated, that is the probability of a patient being assigned to the intervention group, instead of being in the controls, conditional on a set of observed patient covariates (Rosenbaum and Robin 1983). As the propensity score is a probabilityranging from 0 to 1, it is usually estimated by means of logistic regression with the observed covariates being the predictors and receipt of treatment being the dependent variable (Luellen, Shadish, and Clark 2005). Once the propensity score is calculated for each patient, it can be used to adjust for differences either through study design (comparison of effects across groups following different types of matching), or within the analysis by estimation of the treatment effect using stratification or regression (Morshed, Tornetta, and Bhandari 2009).

Selection on unobservables

Instrumental variables

This is a two-stage estimation approach that aims to imitate randomisation of patients by using variables commonly refered to as instruments, which should have a strong effect on treatment assignment but should not be correlated with the outcome (Newhouse and McClellan 1998). More specifically, the estimator uses the instrument(s) to predict the value of a potentially endogenous independent variable (i.e. treatment allocation). These predicted values are subsequently used as a covariate in the outcome model to estimate the magnitude of the effect on the outcome (Angrist, Imbens, and Rubin 1996). Whether this estimation approach gives unbiased estimates depends on the extent to which the instrument used is appropriate and valid (Bound, Jaeger, and Baker 1995).

Difference-in-Differences

Difference-in-differences estimation uses a pre-treatment and a post-treatment group to facilitate comparison between groups by measuring the difference in the outcome for the treated group versus the non-treated (Buckley and Shang 2003). As such, the implementation of the difference-in-differences estimator depends on the availablity of data in both the pre-treatment and the post-treatement period for the treated and non-treated groups. In addition, two important assumptions are required to hold true; first, that the structure of the data is time-invariant, and second that both the treated and untreated groups present a common time trend (Jones and Rice 2009).

Control functions

This is also a two-stage aproach that is used to adjust for biases that arise as a result of selection. This is achieved by incorporating the assignment mechanism in the estimation process, essentially treating it as an omitted variable problem (Navarro 2008). The estimator first uses the joint distribution of the assignment mechanism and treatment to obtain a control function. It is then used to account for endogenous selection in an outcome equation (Blundell and Costa Dias 2008). This method is seen as a variation of the Heckman correction estimator (Heckman 1979).

Regression discontinuity

Regression discontinuity is a method that aims to estimate treatment effects in non-experimental situations where treatment assignment is determined by whether an exogenous variable (the forcing variable) surpasses a threshold (Imbens and Lemieux 2008). The key idea behind this approach is that the comparison of patients with values slightly below the threshold, to the patients slightly above it, is used to identify the treatment effect; as long as the forcing variable is not manipulated, treatment variation near the threshold is thought to be similar to a randomised experiment (Lee and Lemieux 2009).

Review of the literature

Review scope

A structured review of the international English language literature was undertaken in order to identify the currently employed methods for adjustment for confounding arising from selection bias in the health economic evaluation literature . More specifically, the aim was to identify and review health economic evaluation studies employing methods for confounding adjustment, the type of observational data that they use and their overall conclusions.

Literature reviews have sometimes been criticised for being non-transparent with respect to the review methods and selection criteria that they employ, ultimately resulting in biased conclusions (Centre for Reviews and Dissemination 2009). It is of paramount importance therefore that any literature review is carried out in a rigorous mannerThis review was carried out in a systematic way from a methodological point of view seeking to highlight studies that use different methods (and thus broadening the scope of the review) rather than including all applications. As such, it embodies considerable characteristics of a systematic design, in an effort to render the reviewing process more transparent and reproducible. Effort was made to identify as many relevant papers as possible, but as the scope of the review was methodological, this study should not be considered as an exhaustive investigation of the applied literature. The grey literature, that is theses, conference proceedings, internal reports, and non-indexed journals in electronic or print format not controlled by commercial publishing, was not examined systematically.

Search strategy

The studies included in the review were identified using a three-stage process. The first stage involved the search of three generic electronic bibliographic databases namely Ovid Medline, Ovid EMBASE, and Ovid EconLit in order to generate as many papers of potential methodological interest as possible for the years 1989-2009. Because of the nature of health economic evaluation studies, the search strategy at the early stages of the review was designed for Medline and was then applied to the other two databases.

FIGURE 2 HERE

However, there were a number of problems with this approach, mostly relating to the trade-off between the number of appropriate articles identified (i.e. sensitivity), and the number of inappropriate articles (i.e. specificity) eliminated by the search strategy. The search terms used proved very broad, revealing a very large number of studies from epidemiology, medicine, economics and other disciplines, the majority of which were not health economic evaluations. In addition, Medline's categorisation system that allows selection of Medical Sub-Headings (MeSH) terms to aid identification of relevant studies didn't limit the results obtained. Another problem was that most of these results focused on studies that used methods traditionally employed in the medical literature but not methods widely employed in the economics literature. A number of search strategies were subsequently tested and the final search strategy reflects a pragmatic approach, which balances sensitivity and specificity.

Next, the search of two specialised databases namely NHS EED and OHE HEED was carried out. The advantage of searching these databases is that they only include health economic evaluations, which are identified from a wide range of journals in the medical and economic literature. For this reason, the search omitted terms such as "cost" and "effectiveness", which had been used in the main search strategy.

Having identified studies from the above electronic databases, in the second stage the samples were combined and de-duplicated using the bibliographic management software EndNote. Next, all studies underwent thorough screening to ensure that they meet the following criteria:

- Only full health economic evaluation studies using individual participant observational data both for costs and effects were included.

- The sole focus of this review was studies employing statistical and econometric methods that account for confounding arising from selection bias.

- Analytical methods that dealt with issues relevant to the analysis of randomised clinical trials, modelling based studies, etc. were considered beyond the scope of this review and were not included.

The third and final stage aimed to supplement the electronic database searching, by reviewing the bibliographies and the citations of the relevant studies identified by the

search strategy. The additional papers were included in the review sample only if they met the selection criteria laid out above. It should be noted that when it was apparent from the downloaded title or abstract that a study failed on one of these criteria, it was excluded. When it was unclear or if any doubt remained, the full paper was downloaded and examined.

Review process

The principal aim was to present a structured review together with a critical appraisal of the method(s) employed, data used and overall conclusions of the identified studies. One of the authors (DR) reviewed all papers using a checklist of questions (i.e. structured review template) and the remaining authors independently reviewed individual papers.

Bibliographic information for each study and bibliographic connections to other papers were recorded, along with the type of study, the outcome measure used and the country for which the analysis was pertinent. The source of funding for each study was also extracted and documented as eitherthe private sector engaged in the production of medical technologies (industry), the public sector such as governmental agencies, educational institutions, non-governmental organisations and charitable foundations (non-industry), or funding source not clarified (unclear). In addition, the disease and the technologies used were extracted and a categorization by type of technology was included, as well as the health outcomes of each study, the type of the data and the datasets used. Furthermore, abstracted information included the methods used to adjust for confounding, information on whether adjustment was undertaken for costs, effects or summary cost-effectiveness measures (eg. net benefit), whether comparisons with other methods and/or sources of data took place and the estimation methods and software used. Finally, the methods employed to quantify uncertainty and the conclusions as stated by the author(s) were recorded, along with the reviewers' assessment on the ability of the method to provide unbiased cost-effectiveness estimates.

After reviewing the studies using the template described above, the recorded information for each study was stored in EndNote in the form of keywords. This approach has considerable advantages. For example, the recorded information is essentially indexed in a searchable electronic catalogue that can be used to generate a wealth of descriptive statistics. It can also allow the retrieval of identified papers that are relevant to more than one category, for instance a particular disease or type of technology, a particular method, or simply by country.

Review results

Overall, the search strategy yielded approximately 6300 hits in the three generic databases. The identification and assessment of relevant studies is still in progress and thus more papers are likely to be included in the review. Nevertheless, the preliminary results of the review are presented below and reveal that the application of the analytical methods traditionally used for the evaluation of treatment effects, has been very limited in the evaluation of cost-effectiveness. So far, twelve relevant studies have been identified and reviewed (Table 1). Similar proportion of studies were developed based on cohort studies and routinely collected data. The majority of these health economic evaluation studies that have so far employed confounding

adjustment, have focused on multivariable regression models, propensity score analysis, matching, stratification, or combinations of these methods, which only account for observed sources of bias. Adjustment in cost-effectiveness studies when confounding might be due to factors that are unobservable to the investigator still remains a largely undeveloped area. The methods and results of the studies reviewed so far are summarised in Table 2 and discussed in detail below.

TABLE 1 HERE

TABLE 2 HERE

Regression analysis is one of the most frequently employed methods. Windmeijer et al. (Windmeijer et al. 2006) used multivariate regression analysis to adjust for baseline covariates in an attempt to assess the cost-effectiveness of treatments in the Schizophrenia Outpatient Health Outcomes (SOHO) study, a 3-year prospective observational study of the outcomes of antipsychotic treatment for schizophrenia in the outpatient setting. In addition, Knapp et al. (Knapp et al. 2008) employed regression analyses that adjusted for baseline covariates to estimate the incremental cost-utility for patients treated with olanzapine compared with other antipsychotics in the SOHO study. Griffin et al. (2007) conducted a prospective observational study employing data from the appropriateness of coronary revascularisation (ACRE) cohort, combined with other sources such as hospital case notes and general practitioners' and patients' questionnaires among others, to assess the cost-effectiveness of coronary interventions. In their analysis, the authors used matching on appropriateness for treatment and multivariate regression to adjust for potential confounding effects in baseline clinical and demographic characteristics. More recently, DeRidder et al. (2009) evaluated the cost-effectiveness of olanzapine and risperidone for the treatment of schizophrenia in Belgium employing data from a prospective, observational, non-randomized study. In their analysis, DeRidder et al. used a net-benefit regression to control for several patient characteristics resulting in baseline differences between treatment groups. The authors note that after covariate adjustment the difference between risperidone and olanzapine increased, but this increase was not statistically significant.

The hybrid method of **propensity score analysis** appears to increasingly gain popularity. Weiss et al. (2002) used the Medicare Provider Analysis and Review File (MEDPAR) to assess the cost-effectiveness of the implantable defibrillator in a population-based cohort of patients. In their analysis, a multivariable propensity score matching for patient and hospital characteristics was used to match pairs of patients, in which one patient received a defibrillator and the other did not. The comparison of mortality and costs in these matched pairs suggested that the intervention may be still economically acceptable in real-life use (not only in clinical trials). Similarly, Mojtabai and Zivin (2003) used data from the Services Research Outcomes Study (SROS), a survey of 3,047 clients in a random sample of 99 drug treatment facilities across the United States, to evaluate the cost-effectiveness of four treatment modalities for substance abuse. In their analysis, the investigators employed propensity score stratification to control for factors that influence selection into various modalities, such as type and severity of substance disorder and psychiatric co-morbidity. Notably, the results of stratified comparisons of cost-effectiveness ratios were similar to the non-stratified cost-effectiveness analysis findings.

Mitra and Indurkhya (2005) combined data from MEDPAR and the linked Surveillance Epidemiology and End Results (SEER) population based cancer registries to estimate the propensity of patients to receive surgical treatment of their bladder cancer based on their background covariates. They then estimated the net monetary benefit using a regression-based approach that allows modelling the cost-effectiveness while at the same time making adjustment for the propensity of treatment. Their results indicated that balance can be achieved to a significant degree by propensity adjustment and they concluded that propensity score analysis, despite its limitations can yield less biased and more precise results. Another study by Merito and Pezzoti (2006) assessed the costs and effectiveness of starting highly active antiretroviral therapy at different points during the course of HIV infection based on data obtained from the Italian Cohort Naive Antiretrovirals (ICONA), an observational study that recruited HIV-infected adults naive to antiretroviral therapy. In their analysis, Merito and Pezzoti estimated average treatment effects on disease progression and costs of different therapeutic approaches using propensity scores in order to account for selection bias (using separate logit models for treatment assignment). Their results showed that the logit models for propensity of treatment fit reasonably well. Propensity score distributions by strata also confirmed that the balance between comparison groups was largely achieved and the balancing property was satisfied both for patient demographic and clinical characteristics.

In addition, Coleman et al. (2006) compared the cost-effectiveness of utilizing a bolus thrombolytic agent with glycoprotein (GP) IIb/IIIa inhibitor followed by transfer to a tertiary institution for facilitated percutaneous coronary intervention (PCI) or standard of care transfer without primary PCI drugs among patients presenting to a community hospital with ST-segment elevation myocardial infarction (STEMI). In their study, data were derived from a cardiac catheterization laboratory database comprising patients transferred to the authors' institution. In order to ensure that similarities between important demographic characteristics were taken into account, patients receiving primary PCI were matched with those receiving facilitated PCI using propensity scores. Their results suggested that facilitated PCI would be both more effective and less costly.

McClellan and Newhouse (1997) used **instrumental variable** techniques combined with **difference-in-differences** estimation to evaluate the cost-effectiveness of intensive procedures for treating heart attacks among the elderly Medicare beneficiaries and they concluded that their analysis provides a robust design for evaluating the cost-effectiveness of medical techonologies using observational data. Our review also identified a study that attempts to compare the performance of different methods. In their case-study, Polsky and Basu (2006) examined the issue of selection bias in economic evaluations using observational data more extensively and they concluded that the considerable degree of observable bias diminished when covariate adjustment was undertaken. They also noted that the results obtained using linear regression and propensity score analysis were similar. However, when they compared these results with estimates obtained using the instrumental variable approach they found large differences.

Discussion

We aimed to review the current state of the art in the health economic evaluation literature with respect to the application of methods adjusting for confounding when

observational evidence is used for the estimation of cost-effectiveness. The process of identification of relevant studies was not straightforward and while many studies reported seemingly relevant analysis they were often only concerned with costs or effects and rarely extended their analysis to cost-effectiveness. In addition, these analyses rarely aimed to contribute to the development of analytical methods. We have employed a combination of literature searches and citation/bibliography reviews in an effort to capture as many as possible important studies concerning analytical approaches. The results presented here are preliminary and might change with the completion of the review.

The first major limitation of the implementation of the majority of these studies evaluating the cost-effectiveness of medical technologies using observational data constitutes the fact that they mostly focused on bias from observed factors. Even if it is assumed that all relevant observed covariates have been included in the analyses, there is no guarantee that the results will be near the 'true' value because of unobserved confounders that may substantially bias the estimates obtained. The choice of covariates and the choice of the appropriate model were often poorly justified. The second major limitation is the fact that most studies did not attempt to compare their results against estimates obtained using other methods despite the fact that they use rich datasets such as administrative databases, which in theory could provide a large number of observations and a wealth of information, thus allowing the implementation of more sophisticated analyses. On the other hand, the quality of data sources (including routine databases) is crucial in the implementation of appropriate analysis and often varies. The third major limitation is that nearly all studies did not compare their results with estimates obtained from analyses that employ data from randomized clinical trials, which are usually regarded as the 'gold standard'. Trial-based economic evaluations can, in principle, provide the necessary reference case when conducting such comparisons among different data sources. In light of the above one would have expected that these studies would have carried out comprehensive sensitivity analysis, which is not the case.

Limitations also arise from the choice of estimation method and the software used. Currently, estimation using regression analysis, matching and stratification is relatively straightforward and can be done using standard statistical/econometric packages. Implementation of more sophisticated methods such as the combination of **instrumental variable** techniques with **difference-in-differences** estimation employed by McClellan and Newhouse (1997), are less accessible to researchers. In addition, regression analysis, matching and stratification all work well with small numbers of covariates. However, as the number of potential confounders increases the practicality of these methods diminishes. For matching specifically, there are problems relating to statistical power because of overmatching and the effect of the matching variables on the outcome cannot be assessed. In addition, it is assumed that all relevant covariates have been measured. Last but not least, if there isn't significant overlap between the groups on the matching variables then other biases such as regression toward the mean can potentially occur.

The use of propensity score analysis has seen a tremendous growth in the biomedical literature over the last decade (Winkelmayer and Kurth 2004). Recent systematic reviews have assessed the use of propensity score analysis in its various manifestations, by comparing estimates of relationships between exposures and outcomes obtained using propensity score analysis to those obtained from multivariable regression models (Shah et al. 2005) (Stürmer et al. 2006). Despite the potential advantages that this method may offer (Glynn, Schneeweiss, and Stürmer

2006), the results of these reviews indicate that in the vast majority of cases, the view that propensity score analysis is a more robust method for confounding adjustment is ill-founded, and there is little evidence to support the claim that there are significant differences between estimates obtained using propensity score analysis and more traditional methods such as multivariable regression. A potential explanation for this could be that although propensity score analysis can accommodate bias from observed sources, it does nothing to balance unmeasured confounders (Jones and Rice 2009). The selection on observables assumption constitutes an important limitation of propensity score analysis and a number of researchers that applied this method in health economic evaluation have acknowledged this (Mojtabai and Zivin 2003) (Coleman et al. 2006). One could therefore argue that the use of propensity score analysis in economic evaluations as a device for minimising bias may be of little benefit, at least for studies in which the number of observable confounders is small.

Instrumental variable estimation is generally seen as a more reliable method as it can deal with unobserved sources of bias. Nevertheless, its use depends on the availability of an appropriate instrument, something that often may not be guaranteed. For example, Polsky and Basu (2006) were unclear as to whether the exhibited substantial variation in their results obtained using the instrumental variable approach was the result of considerable unobserved bias or merely reflected choice of a weak instrument.

Conclusions

This literature review suggests that studies in the current health economic evalaution literature employing observational data might be considerably biased. To date, different analytical methods that deal with confounding adjustment have been applied to cost-effectiveness studies employing observational data. Nevertheless, most of them are only concerned with known sources of biases, come with their own limitations, and currently there is no general consensus on which is the most appropriate to use. Indeed, in the absence of comparative work (both of methods but crucially against the gold standard of randomised trials), the interpretation of the results obtained from most studies identified by this review should be viewed with caution. The applications presented in this review have exemplified the use of multivariable regression models, propensity score analysis, matching, stratification, or combinations of these methods that can be used for confounding adjustment but further applications are needed to support their wider use. Other econometric methods such as difference-in-differences, control functions as well as discontinuity designs show potential and further research comparing both the performance of these methods as well as different data sources (for data collected both prospectively and retrospectively) is highly desirable.

We would welcome discussion on the following issues:

- Comments on the review methodology and reporting.
- Should Mendelian randomisation applications be added?

- Focus on average treatment effects or heterogeneous treatment effects? Should we report what the studies use?
- Should the comparison against the gold standard (RCTs) be by health area?
- More methodological work of comparative nature is needed but different data requirements for different methods might be a limiting factor. Is there scope for use of simulated data?
- Any other issue that is of particular relevance in your opinion.

References

Angrist, Joshua D., Guido W. Imbens, and Donald B. Rubin. 1996. Identification of Causal Effects Using Instrumental Variables. *Journal of the American Statistical Association* 91 (434):444-455.

Blundell, R., and M. Costa Dias. 2008. Alternative approaches to evaluation in empirical microeconomics. *CEMMAP working paper*.

Bound, John, David A. Jaeger, and Regina M. Baker. 1995. Problems with Instrumental Variables Estimation When the Correlation Between the Instruments and the Endogeneous Explanatory Variable is Weak. *Journal of the American Statistical Association* 90 (430):443-450.

Buckley, Jack , and Yi Shang. 2003. Estimating policy and program effects with observational data: the "differences-in-differences" estimator. *Practical Assessment, Research & Evaluation* 8 (24).

Centre for Reviews and Dissemination. 2009. Systematic reviews: CRD's guidance for undertaking reviews in health care. York: University of York.

Coleman, Craig I., Raymond G. McKay, William E. Boden, Jeffrey F. Mather, and C. Michael White. 2006. Effectiveness and cost-effectiveness of facilitated percutaneous coronary intervention compared with primary percutaneous coronary intervention in patients with ST-segment elevation myocardial infarction transferred from community hospitals. *Clinical Therapeutics* 28 (7):1054-1062.

Collins, Rory, and Stephen MacMahon. 2001. Reliable assessment of the effects of treatment on mortality and major morbidity, I: clinical trials. *The Lancet* 357 (9253):373-380.

De Ridder, Annemieke, and Diana De Graeve. 2009. Comparing the Cost Effectiveness of Risperidone and Olanzapine in the Treatment of Schizophrenia Using the Net-Benefit Regression Approach. *Pharmacoeconomics* 27:69-80.

Glynn, Robert J., Sebastian Schneeweiss, and Til Stürmer. 2006. Indications for Propensity Scores and Review of their Use in Pharmacoepidemiology. *Basic & Clinical Pharmacology & Toxicology* 98 (3):253-259.

Griffin, S C, J A Barber, A Manca, M J Sculpher, S G Thompson, M J Buxton, and H Hemingway. 2007. Cost effectiveness of clinically appropriate decisions on

alternative treatments for angina pectoris: prospective observational study. *BMJ* 334 (7594):624-.
Heckman, J. J. 2008. Econometric causality. *CEMMAP working paper*.
Heckman, James J. 1979. Sample Selection Bias as a Specification Error. *Econometrica* 47 (1):153-161.
Heckman, James J., Hidehiko Ichimura, and Petra Todd. 1998. Matching as an Econometric Evaluation Estimator. *The Review of Economic Studies* 65 (2):261-294.
Hennekens, Charles H., Julie E. Buring, and Sherry L. Mayrent. 1987. *Epidemiology in medicine*. Boston: Little, Brown.
Hoppe, Daniel J., Emil H. Schemitsch, Saam Morshed, Paul Tornetta, III, and Mohit Bhandari. 2009. Hierarchy of Evidence: Where Observational Studies Fit in and Why We Need Them. *J Bone Joint Surg Am* 91 (Supplement_3):2-9.
Imbens, Guido W. 2004. Nonparametric Estimation of Average Treatment Effects Under Exogeneity: A Review. *Review of Economics and Statistics* 86 (1):4-29.
Imbens, Guido W., and Thomas Lemieux. 2008. Regression discontinuity designs: A guide to practice. *Journal of Econometrics* 142 (2):615-635.
Imbens, Guido W., and Jeffrey M. Wooldridge. 2009. Recent Developments in the Econometrics of Program Evaluation. *Journal of Economic Literature* 47 (1):5-86.
Jones, A. 2009. Panel data methods and applications to health economics. In *Palgrave Handbook of Econometrics. Volume 2: Applied Econometrics*, edited by T. C. Mills and K. Patterson: Palgrave Macmillan.
Jones, Andrew, and Nigel Rice. 2009. Econometric evaluation of health policies. In *The Oxford Handbook of Health Economics*, edited by S. Glied and P. C. Smith. Oxford: Oxford University Press.
Kirkwood, Betty R., and Jonathan A. C. Sterne. 2003. *Essential medical statistics*. Malden, Mass.: Blackwell Science.
Knapp, Martin, Frank Windmeijer, Jacqueline Brown, Stathis Kontodimas, Spyridon Tzivelekis, Josep Maria Haro, Mark Ratcliffe, Jihyung Hong, and Diego Novick. 2008. Cost-Utility Analysis of Treatment with Olanzapine Compared with Other Antipsychotic Treatments in Patients with Schizophrenia in the Pan-European SOHO Study. *Pharmacoeconomics* 26:341-358.
Lee, David, and Thomas Lemieux. 2009. Regression Discontinuity Designs in Economics. *SSRN eLibrary*.
Luellen, Jason K., William R. Shadish, and M. H. Clark. 2005. Propensity Scores: An Introduction and Experimental Test. *Eval Rev* 29 (6):530-558.
MacMahon, Stephen, and Rory Collins. 2001. Reliable assessment of the effects of treatment on mortality and major morbidity, II: observational studies. *The Lancet* 357 (9254):455-462.
McClellan, Mark, and Joseph P. Newhouse. 1997. The Marginal Cost-Effectiveness of Medical Technology: A Panel Instrumental-Variables Approach. *Journal of Econometrics* 77 (1):39-64.
Merito, M., and P. Pezzotti. 2006. Comparing costs and effectiveness of different starting points for highly active antiretroviral therapy in HIV-positive patients. *The European Journal of Health Economics* 7 (1):30.
Mitra, N., and A. Indurkhya. 2005. A propensity score approach to estimating the cost-effectiveness of medical therapies from observational data. *Health Econ* 14 (8):805-15.

Mojtabai, R., and J. Zivin. 2003. Effectiveness and cost-effectiveness of four treatment modalities for substance disorders: a propensity score analysis. *Health services research* 38 (1 Pt 1):233.

Morshed, Saam, Paul Tornetta, III, and Mohit Bhandari. 2009. Analysis of Observational Studies: A Guide to Understanding Statistical Methods. *J Bone Joint Surg Am* 91 (Supplement_3):50-60.

Navarro, Salvador. 2008. Control functions. In *The New Palgrave Dictionary of Economics*, edited by S. N. Durlauf and L. E. Blume. Basingstoke: Palgrave Macmillan.

Newhouse, Joseph P., and Mark McClellan. 1998. Econometrics in outcomes research: The use of instrumental variables. *Annual Review of Public Health* 19 (1):17-34.

Polsky, Daniel, and Anirban Basu. 2006. Selection bias in observational data. In *The Elgar Companion to Health Economics*, edited by A. Jones. Cheltenham: Edward Elgar Publishing Ltd.

Rosenbaum, Paul, and Donald Robin. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika* 70 (1):41-55.

Shah, Baiju R., Andreas Laupacis, Janet E. Hux, and Peter C. Austin. 2005. Propensity score methods gave similar results to traditional regression modeling in observational studies: a systematic review. *Journal of Clinical Epidemiology* 58 (6):550-559.

Stürmer, Til, Manisha Joshi, Robert J. Glynn, Jerry Avorn, Kenneth J. Rothman, and Sebastian Schneeweiss. 2006. A review of the application of propensity score methods yielded increasing use, advantages in specific settings, but not substantially different estimates compared with conventional multivariable methods.

Weiss, J. P., O. Saynina, K. M. McDonald, M. B. McClellan, and M. A. Hlatky. 2002. Effectiveness and cost-effectiveness of implantable cardioverter defibrillators in the treatment of ventricular arrhythmias among medicare beneficiaries. *ACC Current Journal Review* 11 (5):73-74.

Windmeijer, F., S. Kontodimas, M. Knapp, J. Brown, and J. M. Haro. 2006. Methodological approach for assessing the cost-effectiveness of treatments using longitudinal observational data: the SOHO study. *Int J Technol Assess Health Care* 22 (4):460-8.

Winkelmayer, Wolfgang C., and Tobias Kurth. 2004. Propensity scores: help or hype? *Nephrol. Dial. Transplant.* 19 (7):1671-1673.

Table 1 – Cost-effectiveness studies reviewed

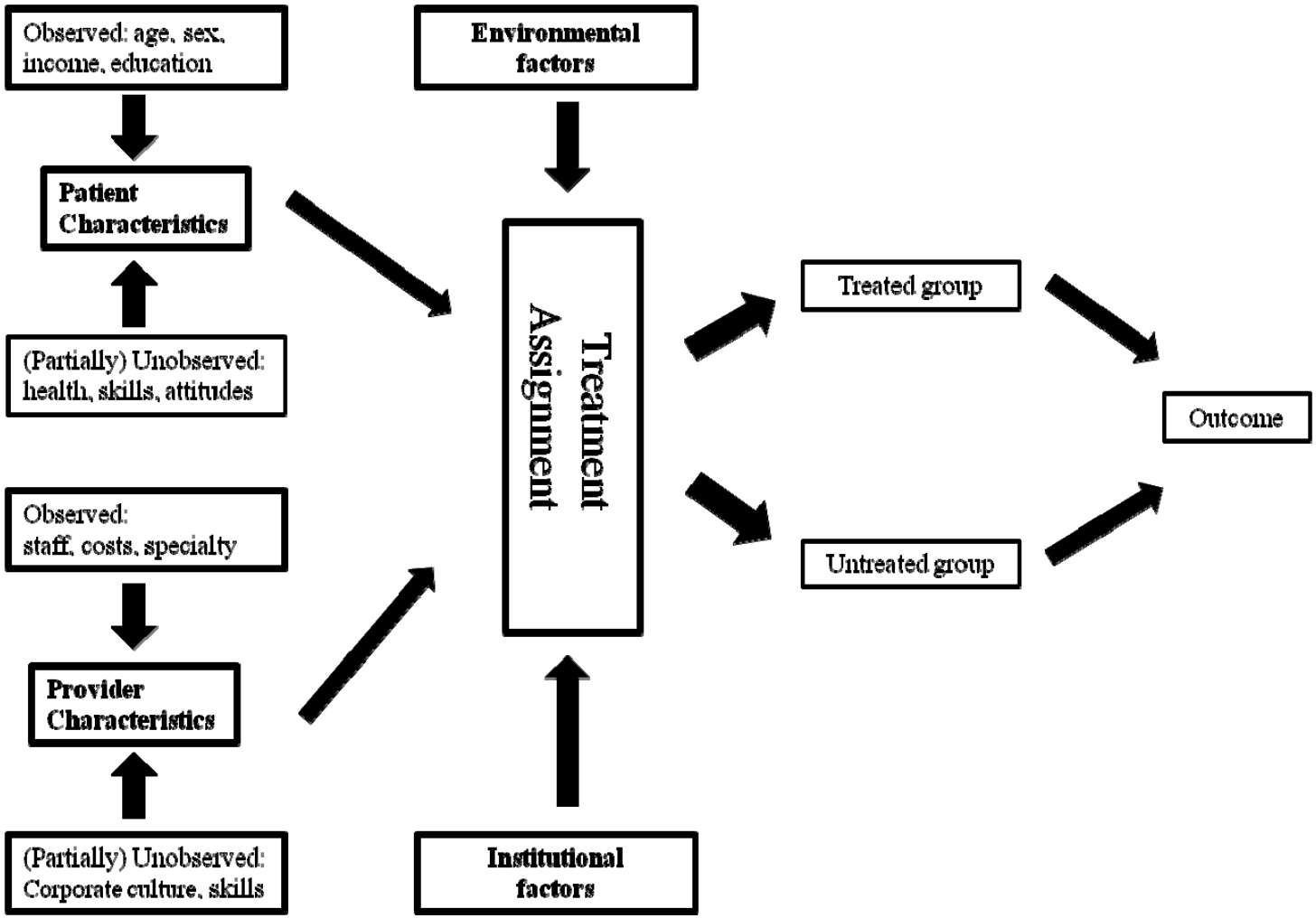
Study	Type	Outcome	Summary CE Measure	Interventions	Data
Coleman et al. (2006)	CEA	Major cardiac events	ICER	Methods for ST-segment elevation myocardial Infarction	Cohort study
DeRidder et al. (2009)	CUA	QALYs	Net-Benefit	Antipsychotic treatments for schizophrenia	Survey
Dhainaut et al. (2007)	CEA/ CUA	Life-Years/ QALYs	ICER	Recombinant human activated protein C	Before-and-after study
Griffin et al. (2007)	CUA	QALYs	ICER	Procedures for coronary revascularisation	Cohort study
Knapp et al. (2008)	CUA	QALYs	ICER	Antipsychotic treatments for schizophrenia	Cohort study
McClellan & Newhouse (1997)	CEA	Deaths avoided	ICER	Catheterization (plus angioplasty or bypass surgery)	Routinely collected
Merito & Pezzoti (2006)	CEA	Progression to AIDS stage or death	ICER / Net-benefit	Highly active anti-retroviral therapies	Cohort study (retrospective analysis)
Mitra & Indurkhya (2005)	CEA	Days survived	Net-Benefit	Treatment of bladder cancer	Routinely collected/ Registry
Mojtabai & Zivin (2003)	CEA	Abstinent case/ case of reduced use	ICER	Treatment modalities for substance abuse	Survey
Polsky & Basu (2006)	CUA	QALYs	ICER	Breast conservation surgery with radiation, mastectomy	Routinely collected
Weiss et al. (2002)	CEA	Life-Years	ICER	Implantable cardioverter defibrillator	Routinely collected
Windmeijer et al. (2006)	CUA	QALYs	ICER	Antipsychotic treatments for schizophrenia	Cohort study

Table 2 – Review of studies

Study	Parameter Adjustment	Analytical Method(s)	Comparison(s)	Software	Handling of Uncertainty	Study Summary
Coleman et al. (2006)	Costs and Effects	Propensity Score Matching	None	SPSS 11	Bootstrapping on ICER	AO/ CRCT-
DeRidder et al. (2009)	Net-Benefit	Regression Analysis	Different regression models	STATA 9	One-way Sensitivity Analysis / CEAC	AO/ CRCT-
Dhainaut et al. (2007)	Costs and Effects	Propensity Score Matching	With estimates from trial-based evaluation	SAS	Bootstrapping on ICER/ CEAC	AO/ CRCT
Griffin et al. (2007)	Costs and Effects	Matching/ Regression Analysis	None	Not specified	One-way Sensitivity, Scenario Analysis/ CEAC	AO-/ CRCT-
Knapp et al. (2008)	Costs and Effects	Regression Analysis	None	Not specified	Bootstrapping on incremental costs and effects/ CEAC	AO/ CRCT-
McClellan, Newhouse (1997)	Costs and Effects	DiD & Instrumental Variables	Least squares estimates of ATEs/ DiD with IV	Not specified	Standard Errors for incremental costs and effects/ Scenario Analysis	AO/ AU/ CRCT-
Merito, Pezzoti (2006)	Costs and Effects / Net-Benefit	Propensity Score Stratification/ Regression Analysis	None	Not specified	Bootstrapping and percentile method on ICER/ CEAC	AO/ CRCT-
Mitra, Indurkhya (2005)	Net-Benefit	Propensity Scores / Regression Analysis	Unadjusted NMB, adjusted NMB and propensity score adjusted NMB	Not specified	CEAC/ Sensitivity Analysis on WTP threshold λ	AO/ AU-/ CRCT-
Mojtabai, Zivin (2003)	Effects	Propensity Score Stratification	Results across the different modalities	Not specified	Bootstrapping on ICER/ Extreme scenario analysis	AO-/ CRCT-
Polsky, Basu (2006)	Costs and Effects	Regression Analysis, Propensity Score Stratification, Instrumental Variables	Unadjusted with adjusted results plus comparison of different methods	Not specified	Bootstrapping on ICER	AO/ AU-
Weiss et al. (2002)	Costs and Effects	Propensity Score Matching	Partial: unadjusted survival results provided	SAS	Not reported	AO-/ CRCT-
Windmeijer et al. (2006)	Costs and Effects	Regression Analysis	Between different time periods (epochs)	Not specified	Bootstrapping on costs and effects/ CEAC	AO/ CRCT-

NMB: Net monetary benefit, ICER: Incremental cost-effectiveness ratio, CEAC: Cost-effectiveness acceptability curve, ATE: Average treatment effect, DiD: Difference-in-differences, AO: adjustment for observable confounders implemented, AO- : adjustment for observable confounders implemented but might be missing important factors, AU: adjustment for unobservables implemented, AU-: weak examination of unobservables (weak instruments etc.), CRCT comparison with RCTs, CRCT- no or partial comparison with RCTs

Figure 1 - Selection into treatment in observational studies



Rovithis et al., Work in progress, HESG 2010, LSE

Figure 2 - Flow chart of the review process