Econometric methods for evaluating the cost-effectiveness of health care interventions using observational data

DIMITRIOS ROVITHIS
GREEN TEMPLETON COLLEGE

Michaelmas Term 2013

Doctor of Philosophy Thesis

ABSTRACT

This thesis explores the use of observational microdata in cost-effectiveness analysis. The application of econometric methods adjusting for selection bias is first reviewed and critically appraised in the economic evaluation literature using a structured template. Limitations of identified studies include lack of good quality evidence regarding the performance of different analytical approaches; inadequate assessment of the sensitivity of their results to violations of fundamental assumptions or variations to crucial estimator parameters; failure to combine the cost and effectiveness outcomes in a summary measure; and no consideration of stochastic uncertainty for the purpose of evaluating cost-effectiveness. Data from the Birthplace national cohort study are used in an attempt to address these limitations in the context of an empirical comparison of estimators relying on regression, matching, as well as the propensity score. It is argued that although these methods cannot address the potential impact of unobservable confounding, a novel approach to bias-corrected matching, combining entropy balancing with seemingly unrelated regression, still has the potential to offer important advantages in terms of analytical robustness. The net economic benefit is proposed as a straightforward way to exploit the strengths of rigorous econometric methodology in the development of reliable and informative cost-effectiveness analyses.

CONTENTS

ACKNOWLEDGEMENTS

The successful completion of this thesis would have not been possible without the support of certain individuals. My greatest gratitude goes to my parents Zafeirios and Georgia. Their love, understanding and encouragement are beyond words and have always been felt despite the long distance. To them I dedicate my thesis. Special thanks also go to my aunt Paraskevi, for always being supportive and comforting when I was worried. In addition, I would like to thank Menelaos Karanasos for being extremely tolerant with my complaints during many difficult moments in Oxford; Leontios Pappas for his frequent phone calls that helped me maintain my sanity, even when this seemed impossible; and Kostas Tsilidis for also lending an ear of support through the years it has taken me to complete the thesis.

My doctoral training was undertaken while I was a health economist at the National Perinatal Epidemiology Unit (NPEU). Partial financial support from the National Institute for Health Research is acknowledged. I am thankful to a number of individuals in Oxford who made a positive contribution to my doctorate. These include Professor Raymond Fitzpatrick, who in his capacity as Director of Graduate Studies, Head of Department and later as supervisor enabled me to work relatively undistracted during the final year of my studies; my colleague in the NPEU, Dr. Elisabeth Schroeder, for answering my questions when I first started

exploring the Birthplace dataset; and Dr. Sarah Wordsworth, who together with Professor Winnie Yip provided me with the opportunity to teach health economics to undergraduate and postgraduate students in Oxford.

My overall training benefited tremendously from my participation as a visiting graduate student in the Programme on Causal Inference of the Harvard School of Public Health. During my stay in Boston, I had the privilege to experience first-hand the mentoring of Professor Miguel Hernán and the scientific vision of Professor James Robins. I thank them both for their hospitality and the training opportunities that they offered me so generously. Gratitude is extended to Professor Eric Tchetgen Tchetgen, Professor Tyler VanderWeele and the other members of the causal inference group for insightful comments. I would also like to acknowledge the very helpful discussions that I had with Professor Joshua Angrist, Professor David Cutler, Professor Guido Imbens, Professor Gary King, Professor John Zupancic, as well as numerous other faculty members, researchers and students while attending the joint BU/Harvard/MIT Health Economics Seminars, the joint Harvard/MIT Econometrics Workshops, and the Applied Statistics Workshops of the Harvard Institute for Quantitative Social Science.

The material presented in the various chapters of the thesis benefited from feedback received, both during seminars that I was invited to give in London, Boston, Athens and Nottingham, as well as from conference presentations, personal communication with other experts, and several training courses that I attended throughout my studies. I am grateful to all colleagues for their questions

and comments, which challenged me to think the issues involved from a variety of perspectives. Of particular interest and value were the suggestions of Professor Willard Manning from the University of Chicago, which concerned potential directions of the review chapter. Professor Richard Grieve from LSHTM also provided an informative discussion based on an early draft of the same chapter during the Health Economists Study Group Meeting held at LSE in January 2010.

Last, but certainly not least, the contribution of three other individuals to my training should be acknowledged. My health economics "supervisors", Professor Stavros Petrou and Dr. Borislava Mihaylova, throughout my studies have taught me an awful lot about certain aspects of academia, and perhaps more importantly, life in general. In addition, during my first year, the Director of the Health Economics Research Centre, Professor Alastair Gray, provided me with the opportunity to attend the short Oxford-based course "Applied Methods of Cost-Effectiveness Analysis" in exchange for the fee of £1,100…

INTRODUCTION

Health care in Europe has seen a great deal of transformation during the last three decades (Mossialos et al., 2002). Following the period of nineteen sixties and nineteen seventies, when health care expenditures presented a rapidly growing trend, health policies throughout the nineteen eighties attempted to contain costs primarily by controlling hospital expenditure and negotiating over physician payment. In the nineteen nineties, the focus shifted to efficiency and markets, with the separation of purchasers and providers, the introduction of incentives, the formation of diagnosis-related groups and the use of resource allocation mechanisms. Nevertheless, despite the reforms, health care expenditure continued to increase in real terms in the majority of European countries (Busse, 2001).

During the last decade, policies have placed particular emphasis on promoting the assessment, accountability and quality of health care (Dixon and Poteliakhoff, 2011). A central development of this period has been the introduction of health technology assessment (HTA), which aims to evaluate in a systematic manner the properties and effects of health care technologies, using explicit analytical frameworks drawn from a variety of research fields (Goodman, 1999). Economic evaluation is an aspect of health technology assessment that has gained considerable prominence in health care decision-making, as it provides an

analytical framework that can aid decisions about which technologies represent an efficient use of health care resources (Sorenson, Drummond, and Kanavos, 2008). The term 'technology' in this context is typically used to refer to any intervention, irrespective of whether this is a medicine, device, procedure, or a complex public health programme. Indeed, the appraisal of public health interventions has generally received very little attention in the economic evaluation literature, with this form of analyses largely assessing the costs and benefits of interventions at the micro level (Kelly et al., 2005).

On the technical front, economic evaluation has evolved considerably as a framework for analysis, with the current state-of-the-art comprising of elaborate mathematical methods originating from decision analysis and operational research (Briggs, Sculpher and Claxton, 2006). These methods have been proven particularly beneficial in synthesising relevant evidence from multiple sources, evaluating the lifetime cost-effectiveness of health care interventions, as well as indicating the need for and value of additional research (Petrou and Gray, 2011). An important area of methodological inquiry has also been the use of microdata for the purpose of economic evaluation (Glick et al., 2007). The main contributor to this trend has been the increased use of the randomised controlled trial in recent years as a vehicle for analysis (Sculpher et al., 2006). This development has largely been a consequence of the perceived superiority that randomised experiments have in terms of internal validity; indeed, when evaluating treatment effects, evidence from randomised controlled trials is seen as the unambiguous gold standard (Collins and MacMahon, 2001). The availability of microdata from trials

has also allowed analysts to use statistical methods to explore the role of individual heterogeneity on cost-effectiveness (Hoch, Briggs and Willan, 2002; Willan, Briggs and Hoch, 2005), and to develop hybrid trial-based decision analytic modelling approaches, which can allow extrapolated analyses with the greatest possible internal validity (Mihaylova et al., 2006).

Notwithstanding its advantages, the randomised trial has important limitations as an evaluation design, particularly in terms of external validity, which can ultimately prevent it from delivering convincing answers for research questions relating to economic decision-making (Buxton et al., 1997). The emergence of large observational datasets such as cohort studies, surveys, disease registries, electronic health records and administrative databases, can provide analysts with the opportunity to exploit microdata for the purpose of undertaking applied cost-effectiveness studies with potentially much more generalisable results (Sculpher et al., 2004). Nevertheless, a key drawback of these studies is the evaluation of treatment effects on outcomes without randomised comparisons, resulting in estimates prone to selection bias (Jones, 2007). Consequently, when observational microdata are employed, association does not necessarily imply causation, preventing in this way the analyst to make credible statements regarding the impact of an intervention on outcomes (Rubin, 1976).

During the last three decades, the development of causal inference methods has been the subject of a vast and rapidly expanding inter-disciplinary literature (Imbens and Wooldridge, 2009). In statistics, Holland's (1986) review paper paved

the way for an ongoing debate on the merits and limitations of what is now seen as the dominant approach for drawing causal inferences - the Rubin Causal Model (RCM). In economics, the applied study from LaLonde (1986) still evokes considerable interest regarding the appropriate way that empirical analysis ought to be undertaken. At the heart of this debate rests the dispute between the statistical paradigm that takes the randomised experiment as the sole basis for establishing causal effects, versus the econometric tradition, which considers this approach incomplete and instead relies on theory and structural equation modeling to directly account for the selection process of individuals into treatment (Heckman, 2008).

Despite the conflicting visions, causal inference methodology has now converged to a great extent in a number of research fields, with the potential outcomes framework serving as a unified foundation (Imbens and Rubin, 2008). In economics, the reduced form strand of programme evaluation is currently widely being used to estimate a variety of average treatment effects (Blundell and Costa-Dias, 2009). A distinctive feature of this literature, from its purely statistical counterpart, is the explicit use of subject-matter knowledge to formulate the research question of interest and guide choice of estimator (Heckman, 2010). In health economics, the practice of cutting-edge programme evaluation methods has already been demonstrated in a variety of thought-provoking applications, reflecting the increasing confidence of analysts in their use (Jones and Rice, 2011). Interestingly, the same analytic approaches have received very little attention in economic evaluation (Sekhon and Grieve, 2012). Indeed, the particular set of

requirements that cost-effectiveness analysis imposes can render the use of econometric methods challenging. Thus, it is not surprising that significant gaps remain in our understanding of analysing observational microdata in this context.

Aims & objectives

The broad aim and contribution of this thesis is to provide a greater understanding of analysing observational microdata for the purpose of addressing selection bias within a bivariate cost-effectiveness framework. In doing so, the thesis explores the interface between causal inference and economic evaluation, by exemplifying the use of analytical approaches that combine developments, not only from economics, but also from methodologically allied sciences such as sociology, political science, epidemiology and health services research. It should be noted that although the scope of the thesis is restricted to studies that employ observational microdata, the methods considered are also relevant to cost-effectiveness studies synthesising evidence. For example, appropriate econometric analysis of microdata can produce estimates that can be used for different input parameters of decision analytical models (Briggs, Claxton and Sculpher, 2006).

The integrated thesis has five specific but interlinking objectives:

(i) Review the conceptual literature on economic evaluation and causal inference

(ii) Appraise the quality of economic evaluations using observational microdata

(iii) Evaluate short-term cost and maternal outcomes of planned birth modality

(iv) Contrast alternative regression approaches for evaluating cost-effectiveness

(v) Extend the comparison to consider traditional and novel matching methods

The way each of these objectives are considered in the thesis and their specific contribution to knowledge is described below. *Chapter 1* sets the scene by presenting the methodological approaches used in the economic evaluation of health care interventions. The chapter also offers an introduction to counterfactual analysis using the potential outcomes framework. In this context, key concepts in cost-effectiveness analysis and the treatment effects literature are formally defined. The limitations of randomised experiments and the potential advantages of using observational data are then discussed. *Chapter 2* provides a synopsis of the non-experimental evaluation methods that can be used to estimate treatment effects, when observational microdata are available. The chapter also identifies which of these analytic approaches are currently employed in the economic evaluation literature and contributes to the growing discussion regarding the scope and scientific quality of the existing evidence, in order to highlight gaps in our knowledge and to consider the future research agenda. *Chapter 3* introduces the Birthplace national prospective cohort study, which evaluated the costs and safety of women planning birth in England. The chapter first offers an overview of the substantive literature in planned place of birth and then provides a description of the main cohort study components including its design, outcome measures, sample size calculations, participating units and trusts, women's eligibility criteria, as well as classification of risk status. *Chapter 4* discusses the rationale behind regression analysis, propensity score methodology, as well as matching, while presenting

formally the identification conditions required for solving the evaluation problem when observational microdata are used. The chapter then describes in detail the analyses undertaken in the empirical part of the thesis. These extend the results of the main Birthplace studies by generating an adjusted incremental cost-effectiveness analysis for the maternal outcome 'normal birth'. More importantly, the empirical investigation contributes to an ever expanding methods literature by taking a consistent approach in comparing traditional estimators such as the ordinary least squares net benefit regression, with novel doubly robust and bias-corrected matching methods that include the combination of entropy balancing with seemingly unrelated regression. Finally, *Chapter 5* provides an in depth report and discussion of the findings. Emphasis is placed on carefully critiquing the analytic approaches employed, with the interpretation of the results obtained taking place in the context of a rigorous assessment of the plausibility of the key assumptions postulated by each method.

CHAPTER ONE

BACKGROUND

1.1 Introduction

Economics involves the study of allocating scarce resources with alternative uses in the presence of infinite human needs and wants (Robbins, 1932). In the context of health care, this concept of scarcity refers to the provision of every potentially effective health care intervention from a resource-limited health care system (Folland, Goodman and Stano, 2010). Difficult decisions in choosing among competing interventions have to be made, with choices involving opportunity costs, that is, health benefits that could have been achieved had the resources been spent on the next best alternative intervention (Guiness and Wiseman, 2011).

Economic evaluation involves the comparison of alternative treatment choices through the estimation of the incremental costs and effectiveness of a health care intervention versus, ideally, all relevant comparators, but usually, against the most common alternative used in practice (Drummond et al., 2005). This comparative nature of economic evaluation reflects the opportunity cost of an alternative health care intervention in terms of outcomes displaced to fund its additional cost (Drummond and McGuire, 2001). Economic evaluation comprises a number of methods for assessing the value offered by alternative health care interventions, each with different scope and suitability. These include cost-effectiveness analysis, cost–utility analysis and cost–benefit analysis (Gold et al., 1996).

The aim of this chapter is to provide a brief overview of the conceptual approaches used in the economic evaluation of health care interventions and to discuss the methodological issues arising when the analyst attempts to measure the causal impact of alternative interventions on outcomes. The chapter is divided in two sections. The first outlines the methods used in economic evaluation and presents analytical issues pertinent to cost-effectiveness analysis. The second offers an introduction to causal inference, including a discussion of the evaluation and selection problems, a presentation of the limitations that arise in trial-based economic evaluations, as well as the rationale for undertaking cost-effectiveness studies using observational microdata.

1.2 Economic evaluation as a framework for analysis

1.2.1 Methods of economic evaluation

Cost–effectiveness analysis evaluates the incremental costs and effectiveness of a health care intervention against a comparator, typically by combining costs and outcomes measured in natural or physical units of effectiveness (for example deaths averted or life years gained) into a summary outcome measure (Gray et al., 2010). However, when an imposed focus on a single outcome dimension is deemed inappropriate or undesirable, cost-effectiveness analysis can also consider the costs and effectiveness of the alternative interventions separately, without aggregating the results. More recently, the cost-effectiveness analytical paradigm has been extended to evaluate the impact of interventions at the health system level. The systems cost-effectiveness (SEC) analysis allows the incorporation of system-wide effects, such as the incentives and responses of different agents to the implementation of a health care intervention (Frank et al., 1999). Cost–utility analysis can be seen as a specific type of cost-effectiveness analysis, which measures and values the effectiveness of an intervention in terms of life extensions combined with improvements in health-related quality of life such as the quality-adjusted life year. The use of quality-adjusted life years renders possible comparisons across different fields of health care and allows changes in quality of life to be traded with survival (Drummond, Aguiar-Ibanez and Nixon, 2006). In contrast with cost-effectiveness and cost-utility analysis, cost–benefit analysis values all costs and effectiveness of a health care intervention in monetary units.

This form of analysis theoretically seems an ideal method for evaluating health care interventions because it allows comparisons across different areas of public policy. Nevertheless, conducting a cost–benefit analysis can be difficult primarily because of objections related to valuing health outcomes in monetary terms (McIntosh et al., 2010).

In cost–effectiveness analysis, the extent to which a health care intervention is deemed worthwhile is determined by comparing its additional cost per extra unit of effectiveness outcome, to some ceiling cost-effectiveness value λ (Drummond et al, 2005). This value is the threshold that reflects the decision-maker's maximum willingness-to-pay for a one-unit increase in the effectiveness outcome. Currently, the level of a threshold value for decision-making purposes is still subject to debate (Culyer et al., 2007). The possible decisions arising from the adoption of a threshold can be illustrated diagrammatically using the cost-effectiveness plane (Black, 1990). In Figure 1, C denotes the average costs and average effects for the comparator, usually an existing intervention, while the slope of the dotted line represents the maximum incremental cost-effectiveness ratio. Each quadrant has a different implication for the decision. When the alternative intervention falls in the lower right quadrant, it is better in both dimensions since it has lower cost and greater effectiveness compared with the comparator. Consequently, it is preferred and is said to dominate the comparator. However, the comparator would dominate any alternative intervention that falls in the upper left quadrant since it will be both more costly and less effective.

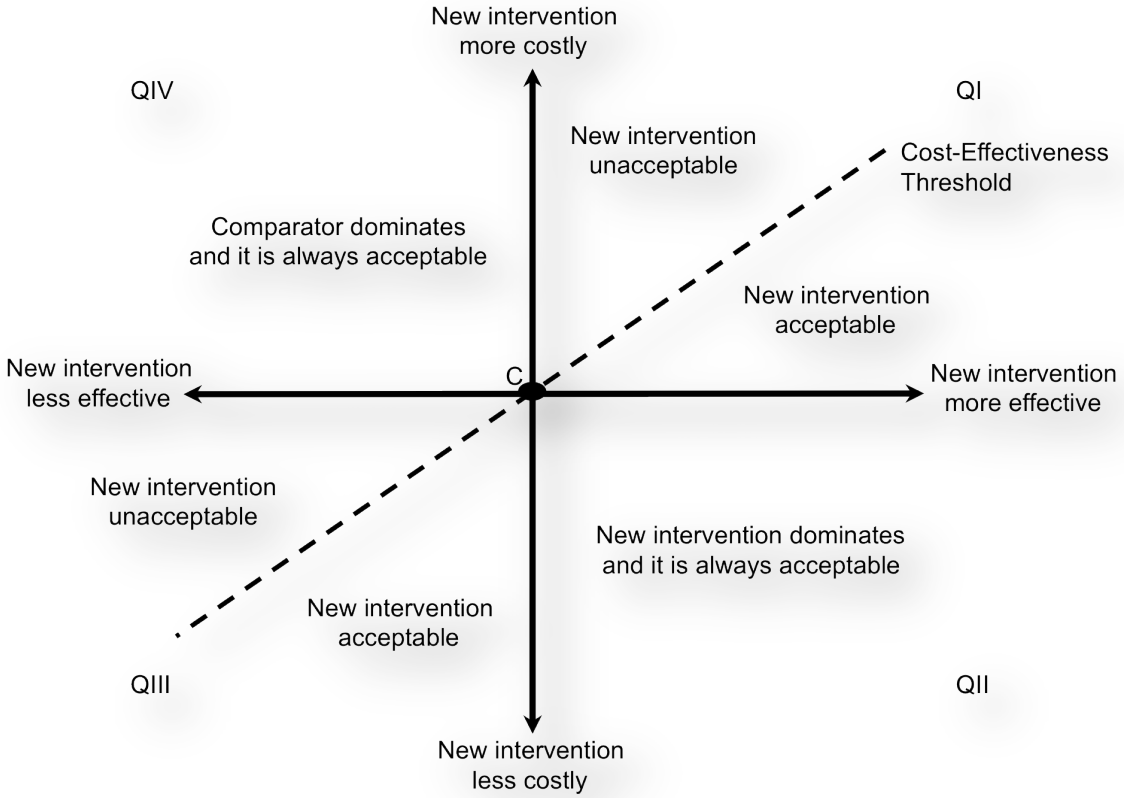


Figure 1.1 – The cost-effectiveness plane. Adapted from Maniadakis and Gray (2000).

An alternative intervention may or may not be considered preferable to the comparator, as it can be less costly but also less effective, or more effective but also more costly. Whether an alternative intervention will be chosen over the comparator, depends on the relation between its cost and effectiveness, that is, its incremental cost-effectiveness ratio (ICER). If a maximum acceptable incremental cost-effectiveness ratio is adopted, then any alternative intervention that falls below the dotted line would be acceptable, that is, preferable to the comparator. For interventions that fall above the dotted line the opposite would be true.

Musgrove and Fox-Rushby (2006) point out that in practice, because of the uncertainty surrounding the estimates of cost and effectiveness, the distinction of interventions to acceptable and non-acceptable will cover a region, the width of which will depend on the confidence intervals around the estimates obtained from the analysis.

1.2.2 Power & sample size in cost-effectiveness analysis

Studies in the biomedical sciences are primarily explanatory in nature, aiming to find indications about the causes of a disease through the detection of a treatment effect using statistical inference (Ioannidis, 2005). As such, their focus is on declarations regarding whether the observed association in the data echoes a true relationship in the population. Power analysis is typically used in these studies to calculate the smallest sample size needed, in order for the analyst to be able to potentially detect an effect of a particular magnitude on a health outcome; alternatively, power analysis may also be employed to determine the smallest magnitude of a treatment effect that can potentially be detected using a particular sample size (Wickramaratne, 1995).

In cost-effectiveness analysis an additional important consideration is the detection of differences between economic endpoints (Briggs and Gray, 1998). Efforts to develop methods that calculate sample size for economic evaluation have evolved concurrently with the establishment of clinical trials as the preferred vehicle for

analysis (Glick et al., 2007). It is now well-documented in the applied literature that the sample size required to power an economic evaluation for detecting differences in economic endpoints, is typically considerably greater than that for a health outcome (Glick 2011a; 2011b).

However, for economic decision-making purposes, of interest is primarily the measurement of the magnitude of an effect and not simply the detection of differences (Jones and Rice, 2011). As such, it has been argued that in cost-effectiveness analysis, estimation for the economic endpoints and the cost-effectiveness summary outcome measure should be preferred over hypothesis testing (Briggs, 2000). The use of estimation over hypothesis testing also has the additional advantage that wide confidence limits around the results will expose low powered studies, even when outcome differences lack statistical significance (Gardner and Altman, 1986). Claxton (1999) goes even further arguing that for economic decision-making, the rules of statistical inference including confidence intervals, p-values, levels of significance, as well as their Bayesian counterparts, are irrelevant, arbitrary, inconsistent with the objectives of a modern health care system, and impose unnecessary costs. He defends his view by stating that if the objective of a health care system is to maximise health gain for a given budget, the decision relating to choice of whether an intervention is cost-effective does not depend on whether differences in the mean net benefit are statistically significant, or fall outside a Bayesian range of equivalence. According to Claxton (1999), this is because "one of the mutually exclusive alternative health care interventions must be chosen and this decision cannot be deferred". Consequently, he argues, "the

opportunity costs of failing to make the right decision based on the mean are symmetrical, and the question of which of the alternatives should be regarded as current practice, is irrelevant". Despite these claims, Nixon and Thomson (2005) point out that it is unlikely for certain differences in cost–effectiveness to be accepted without strong direct evidence. As such, cost-effectiveness studies, ideally, should be designed with sufficient power right from the onset for the economic outcomes (Briggs, 2000).

1.2.3 Summary outcome measures & analysis of uncertainty

Economic evaluations have traditionally assessed the cost-effectiveness of competing health care interventions by calculating the incremental cost-effectiveness ratio. The incremental cost-effectiveness ratio is defined as the ratio of the difference in costs of two alternative health care interventions, to the difference in their effectiveness. Mathematically, let (C_0, C_1) and (E_0, E_1) be the sample means for cost and effectiveness, respectively, under the standard (D=0) and comparator (D=1) health care interventions. The incremental cost-effectiveness ratio is then given by

$$\widehat{\text{ICER}} = \frac{\overline{C_1} - \overline{C_0}}{\overline{E_1} - \overline{E_0}} = \frac{\Delta\overline{C}}{\Delta\overline{E}} \qquad (1.1)$$

In systems cost-effectiveness, the incremental cost-effectiveness ratio can allow consideration of system-level effects and can be obtained by dividing the system effectiveness to system costs (Frank et al., 1999). System effectiveness in this context is defined as the sum of all of the effects produced by health care in a system, whereas system cost is the sum of all direct treatment costs.

The interpretation of the incremental cost-effectiveness ratio estimates obtained from the analysis will depend on the level of confidence that the analyst has in the resulting comparison of costs and effectiveness (Briggs, 1999). In economic evaluations employing microdata, cost-effectiveness estimates are based on a particular sample derived from the population. As such, uncertainty arises from the use of a finite sample to estimate the true population value of costs and effectiveness. This source of uncertainty is commonly referred to as sampling variation or stochastic uncertainty (McCarron, Pullenayegum et al., 2009). Confidence statements regarding comparisons of costs and effectiveness require quantification of the uncertainty arising from sampling variation in the cost-effectiveness estimate. This can be achieved using statistical methods that estimate the incremental cost-effectiveness ratio and calculate its corresponding confidence interval. In the applied literature, two statistical approaches have been the prevailing methods of evaluating uncertainty using the information contained in microdata: Fieller's method and bootstrap procedures (Briggs, O'Brien and Blackhouse, 2002).

Estimation of the confidence interval of the incremental cost-effectiveness ratio using Fieller's method is a parametric approach that calculates an exact confidence interval for the ratio, assuming that the joint distribution of cost and effectiveness is bivariate normal (Willan and O'Brien, 1996). In this case, the central limit theorem, which states that the mean effect of treatment approximates a normal distribution as the sample size increases, ensures that in moderate and large sample sizes the normality assumption is met, provided that the underlying cost and effectiveness distributions do not exhibit substantial skewness. In contrast with Fieller's method, the bootstrap is a category of non-parametric sampling methods, which are free from distributional assumptions and have the advantage of calculating confidence intervals that are potentially robust to deviations from normality (Briggs, Wonderling and Mooney, 1997). Bootstrap procedures create the empirical distribution for the incremental cost-effectiveness ratio by drawing bootstrap samples with replacement from the treatment and comparison groups, in order to estimate the incremental costs and effectiveness for each of these samples. It should be noted that the bootstrap samples are of the same size as the number of individuals in each group. The bootstrap replications are then ordered and the corresponding confidence interval is usually calculated using the percentile method (Campbell and Torgerson, 1999).

Drummond et al. (2005) point out that although the incremental cost-effectiveness ratio summarises both cost and effectiveness in a single statistic, it suffers from a number of limitations. First, the discontinuity of the ratio once the difference in effectiveness among the competing health care interventions approaches zero; in

such a case the incremental cost-effectiveness ratio becomes infinite and thus undefined. Second, the interpretability of the incremental cost-effectiveness ratio on the cost-effectiveness plane; explaining the sign of the incremental cost-effectiveness ratio can be ambiguous, since positive and negative values will have different interpretations in different quadrants. Third, the position of the incremental cost and effectiveness pairs on the cost-effectiveness plane; ordering these pairs when they are scattered over all four quadrants will not be possible and thus the confidence interval may be undefined.

More recently, in order to avoid the problems associated with using statistical methods to estimate the confidence interval for the incremental-cost-effectiveness ratio, analysts undertaking economic evaluation have focused on calculating the incremental net benefit (Stinnett and Mullahy, 1998). The incremental net benefit is defined as the difference in effectiveness minus the difference in costs, both measured in the same units. When monetary units are used, then the net monetary benefit (NMB) is calculated; in contrast, when effectiveness units are used, then the statistic computed is the net health benefit (NHB). The incremental net benefit is a scalar transformation of the incremental cost-effectiveness ratio, with the threshold value λ acting as a "converter" between costs and effectiveness; as such, calculation of the incremental net benefit requires the analysis to be undertaken and presented as a function of λ (Fox-Rushby and Cairns, 2009). Nevertheless, because λ is typically unknown, the incremental net benefit is evaluated for a range of threshold values, which are specified every time as most appropriate for each application (Hoch, Briggs and Willan, 2002). Mathematically,

using the notation introduced earlier, the incremental net monetary benefit and the incremental net health benefit, as well as their respective variances for a particular sample can be defined as

$$\widehat{\mathrm{NMB}}(\lambda)=\lambda\left(\Delta\overline{\mathrm{E}}\right)-\Delta\overline{\mathrm{C}} \tag{1.2}$$

$$\widehat{\mathrm{NHB}}(\lambda)=\Delta\overline{\mathrm{E}}-\frac{\Delta\overline{\mathrm{C}}}{\lambda} \tag{1.3}$$

$$\mathrm{var}\left[\widehat{\mathrm{NMB}}(\lambda)\right]=\lambda^2\ \mathrm{var}\left(\Delta\overline{\mathrm{E}}\right)+\mathrm{var}\left(\Delta\overline{\mathrm{C}}\right)-2\lambda\ \mathrm{cov}\left(\Delta\overline{\mathrm{E}},\Delta\overline{\mathrm{C}}\right) \tag{1.4}$$

$$\mathrm{var}\left[\widehat{\mathrm{NHB}}(\lambda)\right]=\mathrm{var}\left(\Delta\overline{\mathrm{E}}\right)-\frac{\mathrm{var}\left(\Delta\overline{\mathrm{C}}\right)}{\lambda}-\ \frac{2}{\lambda}\ \mathrm{cov}\left(\Delta\overline{\mathrm{E}},\Delta\overline{\mathrm{C}}\right) \tag{1.5}$$

Drummond et al. (2005) point out that the incremental net benefit is a linear expression and has a number of advantages over the incremental cost-effectiveness ratio. For example, larger values indicate a better result; it can take the value of zero because it is formally defined; its variance can easily be obtained as a function of λ; and adjustment for individual case-mix using regression-based methods is possible. Consequently, statistical inferences on the incremental net benefit are more straightforward than inferences on the incremental cost-effectiveness ratio (Willan, and Briggs, 2006). Confidence intervals calculated parametrically using the incremental net benefit have been shown to be equivalent to those obtained using Fieller's method (Heitjan, 2000; Zethraeus and Lothgren, 2000) and can also be obtained using bootstrap procedures (Briggs and Fenn, 1998; Stinnett and Mullahy, 1998; Nixon, Wonderling and Grieve, 2010).

Finally, cost-effectiveness acceptability curves have also been proposed as a way of making confidence statements and summarising the uncertainty surrounding the choice between health care interventions (Van Hout et al., 1994). The cost-effectiveness acceptability curve is a plot of the probability that the intervention is cost-effective as a function of willingness to pay λ. It can be derived parametrically from the joint distribution of incremental costs and effectiveness or through bootstrapping (Briggs, Wonderling and Mooney, 1997). In the latter case, evaluation of the curve is made by looking at the proportion of bootstrap replications that fall below the threshold value (Briggs, Mooney and Wonderling, 1999). This proportion represents the probability that the standard intervention is cost-effective compared to an alternative (Fenwick et al., 2006).

1.3 Causal inference in cost-effectiveness analysis

1.3.1 The aim of evaluative research

According to Jones and Rice (2011), the purpose of evaluative research is the identification and measurement of an intervention's impact on certain outcomes of interest. As noted earlier, in the context of health care, cost-effectiveness analysis seeks to estimate and compare the magnitudes of alternative treatment choices on cost, effectiveness, or summary outcome measures, with a view to provide evidence to inform decision-making (Sorenson, Drummond and Kanavos, 2008). At the micro level, this often means comparing the effect of alternative drugs or

medical devices (Drummond et al., 2005). At the macro level, the analyst may be interested in evaluating the effects of large-scale health care interventions, such as different reimbursement mechanisms (Grieve et al., 2008).

Nevertheless, when evaluating the impact of health care interventions on the outcomes of an individual, associations in the estimated effects do not necessarily imply cause and effect relationships (Jones, 2007). Causality is typically grounded on the concept of ceteris paribus, which refers to the variation of the intervention under investigation, other things being equal (Heckman, 2000). A systematic process of reasoning that can be employed to establish the impact caused by an intervention from the available evidence is causal inference. According to Heckman (2008), this process first defines the causal relationship of interest, subsequently identifies a strategy that isolates the relevant causal effect and finally estimates this effect from sample data.

In recent decades, the evaluation literature has focused on the counterfactual or potential outcomes model of causal inference (Neyman, 1923; Roy, 1951; Cox, 1958; Rubin, 1974). Methodological developments in econometrics have departed from the traditional structural approach to counterfactual analysis and have instead followed the programme evaluation paradigm, a reduced form strand that takes randomised experiments and their surrogates as the sole basis for establishing causal effects when investigating a particular research question (Heckman, 2010). As Angrist and Pischke (2009) point out, "research questions that cannot be answered by any experiment are fundamentally unidentified questions". This

approach to causal inference had as a consequence the evaluation literature to largely focus on the study of selected effects of a specific cause, rather than exploration and understanding of all possible causes of a particular outcome, or detailed estimation of all of its relative effects (Holland, 1986).

Indeed, the potential outcomes paradigm is now dominant in the evaluation literature and has in fact defined the terminology that is used, not only in economics, but also in a range of disciplines including political science, sociology, health services research, and epidemiology (Imbens and Wooldridge, 2009). In this thesis, both substantive arguments developed, as well as empirical applications presented in subsequent chapters, take the potential outcomes framework as their theoretical foundation.

1.3.2 The evaluation problem

Similarly to the rest of the evaluative research, cost-effectiveness studies aim to answer counterfactual questions such as "how would resource use and health outcomes of individuals exposed to one intervention have evolved in its absence, or if they had received an alternative intervention?" This question reflects the fundamental problem of causal inference; at any given point in time, individuals can only be observed under one course of action (Holland, 1986). Counterfactual analysis attempts to explore the causes of situations that have taken place in the past, by resorting to the use of distributions for states that are by definition

unobservable since they never happened (Dawid, 2000). This inherent limitation of all analyses aiming to evaluate the effects of health care interventions can be seen as the well-known problems of missing data in statistics, and omitted variables in econometrics (Rubin, 1976; Heckman, 1979).

Greenland, Robins and Pearl (1999) argue that although certain key features of these distributions will always be empirically unverifiable, the counterfactual model of causal inference establishes a pragmatic theoretical framework, which formulates precise and explicit assumptions required to identify causal effects. The same authors point out that this approach can subsequently provide the basis for the development of quantitative methods that satisfy these assumptions and can recover different aspects of distributions. It is worth emphasising however that, ultimately, causal inferences based on counterfactuals generally depend on assumptions that are not directly testable (Hernán and Robins, 2013).

In the simplest case of a cost-effectiveness study comparing a health care intervention versus no exposure, where the unit of analysis is the individual, one can formulate and investigate causal relationships using the potential outcomes framework. Assuming that an individual *i* is described by a pair of potential outcomes (Y^1, Y^0) in the presence ($D_i=1$) and absence ($D_i=0$) of exposure respectively, then the effect of the health care intervention on an outcome *Y* is given by

$$Y_i^1 - Y_i^0 \qquad (1.6)$$

where Y^1 refers to the outcome of the individual i when in the exposed group and Y^0 denotes the counterfactual, that is, the outcome for the same individual, had this individual not been exposed to the intervention.

Since both exposed and non-exposed states for the same individual cannot be observed, the analyst can reformulate the evaluation problem at the population level, focusing instead on identifying certain mean outcomes (Heckman and Vytlacil, 2007). If microdata on the outcome(s) of interest are available for the individuals in the exposed group, as well as the individuals in the comparison group, the outcomes in both groups can be averaged and then differenced as equation (1.7) demonstrates

$$\mathrm{E}\left[\mathrm{Y}_i^1 - \mathrm{Y}_i^0\right] \tag{1.7}$$

In such a case, an estimate of the effect of the health care intervention can be obtained, with the average outcome for individuals in the comparison group acting as a replacement for the missing counterfactual.

Nevertheless, reformulating the evaluation problem at the population level requires the Stable Unit Treatment Value Assumption (SUTVA) to hold, which only allows partial equilibrium effects and rules out any spill over effects (Jones, 2011). The Stable Unit Treatment Value Assumption has two components and postulates that there is no interference between individuals and there is no variation in exposure (Rubin, 1986). In other words, the values of exposed and

non-exposed outcomes for a given individual are not influenced by the exposure status of other individuals, and only one type of the standard intervention and one of the comparison, exists.

1.3.3 Treatment effect parameters

In the evaluation literature, three average causal effects have been most commonly of interest: the average treatment effect on the treated, the average treatment effect on the non-treated, and the overall average treatment effect (Imbens and Wooldridge, 2009; Blundell and Costa Dias, 2009; Jones and Rice, 2011). In this section, these three average causal effects are defined formally in the context of health care, whereas the conditions required for their identification using specific analytic approaches is left to be discussed in subsequent chapters.

Average Treatment Effect on the Treated

The first treatment effect parameter is the average effect of the health care intervention between the individuals who have been exposed to it. Mathematically, the average treatment effect on the treated (ATT) for the entire population under investigation and its equivalent for a particular sample is defined respectively as

$$\mathrm{ATT}=\mathrm{E}\left[\mathrm{Y}_i^1-\mathrm{Y}_i^0 \mid \mathrm{D}_i=1\right]=\mathrm{E}\left[\mathrm{Y}_i^1 \mid \mathrm{D}_i=1\right]-\mathrm{E}\left[\mathrm{Y}_i^0 \mid \mathrm{D}_i=1\right] \tag{1.8}$$

$$\widehat{\mathrm{ATT}}=\frac{1}{\mathrm{N}_{\mathrm{D}_i=1}}\sum_{i}^{\mathrm{N}_{\mathrm{D}=1}}\left(\mathrm{Y}_i^1-\mathrm{Y}_i^0|\ \mathrm{D}_i=1\right) \tag{1.9}$$

where $N_{D=1}$ is the number of exposed individuals. The term $\mathrm{E}\left[\mathrm{Y}_i^0 \mid \mathrm{D}_i=1\right]$ in (1.8) is the counterfactual that the analyst seeks to estimate. The average treatment effect on the treated is the quantity of the average effect from the health care intervention on an individual arbitrarily chosen from those exposed.

Average Treatment Effect on the Non-Treated

The second treatment effect parameter denotes the average effect from the health care intervention between those individuals who have not been exposed. Mathematically, the average treatment effect on the non-treated (ATNT) for the complete population under investigation, as well as its equivalent for a particular sample is defined respectively as

$$\mathrm{ATNT}=\mathrm{E}\left[\mathrm{Y}_i^1-\mathrm{Y}_i^0|\ \mathrm{D}_i=0\right]=\mathrm{E}\left[\mathrm{Y}_i^1 \mid \mathrm{D}_i=0\right]-\mathrm{E}\left[\mathrm{Y}_i^0 \mid \mathrm{D}_i=0\right] \tag{1.10}$$

$$\widehat{\mathrm{ATNT}}=\frac{1}{\mathrm{N}_{\mathrm{D}_i=0}}\sum_{i}^{\mathrm{N}_{\mathrm{D}=0}}\left(\mathrm{Y}_i^1-\mathrm{Y}_i^0|\ \mathrm{D}_i=0\right) \tag{1.11}$$

where $N_{D=0}$ represents the number of individuals in the non-exposed group. The term $\mathrm{E}\left[\mathrm{Y}_i^1 \mid \mathrm{D}_i=0\right]$ in (1.10) is the counterfactual that the analyst seeks to estimate. The average treatment effect on the non-treated is the quantity of the average

effect of the health care intervention on an individual arbitrarily chosen from those who have not been exposed.

Average Treatment Effect

The third treatment effect parameter represents the average effect of the health care intervention for all individuals, irrespective if someone has been exposed. Mathematically, the overall population average treatment effect (ATE), as well as its sample equivalent, is defined respectively as

$$\begin{aligned} \text{ATE} &= \text{ATT} \times \text{P}(\text{D}_i=1) + \text{ATNT} \times \text{P}(\text{D}_i=0) = \\ &\text{E}\left[\text{Y}_i^1 - \text{Y}_i^0 \mid \text{D}_i=1\right] \times \text{P}(\text{D}_i=1) + \text{E}\left[\text{Y}_i^1 - \text{Y}_i^0 \mid \text{D}_i=0\right] \times \text{P}(\text{D}_i=0) \end{aligned} \tag{1.12}$$

$$\widehat{\text{ATE}} = \frac{1}{\text{N}} \sum_i^{\text{N}} \left(\text{Y}_i^1 - \text{Y}_i^0\right) \tag{1.13}$$

$$\widehat{\text{ATE}} = \frac{1}{\text{N}} \left[\sum_i^{\text{N}_{\text{D}=1}} \left(\text{Y}_i^1 - \text{Y}_i^0 \mid \text{D}_i=1\right) + \sum_i^{\text{N}_{\text{D}=0}} \left(\text{Y}_i^1 - \text{Y}_i^0 \mid \text{D}_i=0\right) \right] \tag{1.14}$$

where $\text{P}(\text{D}_i=1)$ and $\text{P}(\text{D}_i=0)$ are the probabilities of the individual being in the exposed and non-exposed groups. The terms $\text{E}\left[\text{Y}_i^1 - \text{Y}_i^0 \mid \text{D}_i=1\right]$ and $\text{E}\left[\text{Y}_i^1 - \text{Y}_i^0 \mid \text{D}_i=0\right]$ in (1.12) contain the counterfactuals for both components of the average treatment effect that the analyst seeks to estimate. The average treatment

effect is the quantity of the average effect from the health care intervention on an individual arbitrarily chosen from both those exposed and non-exposed.

1.3.4 The selection problem

The selection problem underlies the evaluation problem and arises from the fact that the values of Y^0 or Y^1 that are observable by the analyst may not constitute de facto a random sample of the potential Y^0 or Y^1 distributions (Heckman, 2010). This can be a consequence of factors or processes influencing individuals' assignment into treatment, which in turn may affect the outcome(s) of the study. For example, individuals exposed to a health care intervention will typically be different in a set of characteristics from those who have not been exposed to the intervention. As such, they can be selected into treatment based both on observable factors, as well as factors that are partially or fully unobservable. The former may include individuals' age, sex, education, or income, whereas the latter can consist of personal health, skills, attitudes, as well as environmental or institutional factors. Selection into treatment based on observable and/or unobservable factors renders problematic the separation of the differences between both groups, which are due to selection bias (pre-existing differences before allocation to the intervention), from those that arise only from the intervention itself (Heckman, 2008).

Blundell and Costa Dias (2009) point out that selection bias may confound the treatment effect parameters and ultimately lead to differences between them. In the simplest case, when homogeneous treatment effects are assumed, there is only one impact of the intervention that affects both exposed and non-exposed individuals in the same way. Consequently, all the treatment effect parameters will be indistinguishable. However, in the more realistic case of heterogeneous treatment effects, the average treatment effect on the treated will differ from the overall average treatment effect or the average treatment effect on the non-treated. Mathematically, this can be demonstrated by extending equation (1.7)

$$\begin{aligned}
&E\left[Y_i^1 - Y_i^0 \right] = E\left[Y_i^1 \mid D_i = 1 \right] - E\left[Y_i^0 \mid D_i = 0 \right] \\
&= E\left[Y_i^1 \mid D_i = 1 \right] - E\left[Y_i^0 \mid D_i = 1 \right] + E\left[Y_i^0 \mid D_i = 1 \right] - E\left[Y_i^0 \mid D_i = 0 \right] \\
&= E\left[Y_i^1 - Y_i^0 \mid D_i = 1 \right] + E\left[Y_i^0 \mid D_i = 1 \right] - E\left[Y_i^0 \mid D_i = 0 \right] \qquad (1.15)
\end{aligned}$$

The term $E\left[Y_i^1 - Y_i^0 \mid D_i = 1 \right]$ in the above equation is the average treatment effect on the treated, which is used here for ease of exposition. This treatment effect parameter will only be recovered if the selection bias represented by the term $E\left[Y_i^0 \mid D_i = 1 \right] - E\left[Y_i^0 \mid D_i = 0 \right]$ is zero. Nevertheless, a necessary condition for this is that there are no systematic differences in the non-exposed average outcomes among the exposed and the non-exposed groups.

Angrist and Pischke (2009) note that an important consideration arising from equation (1.15) is the potential direction of selection bias. If individuals in the exposed group are on average healthier than those in the comparison group, then

selection bias will be positive and the effect of the intervention on outcomes will be overestimated. Conversely, if individuals in the comparison group tend to have better health prospects on average than those in the exposed group, then selection bias will be negative and the effect of the intervention will be underestimated. In addition, Heckman, Ichimura, Smith and Todd (1998) emphasise that of considerable interest is also the composition and magnitude of selection bias. In their study, selection bias was decomposed into the following components: differences in the overlap of the observables among the exposed and comparison groups, essentially resulting in the comparison of incomparable individuals; imbalances in the shapes of distributions of the observable individual characteristics, which are represented both among exposed and comparison groups, that is, the common support; and bias resulting from unobservables. Heckman, Ichimura, Smith and Todd (1998) found that the first two components relating to observable characteristics were the most important sources of selection bias, while the third component, although the smallest, still constituted a considerable portion of the selection bias.

1.3.5 Experimental studies: the randomised ideal

In contemporary causal inference, the ideal evaluation design in which selection bias is eliminated is a scientific experiment where the analyst controls the assignment of individuals into treatment (Angrist and Pischke, 2009; Hernán and Robins, 2013). A defining characteristic of experimental evaluation designs is the

randomisation of individuals to the exposed and comparison groups. Assuming that there is no randomisation bias, no comparison group contamination, no drop outs, as well as full compliance, this approach ensures that treatment status is independent of potential outcomes and, on average, the individuals exposed to a health care intervention are not systematically different with regards to observable or unobservable characteristics from their comparison counterparts. Consequently, any statistically significant difference in outcomes between both groups can be thought to belong exclusively to the intervention's effect.

In biomedical research the process of randomisation has traditionally constituted the greatest advantage of clinical trials, allowing any moderate treatment effects to be detected and measured without selection bias in studies of reasonably large size (Collins and MacMahon, 2001). In economics, the use of randomised experiments has been relatively limited, although more recently randomised trials have been gaining prominence in development economics (Duflo, Glennerster and Kremer, 2007). A distinctive characteristic of such trials is the use of background evidence and economic theories that predict which interventions could potentially work. The randomised evaluation of the intervention in this case typically serves as a means to test these theories.

In health economics, cost-effectiveness studies evaluating interventions at the micro level have often focused on trial-based analyses, exploiting the high internal validity of the rigorous randomised study design in order to obtain unbiased treatment effects (Sculpher et al., 2006). Of special interest to health economists

have been pragmatic trials, the large size of which provides the potential to obtain cost-effectiveness results that are more likely to be generalisable to every day clinical practice, and thus to populations and settings for which health care decisions are relevant (Ramsey et al., 2005).

1.3.6 Limitations of randomised studies

Notwithstanding the advantages of randomised experiments in producing unbiased treatment effects estimates, the use of clinical trials as a vehicle for cost-effectiveness analysis presents significant challenges.

First, there are practical problems associated with their use. Such evaluation designs are inherently vulnerable to non-compliance, as well as individuals that drop out of interventions that they do not want, or seek other means to obtain the interventions that they were randomly denied (Heckman, 1992; Heckman and Smith, 1995). Clinical trials may also not be possible to conduct because of cultural, ethical, legal, political and social reasons (Pocock and Elbourne 2000). In addition, the vast number of health care interventions currently in use, together with the continuous development of new ones, exceeds the capacity to fund and undertake clinical trials (Stevens, Raftery and Roderick 2005). The above considerations are particularly pertinent when the aim is the evaluation of health care interventions at the system level.

Second, the comparative nature of cost-effectiveness requires that, ideally, all alternative interventions relevant to the one evaluated should be included in the analysis (Drummond et al., 2005). However, clinical trials usually compare only two health care interventions, with the comparator often being either a placebo, or the most common alternative used in clinical practice. In trials where choice of the comparator intervention is inappropriate, the validity of the cost-effectiveness study can potentially be compromised, limiting in this way the usefulness of their results for economic decision-making purposes (Drummond and Sculpher, 2005).

Third, clinical trials typically employ certain inclusion and exclusion participation criteria and as such the data collected often may be protocol-driven referring to a sample of individuals that is not representative of the entire population. Consequently, both health and economic outcomes will typically not correspond to everyday clinical practice (Soto, 2002). In addition, the results of trial-based cost-effectiveness studies are valid only for the settings in which they are conducted, as practice patterns may vary in different places and may be altered over time (Buxton et al., 1997).

Fourth, the time horizon of cost-effectiveness analysis should be long enough to reflect both the deliberate, as well as the unintentional long-term economic and clinical consequences of the alternative health care interventions being compared (Weinstein et al., 2003). For many cost-effectiveness studies therefore, a lifetime horizon will be deemed as the most appropriate choice but the duration of clinical trials is typically short and insufficient for such purposes (Sculpher et al., 2006).

Collected units of resource use may also be proven incomplete when using these data retrospectively. This is because a clinical trial is usually not designed for evaluating cost-effectiveness, with the units of resources used being country specific and the duration of follow up too limited for economic analysis (Nuijten, 1998). Indeed, because of their often-short duration, the restricted population and the limited generalisability of their results, clinical trials usually aim to offer an indication of an intervention's efficacy, rather than its effectiveness (Black, 1996).

Fifth, the use of clinical trials for policy evaluation may be problematic because these are not conducted within a behaviourally coherent framework of analysis. Heckman (2008) strikes a note of caution suggesting that the implication of this is that evidence about the impact of complex health care interventions generated from randomised experiments, with different participation and eligibility rules, does not cumulate in any interpretable way. Consequently, he argues, no ex ante analysis of policy recommendations can be made.

Sixth, a combined intermediate endpoint specific to the disease may often be more appropriate in evaluating the efficacy of a clinical intervention. Nevertheless, cost-effectiveness studies for decision-making aim to address a different research question; they seek to compare the costs and consequences of health care interventions in real world settings and across a wide range of therapeutic areas (Sculpher et al., 2004). Final endpoints of a more generic nature, such as life-years or quality-adjusted life-years, will therefore be better suited for the analysis, but often these will not be measurable within a clinical trial (O'Sullivan, Thompson

and Drummond, 2005). As Buxton et al. (1997) point out, even pragmatic trials cannot extrapolate from intermediate to final endpoints in certain cases, such as chronic conditions.

Seventh, the appropriate size of a clinical trial is still debated. Despite the extensive work that has been carried out in power estimation, clinical trials are still rarely large enough to accurately measure infrequent outcomes and rare adverse events, or assess long-term consequences (Glick et al., 2007). Evidence from insufficiently large clinical trials can be misleading both for the size and the direction of the treatment effect on outcomes, with prominent studies published in prestigious journals sometimes reaching conclusions that were not in agreement with large observational studies reflecting clinical practice (Collins and MacMahon, 2001; Ioannidis, 2005). In addition, as noted earlier, cost-effectiveness studies should ideally be additionally powered to measure treatment effects in the economic outcomes (Briggs and Gray, 1998). This is because variability in economic endpoints of costs and health-related quality of life is often greater than the variability in clinical endpoints (Drummond, 1998). However, decisions regarding the sample size necessary to measure treatment effects in clinical trials tend to be based on the clinical effectiveness of interventions, rather than their cost-effectiveness (Briggs, 2000). Consequently, the sample size of a trial will often be insufficient to detect significant differences in economic endpoints (Fayers and Hand, 1997). The costs and effectiveness for the majority of health care interventions evaluated will also vary considerably because of the heterogeneous nature of a population. Warren and Normand (2004) point out that in such cases

although the intervention may not be cost-effective across the entire population, it is likely to be cost-effective for part of this population. The same authors also argue that trial-based economic evaluations are typically inadequately powered to explore individual heterogeneity in subpopulations and thus take into consideration heterogeneous effects on outcomes.

1.3.7 Observational studies: the pragmatic alternative

In observational studies, the analyst does not have control over individuals' exposure status. Instead, the characteristics of the intervention and outcome between the exposed and comparison groups are merely observed.

Cohort studies constitute a type of analytic observational study in which individuals are selected in terms of their exposure to an intervention and then are observed over a period of time, with a view to compare the development of their outcomes (MacMahon and Collins, 2001). Their time horizon can range from a few hours, as was the case with the Birthplace cohort study that was used for the empirical part of this thesis (Hollowell et. al, 2011), to several years, in the case of large longitudinal cohort studies such as the one undertaken in the context of the Women's Health Initiative (Prentice et al., 1998). Cohort studies can be distinguished in prospective and retrospective designs, with the distinction based on whether the outcomes of interest and exposure of individuals to the intervention(s) under investigation existed when the study was started. According

to Hennekens and Buring (1987), in prospective designs the outcomes have not occurred at the time the study began, although the exposure(s) could or could not have happened, while in retrospective designs, both the outcomes, as well as the exposure(s) have already happened when the study started. Prospective designs have the advantage that collect most relevant information and hence have the potential to minimise selection biases (Delgado-Rodríguez and Llorca, 2004). Nevertheless, they may be proven very costly since they often involve the recruitment of large number of individuals, which sometimes may need to be followed-up for many years in order to observe the development of outcome(s) under investigation (Sica, 2006).

During the last decade an increasing number of retrospective observational studies have been using data from electronic sources (Shi et al., 2007). Routinely collected data in administrative databases and electronic health records have evolved in terms of quality, increasingly including richer and more complex information (Tannen, Weiner and Xie, 2009). Routinely collected data share a set of characteristics, such as comprehensive and obligatory regular continuous or periodic collection; use of standard definitions for the whole population group of interest; and collection at national or regional level, including more than one centre depending on the representativeness of the sample (Raftery, Roderick, and Stevens 2005). For example, a wealth of data is routinely collected for administrative purposes as part of health insurance arrangements. These data are used extensively in observational research and administrative datasets employed by analysts often include tax records, reimbursement and claims databases, as well as

registries of birth, death and cancer cases (Meenan et al., 2002). Electronic health records also appear to gain momentum since they constitute fully functional information systems, which usually contain data that are better suited for research purposes (Friedman, 2006).

Routinely collected data of any form can constitute a valuable source of information that may be used in population health assessment, clinical audit studies, analytical epidemiological studies, and of course the evaluation of health care interventions. Bain, Chalmers and Brewster (1997) point out that routinely collected data offer some distinct advantages over other sources of data. These comprise of relatively inexpensive and less time-consuming collection, coverage of larger populations consisting of millions rather than thousands of observations and prolonged periods of time. Administrative data in particular offer improved availability of data to analysts, enhanced provision of information relating to hard-to-reach and socially disadvantaged groups of the population, as well as reduced vulnerability to reporting bias (Jones, 2011). Electronic health records on the other hand offer better standardisation of data collection, provision of expanded clinical detail on a spectrum of conditions and the ability to update information in real time (Sequist and Bates, 2009). Birnbaum et al. (1999) point out that even when the available routinely collected data are not adequate on their own, they can still form an important base with which other data can be linked using personal identification numbers. In this way, different data sources can be combined to produce a comprehensive dataset, which can contain all the variables that are of interest to analysts. It is not surprising therefore that the use of routinely collected

data to conduct observational studies for health technology assessment purposes is increasingly being seen as a way of reducing current dependence on clinical trials (Williams et al. 2003).

At this point, it should be noted that observational data are also subject to several of the general problems that often plague microdata collected as part of randomised controlled trials. For example, Hennekens and Buring (1987) point out that data from cohort studies may be prone to bias associated with lack or loss of follow-up that may happen either after following individuals for many years or because of late entry. This problem is also commonly referred to as censoring (Willan and Briggs, 2006; Glick et al., 2007). On the other hand, despite the fact that the quality of data routinely collected in electronic health records and administrative databases appear to increasingly improve, these are typically not being generated specifically for research purposes (Meenan et al., 2002). As such, they can be subject to considerable missing values and measurement errors, with the former often being a consequence either of a lack of collection or a lack of documentation (Wells et al., 2013), while the latter usually resulting from errors in the coding of variables, as well as the inaccurate collection of information that leads to overstating or understating the true value of the measurement (Cox and Koutroumanos, 2010).

1.4 Concluding remarks

Although some of the problems faced in clinical trials can also be present in studies using observational microdata, such studies have more potential to satisfy the particular requirements that the cost-effectiveness evaluative framework imposes. Indeed, the absence of randomised evidence for each and every intervention available de facto means that the use of observational data will often constitute the only alternative (Buxton et al., 1997). Analysts can exploit the wealth of the available information in observational datasets to inform an economic evaluation, potentially increasing in this way the external validity of the analysis. In addition, these data may be used retrospectively to devise more flexible study designs, which can compare the value of a broad range of interventions, rendering possible in this way the investigation of a variety of research questions. The availability of large sample sizes which is more often a feature of observational datasets may also render feasible the reliable assessment of the cost-effectiveness of competing interventions in different subpopulations and consequently the exploration of heterogeneous effects. Finally, the use of large longitudinal observational datasets may allow the analyst to take into consideration a longer time horizon and potentially explore the lifetime cost-effectiveness of health care interventions without relying on summary evidence (Cutler, 2007).

CHAPTER TWO

LITERATURE REVIEW

2.1 Introduction

The present chapter aims to provide a conceptual overview of how different non-experimental methods attempt to solve the evaluation problem in the presence of non-random selection of individuals into treatment. In addition, the chapter reviews and critically appraises the use of such analytic approaches in the economic evaluation literature. The chapter is organised as follows. The next section introduces the econometric methods that can be used for impact evaluation, when observational microdata are available. Section 3 identifies which of these methods are currently used in published economic evaluation studies employing observational microdata. Section 4 elucidates issues surrounding the use

of these methods within the cost-effectiveness evaluative framework. Section 5 summarises the findings of the review and presents some concluding remarks.

2.2 Non-experimental evaluation methods

Over the years, a number of econometric methods that deal with selection bias have been employed in the programme evaluation literature, depending on whether the source of bias is observed or not (Imbens and Wooldridge, 2009). Analytical approaches include matching, regression analysis, propensity scores, instrumental variables, regression discontinuity designs, difference-in-differences and control functions. The key idea behind these methods is the construction of the counterfactual outcome, when the evaluation problem is the measurement of a treatment effect in the presence of non-random selection into treatment (Jones, 2011). Non-experimental evaluation methods have the potential to estimate a single average effect, or look into the heterogeneity of individuals' responses to the intervention of interest, depending on the nature of the research question, the richness and type of the available data, as well as the postulated model for outcome and selection processes (Blundell and Costa Dias, 2009).

Detailed technical exposition of these methods and examples of their applications in a range of fields can be found elsewhere (Blundell and Costa Dias, 2009; Imbens and Wooldridge, 2009; Jones, 2011; Heckman, 2010; Jones and Rice, 2011). Below, a brief overview of these methods is given, with Table 1

summarising the key advantages and disadvantages of each method, as well as the principal assumptions required to hold true for identification.

2.2.1 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) (Gujarati, 2003). What is typically estimated in this form of analysis is the conditional expectation, that is, the average value of the outcome variable when all but one independent variable is held constant. In addition, in linear regression models, the coefficient on the explanatory variable can also be thought of as representing its marginal effect 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. This method will allow unbiased estimation of treatment effects on health outcomes and costs, only if the model is correctly specified and all confounding factors are measured. Furthermore, by including interaction terms and/or by employing a non-linear model structure, the analyst can explore the influence of heterogeneous treatment effects.

Regression methods have also been extended to address selection bias in the estimation of treatment effects when using longitudinal or panel data (Jones, 2011). Fixed and random effects estimators can cope with individual heterogeneity arising from unmeasured covariates, when this heterogeneity is time-invariant and

potentially correlated with independent variables. Fixed and random effects estimators differ in the assumptions that they impose about the individual specific effect; the former allows for the individual specific effect to be potentially correlated with the independent variables, while the latter assumes that these effects are random variables distributed independently of the other covariates. Choice between the two estimators can be guided by the Hausman specification test (Hausman, 1978); if the assumption that the individual specific effect is uncorrelated with the independent variables, the random effects estimator should be preferred as it is more efficient.

2.2.2 Matching

The objective of this category of methods is to make comparable (i.e. balance) the distributions of confounders in the exposed and comparison groups (Todd, 2008). This is achieved by artificially creating groups that share the same characteristics, through the pairing of an individual in the exposed group, with another in the comparison group. The pairing of individuals from the different groups can be done using a range of observed confounding variables, including personal characteristics, disease risk factors, or environmental influences. Matching methods typically comprise of both exact, as well as inexact approaches, each of which can be used as a non-parametric standalone solution, or can be used as a pre-processing method before parametric analysis (Imbens, 2004).

Different matching estimators essentially reflect different weighting techniques for constructing the matched counterfactual outcome for each exposed individual (Morgan and Harding, 2006). These include exact matching, inverse probability weighting, nearest neighbour matching (with or without replacement), caliper/radius matching, stratification, as well as kernel and local linear matching. The use of a number of other matching methods such as optimal and full matching, coarsened exact matching, entropy balancing and genetic matching, has also been exemplified (Stuart, 2010). For inexact approaches, different metrics have been proposed in measuring the proximity of similar individuals, with popular choices including the Mahalanobis distance or the difference in propensity scores (Zhao, 2004). Matching can be used with cross-sectional data, as well as longitudinal data if combined with difference-in-differences (Heckman, Ichimura and Todd, 1997).

2.2.3 Propensity score analysis

Propensity score analysis employs different analytical methods into a two-step approach. First, the propensity score is obtained, that is the probability of an individual being assigned to the exposed group instead of the comparator, conditional on a set of observed covariates (Rosenbaum and Rubin, 1983). It is usually estimated by means of logistic or probit regression, with the observed covariates being the predictors and receipt of treatment being the dependent variable. The estimated propensity scores are subsequently used either with regression analysis, or matching methods (Imbens, 2004).

2.2.4 Difference-in-differences

Difference-in-differences is an analytic approach the implementation of which depends on the availability of data for at least two time periods, as well as for an exposed and a comparison group (Meyer, 1995; Heckman, Ichimura, Todd, 1997; Abadie, 2005). This method estimates the effect of an intervention on outcomes by calculating a double difference; one across the groups of exposed and comparison individuals, and one over time between a pre-intervention and a post-intervention period. As such, it has the potential to remove effects both due to time-invariant individual characteristics, as well as effects arising from other unobserved factors and processes that may develop over time.

Difference-in-differences is a simple and easily implemented method, which however can only identify causal effects if the analyst invokes a parallel paths assumption. This assumption postulates that confounding factors present a common time trend over the course of the treatment, and both exposed and comparison groups do not exhibit systematic composition changes. In this way, it is ensured that the counterfactual trend for both the exposed and comparison individuals is the same. The parallel path assumption can only be relaxed if rich microdata are available. In such a case, the analyst can use a regression model to condition on available confounding covariates.

2.2.5 Instrumental variables

The use of instrumental variables to estimate causal effects in the context of the potential outcomes framework has also been established recently (Imbens and Angrist, 1994; Angrist, Imbens, Rubin, 1996; Heckman, 1997; Staiger and Stock, 1997). Instrumental variables estimate average causal effects through a two-stage approach that uses variables commonly referred to as instruments, which have a strong effect on treatment assignment and influence outcomes only through the treatment. More specifically, the chosen estimator first predicts the value of the potentially endogenous independent variable as a function of the instrument(s) and other covariates. The predicted value is subsequently used as a covariate in the outcome model to estimate the magnitude of the effect on the outcome. Most models that are linear in parameters are typically estimated using two-stage least squares, although estimation through generalised methods of moments (GMM) is generally more efficient and can be particularly advantageous in large samples when heteroskedasticity is present (Baum, Schaffer and Stillman, 2003). Whether instrumental variable approaches give unbiased treatment effect estimates, largely depends on the extent to which the instrument used is relevant and valid, that is, strongly correlated with the endogenous regressor and at the same time orthogonal to the errors.

In recent years, there has been a growing interest in instrumental variable analyses using genetic instruments (Wehby, Ohsfeldt and Murray, 2008). This approach, which is also known as Mendelian randomisation, exploits the random assignment

Table 2.1 – Overview of evaluation methods

Method		Description	Main Assumptions	Advantages	Disadvantages	
Matching	*Interval*	Individuals are grouped in strata according to the same level of the identified confounders. Analysis is then carried out in each stratum within which the confounding variables remain constant and summarised across the strata.	Unconfoundedness: all appropriate confounding covariates are observed and used. Common Support: the subspace of individual characteristics that is represented both among exposed and comparison groups.	Creation of strata is an uncomplicated procedure. Straightforward calculation of average treatment effects in homogenous groups. No parametric assumptions necessary. Often five strata can remove most bias.	Confounding from unobserved sources of bias invalidates results. Stratifying individuals into a large number of groups can be infeasible and potentially limit the interpretation of the results obtained.	Can recover ATE, ATT and ATNT depending on implementation approach.
	Exact	Creates exposure-comparison groups that share identical characteristics. This is achieved by pairing individuals in the groups according to a number of confounders.		Allows direct control of confounders in an intuitive and straightforward implementation. Free of parametric assumptions and thus more robust than ordinary least squares regression. Coarsened exact matching is a variant that adds more flexibility by coarsening the data first and then exact matches on these coarsened data.	May lead to regression toward the mean and overmatching. Confounding can still arise from unobserved sources of bias. Dimensionality problem: Difficult to find (exact) matches, when the number of potential confounders is large. Greedy methods can limit statistical power considerably. Assessing the impact of matched variables on outcomes is not possible.	
	Inexact	Pairing process of all the comparison individuals that are sufficiently "close" to a given exposed counterpart. Distance of individuals is determined in terms of their vector of observed covariates.		More efficient approach that can be used when exact matching is not feasible. Very flexible. Can be used with different distance measures, sampling with or without replacement, multiple neighbours, as well as within a maximum tolerated distance.		
	Adaptive	An example is Genetic matching, which uses an evolutionary search algorithm to iteratively check and improve covariate balance.		Contains propensity score and Mahalanobis matching as special cases. The algorithm automatically converges to the most appropriate distance measure. The loss function that the algorithm minimises can include many forms of imbalance.	Similarly to other matching methods it is vulnerable to hidden bias. The analyst must specify in advance the loss function that measures covariate balance.	

Method		Description	Main Assumptions	Advantages	Disadvantages	
Regression analysis	*Cross-sectional data*	Evaluates the relationships between the outcome and treatment by quantifying the level of change of outcome resulting from treatment.	Unconfoundedness: all appropriate confounding variables are observed and used. Common Support: the subspace of individual characteristics that is represented both among treated and comparison groups. Additional assumptions for the model fitted.	Simple, well established in the literature and available in all software packages. Takes into account several covariates simultaneously and allows easy assessment of treatment effects from individual covariates, as well as interactions. Functional form/homogeneity assumptions testable.	Does not address selection on unobservables. Detection of interactions will largely depend on power and might be an artefact. Sensitive to parametric assumptions of the model. Assumes model is correctly specified.	Can recover ATE, ATT and ATNT.
	Longitudinal data		Fixed effects: individual specific effects are time-invariant and potentially correlated with the independent variables.	Can account for unobserved confounding variables. Choice between fixed and random effects estimators can be guided by the Hausman specification test.	Consistent but potentially inefficient. Unobserved confounders assumed to be time-invariant.	
			Random effects: individual specific effect is uncorrelated with the independent variables.		More efficient than fixed effects but potentially inconsistent.	
Propensity score analysis		A balancing score, which is the probability of an individual being assigned to the exposed group instead of being in the comparison, conditional on a set of observed confounders. Can be used with matching methods, as well as regression analysis.	Unconfoundedness: all appropriate confounding variables are observed. Common Support: the subspace of individual characteristics that is represented both among treated and comparison groups. Subject to further assumptions depending on whether it is combined with regression-based methods or matching.	Can potentially balance confounders between exposed and unexposed individuals more efficiently. No dimensionality problem: models outcome on the balancing score rather than all confounders. Flexible: depending on the implementation can require less restrictive (parametric) assumptions. Can recover ATE, ATT and ATNT depending on implementation approach. Plug-in programmes available in most software packages.	If treatment impact differs across exposed individuals, restricting to Common Support may change parameter being estimated and thus unable to identify ATT. Suffers from most limitations associated with matching. Use within a standard regression framework subject to the relevant model assumptions. Unclear which propensity score estimator gives the most efficient and least biased results.	

Method	Description	Main Assumptions	Advantages	Disadvantages
Difference-in-differences	A before and after design that uses an exposed and a comparison group to facilitate contrast between groups, by measuring the difference in the outcome before and after implementation of treatment for both groups.	Observed or unobserved confounding factors are fixed. Composition of exposed and unexposed groups presents a parallel path. Unconfoundedness and common support will be required if combined with other control strategies.	Popular method that accounts for time-invariant unobserved bias. Able to detect small treatment effects. Flexible: can be combined with other (non/semi) parametric methods to improve accuracy and relax some assumptions. Straightforward implementation with most software packages.	Not robust to unobserved temporary individual-specific components affecting treatment assignment. Requires longitudinal data or repeated cross-sections and can only recover ATT. Parallel path assumption will most often be implausible. Extensions and combinations with other methods to overcome the standard assumptions can complicate the analysis.
Instrumental Variables	Two-stage approach using instruments with strong effect on treatment assignment (relevant) but not correlated with the outcome of the untreated (valid), to estimate the magnitude of the effect on the outcome.	Instrument only influences outcomes through treatment and is independent of the unobserved confounders For heterogeneous effects monotonicity must also hold: The effect of the instrument on treatment status must be in the same direction for all individuals.	Deals with both observables and unobservables. Well-established method with many extensions and a large evidence base, particularly in economics. Can recover ATE, ATT and ATNT when homogenous effects. ATT and LATE when heterogeneous. Can exploit genetic variables as instruments to achieve randomisation. Conventional instrumental variable analysis available in many software packages.	Instruments satisfying the assumed properties are scarce. The validity assumption is not directly testable. When an instrument is weak, the estimates obtained are poor in terms of bias, standard errors and confidence intervals. Non-linear relationships can considerably complicate the analysis. With heterogeneous treatment effects, inference is limited only to individuals whose treatment status is affected by the instrumental variable (compliers). The instrumental variable estimate is only consistent not unbiased, that is, it performs well only in large samples.

Method	Description	Main Assumptions	Advantages	Disadvantages
Regression Discontinuity	Determines individuals' assignment to treatment by whether an exogenous "forcing" variable surpasses a cut-off point. It then identifies the treatment effect by comparing individual values before and after the cut-off point.	Unconfoundedness (for sharp design): all appropriate confounding covariates are observed and used. Continuity: individuals just above and below the cut-off need to be comparable, requiring them to have similar average potential outcomes independent of treatment receipt. No precise manipulation of the forcing variable around the cut-off point.	Has the potential to accommodate unobservable sources of bias. Conceptually straightforward since assignment to treatment solely based on a cut-off point with treatment effect measured by the value of the discontinuity. Can exploit graphical methods to identify any discontinuity. Versatile treatment assignment: Sharp (deterministic) and fuzzy (stochastic) designs. The former can recover ATE and the latter ATT and LATE. High degree of internal validity where applicable.	Relatively unknown to researchers and evidence on its performance limited. Continuity assumption not formally testable. Offers low external validity. Good data must be available in the neighbourhood around the discontinuity. Other changes at the same cut-off point may potentially affect the outcome. The design may be invalidated if the forcing variable is exactly manipulated.
Control Functions	A two-stage technique that controls directly for the part of the error term in the outcome equation that is correlated with the treatment indicator, using an explicit model of the treatment assignment process.	The control function is a function of observed confounders as well as treatment assignment and should be identified. Once one conditions on the control function the endogeneity problem is resolved.	Deals with unobserved confounding variables. Generalised to deal with selection problems in many different settings. Can recover ATE, ATT and ATNT. Estimates correlation between unobservables in the two equations and can be extended to multiple treatments. Can be used for random coefficient models (i.e. models where unobserved heterogeneity interacts with endogenous explanatory variables). Straightforward estimation of certain non-linear models with endogenous explanatory variables Available in most standard software packages.	Effort to identify more sources of selection requires the introduction of more structural assumptions. Distributional assumptions are required when parametric and semi-parametric implementation approaches are used. Requires exclusion restrictions (i.e. no direct effect of the instrument on the dependent variable or any effect going through omitted variables) in order to obtain convincing estimates in the absence of functional form assumptions.

ATE: Average Treatment Effect, ATT: Average Treatment Effect on the Treated, ATNT: Average Treatment Effect on the Non-Treated, LATE: Local Average Treatment Effect

of an individual's genotype from the parental genotypes to make causal inferences (Lawlor et al., 2008).

2.2.6 Regression discontinuity

Regression discontinuity is a versatile method that recently has begun receiving increasing attention (Hahn, Todd and Van der Klauuw, 2001; Imbens and Lemieux, 2008; Lee and Lemieux, 2010). The key idea behind this method is that treatment assignment is determined by whether an exogenous "forcing" variable surpasses a specified threshold. The comparison of individuals with values marginally below the threshold, to the individuals marginally above it, is used to identify the treatment effect. As long as the forcing variable is not exactly manipulated, treatment variation near the threshold is thought to be similar to a randomised experiment. A central assumption required to hold true for identification using this method is continuity. This assumption postulates that the individuals slightly above and below the threshold need to be comparable, with similar average potential outcomes independent of treatment receipt.

Regression discontinuity designs can be distinguished in "sharp" and "fuzzy". The "sharp" design has a selection on observables interpretation since it handles allocation to treatment deterministically, with the discontinuity precisely determining treatment. In the "fuzzy" design, assignment to treatment is stochastic, with the

discontinuity being highly correlated with treatment and the assignment used as an instrumental variable.

2.2.7 Control functions

Control functions represent a more broad two-stage technique that is used by economists to adjust for bias arising from sample selection and endogeneity problems (Garen, 1984; Wooldridge, 1997). Control functions incorporate the treatment assignment mechanism in the estimation process by controlling directly for the part of the error term in the outcome equation that is correlated with the treatment indicator. This is achieved using an explicit model of the treatment assignment process. A widely used estimator, which is seen as a particular case of this method, is the Heckman correction estimator (Heckman, 1979).

The control function approach is the closest to the structural approach, which estimates jointly the outcome and treatment of interest (Reiss and Wolack, 2007). It can rely on the same identification conditions with instrumental variables and as such, in the standard case where an endogenous explanatory variable is linear, the control function approach leads to the traditional two-stage least squares estimator. Nevertheless, the second stage need not be necessarily a linear regression, but can also be probit, logit, poisson or any generalised linear model. Latest developments in control function methodology have considered semi-parametric and non-parametric approaches relaxing distributional assumptions (Blundell and Powell, 2001; 2004).

2.3 Review of the economic evaluation literature

2.3.1 Eligibility criteria & identification strategy

A review of the international English language literature was undertaken. The eligibility criteria for inclusion required studies to be full economic evaluations as defined by Drummond et al. (2005) and use observational microdata referring to the same population for both the cost and effectiveness outcome. The review placed particular emphasis on identifying methods than including all applications. As such, only studies that demonstrated some modification in the methodology addressing selection bias were included. In addition, studies should use an econometric method to adjust at least one of the cost, effectiveness and cost-effectiveness outcomes.

The studies reviewed were identified using a four-stage process. First, three generic electronic bibliographic databases, namely MEDLINE, EMBASE, and Econlit were searched through the OvidSP interface in order to generate as many papers of potential methodological interest as possible for the years 1990-2010. Second, additional searches for the same time period were carried out in three specialised databases: NHS EED, HEED and CEA Registry. The search strategy, which was adapted for each database, combined and interacted the terms "cost*", "effect*", "benefit*", "cost-effective*" and "cost-benefit*", with "matching", "stratification", "regression*", "propensity score*", "instrumental variable*", "difference in difference*", "control function" and "discontinuity". Third, the database searching

was supplemented by communicating with other experts. The expert communication involved sending a brief outline of the review objectives, together with a list of key publications to individuals working in similar research areas. Colleagues were requested to suggest further published, unpublished or work-in-progress research for inclusion in the review. Experts were identified through the literature, known contacts and posting on relevant online discussion lists. Fifth, an examination of the references and the citations of all eligible studies was undertaken, with a view to identify further papers that were not already captured during the previous stages.

2.3.2 Review process

In epidemiology, a checklist of items that should be included in studies reporting observational research has been established (von Elm et al., 2007). This initiative aims to help analysts better assess the strengths and weaknesses of the research findings reported in the medical literature by improving the quality of reporting in studies employing observational data. In the absence of a similar methodological inquiry for economic evaluation, all papers were reviewed using a custom-made structured template, the development of which was informed by the conceptual review. A copy of this template is presented in Table A2 for illustrative purposes. The structured template aimed to extract factual information from each study in a comprehensive and rigorous manner, as well as critically appraise important aspects of their

methodology. It comprised of three parts: "General Information", "Analytical Approaches" and "Reviewer's Assessment and Comments".

More specifically, the first part recorded general characteristics such as bibliographic information, the type of economic evaluation undertaken, whether a summary cost-effectiveness outcome measure was used, as well as the interventions evaluated. In the second part, the template focused on extracting the method(s) adjusting for selection bias, the estimation techniques employed and whether adjustment was undertaken for costs, effectiveness or cost-effectiveness. Any comparisons with other methods or studies, the types of uncertainty evaluated and the authors' conclusions with respect to the ability of methods to adjust for selection bias were also extracted. Finally, the third part recorded information relating to the justification of choice of method and the specification(s) used, as well as whether any relevant tests or graphical analyses were carried out. A reviewer's assessment concerning potential weakness of the study was also included. This was based on the extracted information as these were provided by the authors and placed particular emphasis on assessing the plausibility of the assumptions postulated by the analytical method employed.

2.3.3 Results

A schematic diagram of the overall identification and review process is presented in Figure 2.1. The original search strategy yielded a total of 6,647 unique studies. Additional independent search strategies in the three specialised databases returned

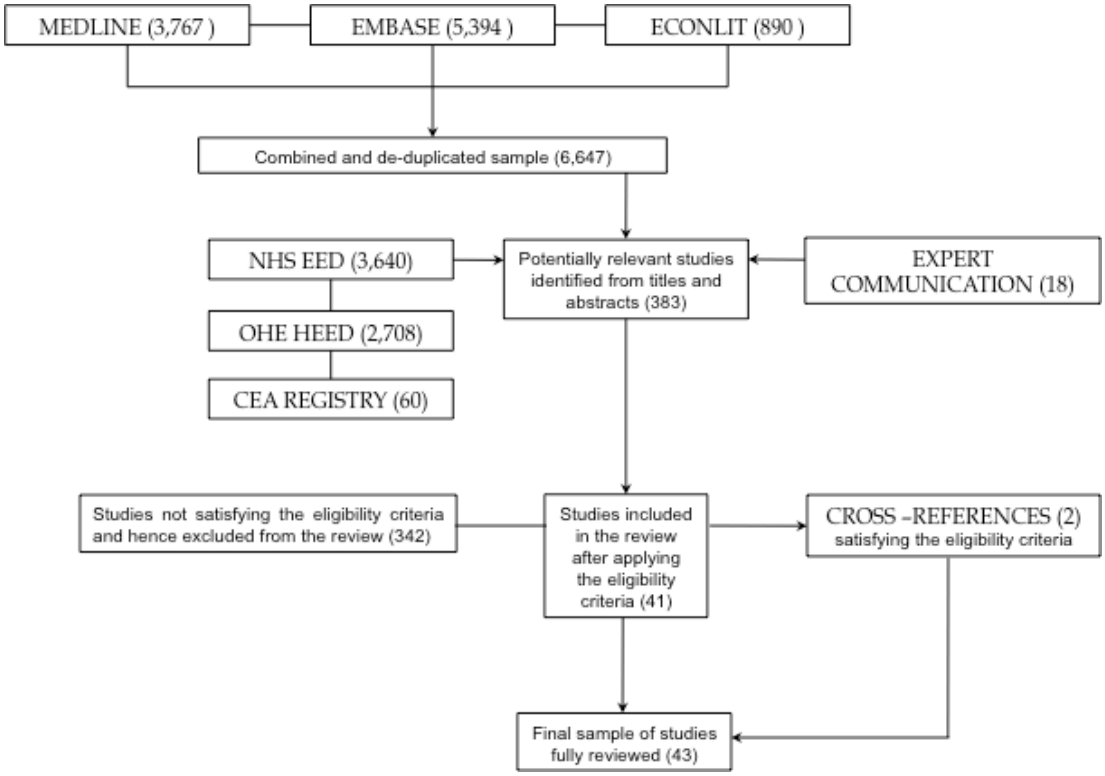


Figure 2.1 – Schematic diagram of the identification strategy and review process.

3,640 studies in NHS EED, 2,708 in HEED and 60 in the CEA Registry. Requests for studies from other experts yielded a further 18 studies. All studies underwent a screening process to ensure that they met the eligibility criteria of the review. It should be noted that when it was apparent from the title or abstract that a study failed on any of these criteria, it was discarded. When it was unclear or if any doubt remained, the full paper was examined. Following the review of titles and abstracts, full text copies were obtained for 383 potentially relevant studies. After assessment, 342 were excluded from the review because they did not meet the eligibility criteria. Cross-reference checks of the selected studies yielded another 2 relevant studies that

satisfied the eligibility criteria. The final sample of the studies fully reviewed using an equivalent number of structured templates comprised of 43 studies.

Table 2.2 provides a summary of the results of the review. As it can be seen, around a third of the economic evaluations included in the review did not report a summary outcome measure, whereas half of the studies failed to report in a precise and transparent manner information relating to sampling uncertainty for cost-effectiveness. A relatively small percentage of studies evaluated multiple interventions but the majority did not consider explicitly the issue of how these should be handled in the analysis and relied on pairwise comparisons of interventions. In terms of analytical methods currently employed to adjust for selection bias, the review identified five broad categories, which were mostly applied in sample sizes of 5000 individuals or less. These include different methods matching on individual covariates, some form of regression analysis using cross-sectional or longitudinal data; propensity score analysis either through matching, or regression modelling using the propensity score as a covariate; difference-in-differences approaches; and instrumental variables analysis. Solutions based on the propensity score dominated the sample of the reviewed studies. The majority of studies failed to adequately assess the assumptions postulated by each method. For example, studies relying on matching, regression and propensity score analysis usually justified the selection on observables assumption by providing a simple description for the confounders adjusting. Overlap between treatment groups was explored mostly through standard tests that assess heterogeneity, although in studies where regression analysis was employed, balance between groups was sometimes investigated using models that

Table 2.2 – Main characteristics of reviewed studies

Type of study

Cost-effectiveness analysis (70%) Cost-Utility Analysis (30%)

Type of journal

Statistics/Econometrics (7%)	Health Economics (19%)	Health Services (30%)	Medical (39%)	Working Paper (5%)

Year of publication

1990 – 2000 (14%) 2001 – 2010 (86%)

Type of intervention

Surgical (37%) Medical (33%) Rehabilitation (5%) Public Health Policy (14%)

Diagnostic (2%) Preventative (9%)

Number of interventions

Two (74%) Three or more (26%)

Sample size

100 – 1000 (33%) 1001 – 5000 (29%) 5001+ (33%) Not reported (5%)

Summary outcome

None (35%) Incremental Cost-Effectiveness Ratio (51%) Net Benefit (14%)

Evaluation of uncertainty on summary outcome

Yes (43%) No (6%) Partial/Unclear (51%)

Method addressing selection bias

Regression Analysis (28%)	Covariate Matching (7%)	Propensity Scores (49%)	Instrumental Variables (7%)	Difference-in-Differences (9%)

Assessment of methods' assumptions

Detailed (9%) Partial or None (91%)

Comparison of analytical methods

Yes (23%) No (77%)

Effort to contrast findings with other studies

Yes (47%) No (53%)

assessed whether significant interactions were present between treatment and covariates. Studies employing some form of matching assessed covariate balance post-matching by comparing means in the resulting groups. Economic evaluations relying on difference-in-differences approaches typically assessed the parallel path assumption by comparing pre-existing time trends, whereas analyses exploiting the use of instruments to achieve quasi-randomisation mostly assessed the relevance of the instrument but not its validity. Studies using parametric models rarely assessed functional form assumptions through formal statistical tests and no studies employed any graphical analysis for visual inspection. Finally, almost half of the reviewed studies attempted to contrast their findings with those obtained from other studies, while a small percentage directly compared the results that different methods produced.

2.4 Discussion

Economic evaluations often employ observational data. Econometric methods can only adjust for selection bias by relying on assumptions that should approximately be met. Although these assumptions are mostly untestable, their credibility in a particular setting can be assessed and analysts engaging in applied economic evaluation should always undertake extensive checking procedures to confirm the robustness of their results. In fact, this process may often require more effort than the estimation of the treatment effect itself.

The review revealed that this is typically not the case. The published economic evaluation literature currently routinely applies econometric methodology without carefully considering whether the key assumptions under which these methods operate hold. For example, reviewed studies making use of methods that assume selection on observables rarely used findings from prior research or expert advice to establish the causal pathway between interventions and outcomes. These studies also did not report the conduct of observational 'placebo' tests (Sekhon, 2009), or sensitivity analyses such as Rosenbaum's bounds simulating the likely presence and impact of unobserved confounders (Rosenbaum, 2002). Similarly, alternative specifications considering varying sets of covariates or different functional forms in regression models were only reported in the few studies published in statistical and health economics journals, aiming to offer a more methodological treatment of the evaluation. In addition, in studies where individuals between treatment groups are not comparable, parametric methods may extrapolate beyond the support provided by the available data (Sekhon and Grieve, 2009). The literature currently is dominated by studies that do not adequately assess the 'common support' assumption. For instance, although no statistically significant interactions between treatment and covariates may be identified, imbalances may still remain (Mihaylova et al., 2010). Graphical analysis such as histograms or smoothed density plots can be very effective in detecting areas of insufficient overlap and can complement statistical tests.

Matching methods can act as a data pre-processing stage before inference (Ho et al., 2007). This approach allows the analyst to consider problems of imbalance in individual covariates and assess overlap between groups in a direct and explicit

manner, both before and after adjustment (Stuart, 2010). Authors of economic evaluations employing some form of matching often failed to report details regarding the type of matching performed and generally did not consider different methods in their analyses. Alternative matching procedures, as well as different values of key estimator parameters can have important implications with regards to the bias-efficiency trade-off inherent in all these methods and may also result in varying interpretations of the treatment effect reported (Imbens, 2004). Studies using matching reported balance tests relying on mean differences, as well as standardised differences. The latter will typically be preferred as they are not affected by sample size and therefore can be used for comparisons between treatment groups that contain different numbers of individuals (Flury and Riedwyl, 1986). Nevertheless, solutions relying on the propensity score require post-match balance in the entire distribution of individual covariates. As such, higher moments including variance, skewness, and kurtosis, as well as cross-moments such as the covariance, should ideally be examined (Ho et al., 2007). This was something that nearly all relevant studies failed to report. For continuous covariates, graphical analyses using quantile–quantile plots and side-by-side boxplots, or Kolmogorov–Smirnoff non-parametric tests of the equality of distributions can consider the full covariate distribution and thus be more informative (Austin, 2009b; Sekhon and Grieve, 2009).

For studies relying on exclusion restrictions to account for unobservables, it is crucial for the analyst to determine whether these are relevant and valid in a particular setting (Blundell and Costa Dias, 2009). In a purely statistical context, random assignment in a trial meets all the required assumptions and as such it is an exclusion

restriction by definition. In contrast, in observational studies, particularly in socially behavioural settings, choice of convincing exclusion restrictions will often be less straightforward and will have to be justified on qualitative rather than empirical grounds (Heckman, 2008). For economic decision-making, strong exclusion restrictions will require the analyst to go through the challenging process of crafting plausible natural experiments, which exploit extensive demand and supply side information to construct variables that can induce strong external variation in the treatment assignment of individuals (Angrist, Imbens and Rubin, 1996). In their quest for finding an instrument that satisfies the assumed properties laid out above, studies included in this review often employed a well-known tactic in economics, which involves exploiting the use of geographical variables. Nevertheless, the externality of an instrument does not necessarily also assure exogeneity; that is it does not automatically fulfill the orthogonality condition required for consistent estimation in the instrumental variable context (Deaton, 2010). Indeed, the economic evaluation by Polsky and Basu (2006) should act as a reminder that the performance of an instrument will not always be guaranteed.

At this point, it is important to stress that the application of different methods will ultimately depend on the availability of data. Econometric methods are all data driven, being applicable only in situations where relevant microdata can be accessed to support them and as long as the analysis takes advantage of their availability. Administrative data can potentially provide the analyst with the ability to link information from multiple databases creating datasets containing more complete data on individuals over time, additional background and demographic variables, as well as

data on participants and non-participants (Hotz et al., 1997). In addition, routinely collected information is increasingly shifting from data related to processes of care and patient outcomes such as mortality and morbidity, to data related to more complex measures of health status (Hutchings, 2005). These considerations can expand the range of non-experimental methods that can be used to measure treatment effects.

A key aspect of the review appraisal was to consider whether a comparison of cost-effectiveness estimates with existing evidence was attempted, or whether studies explored the sensitivity of their results to alternative methods. The motivation for the former rests on the fact that estimates from other relevant studies, when available, can potentially offer a prior indication regarding the direction of the treatment effect. This is particularly true for evidence generated from randomised trials, which in principle can constitute an important benchmark for learning about non-experimental methods (LaLonde, 1986; Smith and Todd, 2005). Unfortunately, when comparisons of this kind were attempted in the studies reviewed, these were mostly qualitative in nature, relating to overall conclusions or comparing only certain outcomes such as costs, survival or hospitalisations. Some economic evaluations restricted the scope of their comparisons to those across methods. Such comparisons can also act as sensitivity analysis when the availability of data allows the use of alternative analytical approaches, which rely on different assumptions and have the potential to exhibit variable performance in different settings. For example, in the econometrics literature, choice among estimators that rely on the selection on observables assumption is normally warranted on small sample arguments (Imbens and

Wooldridge, 2009). Currently, the embryonic nature of such evidence in the studies reviewed does not allow any firm conclusions to be drawn regarding the relative ability of different methods or their combinations to reduce selection bias in the context of cost-effectiveness. However, what is clear is that choice of method may not only influence estimates, but can also fundamentally alter conclusions (Polsky and Basu, 2006).

In addition, no economic evaluations identified by this review employed any 'doubly robust' approaches, which typically involve the use of regression analysis in combination with some form of weighting. For example, Robins and colleagues (1994) proposed the use of the inverse propensity score to weigh a regression model, offering in this way additional protection against misspecification. More recently, doubly robust estimation has been extended to instrumental variable analysis (Uysal, 2011; Okui et al., 2012). Another strand of this type of research that gets increasing attention in the econometrics literature is the use of regression analysis after matching. This is a 'bias-correction' solution that has been shown to correct for remaining finite sample bias, while potentially also making violations of functional form assumptions less consequential (Abadie and Imbens, 2011; Iacus, King and Porro, 2011). Given the greater potential for misspecification that arises from the consideration of economic and clinical endpoints in the analysis, the development of such approaches for evaluating cost-effectiveness and their comparison with standalone solutions is highly desirable.

In economic evaluation for decision-making, three additional issues merit attention (Drummond et al., 2005). First, incremental costs and effectiveness should be combined in a summary outcome measure. Second, the analyst must quantify and evaluate the sampling variability in this cost-effectiveness estimate. Third, the analysis must ideally take into consideration all relevant comparators. The review revealed that a number of studies did not combine incremental costs and effectiveness. Summary outcome measures are used by decision-makers to help make policy recommendations on the allocation of resources for competing health care interventions (National Institute for Health and Clinical Excellence, 2008). In the absence of a summary outcome measure, evaluating sampling uncertainty for the purpose of cost-effectiveness will not be possible. In addition, there seems to be a lack of transparency in the reporting of such information. For example, reviewed studies failed to report which bootstrap method was used to construct the reported confidence intervals, or did not provide any justification for the number of replications employed. The use of multiple comparators in an economic evaluation also raises the question of how these should be handled in the econometric analysis. Some studies identified by the review have shown that in regression analysis the use of multinomial choice models can act as an alternative to pairwise comparisons of interventions. Although these approaches have also been exemplified in the econometrics literature for propensity score matching (Lechner, 2001) and doubly robust methods (Uysal, 2012), no such extensions in the context of cost-effectiveness were identified.

This review is subject to certain caveats, which must be acknowledged. First, the conclusions of this review do not apply to all economic evaluations that use observational data. Decision analytical modelling-based studies, as well as studies employing hypothetical data or summary evidence for costs and effectiveness were considered beyond the scope of this review and were excluded. In addition, methods dealing with issues relevant to missing and censored data were also not included. Second, the review should not be considered an exhaustive investigation of the applied economic evaluation literature employing observational microdata. Nevertheless, a four-stage identification process ensured that as many studies as possible exemplifying modifications of analytical approaches were captured. Finally, it should also be acknowledged that only one reviewer carried out the review of studies. As such, although a structured template was used in an attempt to streamline the review process and render the appraisal of studies more rigorous, the categorisation of the collected information and the interpretation of the findings presented here may be subject to a certain degree of subjectivity.

2.5 Concluding remarks

Estimation of treatment effects in economic evaluation involves considerable challenges when observational data are used. Current limitations in substance cost-effectiveness analysis include inadequate assessment of the credibility of fundamental assumptions; absence of good quality evidence regarding the sensitivity of results to

different analytical approaches or variations in crucial estimator parameters; failure to combine incremental costs and effectiveness in a summary outcome measure; and no consideration of sampling uncertainty for the purpose of evaluating cost-effectiveness. At this point, it is worth noting that Kreif and colleagues (2013) also undertook a similar literature review. Their study was carried out concurrently with the review presented in this chapter and employed a checklist with explicit pre-specified criteria in order to appraise the quality of the applied cost-effectiveness literature. The conclusion of that independent study is in agreement with a key outcome of this structured review; the majority of economic evaluations employing observational microdata rarely assess the plausibility of the main assumptions postulated by analytical methods.

In light of the findings, future work should exemplify analyses that explicitly acknowledge related issues and address them in a convincing manner. The empirical part of the thesis uses data from a large cohort study to address some of the methodological limitations identified by this chapter. This is achieved in the context of a comparison of estimators relying on regression, matching, as well as the propensity score. Emphasis is placed on the combination of both traditional and novel matching approaches with a bias-correction stage that uses flexible bivariate modeling to consider the joint uncertainty in the estimates of the incremental cost-effectiveness.

CHAPTER THREE

COHORT STUDY

3.1 Introduction

The Birthplace national prospective cohort study is a landmark observational study that compared perinatal and maternal outcomes by planned place of birth for women in England and offers a unique opportunity to explore the use of appropriate non-experimental methods in the context of cost-effectiveness analysis. The present chapter introduces the cohort study, by providing a brief description of the background literature, study design, outcome measures, sample size calculations, participating units/trusts, women's eligibility criteria, classification of risk status, and complicating conditions. The material of this chapter aims to provide context

regarding the dataset which is used in the subsequent empirical work and is mostly reproduced from the original reports, in which the reader is referred for more detailed information regarding the Birthplace cohort study and the Birthplace in England research programme (Hollowell et al., 2011; McCourt el al., 2011; Schroeder et al., 2011).

3.2 Substantive literature

Current health policies in England regarding places where women can give birth require the National Health Service to provide a choice of locations (Campbell and Macfarlane, 1994; Hall, 2003; Department of Health, 2004). As such, for women who are at low risk of complications at the start of care in labour, maternity care is offered in four settings (Department of Health, 2007). These include care in an obstetric unit, a midwifery unit on the same site or geographically separate from the hospital obstetric unit, and at home. This choice-based approach has been supported by poor evidence regarding the quantification of outcomes in the different planned places of birth, while economic evidence related to care in each of these settings is also inconclusive (National Institute for Health and Clinical Excellence, 2007).

More specifically, with regards to clinical outcomes, according to Hollowell et al. (2011) who reviewed the relevant literature "a Cochrane systematic review of home versus hospital birth identified only one randomised controlled trial which included

eleven women and was unable to detect any differences in safety or other outcomes between the two settings (Olsen and Jewell, 1998). A meta-analysis of six observational studies examined perinatal outcomes for 24,092 'low risk' women and their babies (Olsen, 1997). No difference was observed for perinatal mortality. There was evidence that women planning birth at home had a lower risk of induction, augmentation, instrumental vaginal birth, caesarean section, episiotomy, severe perineal lacerations and that their babies were less likely to have low Apgar scores".

Hollowell and colleagues (2011) also point out that "the results of a few other observational studies comparing home births with birth in an obstetric unit have also been published. A retrospective cohort study from the Netherlands using routine data from over 500,000 women found no evidence of a difference in perinatal mortality or morbidity between 'low risk' women who planned to give birth at home and 'low risk' women who planned to give birth in hospital (de Jonge et al., 2009). Canadian and Swedish studies of planned home births compared to planned hospital births for 'low risk' women also showed no difference in perinatal mortality (Janssen et al., 2009; Lindgren et al., 2008). Lower rates of obstetric interventions were observed in the planned home birth group for both studies. However, both studies included fewer than 20,000 births and lacked statistical power to demonstrate differences in rare but important adverse outcomes. A study using data from England and Wales attempted to quantify the intrapartum-related perinatal mortality rates for booked home births from 1994 to 2003 using routine statistics (Mori, Dougherty and Whittle, 2008). However, the data available were of poor quality for this comparison and highlighted the need for a more accurate quantification of the risks associated with each planned

place of birth. A recent meta-analysis found planned home births, compared to planned hospital births, were associated with less medical intervention, had a similar perinatal mortality rate and an increased neonatal mortality rate (Wax et al., 2010). This study has been criticized for failing to report the assessment of the quality of the studies included (Gyte et al., 2010)".

In addition, in their report Hollowell et al. (2011) highlight the fact that "a different Cochrane systematic review compared birth in alternative birth settings with conventional institutional settings (obstetric units) (Hodnett et al., 2010). The review included nine randomised controlled trials and 10,684 women, and the alternative birth settings studied were most similar to alongside midwifery units. Alternative birth settings were associated with an increased likelihood of spontaneous vaginal birth, increased maternal satisfaction and fewer medical interventions during labour and birth. There was no association between birth setting and severe perinatal morbidity or mortality. Also, there was no association between birth setting and serious maternal morbidity or mortality. However, it is likely that the review was underpowered to detect any differences in rare but important severe adverse perinatal and maternal outcomes. No trials of freestanding midwifery units were included in this review. Prospective observational studies have shown a lower rate of intervention during labour for births planned in freestanding midwifery units (Walsh and Downe, 2004; National Institute for Health and Clinical Excellence, 2007)."

As regards the availability of cost and cost-effectiveness evidence, Henderson and Petrou (2008) recently carried out a structured review of the economic literature

relating to home births and birth centers, which revealed that this is even more scarce than the available clinical evidence. More specifically, the review identified only two economic studies evaluating home births, and only six studies examining the economic implications of birth centers. Henderson and Petrou (2008) emphasize that the findings highlighted the fact that the studies relating to birth centers produced conflicting evidence. For example, "some reported that outcomes in birth centers were better than in hospitals and costs were lower, whereas other studies reported that outcomes in birth centers were not significantly different from standard hospital care and costs were higher". The authors concluded that differences in results among studies may be attributed to differences in health care systems, differences in methods used, differences in costs included (costs may differ from those faced in the NHS and place of birth comparisons do not reflect current clinical practice in England), the differences in interventions being examined, as well as the probability of selection bias in the non-randomised studies.

Of particular importance in any non-randomised clinical and economic studies is the consideration of how this can provide credible estimates in the potential presence of selection bias. In general, studies identified a number of potential confounders that could address selection bias. These included maternal age, ethnicity, education, understanding of English, employment, religion, marital or partner status, body mass index in pregnancy, index of multiple deprivation score, area of residence (urban versus rural), parity, gestational age at birth, smoking status, drug and alcohol use during pregnancy and NHS trust size. Table 3.1 clarifies the issue of confounders in the context of the Birthplace cohort study and their use in this thesis.

Table 3.1 – Potential confounders

Variable	Cohort study	Analysis
Maternal age	Collected as continuous	Categorical as groups
Ethnicity	Collected as continuous	Categorical as groups
Education	Not collected	N/A
Understanding of English	Collected as categorical	Categorical
Employment	Not collected	N/A
Religion	Not collected	N/A
Marital or partner status	Collected as categorical	Categorical
Body mass index in pregnancy	Collected as continuous	Categorical as groups
Area of residence (urban versus rural)	Not collected	N/A
Index of multiple deprivation score	Collected as continuous	Categorical as quintiles
Parity	Collected as continuous	Categorical as groups
Gestational age at birth	Collected as continuous	Categorical as groups
Smoking status	Not collected	N/A
Drug and alcohol use during pregnancy	Not collected	N/A
NHS trust size	Not collected	N/A

Notes: The table displays potential confounders identified in the substantive literature. The second column considers how these were collected and measured in the cohort study, while the third column how these were analysed in the empirical part of the thesis. N/A: Not Applicable.

3.3 Primary objective

As discussed in the first chapter, a randomised experiment is considered as the ideal evaluation design for solving convincingly the evaluation problem in a particular research question. However, undertaking a randomised controlled trial for evaluating the safety of different planned places of birth would neither be feasible, nor perhaps ethical. Birthplace was conceived and implemented with a view to produce comprehensive high quality evidence regarding the risks and benefits associated with planning to give birth in different settings. It represents a classic example of a research question where observational studies constitute the only realistic alternative for generating the required evidence for decision-making.

Hollowell and colleagues (2011) state in their report that the principal aim of the Birthplace cohort study was "to compare intrapartum and early neonatal mortality and specific neonatal morbidities for different planned places of birth, for babies of women judged to be at 'low risk' of complications at labour onset". According to the authors "the definition of the primary objective allows for ambiguity in the definition of risk status for women entering the cohort". This is because "women judged to be at 'low risk' of complications at labour onset could be misinterpreted" with "the classification of women into 'low risk' or 'higher risk' for women entering the cohort would in reality have been assessed at the last episode of antenatal care, which may have been weeks or minutes before the onset of labour". As such, Hollowell et al. (2011) point out that "the definition of risk is not accurately at the time of 'labour onset' but at some point prior to labour onset" and clarify that the definition of risk

status that was ultimately operationalised was "... for women judged to be at 'low risk' of complications prior to the onset of labour".

3.4 Study design

Birthplace was a prospective cohort study with planned place of birth at the start of care in labour as the intervention and a range of perinatal, maternal and economic outcomes. Four groups of women were included based on their planned place of birth at the start of care in labour. This included home; freestanding midwifery unit alongside midwifery unit and obstetric unit. Hollowell et al. (2011) note that "women were included in the group in which they planned to give birth at the start of care in labour regardless of whether they were transferred during labour care or immediately after the birth. In some trusts, women are able to wait until the start of care in labour at home to decide whether they would prefer a planned home birth or to go to a midwifery or obstetric unit. These women were included in the study in the setting where they decided to receive labour care, reflecting their decision in early labour regarding planned place of birth".

3.4.1 Planned places of birth

Throughout the thesis reference is made to births planned in units or trusts. Units refer to births planned in midwifery or obstetric units. Similar to the report by

Hollowell and colleagues (2011), the term "trusts" is used to describe births planned at home because home birth services are delivered within National Health Service trusts. Hollowell et al. (2011) also define each of the planned birth settings as follows:

Planned home births: "A birth which occurs for a woman who, at the start of care in labour, intended to give birth at home and who received care from a midwife during established labour at home, regardless of where the woman actually gives birth. This includes women who make their final decision about planned place of birth during labour".

Planned freestanding midwifery unit (FMU) births: "A birth which occurs for a woman who, at the start of care in labour, intended to give birth in a freestanding midwifery unit and who received care from a midwife during established labour in a freestanding midwifery unit, regardless of where the woman actually gives birth. Freestanding midwifery units are defined as being on a separate geographical site from an obstetric unit and transfer will normally be by ambulance or car".

Planned alongside midwifery unit (AMU) births: "A birth which occurs for a woman who, at the start of care in labour, intended to give birth in an alongside midwifery unit and who received care from a midwife during established labour in an alongside midwifery unit, regardless of where the woman actually gives birth. Alongside midwifery units are defined as being in the same building or on the same geographical site as an obstetric unit and transfer will normally be by trolley, bed or wheelchair".

Planned obstetric unit (OU) births: "A birth which occurs for a woman who, at the start of care in labour, intended to give birth in an obstetric unit and who received care from a midwife during established labour in an obstetric unit".

3.4.2 Choice of birth modality

Given the sparseness of the substantive literature concerning planned place of birth, it is not surprising that currently there are huge gaps in our understanding about the mechanism for selection of women in different birth modalities. The Birthplace in England research programme through a series of qualitative case studies provided some evidence with regards to how women make the choice of birth modality. In their report, McCourt el al. (2011) reveal that "there were variations in the number of women who had practical access to the full range of birth settings within their locality, as most women did not see travelling over a long distance in labour as a realistic choice. Choice was influenced by geographical, organisational, service culture and provider factors. Some women were not aware that choice of birthplace was possible, and lacked sources of evidence-based information on which to base choices. Women's views of safe care were influenced by what was locally on offer, their previous experience and that of other women that they knew. The prospect of intrapartum transfer was a major consideration when women made a decision around place of birth, and women often cited concerns about transfer distance as reasons for planning labour in hospital".

3.4.3 Primary outcome

The primary outcome used in the cohort study is described in the report produced by Hollowell et al. (2011). This is "a composite of stillbirth after the start of care in labour; early neonatal death (within seven days); neonatal encephalopathy defined as either a clinical diagnosis of neonatal encephalopathy or signs of neonatal encephalopathy; meconium aspiration syndrome; brachial plexus injury; and fractured humerus or clavicle".

Hollowell and colleagues (2011) explain that "a clinical diagnosis of neonatal encephalopathy was defined as either a clinical diagnosis of neonatal encephalopathy, or a clinical diagnosis of isolated seizures without a known cause other than perinatal asphyxia. Signs of neonatal encephalopathy was defined as admission to neonatal unit within forty-eight hours of birth for at least forty-eight hours with signs consistent with a diagnosis of neonatal encephalopathy: a) receipt of parenteral or tube feeding or receipt of supplemental oxygen or respiratory support; and b) absence of meconium aspiration, suspected or confirmed sepsis or other diagnosis consistent with feeding difficulties or need for respiratory support".

According to Hollowell et al. (2011) "a composite outcome was chosen to give the study more power to detect differences in safety between planned places of birth compared with a single outcome, which would have a lower incidence. Using a composite outcome could provide misleading results if planned place of birth affects different contributing outcomes in different ways. For example, if the effect of

planned place of birth in a particular setting decreased deaths but resulted in increased morbidity there might be little or no difference observed in the primary outcome, even though deaths were being prevented in one setting. The likelihood of this occurring was unknown but the increased statistical power of using a composite outcome outweighed the alternative approach of substantially increasing the sample size to address individual components of the primary outcome".

3.4.4 Secondary outcomes

Birthplace also evaluated a number of other outcomes, and these are similarly described in the report by Hollowell and colleagues (2011). More specifically, perinatal outcomes included "stillbirth after the start of care in labour; early neonatal death (within 7 days); neonatal encephalopathy defined as either a clinical diagnosis of neonatal encephalopathy or signs of neonatal encephalopathy; meconium aspiration syndrome; brachial plexus injury; fractured humerus; fractured clavicle; fractured skull; cephalohaematoma; cerebral haemorrhage; early onset neonatal sepsis; kernicterus (severe bilirubin encephalopathy); seizures; neonatal unit admission; apgar score less than seven at five minutes; breastfeeding initiation". Maternal outcomes focused on "mode of birth: spontaneous vertex birth, vaginal breech birth, ventouse delivery, forceps delivery, intrapartum caesarean section; normal birth defined as a birth with none of the following interventions: induction of labour, epidural or spinal analgesia, general anaesthetic, forceps or ventouse,

caesarean section, episiotomy, third or fourth degree perineal trauma, blood transfusion, admission to an intensive therapy unit or high dependency unit or specialist unit, maternal death (within 42 days of giving birth); maternal interventions in labour: syntocinon augmentation, immersion in water for pain relief, epidural or spinal analgesia, general anaesthetic, active management of the third stage of labour, and episiotomy". At this point, it is important to stress that the outcome used in the empirical part of the thesis is the maternal outcome 'normal birth' as defined above.

3.4.5 Economic outcome

A further aim of Birthplace was enriching the provision of cost-effective decision-making in the English National Health Service. Using microdata collected from the cohort study, Schroeder et al., (2011) evaluated the cost-effectiveness of different planned places of birth. The cost-effectiveness analysis was carried out from a health payer perspective and as such, only direct costs to the English National Health Service were included. The time horizon was the duration of follow-up of the Birthplace cohort study, which identified women at the start of their care in labour, and was completed "when the intrapartum care for both mother and baby ended, irrespective of whether this was at home or discharge into post-natal care in a midwifery unit or hospital".

The economic outcome was constructed using a rigorous costing approach. For resource use items, Schroeder and colleagues (2011) "devised detailed structured data collection forms in order to capture all possible National Health Service resources used in the care of the mother and baby during the period between admission and discharge in midwifery units and hospitals". The authors explain that "both top down and bottom up methods were subsequently employed to identify relevant costs" and that "total costs were generated by apportioning to each woman the unit costs for key resource items, episodes or procedures, according to the actual place of labour and birth, as well as the duration in hours that was spent there". Schroeder et al., (2011) also note that "although standard post-natal care was not included, if higher-level care following the birth was required for the mother, the baby, or both, this was taken into consideration".

3.4.6 Sample size

Hollowell and colleagues (2011), in their report point out that "major perinatal and maternal morbidity are rare in women judged to be at "low risk" of complications prior to the onset of labour. The incidence of neonatal encephalopathy at term is approximately 1.8 per 1,000 live births. However, the incidence of intrapartum stillbirth after labour onset, early neonatal death and other related neonatal morbidity at term for babies of women at "low risk" of complications prior to the onset of labour is much less certain. A reasonable estimate of the incidence of the composite primary

outcome is 3.6 per 1,000 births. As the vast majority of data on neonatal morbidity are from obstetric units, this estimate is assumed to be the incidence of the primary outcome in obstetric units. In order to have adequate power to detect clinically important differences in outcome that are associated with planned place of birth, the study needed to collect data on at least 20,000 "low risk" women planning to give birth in an obstetric unit, at least 17,000 women planning to give birth at home and at least 5,000 women planning to give birth in each type of midwifery unit".

According to Hollowell et al. (2011) "the study aimed to collect data on at least 85% of all eligible women planning birth at home over approximately sixteen months, which it was estimated to be 17,000 women. With data from 17,000 planned home births, the study would be able to detect an increase in the incidence of the primary outcome from 3.6 per 1,000 births in obstetric units to 5.7 per 1,000 for planned home births, with a 5% two-sided level of significance and 82% power. Alternatively, the study would be able to detect a reduction in the incidence of the primary outcome from 3.6 per 1,000 births in obstetric units to 2.0 per 1,000 births for planned home births, with a 5% two-sided level of significance and 80% power".

Hollowell et al. (2011) go on to explain that "data collection was planned for at least six months in each type of midwifery unit, which would allow a minimum of 5,000 women from each type of unit to be included. Freestanding midwifery units and alongside midwifery units were to be analysed separately when being compared to obstetric units. With 5,000 women included from each type of midwifery unit, the study would be able to detect an increase in the incidence of the primary outcome

from 3.6 per 1,000 births in obstetric units to 6.8 per 1,000 in midwifery units, with a 5% two-sided level of significance and 80% power. Alternatively, the study would be able to detect a reduction in the incidence of the primary outcome from 3.6 per 1,000 births in obstetric units to 1.2 per 1,000 births in midwifery units, with a 5% two-sided level of significance and 80% power. With these sample sizes, assuming 80% power and a 1% level of significance the study would be able to detect similar or smaller relative differences in more common serious outcomes of maternal morbidity amongst women at "low risk" of complications. For example, for blood transfusion affecting approximately 0.5% of women, the detectable relative differences would be similar; and for 3rd and 4th degree perineal trauma experienced by 1.2% of women, the detectable relative differences would be smaller due to the higher control group event rate". It is important to note that no such calculations were carried out for the economic outcome, described in the previous section.

3.4.7 Participating trusts/units

The aim of the Birthplace study, as Hollowell and colleagues (2011) state, was "to collect data in every National Health Service trust in England providing home birth services; every freestanding midwifery unit as well as alongside midwifery unit in England; and a stratified random sample of 37 obstetric units." In addition, in their report, the same authors indicate that "eligible trusts and units were identified using data from a national mapping survey of all National Health Service trusts providing

maternity care in England" and that "midwifery units that opened during the study period were also invited to participate".

Hollowell et al. (2011) also provide more details on how the stratified random sample of obstetric units was selected. More specifically, "the sample stratified by unit size (<2600 births, 2600-4850 births and >4850 births per year) and geographic location (northern England or southern England). Data from the Department of Geography at the University of Sheffield were used to define northern and southern England. For any sampled obstetric unit that declined to participate, another unit randomly selected from within the same stratum replaced it. The method of sampling was such that each obstetric unit in England had approximately the same probability of selection (~37/180). The study aimed to include close to 100% of eligible women from each obstetric unit over a three-month period thus giving each eligible woman the same probability of being included in the sample. The aim was for each participating unit to collect data prospectively for all eligible births within a defined study period. In practice, it was not possible to collect data over the same time period and for the same duration for each trust and for each unit type. The number of trusts and units changed during the study period as trusts merged, units opened and units were closed and as such some of them were replaced by resampling from within the same stratum".

3.4.8 Women's eligibility

According to Hollowell et al. (2011) "all women who were attended by an National Health Service midwife during labour in their planned place of birth, for any amount of time, were eligible for inclusion in the study with the exception of: a) women who had a caesarean section before the start of labour; b) women who presented in labour before 37 weeks and 0 days gestation; c) women with a multiple pregnancy; and d) women who were unbooked (i.e. had received no antenatal care). Stillbirths occurring prior to the start of care in labour were excluded".

3.4.9 Risk-status classification

Hollowell et al. (2011) also explain that "in order to make meaningful comparisons between the planned places of birth, it was necessary to define women as being known to be at low risk or higher risk of complications prior to the onset of labour using standard criteria applied across all participating centres. Women were classified as low risk if, immediately prior to the onset of labour, they were not known to have:

a) Any of the medical conditions or situations listed in the National Institute for Health and Clinical Excellence Intrapartum Care guidelines that result in increased risk for the woman or baby during or shortly after labour, where care in an obstetric unit would be expected to reduce this risk.

b) Other medical conditions or situations not listed in the National Institute for Health and Clinical Excellence guidelines considered to confer an increased risk such that care in an obstetric unit would be expected to reduce the risk. These included, but were not limited to: a known fetal anomaly; reduced fetal movements; obstetric cholestasis; cervical suture, cervical fibroid; low lying placenta; previous 3rd/4th degree tear; female genital mutilation; symphysis pubis dysfunction; recurrent urinary tract infections; current or recent malignancy; Crohn's disease; sarcoidosis; pneumothorax".

3.4.10 Complicating conditions

In their report, Hollowell and colleagues (2011) highlighted the fact that "women were assessed by the attending midwife for any risk factors present when they started labour care in their planned place of birth. New risk factors identified at this point could not affect the woman's planned place of birth and hence did not affect the woman's classification of risk status prior to the onset of labour. Any conditions identified at this time are referred to as complicating conditions at the start of care in labour. These data were collected to enable assessment of the homogeneity of the low risk groups and included prolonged rupture of membranes (>18 hours), meconium stained liquor, proteinuria (1+ or more), hypertension, abnormal vaginal bleeding, non-cephalic presentation, abnormal fetal heart rate, or other complications. Some of the categories used for this intentionally had a lower risk threshold than criteria used

in clinical guidelines (e.g. meconium stained liquor rather than significant meconium staining and prolonged rupture of membranes >18 hours rather than >24 hours)".

3.5 Concluding remarks

Planning birth in different settings is an important area of research, which however has received relatively little attention in the substantive literature. Studies currently exploring this research theme are plagued with significant gaps in their methodology and particularly in aspects relating to how women choose location. As such, the causal pathway between planning birth in different settings and perinatal outcomes is not yet firmly established. This raises serious concerns that estimates obtained from existing studies may be considerably subject to selection bias.

The Birthplace national prospective cohort study represents a herculean effort in the substantive area concerning the organisation of maternity services. The rigorous prospective study design, together with qualitative work undertaken in the context of the wider Birthplace in England research programme, attempted to address some of the problems outlined above. The cohort study also offers a range of outcomes for evaluating the short term cost-effectiveness of alternative planned places of birth and the empirical part of the thesis extends both applied and methodological aspects of the substantive cost-effectiveness analysis of Schroeder and colleagues (2012).

CHAPTER FOUR

ANALYTICAL METHODS

4.1 Introduction

This chapter discusses the analytical methods employed in the empirical part of the thesis and introduces formally the identification conditions required to solve the evaluation problem when observational microdata are used. The next section presents the key assumptions that regression analysis and matching procedures rely on. Section 3 discusses the rationale behind regression analysis, while section 4 considers the role of propensity score. Section 5 gives a more detailed overview of matching as a method to construct counterfactual outcomes. Section 6 laids out the analyses undertaken in

the context of an empirical case study using the Birthplace national prospective cohort study. Finally, section 7 concludes.

4.2 Identification conditions

Identification of average treatment effects and causal interpretation of the estimates obtained using regression and matching methods will depend upon the analyst invoking two crucial assumptions.

The first is unconfoundedness, which in general is also known as ignorability (Rosenbaum and Rubin, 1983), selection on observables (Barnow, Cain and Goldberger, 1980), conditional independence (Lechner, 2002), no hidden bias (Rosenbaum, 2002), exogeneity (Imbens, 2004) and conditional exchangeability (Hernán and Robins, 2013). Following Imbens (2004), if the observations of each individual are described by a vector of exogenous confounding covariates X_i, the individual outcome distributions Y_i, and an indicator D_i that takes the value of one if the individual has been exposed to the health care intervention under investigation, and zero otherwise, then unconfoundedness can be defined as

$$(Y^1, Y^0) \perp D \mid X \tag{4.1}$$

The unconfoundedness assumption postulates that there is an exogenous vector of confounding covariates X that jointly affects the potential outcomes of an individual, as well as its exposure to an intervention. This vector of confounding covariates is observed by the analyst and after conditioning on X, treatment assignment of an individual is said to be "as good as random", with potential outcomes being independent of treatment status. Rosenbaum and Rubin (1983) have also demonstrated that when it is valid to condition on the covariates vector X, unconfoundedness is not violated when instead the analyst chooses to condition only on the propensity score. Unconfoundedness given the propensity score is defined as

$$(Y^1, Y^0) \perp D \mid p(X) \tag{4.2}$$

Stuart (2010) notes that although the plausibility of the unconfoundedness assumption in different applications may be questionable, often it will be not be as restrictive as it may initially seem. This is because the only unobserved confounding variables of concern will be those that are not related to the observed ones. As such, she argues that if the analysis conditions on the observed covariates, then differences due to unobserved covariates correlated with those that are observed, will also be accounted for. Consequently, the unconfoundeness assumption will become less limiting when a large number of confounding variables, which are observed by the analyst, potentially affect selection into treatment.

The second assumption required is often referred as the common support, overlap or positivity (Hernán and Robins, 2013) and is defined as

$$0 < P(D = 1 \mid X) < 1 \tag{4.3}$$

The above equation postulates that in order for the analyst to estimate the difference in mean outcomes for each value of the vector of covariates *X*, there is a positive probability that both exposed and comparison individuals for all values of *X* exist (Heckman et al., 1998; Smith and Todd, 2001). In other words, sufficient overlap between the exposed and comparison subpopulations is required in terms of confounding characteristics, in order to ensure that each exposed individual has a similar comparison individual. Moreno-Serra (2007) points out that the common support assumption refers to the joint distribution of the exposure variable and covariates, which implies that conditional on *X*, there should be other unobserved variables affecting the allocation of individuals into treatment, thus preventing *X* from being a perfect predictor of treatment assignment.

Nevertheless, if both the unconfoundedness and the common support assumptions hold, then allocation of individuals into treatment is said to be strongly ignorable (Rosenbaum and Rubin, 1983). Following Stuart's (2010) reasoning, under strong ignorability, these unobserved variables affecting the allocation of individuals into treatment are not correlated with potential outcomes. In such a case, the potential outcomes of exposed individuals have the same distribution with the counterfactual potential outcomes that comparison individuals would have observed in the absence

of the intervention and if there is no selection on unobservables, that is, no bias arising from omitted variables, then by conditioning on *X*, all systematic differences in the outcomes of exposed and non-exposed individuals can be attributed exclusively to the intervention under investigation (Moreno-Serra, 2007).

4.3 Regression analysis

The premise of applied econometrics is the study of the relationships between outcomes and factors potentially explaining them, with the analyst typically interested in quantifying the level of change in a dependent variable, in terms of a given level of change in an independent variable (Kennedy, 2009). Regression-based models in their most archetypal form estimate the conditional expectation (i.e. the average value) of an outcome variable arising from the corresponding change in the predictor, when all other factors are held fixed:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \tag{4.4}$$

A regression model that is linear in parameters will often be proven adequate in evaluating the impact of an intervention on outcomes (Woolridge, 2009). Linearity in this context implies that a one-unit change in the independent variable has the same effect on the outcome variable. In addition, in linear regression models, the coefficient on the explanatory variable can also be thought of as representing its

marginal effect on the outcome under the statistical model used (Cameron and Trivedi, 2005). The choice of the appropriate model usually depends on the available data, as well as the outcome under evaluation. A model for binary dependent variables based on the linear regression model is known as the linear probability model (Angrist and Pischke, 2009). In this case, the interpretation of the estimates obtained will be different, since what is estimated is the probability of an event occurring. In linear regression models, where the error terms have zero expectation and are uncorrelated and homoscedastic, the best linear unbiased estimator of the coefficients, in terms of the lowest possible mean squared error, is given by the ordinary least squares estimator (Gujarati, 2003). An alternative estimation approach is maximum likelihood, which specifies the joint probability of observed set of data and finds the parameter values that are most likely (Jones, 2007).

Regression analysis can be used in cost-effectiveness evaluation as a control strategy to adjust for potential confounding factors. Two regression approaches have been exemplified in the economic evaluation literature; the net benefit regression (NBR) and seemingly unrelated regression (SUR). Both rely on the net benefit statistic, which was formally introduced in the first chapter. Net benefit regression was first proposed by Hoch, Briggs and Willan (2002) and combines cost and effectiveness data with pre-specified willingness to pay values, to generate a net benefit for each individual. This statistic can subsequently be used as the dependent variable in a linear regression model, where the coefficient of the treatment indicator variable will provide the analyst with the difference between the extra benefits and extra costs (i.e. the incremental net benefit).

The second regression approach is known as seemingly unrelated regression and was originally proposed by Zellner (1962). Seemingly unrelated regression can be conceptualised as a type of structural model, where each member of a set of observed endogenous variables is a function of a set of observed exogenous variables and a unique random error term. The method relies on the use of a system of equations that may be related not because they interact, but because their error terms are correlated. In cost-effectiveness analysis, Willan, Briggs, and Hoch (2004) have shown that this regression approach is a more general method than the net benefit regression, which exploits the estimated variance-covariance matrix to produce estimates that subsequently can be used in either an incremental cost-effectiveness ratio, or an incremental net benefit context. Analysts employing this approach can choose to use different set of covariates for cost and effectiveness. In addition, estimation using feasible generalised least squares also has the potential to improve efficiency (in terms of precision) compared to simple ordinary least squares, if different sets of covariates are used in each equation (Cameron and Trivendi, 2005). These efficiency gains will increase when the disturbances are highly correlated and the independent variables for cost and effectiveness are not highly correlated (Willan, Briggs and Hoch, 2004).

4.4 The propensity score

Experience from the evaluation literature suggests that the unconfoundedness assumption will be implausible when few covariates are used in the regression model, since for a considerable number of empirical investigations a large number of observable characteristics will be required (Jones and Rice, 2011). Nevertheless, the sparseness of information in the data at hand, typically resulting from the finiteness of a sample, will often mean that when the analyst uses several confounding variables (i.e. multiple dimensions), overlap of individuals will be challenging. This is because the greater the number and the wider the range of the vector of covariates X that the analyst chooses to employ, the more difficult it will be to find similar comparison individuals. This problem is commonly referred as the curse of dimensionality (Rosenbaum and Rubin, 1983).

A popular solution to the curse of dimensionality has been the use of the propensity score. The propensity score approach changes the focus of the analysis from the conditional expectation of an outcome, to the conditional probability of an individual being assigned to the treatment group instead of the comparator (Angrist and Pischke, 2009). The propensity score is a scalar representing the probability of an individual assigned into treatment, conditional on an exogenous vector of confounding covariates X. It summarises all the relevant information comprising the vector X and it is effectively a balancing score, which ensures that for a given value of the propensity score variable, the distribution of X will be the same for the exposed and comparison individuals. In this way, the analyst can reduce dimensionality by

matching or regressing directly on a single dimension (i.e. the propensity score variable), instead of the entire vector of covariates *X*. Mathematically, the propensity score is defined as

$$P(X) = P(D=1 \mid X) \tag{4.5}$$

Propensity score analysis involves a two-stage setup (Imbens, 2004). The first stage involves calculation of the propensity score for each individual. In contrast with randomised controlled trials where the true propensity score is known and fixed by design, in observational studies it is always unknown and has to be estimated from the observed data (Rosenbaum and Rubin, 1983). Since the propensity score represents the conditional probability of an individual assigned into treatment, its estimation can be undertaken with any models relating a binary variable to the vector *X*. The most common method used for estimating the propensity score is a logistic, probit or linear probability model, with the observed covariates being the predictors and receipt of treatment being the dependent variable. Non-parametric approaches can also be employed but so far the literature has favoured the former since the latter can lead back to dimensionality problems (Blundell and Costa Dias, 2009; Stuart, 2010).

In the second stage, the propensity score can be used using two approaches (Rosenbaum and Rubin, 1983). The first involves its inclusion as a covariate in a regression outcome model. Austin (2009) points out that the idea underpinning this approach is that exposed and comparison individuals with the same propensity score have the same distribution of measured confounders. As such, what is compared in

this case is the treatment effect between exposed and comparison individuals with the same propensity score. Alternatively, the inverse propensity score can be used to weigh a regression outcome model without covariates. This approach is known as inverse probability (of treatment) weighting and can effectively be seen as a matching approach that re-constructs an experimental dataset, by altering the treatment assignment patterns in the same fashion that an unrepresentative sample is reweighted (Morgan and Harding, 2006). When this approach is employed, the exposed individuals are matched with non-exposed counterparts inversely weighted for the distance in terms of their predicted conditional probabilities of receiving treatment (Jones and Rice, 2011). It should be noted that when inverse probability weighting is employed, the estimation approach used for the propensity score will considerably influence the reweighting process as it relies on both stages on parametric modeling. Inverse probability weighting in econometrics has been used predominantly to deal with survey non-response and attrition, but the same approach can also be applied to the estimation of treatment effects (Jones, 2006).

4.5 Matching

In broad terms, the idea of matching is to act as a surrogate for randomisation by constructing a comparison group containing the missing counterfactual information (Imbens and Wooldridge, 2009). This is typically achieved by pairing to each exposed individual a counterpart in the comparison group that is similar in terms of a vector of

observable characteristics. This vector of observables, which must not be influenced by the intervention itself, should affect simultaneously both the exposure status, as well as the outcomes of the individual (Todd, 2008). If these conditions hold, matching ensures that comparisons will only be made among groups of individuals that are equivalent (in a statistical sense).

Non-parametric matching estimators have the advantage that they do not impose any assumptions regarding the specification of the mechanism that allocates individuals into treatment, the outcomes of individuals or the unobserved term (Imbens, 2004). As such, if the groups are exactly balanced, the vector of observable confounding characteristics will be unrelated to the treatment variable and thus a simple difference in means on the matched data can provide the analyst with the causal effect (Abadie and Imbens, 2002). Different matching estimators essentially represent different techniques for deriving the weights that construct the matched counterfactual outcome for each exposed individual (Todd, 2008). Nevertheless, irrespective of the method chosen to match individuals, the aim is always the construction of a weighted balanced distribution of covariates across pairs of exposed and comparison individuals (Stuart, 2010). Estimators in this category of methods generally vary in two components; first, the number of comparison individuals used for each exposed individual that needs to be matched; second, the method employed to weigh for multiple comparison individuals, if more than one is used to match each exposed individual (Morgan and Harding, 2006).

Exact matching pairs to each exposed individual, a counterpart in the comparison group that is identical in terms of a vector of observable confounding characteristics *X*. Constructing pairs of individuals using exact matching typically involves the use of individuals from the comparison group as matches only once. During the last three decades, matching estimators have evolved to make more efficient uses of data, establishing in this way the wider applicability of matching as a method to draw causal inferences (Imbens and Wooldridge, 2009). A common feature of these inexact estimators is the consideration in the pairing process of all the comparison individuals that are sufficiently "close" to a given exposed individual. Distance of individuals (also known as proximity or closeness) is determined in terms of their vector of observed covariates, with outcomes usually weighted according to their degree of similarity.

A number of other measures of distance have also been used to construct weights (Zhao, 2004). The Mahalanobis approach collapses the different covariates into a single scalar metric. The distance in this metric is a generalisation of the standardised distance from the origin of an n-dimensional space, to a point where the coordinates characterise the X values for a particular individual (Cochran and Rubin, 1973; Rubin, 1979; Rubin, 1980). The Mahalanobis distance metric scales differences in the observed vector of covariates X by the inverse of their covariance matrix in the sample. In this way, when individual patterns are identified and analysed, the Mahalanobis metric allows the analyst to take into account the correlation between the covariates in the vector X. Moreno-Serra (2007) notes that the Mahalanobis metric has the advantage that incorporates the Euclidean distance, with the influence of the latter for a particular covariate depending on the precision with which that

covariate is measured; the higher the precision with which the covariate is measured in the sample, the higher the weight its corresponding distance will be given in the calculation of the Mahalanobis distance.

Stuart (2010) points out that in the presence of high dimensionality in the vector *X*, the use of exact and Mahalanobis distance measures may be difficult. Both measures, she argues, have been shown to perform well mostly in situations in which the number of observable characteristics is small. For the Mahalanobis metric in particular, it has been suggested that good performance is achieved when the analysis limits the number of covariates to fewer than eight (Rubin, 1979; Zhao, 2004). Its performance diminishes in large numbers of covariates, or in applications in which the covariates are not normally distributed (Gu and Rosenbaum, 1993). In addition, the Mahalanobis distance is best suited in measuring differences between ellipsoidally symmetric distributions (Rubin, 2006). In cost-effectiveness applications, it is typical for the covariates in *X* to have nonellipsoidally symmetric distributions and as such, Sekhon and Grieve (2012) argue that the Mahalanobis distance may often constitute an inappropriate choice.

An alternative approach has been the use of the difference in the propensity score as the distance measure to construct weights (Rosenbaum and Rubin, 1983). Matching estimators are now used predominantly in combination with the propensity score, typically leading in a semi-parametric two-stage setup (Imbens, 2004). Once the first stage is undertaken for all individuals, the analyst must choose a matching method that will pair the comparison individuals to their exposed counterparts using their

estimated propensity scores (Todd, 2008). A number of matching estimators relying on the propensity score have been exemplified as alternative ways for constructing weights.

Nearest neighbour (greedy) matching constructs the matched counterfactual for the exposed individuals through their random ordering and the subsequent selection for each of them of the closest comparison individual in terms of the propensity score (Rubin, 1973; Smith and Todd, 2005). Nearest neighbour matching can be distinguished in different versions. First, the estimator can sample with or without replacement. When no replacement is used, the estimator considers a comparison individual only once and then excludes this individual from the pool after matching takes place. In contrast, sampling with replacement allows for one comparison individual to serve as the match for more than one exposed individual. This is achieved by returning the comparison individual to the pool after a match so it can be matched again to another exposed individual. It should be noted that the order in which exposed individuals are paired in simple nearest neighbour matching, may ultimately affect the quality of the matches and thus of the results obtained. Stuart (2010) points out that the use of replacement renders the order in which individuals are paired irrelevant. However, she also strikes a note of caution stating that the ordering process in this case is no longer independent, because some comparison individuals are used more than once and this should be taken into account during the stage of inference. Second, the estimator can match on the basis of one or multiple neighbouring individuals, a variant of which is also known as ratio matching (Smith, 1997; Rubin and Thomas, 2000). In its simplest form, the nearest neighbour matching

estimator pairs individuals in the exposed group with those in the comparison group, on a one-to-one basis. More elaborate versions introduced some flexibility by allowing the analyst to employ weights equal to k-to-one for matched comparison individuals, where k is the number of matches selected for each exposed individual. Matching using multiple nearest neighbours can be desirable if the analyst wants to take advantage of the additional information that has available and thus obtain more precise counterfactual estimates.

Although nearest neighbour matching has the advantage of potentially yielding better precision, Dehejia and Wahba (2002) demonstrated that in finite samples this form of matching raises the prospect of greater bias. This is because when there is insufficient overlap, nearest neighbour matching can pair exposed individuals to comparison counterparts that have very different propensity scores, despite defined as closest neighbours. In such cases, the same authors have shown that the quality of matches will be poor, preventing the best possible balance to be obtained. Caliper matching is a modification of nearest neighbour matching, which gives to the selected comparison individuals equal weight and if there are no comparison individuals within a tolerated distance defined by the analyst (the caliper), then the nearest available neighbouring comparison individual is used (Cochran and Rubin, 1973). Radius matching is a close alternative that pairs multiple comparison individuals within a maximum distance (the radius) from the exposed individual, when there is more than one similar comparison individual (Dehejia and Wahba, 2002). In radius matching, the counterfactual outcome is the average outcome of all comparison individuals within the radius. According to Moreno-Serra (2007), both of these versions of nearest neighbour

matching impose a common support condition, with the choice of a caliper or radius being a matter of personal choice, and reflecting the trade-off between obtaining lower bias, in exchange for using a narrower maximum distance. He also points out that smaller tolerated distances are generally preferred because they exclude as few exposed individuals as possible. Rosenbaum and Rubin (1985b) have investigated this issue and generally recommend a caliper of 0.25 standard deviations of the propensity score.

Stratification, also known as blocking, subclassification and interval matching, is an approach in which the exposed and comparison individuals are divided into strata according to their observable characteristics (Cochran, 1968; Rubin, 1977; Rosenbaum and Rubin, 1984; Rosenbaum and Rubin, 1983). The analyst then calculates the treatment effect within each stratum and subsequently obtains the overall treatment effect by aggregating across the strata using weighting if the number of individuals in each stratum is different. The idea here is that the strata act as the equivalent of matched pairs, with weighting simply averaging the treatment effect estimates in the different strata in order to obtain the average treatment effect parameter of interest. Most analysts now, by convention, use five to ten strata since Rosenbaum and Rubin (1985a) have demonstrated that stratifying individuals in five strata based on the propensity score, removes most of the bias in the estimated treatment effect.

Kernel matching (also known as kernel weighting and kernel density matching) offers another approach in constructing the counterfactual outcome for the exposed

individuals (Heckman, Ichimura and Todd, 1997; Smith and Todd, 2005). This matching algorithm constructs counterfactuals from a weighted average of the outcomes of all or multiple comparison individuals, with the weights generated using a kernel function. More specifically, the influence of a comparison individual to the counterfactual outcome will depend on a combination of its distance from the exposed individual and the chosen bandwidth for the kernel function (the smoothing parameter). The evaluation literature has exemplified the use of a number of different kernels including Gaussian, Uniform, Epanechnikov, Biweight and Tricube (Guo and Fraser, 2010). Distance is typically measured in terms of propensity scores, with the estimator placing more weight on those comparison individuals with propensity scores closest to the exposed individual (Imbens, 2004). Heckman, Ichimura and Todd (1997) have also demonstrated a generalised version of kernel matching known as local linear matching, which can be particularly advantageous when the propensity score takes extreme values.

4.6 Combined & novel approaches

Different matching methods can be used as standalone solutions or in combination. For example, Rosenbaum and Rubin (1985b) have shown that in finite samples, covariate imbalance can be improved by combining the propensity score with another metric measuring proximity, such as the Mahalanobis distance. Rubin and Thomas (2000) have also exemplified the use of Mahalanobis matching within propensity

score calipers. Irrespective of the chosen approach, ultimately, it is important for the analyst to assess the balance achieved. If the result is not satisfactory, the specification of the propensity score model must be revised, and/or a different distance metric chosen until the best possible balance is achieved (Rosenbaum and Rubin, 1984). Iacus, King and Porro (2011) point out that when matching is inexact, a parametric model must be used to control for remaining imbalances in the covariates across the groups compared, a process that Abadie and Imbens (2011) term bias-corrected matching. Stuart (2010) also notes that the idea behind this approach is similar to double robustness proposed by Robins and colleagues. This refers to the combination of inverse probability weighting with parametric regression modeling, which results in an estimator that is said to be "doubly robust" (Robins, Rotnitzky and Zhao, 1994; Wooldridge, 2007). Bang and Robins (2005) point out that doubly robust estimators have the advantage that as long as either the propensity score model or the outcome regression models are correctly specified, the effect of the exposure on the outcome will be correctly estimated.

Achieving balance between the exposed and comparison groups is an important task of any matching exercise. This can only be achieved by re-estimating the propensity score until the best possible balance is reached in the individual covariates. Ideally, balance should not only concern the covariate means, but also higher moments including variance, skewness, and kurtosis, as well as cross-moments such as the covariance (Ho et al., 2007). Diamond and Sekhon (2005) point out that achieving the desired degree of balance on such a broad range of terms can be problematic, since usually it will not be straightforward how to modify the propensity score model, or the

distance metric. For this reason, the same authors propose the adaptive method of genetic matching as an alternative method to pair individuals, which treats the matching based on both the Mahalanobis distance, as well as the propensity score as special cases. The argument here is that genetic matching is much broader in scope and thus advantageous because it uses an evolutionary algorithm that searches over the space of distance metrics, in order to find the best metric for assigning a weight to each baseline covariate and optimise covariate balance (Sekhon and Mebane, 1998). In this thesis, two novel matching methods that address some of the problems described above and have not been used in cost-effectiveness analysis before, are explored and contrasted with traditional matching estimators in the context of an empirical case study. The novel methods considered include coarsened exact matching, as well as entropy balancing and are presented in more detail below.

Coarsened exact matching

Coarsened exact matching has been proposed as an alternative approach to perform exact matching using a two-stage approach. According to Iacus, King and Porro (2012), first each covariate is recoded, with a view to group substantively indistinguishable values and assign them the same numerical value. Exact matching is then undertaken on the recoded data in order to establish matches and discard observations for which appropriate matches could not be found. The same authors point out that the recoded data are also discarded, with the uncoarsened values of the matched data retained for subsequent analysis. Similarly to other matching methods,

after the matching stage takes place the analyst can use a simple difference in means or any statistical model to obtain the treatment effect.

Iacus, King and Porro (2011) emphasize that coarsened exact matching belongs to a new monotonic imbalance bounding class of matching methods that generalises and extends the only existing equal percent bias reducing class, in which propensity score matching methods belong. The same authors point out that the main advantage of coarsened exact matching is the fact that the analyst selects beforehand the balance desired between the exposed and comparison groups and produces a matched sample size ex post. As such, they argue, the analyst avoids the repetitive practice of checking covariate imbalance and re-estimating the propensity score until a satisfactory degree of balance is achieved. Iacus, King and Porro (2011) also highlight the fact that unlike equal percent bias reducing methods, in monotonic imbalance bounding approaches, when the analyst alters the parameters of a covariate, the maximum imbalance for the other covariates is not affected.

Coarsened exact matching can be used either as a standalone solution, or before other matching methods are considered. In the latter case, Blackwell et al. (2009) point out that the matching scheme subsequently undertaken inherits several of the properties of coarsened exact matching. Other advantages of coarsened exact matching according to Iacus, King and Porro (2009) include the bounding of the degree of model dependence and the treatment effect estimation error, elimination of the need for a distinct technique to enforce common support, robustness to measurement error, complete automation, as well as rapid computation with large datasets.

Entropy balancing

As discussed in Hainmueller (2012), entropy balancing uses a maximum entropy-reweighting scheme that optimises a set of weights, such that the exposed and the matched comparison groups satisfy multiple balance constraints. These balance constraints are defined beforehand and involve exact balance on different moments of the covariate distributions such as the means, variance and skewness. For binary variables, Hainmueller and Xu (2011) emphasise that adjusting only for the mean is sufficient to also automatically match variance and skewness. Similarly to other matching estimators, the weights that are produced using entropy balancing can be used in conjunction with any statistical model to analyse the processed data.

Hainmueller (2012) argues that entropy balancing can be conceptualised as a generalisation of inverse probability weighting and provides an in-depth technical discussion of how this approach can balance the distributions of covariates between the treatment and comparison groups. In brief, unlike weighting using the inverse propensity score, entropy balancing performs the adjustment using a reverse approach that essentially estimates the weights directly from the chosen balance constraints. The advantage in this case, he claims, is that the analyst avoids undertaking the repetitive process of balance checking. This is because instead of first estimating the propensity scores using a binary choice model (typically logistic or probit) and then evaluating whether the resulting balance in the covariate distributions from the estimated weights is sufficient, entropy balancing computes weights that adjust for known sample distributions by building covariate balance

directly into the weights using a pre-defined set of balance constraints, which suggest that the sample moments in the reweighted comparison group exactly match the analogous moments in the treatment group. The weighting process computes weights that are optimised to meet the imposed constraints, while at the same time they result in an entropy distance that makes them as similar as possible to base weights. According to Hainmueller (2012), this approach produces not only improved balance in the covariate distribution since the inclusion of balance constraints for the moments of all confounders rules out the possibility that balance decreases on any of the specified moments, but also retains full information for subsequent analysis.

4.7 Empirical case study

4.7.1 Research question

The substantive part of the empirical investigation builds on the background literature presented in *Chapter 3*. However, given the focus of the thesis on comparing econometric methods for the purpose of evaluating the cost-effectiveness of health care interventions using observational data, for pragmatic reasons, the analyses described below were restricted only on one outcome of the Birthplace cohort study. Indeed, Hollowell et al. (2011) have pointed out that the empirical work undertaken in the area of planned place of birth has so far focused on perinatal

outcomes, with some sparse evidence also indicating that there is a higher likelihood of a normal birth with less intervention for healthy women who plan to give birth at home or in a midwifery unit compared with planned obstetric unit births. As such, the maternal outcome 'normal birth', defined by Hollowell and colleagues (2011) as "a birth without induction of labour, epidural or spinal analgesia, general anaesthetic, forceps or ventouse, caesarean section, episiotomy, third or fourth degree perineal trauma, blood transfusion, admission to an intensive therapy unit or high dependency unit or specialist unit, and maternal death within 42 days of giving birth", is used here as the effectiveness outcome to further explore the research question concerning whether planning birth in different settings in England is cost-effective in the short term. This is achieved in the context of an adjusted analysis, which builds on the one exemplified by Schroeder and colleagues (2011).

4.7.2 Study design & data

Detailed microdata were obtained from Birthplace, which collected data from every National Health Service trust in England. The collection of data included home birth services, freestanding midwifery units, alongside midwifery units, and obstetric units. The study population of the analyses presented here comprised of eligible women at 'low risk' of complications prior to the onset of labour and similarly to the other analyses undertaken with Birthplace, it was restricted to women for whom there were no missing data with respect to the outcome and potential confounders. An

'intention-to-treat' approach was employed, with women analysed in the group in which they planned to give birth at the start of care in labour, irrespective of whether they were transferred during labour or immediately after birth. No individual women were identifiable in this case study.

Two-way comparisons of birth modalities were undertaken, a decision made in light of Rubin's observation that in the context of the propensity score, estimation of treatment effects using one model can be deceptive than estimation in the two-group setting (Rubin, 1997). This is because the model being used to compare one pair of groups (for example, planned birth at home compared with an obstetric unit) is affected by the data from the third and fourth group (planned birth in a freestanding and alongside midwifer unit), which probably have covariate values that differ from those in either one of the other two groups being compared. In the two-way comparisons presented in this case study, the obstetric unit group, in accordance with Hollowell et al. (2011), was chosen as the comparison intervention because it contained the largest number of eligible births and as such, its use as a reference group maximised statistical efficiency.

Similarly to the cost-effectiveness study undertaken by Schroeder et al. (2011), the time horizon of the analysis consisted of the duration of follow-up of the Birthplace cohort study and the perspective of the evaluation was that of a health system including only direct costs to the National Health Service. Detailed information regarding the approaches used for costing resource use items is reported in the cost-effectiveness analysis undertaken by Schroeder et al. (2011). With regards to

presenting the cost-effectiveness of planning birth in different settings, an explicit choice was made to use the net benefit as a summary outcome measure, despite the fact that the incremental cost-effectiveness ratio is the recommended approach for summarising costs and effectiveness (National Institute for Health and Clinical Excellence, 2008). The incremental net benefit avoids the problems associated with ratio statistics since it is a scalar transformation of the incremental cost-effectiveness ratio (Drummond et al., 2005). As regards choice of a threshold value, currently there is a paucity of revealed and stated preference literature on the economic value placed on the unit of health gain, namely an additional normal birth, with no contingent valuation studies or discrete choice experiments in this area. Given the absence of an empirical estimate of the threshold willingness-to-pay (lambda), the analyses were carried out for a range of values, the chosen levels of which relied on values that were already reported in relevant literature (Schroeder et al., 2012; Kreif et al, 2012a; Kreif et al, 2012b).

4.7.3 Baseline cost-effectiveness analysis

Net benefit regression

Regression analysis using the net benefit statistic has some important advantages for analysts undertaking cost-effectiveness analysis using microdata. According to Hoch (2009) these include the direct application of well-established econometric methods

that can analyse observational data, performing easily model fit diagnostics, and straightforward calculation of confidence intervals for the purpose of assessing the stochastic uncertainty in the incremental net benefit estimate. The analyst can also explore the heterogeneity of individuals by including in the regression model a vector of individual characteristics, while the cost-effectiveness when treatment effects are heterogeneous can be estimated for different subgroups by adding appropriate interaction terms. The advantages of net benefit regression have already been exemplified in numerous studies in the cost-effectiveness literature employing observational microdata and this approach is used here as a benchmark method (Shih, Bekele and Xu, 2007; Soegaard et al., 2007; DeRidder and De Graeve, 2009).

For each combination of planned places of birth, linear regression models estimated using ordinary least squares were first fitted to the data to obtain a benchmark incremental net monetary benefit:

$$\lambda e_i - c_i = \beta_0 + \beta_1 TX_i + \varepsilon_i \qquad (4.6)$$

where λ is the threshold willingness to pay, e_i represents the binary 'normal birth' effectiveness outcome of the cohort study, c_i is the total cost for woman i, TX_i denotes a binary treatment indicator for planned place of birth, and ε_i is a stochastic error term with $\varepsilon_i \sim N(0, \sigma^2)$. A different model adjusting for a large number of potential confounding variables that were available in the cohort study was then estimated:

$$\lambda e_i - c_i = \beta_0 + \beta_1 TX_i + \sum \beta_j X(j) + \varepsilon_i \qquad (4.7)$$

This model adjusted for maternal age, ethnic group, understanding of English, marital or partner status, body mass index in pregnancy, index of multiple deprivation score, parity, as well as gestational age at birth and provided the baseline results of the adjusted cost-effectiveness analysis. All explanatory variables were analysed as unordered categorical variables in line with the studies by Hollowell et al. (2011) and Schroeder et al. (2011). The analysis subsequently estimated binary probit models using all these potential confounders in order to obtain the conditional probability P_i of a woman being assigned into each group. The use of the propensity score here builds upon previous work in econometrics and cost-effectiveness analysis, which has highlighted its advantages on efficiency and parsimony arguments (Indurkhya, Mitra and Schrag, 2006; Mitra and Indurkhya, 2005). The estimated propensity score was inserted in a linear regression model as a covariate in order to obtain the incremental net monetary benefit:

$$\lambda e_i - c_i = \beta_0 + \beta_1 TX_i + \beta_2 P_i + \varepsilon_i \qquad (4.8)$$

Hoch et al. (2002) have shown that one can determines the cost-effectiveness of a new treatment by simply considering the following hypothesis test:

$$H_0: \beta_1 \geq 0 \quad \text{vs} \quad H_1: \beta_1 \leq 0 \qquad (4.9)$$

If the null hypothesis is not rejected, then one can conclude that the new treatment (in this case the alternative planned place of birth) is cost-effective.

A number of statistical tests were applied to the baseline cost-effectiveness results from equations 4.7 and 4.8 using inbuilt functions from the software package Stata version 11.2 (StataCorp, 2009). The Breusch-Pagan test was used to test for heteroskedasticity of unknown form (Gujarati, 2003), while the Regression Specification Error Test (RESET) explored whether the regressions are misspecified (Ramsey, 1969) Finally, the Shapiro-Wilk W test assessed the normality assumption (Shapiro and Wilk, 1965; Royston, 1992).

4.7.4 Comparison of methods

Seemingly unrelated regression

In contrast with net benefit regression, the use of seemingly unrelated regression so far has not been widespread outside the context of trial-based economic analyses, despite the fact that it can potentially provide additional advantages for the purpose of evaluating cost-effectiveness (Willan, Briggs and Hoch, 2004). Recent advances in the methodological literature have instead focused on the use of independent models for costs and effectiveness, with the joint uncertainty in the incremental cost-effectiveness to be accounted using non-parametric bootstrapping (Kreif et al., 2012a). In addition, as the literature review revealed, the use of the propensity score in the context of seemingly unrelated regression has not been considered yet.

Here, a two-equation seemingly unrelated regression model was used, where both equations for costs and effectiveness are normal linear regressions. The same potential confounders that were employed in the net benefit regression were included in the model both individually, as well as after they were summarised by the propensity score. The analysis also considered specifications where one equation used all potential confounders, while the other employed the propensity score. Maximum likelihood was used in all cases to estimate the seemingly unrelated model, assuming a joint normal distribution for the error terms.

The specification for the clinical endpoint was motivated from the fact that the linear probability model provides a direct way of obtaining the marginal effect, unlike non-linear models such the probit. Angrist and Pischke (2009) argue that although non-linear models may provide a better fit to the data than a linear model, this in practice probably matters little for marginal effects. In addition, the same authors point out that a number of decisions come into play when non-linear models are used, unlike ordinary least squares, which constitutes a standardised approach in obtaining marginal effects.

Matching estimators

An extensive matching exercise extended the analysis by comparing both traditional matching estimators, as well as innovative algorithms. The motivation for using several different traditional matching methods rested on the findings of the literature

review, which revealed that a comprehensive comparison of matching methods in the context of cost-effectiveness has not taken place yet. In addition, novel approaches such as coarsened exact matching entropy balancing and genetic matching have been shown to have distinct advantages over traditional matching estimators. For example, the genetic matching solution proposed by Diamond and Sekhon (2005) focuses on achieving covariate balance based on an optimisation problem that strives to achieve the most optimal results. Genetic matching in economic evaluation has already been compared extensively against traditional matching estimators in a series of studies undertaken by Sekhon and Grieve (2012), Radice et al. (2012) and Kreif et al. (2012b). As such, the focus in this comparative case study was on coarsened exact matching and entropy balancing, both of which had not been considered in cost-effectiveness analysis before. Emphasis was also placed on doubly robust and bias-corrected implementations of the chosen estimators, which as the literature review revealed, up to very recently their use has not been exemplified in cost-effectiveness analysis.

Coarsened exact matching employed Scott's rule for the binning algorithm that generated the cutpoints used for the coarsening, whereas entropy balancing used a tolerance of 0.005. Other estimators considered include inverse probability weighting; propensity score nearest neighbour matching with replacement, as well as without; ratio matching with replacement using the three nearest neighbours in terms of their propensity score; caliper matching without replacement and radius matching with replacement, both using a maximum distance of 0.25 standard deviations of the probit of the propensity score; kernel matching using both Tricube and Gaussian kernels with a bandwidth of 0.06; and finally matching using the Mahalanobis distance on its

own, as well as combined with the propensity score. Since the debate regarding choice of estimator parameters such as caliper width and bandwidth currently remains inconclusive (Imbens, 2004), the values used in this matching exercise relied on choices that are typically made arbitrarily by analysts in the causal inference literature (Imbens and Wooldridge, 2009).

The user-written programmes -CEM- (Blackwell, Iacus, King, Porro, 2012), -EBAL- (Hainmueller and Xu, 2011) and -PSMATCH2- (Leuven and Sianesi, 2003) for the software package Stata version 11.2 (StataCorp, 2009) were first used to match women between the different groups. All methods matched women in terms of the potential confounders introduced in the baseline analyses. For coarsened exact matching, categorical variables were collapsed into binary variables. For all other matching methods, the estimated propensity score was used to match women. Inference was carried out using the seemingly unrelated regression solution described above. The parametric model adjusted for remaining imbalances and was appropriately weighted to reflect the contribution of each matching scheme in the calculation of the treatment effect.

Differences in women's pre-treatment characteristics were assessed, between the non-obstetric and the obstetric unit groups, as well as before and after matching, using the standardised percentage difference. This measure is the percentage difference of the sample means in the exposed and comparison sub-samples as a percentage of the square root of the average of the sample variances in the exposed and comparison groups (Rosenbaum and Rubin, 1985b). In addition, the McFadden

pseudo R^2 from probit estimation of the propensity score on all potential confounders was calculated for the original and the matched samples (Sianesi, 2004), while the p-value of the likelihood ratio test after matching tested the hypothesis that the regressors are jointly significant (Smith and Todd, 2005). Finally, graphical diagnostics were also employed. These involved assessment of common support through examination of the distribution of the propensity scores in the matched samples, as well as plots of the standardised percentage differences, which provide a summary of the balance achieved in individual covariates (Stuart, 2010).

4.7.5 Sensitivity analysis

A treatment effects model (Maddala, 1983) was first used to test whether the choice of planned place of birth in the baseline net benefit model is subject to selection bias. A two-step estimation process was employed, with a probit model for confounders related to the planned place of birth used at the first stage, and a linear regression model with the inverse Mill's ratio as one of the explanatory variables in the net benefit regression at the second stage. Evidence of selection bias can be found from having a statistically significant coefficient associated with the inverse Mill's ratio (Wooldridge, 2009). In an effort to avoid perfect multicollinearity, the estimation considered all potential confounders in the selection regression, but omitted some of them in different specifications of the outcome model. However, it should be stressed that in all cases no exclusion restriction was used for identification.

A second approach to sensitivity analysis explored the impact of unknown unobserved confounders on the clinical outcome. Rosenbaum's bounds use a gamma parameter, which measures the degree of departure from a study that is free from hidden bias (Rosenbaum, 2002). The idea behind this approach is to control the odds of receiving an intervention as a test of how much the estimated treatment effects may vary in the outcome of interest (Berk, 2004). Women were first propensity score matched using -PSMATCH2- in order to make the treatment groups comparable. A caliper of 0.25 standard deviations of the probit of the propensity score was imposed. Next, the user-written programme -MBOUNDS- was employed to compute Mantel-Haenszel bounds for the binary effectiveness endpoint (Becker and Caliendo, 2007). Several values of gamma were considered in order to determine whether the baseline results are robust to a range of selection biases.

In a third piece of sensitivity analysis, alternative specifications of net benefit regression and seemingly unrelated regression models were simulated. Such simulations serve as a bridge between the baseline analysis and various scenarios, giving a sense of how sensitive the results can be when known potential confounders considered in the baseline analysis are assumed unobserved and thus the unconfoundedness assumption is violated. For example, the incremental net monetary benefit was estimated assuming that body mass index in pregnancy was an unmeasured covariate. Similarly, incremental net monetary benefit estimates were obtained assuming that parity information was not recorded and maternal age remained undocumented. A more extreme scenario, where body mass index, index of multiple deprivation score and maternal age were assumed unobserved, was also

explored. It should be noted that in the context of seemingly unrelated regression, the simulated specifications considered the omission of the aforementioned confounders only in the cost or the effectiveness equation.

In a final set of sensitivity analysis, a number of scenarios exploring the imposition of common support restrictions using differential caliper widths were considered. More specifically, two of them were based on matching on the probit of the propensity score using calipers of width equal to either 0.1 or 0.6 of the standard deviation of the probit of the propensity score, and five of them based on fixed caliper widths of 0.005, 0.01, 0.02, 0.03, 0.05, and 0.1 on the propensity score scale. The arbitrary choice of crucial estimator parameters, although unavoidable, it can have important practical implications for the analysis. For example, nearest neighbour matching without replacement, but within a specified distance, constitutes a very common implementation of propensity score matching (Austin, 2008). Currently, there is a paucity of research aiming to determine the most appropriate caliper width for estimating average treatment effects when using this matching approach (Austin, 2008; 2009a). The literature review also revealed that in the cost-effectiveness literature specifically, so far, there has been no consistency in the calipers that have been used for matching individuals.

4.7.6 Subgroup analysis

A different type of selection bias that can take place even when the unconfoundedness assumption holds, is related to the aggregation of individuals across heterogeneous subpopulations (Blundell and Costa Dias, 2009). Heterogeneity in treatment effect is the non-random, explainable variability in the direction and magnitude of treatment effects for individuals within a population, which arises from characteristics that can potentially modify the effect of an intervention on outcomes (Varadhan and Seeger, 2013). This is because differences in individual characteristics will typically lead to heterogeneous responses to treatment. Hollowell et al. (2011) point out that complicating conditions identified at the start of care in labour were more common among low risk women in the planned obstetric unit group, suggesting that the groups were not homogeneous with regard to risk. In order to explore the role of parity and risk status of women in treatment effects, a series of pre-specified exploratory subgroup analyses was undertaken, similarly to those exemplified in the study by Schroeder and colleagues (2012). These repeated the baseline analysis by parity subgroup and for women without complicating conditions at the start of care in labour. A seemingly unrelated regression model adjusting for the potential confounders described earlier was used to obtain estimates of incremental costs, incremental effectiveness and incremental cost-effectiveness in different subgroups.

4.8 Concluding remarks

Several methods now exist in econometrics that can measure treatment effects. These can readily be applied in health economic evaluation using the net benefit framework. Nevertheless, the debate regarding the ability of different techniques to adjust for selection bias is currently inconclusive. As such, the empirical part of the thesis presents a comprehensive comparison of regression and matching methods in an attempt to contribute to the growing evidence-base of this literature. The analyses undertaken exemplify innovative approaches that address in an explicit manner some of the particular challenges associated with the use of these methods in cost-effectiveness analysis, as these were identified by the literature review. This is achieved in the context of a case study aiming to evaluate certain aspects of planned place of birth in England. Nevertheless, the methods employed depend on a number of assumptions for recovering the treatment effect. The thesis attempts to explore the plausibility of these assumptions by considering the prior scientific literature, as well as relevant expert opinion, while at the same time using an extensive range of statistical tests, graphical diagnostics, and sensitivity analyses.

CHAPTER FIVE

RESULTS & DISCUSSION

5.1 Introduction

This chapter offers an in depth discussion of the results obtained. The next section first reports the key substantive results of the case study. The section also summarises the findings from the comparisons of regression and matching methods. Section 3 presents a careful critique of the methods employed, with the emphasis placed on assessing the credibility of the key assumptions on which these methods rely. These include the unconfoundedness and common support assumptions, good covariate balance after matching, assumptions of the functional form, as well as the assumption of homogenous treatment effects. Section 4 summarises and concludes.

5.2 Baseline findings

The total number of observations in the cohort study comprised of 79,774 eligible women, of whom 64,538 were classified as 'low risk'. In each planned place of birth setting, the completeness of data was over 95% for 'low risk' women and only 3.9% of births excluded from the analyses because of missing data. After these amendments, the observations were limited to 62,036 women, of whom 18,847 planned to give birth in an obstetric unit, 16,187 planned to give birth at home, 10,971 planned to give birth in a freestanding midwifery unit and 16,031 planned to give birth in an alongside midwifery unit.

While the probit models in this case are primarily used as an analytical tool to estimate the conditional probability of women receiving treatment, they can also shed light on inequalities in planned places of birth. Indeed, as it can be seen in the Tables A7-A9 in the Appendix, for all comparisons, the coefficients indicate that a number of covariates are important predictors of treatment status. For example, most deprived women and women aged between 35 and 40 are more likely to plan birth at home, as are women with one or more previous pregnancies. Younger women, as well as women with one or more previous pregnancies are more likely to plan birth in a freestanding midwifery unit, while women below 24 years and women in the age range 30-60 are less likely to plan birth in an alongside midwifery unit. Nulliparous and multiparous women also have a higher chance of planning birth in an alongside midwifery unit, as have women with some knowledge of English and women at 37

weeks of gestation. On the other hand, single/unsupported women, with some or no knowledge of English and over 30 years old, are less likely to plan birth in a freestanding midwifery unit. In addition, women belonging to Indian/Bangladeshi, Pakistani, or Black African ethnic groups are less likely than the reference group (White) to have planned birth at home and in a freestanding or an alongside midwifery unit. Finally, women from areas with an index of multiple deprivation score of more than 8.32 are also more likely than those in areas with a score between 0.37 and 8.31 to plan birth in a freestanding midwifery unit, while the opposite is true for women planned birth in an alongside midwifery unit. Yet, all this serves to show that there is likely selection bias on the observables.

Table 4.1 offers a summary of the results obtained for incremental costs, incremental effectiveness and incremental net monetary benefits for each group comparison, with the results from all the specifications considered available in the Appendix. The first column of the table provides a conventional cost-effectiveness analysis of the data, while the rest of the columns provide the estimates obtained from the different regression models adjusting for observable characteristics. Women both in the home group, as well as in the freestanding and alongside midwifery unit groups have significantly lower costs and a much higher chance of having a normal birth, suggesting that these alternative interventions dominate the option of planning birth in an obstetric unit. For all specifications, the confidence intervals of the incremental net monetary benefits although fairly wide, they contain only positive values, a finding which in the context of net benefit regression suggests that planning birth in a setting outside obstetric units is cost-effective when decision makers value an extra normal

Table 4.1 – Baseline incremental estimates

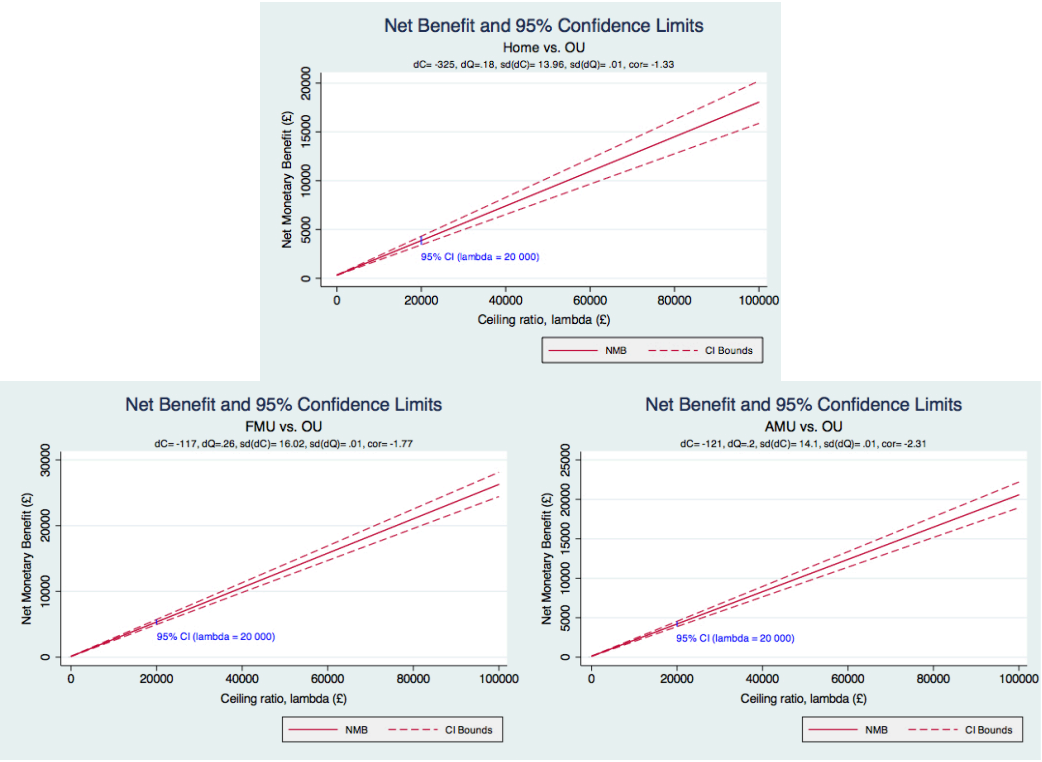
	Unadjusted	NBR	SUR
Home vs. Obstetric Unit			
Cost	-565	N/A	-325
CI	-591 – -538	N/A	-353 – -296
Effect	0.25	N/A	0.18
CI	0.23 – 0.27	N/A	0.16 - 0.20
NMB	5,614	3,869	3,869
CI	5,203 – 6,025	3,415 – 4,322	3,429 – 4,309
Freestanding Midwifery Unit vs. Obstetric Unit			
Cost	-196	N/A	-118
CI	-229 – -163	N/A	-149 – -86
Effect	0.29	N/A	0.26
CI	0.27 – 0.30	N/A	0.25 – 0.28
NMB	5,893	5,349	5,349
CI	5,552 – 6,235	5,011 – 5,687	4.971 – 5,727
Alongside Midwifery Unit vs. Obstetric Unit			
Cost	-170	N/A	-121
CI	-200 – -141	N/A	-149 – -93
Effect	0.22	N/A	0.21
CI	0.20 – 0.23	N/A	0.19 – 0.22
NMB	4,496	4,211	4,211
CI	4,164 – 4,829	3,884 – 4,537	3,878 – 4,544

Notes: Table presents cost, effect and NMB differences, as well as associated 95% confidence intervals. NBR: net benefit regression; SUR: seemingly unrelated regression; N/A: not applicable; NMB: net monetary benefit; CI: confidence interval. NBR and SUR adjusted for maternal age, ethnic group, understanding of English, marital or partner status, body mass index in pregnancy, index of multiple deprivation score, parity, as well as gestational age at birth.

birth (or the probability of having a normal birth as it is the case in seemingly unrelated regression) more than £20,000. The absence of zeros in the computed confidence intervals also indicates that the data provide a definitive conclusion about whether planning birth in a non-obstetric unit is a cost-effective option. Figure 5.1 also reveals that the economic conclusions are not sensitive to different willingness-to-pay values and that planning birth in a non-obstetric unit is cost-effective for women regardless of the willingness-to-pay for the probability of having a normal birth.

A number of points regarding the performance of the different regression approaches can also be made. First of all, both net benefit regression, as well as seemingly unrelated regression reduced the magnitude of the coefficients considerably, albeit to more or less the same extent. This was the case for all the comparisons undertaken. Second, the confidence intervals in all cases exhibited considerable overlap, indicating that the amount of uncertainty arising from sampling variation is comparable across approaches and specifications. Third, the results clearly show that although seemingly unrelated regression indeed has the potential to yield more efficient estimates than the net benefit regression, these gains may be relatively modest. Notably, as it can be seen in Tables A16-A18, the performance of the regression approaches continued to be similar in the alternative specifications considered. For example, summarising the information from several covariates into a propensity score does not seem to produce gains in terms of bias reduction, with the point estimates from the covariate adjusted models only being marginally different from those adjusting using the propensity score. The proclaimed advantage of propensity score analysis in terms of improved

Figure 5.1 – Net benefit estimates for varying willingness-to-pay values



efficiency was also not observed in either the net benefit regression or the seemingly unrelated regression, with the confidence intervals being wider in nearly all cases. This finding was also consistent in all simulated specifications and for all the willingness-to-pay values considered. Further analysis employing the propensity score in only one of the equations of the seemingly unrelated regression model did not result in any sizeable gains in terms of efficiency.

The findings from the matching exercise reveal that after correcting for remaining imbalances using parametric modeling, the different matching methods generally produced estimates (presented in detail in Tables A11-A13 in the Appendix) that are

very similar to those obtained from the regression approaches. In addition, with the exception of Mahalanobis matching and nearest neighbour matching with replacement, the rest of the bias-corrected estimates remained reassuringly similar to each other, while they also presented a similar degree of uncertainty. Considerable overlap was particularly observed in the confidence intervals produced by entropy balancing, inverse probability weighting, nearest neighbour matching, caliper and radius matching, as well as kernel matching. The results remained robust when different caliper widths were considered in the context of caliper matching, with the estimates and their associated confidence intervals being virtually indistinguishable both to each other, as well as to those obtained from the baseline analysis. Finally, all findings were consistent for higher ceiling ratio values.

The similarity of the results obtained should not be regarded surprising since as the sample size increases, the more different estimators will resemble comparisons of exact matches (Smith, 2000). However, it is important to stress that despite their commonality, the reported estimates may not be in all cases directly comparable because certain methods such as caliper matching can change the estimand. Indeed, the success of matching methods in obtaining credible causal effects in a particular context depends on two fundamental considerations (Imbens, 2004; Stuart, 2010). First, different matching procedures by definition have the ability to recover different treatment effect parameters. For example, approaches that discard individuals cannot obtain the average treatment effect. In addition, the treatment effect parameter that can be recovered also depends every time on the degree of overlap that can been achieved for a particular (sub)population. If overlap is adequate for all individuals,

then the average treatment effect can be estimated. However, when the analyst has to enforce common support restrictions in order to improve the quality of matches, then only the average treatment effect on the treated can be estimated. In case common support restrictions are also placed on the exposed group, then the estimated treatment effect will only be relevant for the subpopulation consisting of the exposed individuals that were matched and not discarded.

In addition, all different matching procedures are inherent to a bias-efficiency trade-off. Methods that use the most similar individual to construct the matched counterfactual will generally minimise bias, but they will ignore a lot of information from the sample because a lot of non-exposed individuals will be discarded during the pairing process. Therefore, any reduction in bias will come at the expense of lower efficiency, that is, decreased precision caused by a higher variance. Conversely, methods that use multiple individuals can be more efficient since they exploit a larger quantity of information from the pool of non-exposed individuals. Nevertheless, matching using multiple individuals also has the potential to lead to increased amounts of bias, which will typically result from poorer matches. Popular matching methods that reflect the above trade-offs include caliper and kernel matching. For both of these methods, the robustness of the results obtained depends on choosing an appropriate caliper width for the former, as well as a kernel function and bandwidth for the latter. Knowing a priori what parameter levels are reasonable to achieve good balance will often be difficult and as Imbens (2004) points out choice will generally rest on personal judgement. Overlap, common support restrictions and covariate balance are topics that are explored in detail in the next section.

5.3 Critique of methods

5.3.1 Assumption of unconfoundedness

The extent to which these estimates reflect causal statements depend primarily on whether the fundamental assumption of unconfoundedness holds (Imbens and Wooldridge, 2009). Because in the majority of cases the exact factors that govern selection into treatment will be unknown, it is recommended that analysts exploit all the variables that are observed and at the same time potentially relevant (Imbens, 2004). A statistical covariate elimination approach that is typically used by analysts to determine the most appropriate specification for a model was thus not used here as the goal was not to obtain a parsimonious model, but rather to adjust for as many potential confounders as possible, which expert opinion and previous research has indicated as important. For example, age, body mass index and index of multiple deprivation are basic demographics that are almost always considered as potential confounders in all type of studies (Hennekens and Buring, 1987). Other potential confounders, such as parity and gestational age at birth are always related to perinatal outcomes (Macfarlane, 2000), and in this particular case they were also clearly related to the exposure (Brocklehurst, 2011).

Nevertheless, the results presented above essentially suggest that planned place of birth in non-obstetric settings always confers cost savings and striking clinical benefits. Consequently, they warrant careful interpretation. Indeed, the results from the

RESET tests offered indication for misspecification in all the net benefit regression models comparing alternative planned places of birth. As such, one might conclude that the estimates presented in Table 4.1, as well as the Appendix are biased, a finding which appears to be at odds with the rigorous prospective design and the comprehensive data collection of the Birthplace cohort study. The first point to note with regards to this, is that although self-selection of women into the cohort study was not a concern as womens' consent to participate was not required, selection bias can in fact still arise from endogeneity attributed to measurement errors in key explanatory variables. For example, body mass index was not always recorded and detailed data about the socio-economic status of individual women were not collected as part of the cohort study (Hollowell et al., 2011). For the latter, an index of multiple deprivation score was used instead, which although it is based on the socio-economic characteristics of the area where the woman lives, it merely summarises a number of economic and social indicators into a single metric (McLennan et al., 2010).

In addition, although sparse, the extant substantive literature provides some indication that it is still likely the results are plagued by selection bias because planned place of birth may be correlated to confounding factors that were not collected, and which influence choice of women regarding birth modality (Janssen et al., 2009; Ahsan, Li and Streatfield, 2007). For example, other things being equal, if drug and alcohol use during pregnancy do lead women to plan birth in an obstetric unit and to lower probability of having a normal birth, then the analysis without taking into consideration this confounder will bias upwardly the effect of planned place of birth on normal birth. On the other hand, if drug and alcohol use force women to

plan birth in an obstetric unit and result in higher costs, then the effect of planned place of birth on costs will be underestimated. Another important confounder that remained unmeasured and which could also impart bias is education. For instance, other things being equal, if higher education facilitates a better assessment of available information encouraging women to plan birth at home, while at the same time increases the probability of women having a normal birth, the effect of planned place of birth on normal birth will be downwardly biased. In the same line of thinking, if other things being equal education leads to better-informed choices prompting women to plan birth at home and thus lower costs, then the effect of planned place of birth on costs if education is not considered will be upwardly biased. Other unmeasured potential confounders that may also be associated with the treatment assignment and the cost or effectiveness endpoint in this context include employment, smoking status, religion, NHS trust size, as well as type of area of residence (urban versus rural). Such characteristics are difficult to be collected and stored in a way that analysts can readily access without compromising confidentiality. Most importantly, the distributions of such unobservable characteristics are likely to differ across demographic and socioeconomic groups, compromising the ability to construct an analyst's perception regarding a woman's choice of planned birth modality.

The above concerns were first investigated using a Heckit-type econometric model. This is essentially a control function approach that aims to test and potentially correct for different selection biases in observational studies by estimating the conditional probability of receiving treatment. The key difference of this approach from propensity score analysis is that it does not assume that selection is random

conditional on covariates; rather it sees selection as a non-random process that needs to be explicitly recognised and modelled (Guo and Frazer, 2010). Wooldridge (2009) points out that although one can use Heckit-type models to routinely test for the presence of selection bias, identification will be problematic if the same covariates are included in both the selection and outcome regressions, since this will rely exclusively on functional form and distributional assumptions. Here, different subsets of observable characteristics were considered in the outcome regression. In all cases, despite the fact that the inverse Mill's ratio was not found to be statistically significant and hence offer evidence in support of selection bias, the estimates obtained were implausible and very different to those of the baseline analyses. This finding casted serious doubts regarding their credibility. For example, Table A14 in the Appendix reveals that for specifications omitting the index of multiple deprivation score, the incremental net benefit for the comparison of planned birth at home versus in an obstetric unit is clearly considerably higher from the estimate obtained in the conventional (unadjusted) cost-effectiveness analysis. Similarly, the incremental net benefit for the comparison of planned birth in an alongside midwifery unit versus an obstetric unit is almost half of the adjusted estimate, suggesting that the latter is upwardly biased. In fact, these imprecise estimates are probably an artefact arising from high multicollinearity between the inverse Mill's ratio and the potential confounders included in the outcome regression. The results of the Heckit models highlight a finding that is empirically well-documented in the economics literature and which emphasises that the estimates obtained from this approach will rarely be convincing in the absence of variables that can induce exogenous variation in the treatment assignment of individuals (Blundell and Costa Dias, 2009). In a health

economics context, these may be constructed using characteristics of health care professionals and their practices, as well as detailed geographic data for patients and providers (Jones, 2011). For example, for this case study one could exploit supply side information about the provision of maternity services such as the number of midwives in NHS trusts, in order to construct an instrument that could potentially be strongly correlated with choice of planned place of birth but not with normal birth. However, such information was not recorded in the Birthplace cohort study and consequently the analysis had to rely exclusively on the unconfoundedness assumption.

In the absence of a strong exclusion restriction, further sensitivity analysis was undertaken in an attempt to explore hidden biases. First, Rosenbaum's bounds (Rosenbaum, 2002) explored the likely impact of deviations from the selection on observables assumption on normal birth. The propensity score matching of women resulted in a significant positive treatment effect and as the output in Table A15 indicates, the results were not sensitive to possible deviations from the identifying unconfoundedness assumption, with the treatment effect being in all cases strongly significant similarly to the baseline analysis. The results obtained from Rosenbaum's approach to sensitivity analysis although based on a mathematically neat methodology, in practice require careful interpretation. For example, a gamma value of 1 indicates that the study is free of hidden bias. In contrast, a gamma value of 10 does not state that unobserved bias exists, but that the confidence interval for the effect would include zero if an unobserved variable caused the odds ratio of treatment assignment to differ between the treatment and comparison groups by 10. Ultimately, as Becker and Caliendo (2007) also point out Rosenbaum's approach cannot directly justify the

unconfoundedness assumption and as such it cannot state whether this assumption holds or not for the used data, the chosen covariates and the specification of the propensity score. Robins (2002) also emphasises that in order this approach to be truly informative, one must have access to expert opinion, which can provide a plausible range for the value of the sensitivity parameter gamma. The analyses undertaken here should be regarded as exploratory since no consensus could be obtained among the experts consulted regarding appropriate choice of gamma values.

Of more informative nature are perhaps the findings from the alternative regression specifications. The chosen covariates reflect important substantive predictors of treatment assignment (maternal age and parity), as well as variables that were subject to measurement errors (body mass index and index of multiple deprivation score). The results in Tables A16-A18 clearly raise a number of important points about the sensitivity of the adjusted incremental net monetary benefit estimates. For women planning to give birth in an obstetric unit versus any other setting, the incremental net monetary benefit estimate, when parity is unobserved, is almost the same with the unadjusted estimate for the baseline willingness-to-pay value. Moreover, a greater number of unobserved covariates does not always lead to more biased estimates. For example, in the absence of body mass index, index of multiple deprivation, as well as maternal age, the incremental net monetary benefit estimate is lower than that of the estimate when body mass index and maternal age are unmeasured, and when only body mass index is unobserved. This finding, which was consistent for all willingness to pay values, concerned the comparisons of women planning to give in an obstetric unit, versus women planning to give birth at home and in a freestanding midwifery

unit. In addition, the incremental net monetary benefit estimates are in nearly all cases lower than the corresponding unadjusted estimate in Table 4.1. This provides further evidence that the unadjusted incremental net monetary benefit estimates are indeed upwardly biased, because of the systematic differences between the different groups. The results also suggest that although the adjusted incremental net monetary benefit estimates are for the most part less biased, they tend to be similarly sensitive to the assumption of unconfoundedness for all willingness to pay values. The flexibility of seemingly unrelated regression takes the above simulations a step further by allowing the impact of the above scenarios to be considered separately for costs and effects. Interestingly, as Table A19 shows, the omission of confounders in either the cost or the effectiveness equation produces similar results in terms of incremental net monetary benefit. This was also the case for higher ceiling ratios (results not presented here). The findings from the seemingly unrelated regression simulations also reaffirm those of the net benefit regression simulations outlined above.

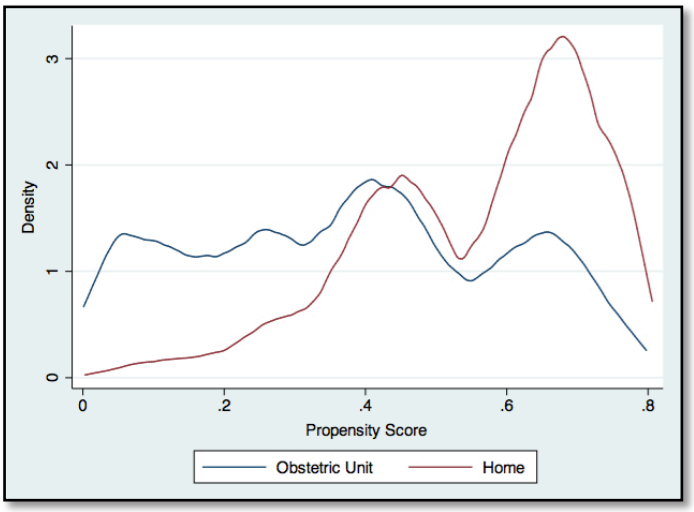
In conclusion, as already explained earlier in the thesis, all the regression and matching methods employed here rely exclusively on the unconfoundedness assumption. Indeed, although it has been pointed out in the causal inference literature that this assumption may not be as limiting as it seems -especially when a large number of potential confounders have been included in the analysis- it is ultimately an assumption that is untestable. As such, despite the extensive sensitivity analyses undertaken as part of this empirical case study, it is important to stress that concerns regarding the presence and impact of unobserved confounding in the estimates reported could not be fully resolved.

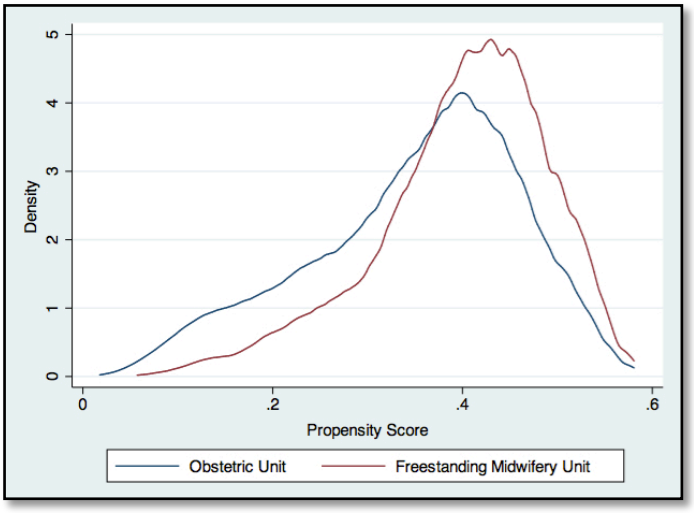
5.3.2 Assumptions of common support & covariate balance

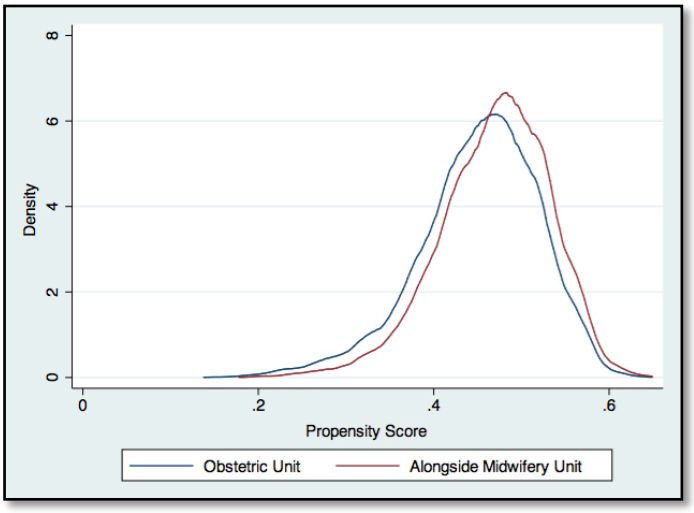
Overlap in terms of observable characteristics is the second condition that is required for reliable identification of average treatment effects using regression and matching methods. Indeed, in most empirical applications, a trade-off between the plausibility of the unconfoundedness assumption and adequacy of the common support condition might be involved. This is because deciding to include more covariates in the model may render more difficult for the analyst finding individuals who share all these characteristics between the groups compared (Stuart, 2010). In addition, matching methods require that covariates between the exposed and comparison groups be balanced after matching takes place (Sekhon and Grieve, 2012).

Figure 5.2 shows the distributions of the propensity scores before matching for women in the treatment groups and their counterparts in the obstetric unit group. The region of overlap is considerable, despite density differences and a right-skewed distribution in the case of women planning birth at home. The corresponding distributions of the propensity score after each matching method was implemented can be found in Figures A1-A3 in the Appendix. Evidently, matching reduced the density differences in most cases, while at the same time maintained the substantial overlap observed before matching. Notable exceptions included nearest neighbour matching without replacement and caliper matching in the comparison of home versus obstetric unit, where both maintained overlap but only marginally reduced density differences.

Figure 5.2 – Distribution of the propensity score







The number of observations lost in the home, freestanding and alongside midwifery unit groups due to common support can also provide a basis with which one can evaluate the different matching methods. As it can be seen in Tables A24-A26, coarsened exact matching pruned a significant amount of treated in all comparisons. Similarly, caliper matching lost a great deal of treated observations in the home versus obstetric unit comparison. Nevertheless, in all other cases losses were zero, confirming that matching in this case study generally does not pose a problem in this regard. Indeed, the common support diagnostics and balancing test results (discussed below), indicate that the differences in the observables between the treatment and obstetric unit groups have the potential to be reduced without the need to prune a large number of observations from the sample due to a lack of overlap. The region of common support after matching also includes virtually the entire sample, which means that the estimated treatment effect does not have to be redefined in the majority of cases.

In addition, the goal of any matching exercise is to balance the covariates in a way that selection on observables is decreased as much as possible. Here, both overall and individual measures of imbalance were considered. Caliendo and Kopeinig (2008) note that a difference reduction below five percent can generally be seen as sufficient. The results reported in Tables A24-A26 are in the vast majority of cases below this value. More specifically, before matching the average standardised percentage difference in all comparisons was considerably high. The majority of matching methods reduced this difference dramatically, although some more than others. Entropy balancing eliminated the difference in all cases, while inverse probability

weighting and Mahalanobis matching offered the next best performance, with the groups exhibiting on average negligible levels of mean differences. Interestingly, the inclusion of the propensity score as a covariate in Mahalanobis matching did not translate in further gains. For nearest neighbour matching, the imposition of a caliper and (separately) of replacement, as well as the use of multiple neighbours in radius matching, enhanced considerably its performance in the home versus obstetric unit comparison, but did not materially affect the mean difference in subsequent comparisons. An important finding was also the fact that Gaussian kernel matching performed in all cases less well than tricube kernel matching, while coarsened exact matching not only performed worse than all the other methods in the comparisons of freestanding and alongside midwifery units versus an obstetric unit, but even increased the average standardised percentage difference in the latter.

More detailed evidence on the quality of matching is provided from the plots of the standardised percentage differences for individual covariates, which can also be found in Figures A4-A6. The plots confirm that before matching, women planning to give birth in one of the non-obstetric unit groups presented severe imbalances in a number of covariates, from women planning to give birth in an obstetric unit. These imbalances concerned different covariate categories for different comparisons, with most prevalent however those related to maternal age, parity, ethnic group and index of multiple deprivation score. As it can be seen, in all comparisons entropy balancing achieved perfect balance, followed by inverse probability weighting and Mahalanobis matching, which also minimised differences in nearly all covariates. Nearest neighbour matching in all cases performed poorly in a large number of covariates

albeit to a varying extent. The use of replacement in the home versus obstetric unit comparison improved considerably its performance, yielding covariate balance comparable to that of ratio matching. In the same comparison, radius matching achieved balance in covariates very similar to that of tricube kernel matching. Caliper and Gaussian kernel matching performed better than nearest neighbour matching, although the former failed to achieve balance in a handful of covariates such as those representing parity categories. In the freestanding midwifery unit versus obstetric unit comparison, for nearest neighbour matching, neither the use of replacement nor the imposition of a caliper seemed to materially affect the balance achieved in the majority of covariates. This was also the case when using multiple neighbours as in ratio matching. In addition, radius matching achieved better balanced in comparison to caliper matching in most covariates. Tricube kernel matching comfortably outperformed in nearly all covariates, both nearest neighbour matching, as well as its counterpart using a Gaussian kernel. Finally, in the alongside midwifery unit versus obstetric unit comparison, the use of replacement in nearest neighbour matching introduced imbalance to some covariates after matching but compensated it by improving balance in others. Imposing a maximum tolerated distance through the use of a caliper achieved virtually the same balance in all covariates, while matching on three neighbours as opposed to one actually worsen balance in the majority of covariates. Tricube kernel matching improved balance in most covariates compared to nearest neighbour matching, but this trend was reversed when a Gaussian kernel was used to match women.

The robustness of the findings reported above was assessed using differential common support restrictions in the context of caliper matching. Evidently, as Table A27 shows, the choice of caliper influences the average standardised difference in the home versus obstetric unit comparison. This exhibits a maximum variation slightly in excess of 2.5 percent, with Figure A13 also revealing that the covariates most sensitive to choice of caliper are related to parity, index of multiple deprivation score and body mass index. In contrast, in the comparisons of freestanding and alongside midwifery unit versus obstetric unit reported in Tables A28 and A29, the average standardised difference remained in all cases constant, while the plots in Figures A14 and A15 demonstrate that all calipers produced nearly identical balance in individual covariates. With regards to overlap, the densities of the propensity score distributions in the home versus obstetric unit comparison after matching (shown in Figure A10) are more dissimilar when a wider caliper width is used. In addition, near perfect overlap was achieved when the fixed calipers of 0.005 and 0.01 on the propensity score scale were used. However, this was not the case for the other two comparisons, where Figures A11 and A12 illustrate that for all caliper widths the densities are virtually identical. Finally, Table A27 presenting overall measures of imbalance also makes clear that the choice of caliper affects the number of women in the home group that are outside the area of the imposed common support. The percentage of treated observations discarded ranged between 11.79 and 29.68, suggesting that the sample average treatment effect for the treated cannot be estimated. A fixed caliper of 0.005 yields the lowest average standardised difference but discards most observations. In contrast, a fixed caliper of 0.1 prunes the smallest number of women but at the same time produces the highest average standardised difference in the

observed range. As expected, this trade-off is also present when the caliper width is calculated as a function of the probit of the propensity score. In addition, the two more strict tolerated distances yield similar reductions in terms of average standardised difference, with the caliper of 0.1 standard deviations of the probit of the propensity score decreasing it to 3.2 percent, while the fixed caliper of 0.005 further dropping it to 2.9 percent. These two caliper widths also discarded a similar number of observations. Unsurprisingly, the opposite was true for the two most relaxed caliper widths. Tables A28 and A29 reveal that the trade-off between caliper width, average standardised difference and number of women pruned is also present in the other two comparisons, albeit significantly weakened. More specifically, in the comparison of planned birth in a freestanding midwifery unit versus obstetric unit, most caliper widths did not lose any treated observations, with the exception of the fixed 0.005 and 0.01 calipers, both of which lost a rather negligible amount. In the comparison of planned birth in an alongside midwifery unit versus an obstetric unit, only the most strict caliper widths lost some women planning birth in an alongside midwifery unit, but in all three occasions the number was trivial.

5.3.3 Assumptions of the functional form

As well as the unconfoundedness and the common support assumptions, parametric methods also have to rely on assumptions of the functional form. The Breusch-Pagan test indicated the presence of heteroskedasticity in the net benefit regression models

and as such, all the analyses were replicated using White-corrected heteroskedasticity-robust standard errors (Huber, 1967; White, 1980). The net benefit regression models also failed the RESET test at five percent significance level. As discussed above, this can be seen as evidence of misspecification arising either from omitted variables, or due to the existence of non-linearity. Since the analyses included all potential confounders that were collected as part of the cohort study, efforts focused on testing supplementary specifications using squared terms and interactions. The results failed to pass the RESET test even after the additional regressors were introduced and as is sometimes the case with the RESET test, there is no obvious resolution to the problem in this particular context. Last, but not least, the results were not found to be normally distributed using a Shapiro-Wilk W test at the five percent significance level. Nevertheless, linear regression models are generally seen as robust against deviations from normality (McGuinness, Bennett and Riley, 1997). As a sensitivity check, bootstrap estimates for the standard errors across the baseline models were obtained, an approach that does not make any distributional assumptions about the observed data. Commonly used numbers of 1,000, 5,000 and 10,000 replications were tested. The results indicated that differences in the standard errors between standard and bootstrapped regression models were relatively small. As such, only the confidence intervals obtained from the former are reported in the thesis. At this point, it is important to stress that the application of parametric methods is largely a matter of convenience, since the functional form assumptions on which they rely are typically unknown and in practice can only be approximately met (Hernán and Robins, 2013). In this case study, because of the very large sample size

of the cohort study and by virtue of the central limit theorem, violations of these assumptions are not particularly problematic.

The seemingly unrelated regression approach considers costs and effectiveness separately. Visual inspection in Figure A19 illustrates that the distribution of the cost outcome exhibits a considerable degree of skewness. A generalised linear regression model with a gamma distribution and identity link function was estimated using maximum likelihood as an alternative way to control for potential confounders and estimate adjusted incremental costs. This method may be seen as more appropriate in handling the skewness without the need to transform the outcome (Basu and Manning, 2009). As regards the binary effectiveness outcome, this is also by definition not normally distributed. Although the linear probability model may typically be adequate for the analysis of binary limited dependent variables, in certain cases it can be subject to caveats. These include its inherently heteroscedastic nature, which can be highly and difficult to correct. Again, a robust covariance matrix was used to account for this heteroskedasticity and obtain as accurate standard errors and confidence intervals as possible (Huber, 1967; White, 1980). Another limitation of the linear probability model is that there may be cases that this will provide estimates outside the range that a probability requires (Angrist and Pischke, 2009). Despite the fact that the predicted values here all fell comfortably between 0 and 1, in order to explore the likely impact of the chosen specification for this equation of the seemingly unrelated model, the marginal effect from a separate probit regression for the effectiveness endpoint was obtained to test the robustness of the results.

Reassuringly enough, as it can be seen in Table A23 the extra analyses did not materially affect the baseline conclusions of the case study. The results indicate that in all cases, the linear model in seemingly unrelated regression constitutes a very good approximation. It is worth pointing out that in cases of extreme discrepancies or implausible estimates arising from the use of a linear probability model, a probit model can instead be used. Bhatacharrya (2004) has shown that this approach in the context of seemingly unrelated regression may even result in some potential efficiency gains for parameters of the limited dependent variable equations but not for parameters of the linear equations. Other Bayesian bivariate approaches for evaluating cost-effectiveness may also result in further efficiency gains, since they can allow the simultaneous estimation of cost and effectiveness using different distributions more appropriate for the data at hand (Nixon and Thomson, 2005; Manca and Austin, 2008). Nevertheless, such approaches can considerably complicate the analysis and one should always be aware of the trade-offs, both between bias and efficiency (Imbens and Wooldridge, 2009), as well as analytical rigour and practical relevance for decision-making purposes (Rovithis, 2009).

A further goal of matching methods is to make violations of assumptions concerning modeling decisions less consequential (Ho et al., 2007). In this context, the McFadden pseudo R^2 and the likelihood ratio test for the joint insignificance of all the regressors from the probit models can provide another indication about the quality of matching. Here, the pseudo R^2 was estimated both in the raw sample prior to matching, as well as on the matched sample using the weights generated from each matching method. Since the pseudo R^2 can be seen as a broad measure of the extent

to which the variation in a sample is explained by the vector of the relevant covariates, once the sample is matched conditioning on this vector, the pseudo R^2 on the matched sample should be lower than in the unmatched case. Similarly, the covariates should no longer be jointly significant after matching. That is, the likelihood ratio test for the joint insignificance of all the regressors should be rejected before matching but not after the matching process takes place. Clearly, the results in Tables A24-A26 demonstrate that in contrast with the pseudo R^2 value obtained from the model on the original unweighted sample, in all cases the value of the pseudo R^2 from the model on the matched sample is almost zero. This suggests that together, the observed covariates after matching generally explain very little or no fraction of treatment propensity. Indeed, the results from the likelihood ratio test confirm that in the majority of cases the joint significance of covariates is rejected (i.e. the p-values of the likelihood ratio are insignificant). In addition, Tables A27-A29 reveal that these findings were also generally consistent for differential calipers, with the pseudo R^2 notably remaining constant in the comparisons of freestanding and alongside midwifery unit versus obstetric unit. Only in the comparison of home versus obstetric unit a more relaxed caliper width resulted in a higher pseudo R^2 value.

5.3.4 Assumption of homogeneous treatment effects

The estimates obtained from the simulated specifications revealed that although the results are generally robust to the baseline specification employed, considerable

Table 4.2 – Incremental estimates from subgroup analysis

	Home vs. Obstetric Unit		
	No complications	Nulliparous	Multiparous
Costs	-273	-313	-353
CI	-303 – -243	-373 – -253	-381 – -326
Effects	0.21	0.22	0.16
CI	0.19 – 0.23	0.18 – 0.25	0.14 – 0.19
NMB	4,428	4,607	3,504
CI	4,039 – 4,818	3,929 – 5,285	2,924 – 4,083
	Freestanding Midwifery Unit vs. Obstetric Unit		
	No complications	Nulliparous	Multiparous
Costs	-51	-73	-154
CI	-84 – -17	-126 – -21	-190 – -117
Effects	0.23	0.29	0.23
CI	0.21 – 0.25	0.27 – 0.31	0.21 – 0.25
NMB	4,655	5,951	4,712
CI	4,265 – 5,044	5,425 – 6,478	4,171 – 5,254
	Alongside Midwifery Unit vs. Obstetric Unit		
	No complications	Nulliparous	Multiparous
Costs	-55	-82	-151
CI	-84 – -26	-128 – -37	-181 – -121
Effects	0.18	0.21	0.19
CI	0.16 – 0.19	0.19 – 0.24	0.17 – 0.21
NMB	3,555	4,350	3,999
CI	3,210 – 3,900	3,893 – 4,806	3,514 – 4,485

Notes: Adjusted cost, effect and NMB differences, as well as associated 95% confidence intervals for women with no complications at the start of care in labour and women by parity. NMB: net monetary benefit (λ=£20,000); CI: confidence interval.

sensitivity was evident when there was omission of parity. The estimates from the subgroup analysis are reported in Table 4.2 and can be interpreted in a manner similar to those obtained from the baseline analysis. As it can be seen, in all cases the alternative interventions are cost-effective and again dominate the option of planning birth in an obstetric unit. At the same time, the estimates also provide evidence for heterogeneity in the treatment effects. For example, for nulliparous women and women with no complications at the start of care in labour, the associated cost savings are fairly similar. However, they are all reduced when compared to the baseline estimates, with the magnitude of the coefficient being considerably lower for women planning birth in a freestanding and an alongside midwifery unit. Interestingly, the opposite is true for multiparous women, the cost differences of which are only marginally higher from the baseline estimates. A striking result is also the magnitude of the disparity for these same women, which in some cases is more than double from that of women with no complications and nulliparous women.

The picture is somewhat different with regards to clinical effectiveness, where the results exhibit less variability. Clearly, multiparous women face a slightly lower probability of having a normal birth in non-obstetric units compared with the baseline, while the opposite is true for nulliparous women. Women with no complications at the start of care in labour have a higher chance of normal birth compared with baseline when planning birth at home, but not when planning birth in a freestanding or an alongside midwifery unit. Furthermore, nulliparous women have a better chance of a normal birth compared with multiparous women, the difference of which in the case of planning birth at home and in a freestanding midwifery unit is rather

substantial. Planning birth in an alongside midwifery unit also results in fairly similar probabilities of normal birth for both the women with no complications, as well as those in the parity subgroups. In contrast, for women with no complications, planning birth in a freestanding midwifery unit produces the same effect size with multiparous women, while the opposite is true when planning birth at home, with the probability of normal birth being almost the same with that of nulliparous women.

5.4 Concluding remarks

The case study presented here provides robust evidence on the short-term cost-effectiveness of planned birth in alternative settings. An important feature of all the analyses undertaken was the use of an intention to treat approach. In this type of approach, data for each individual are kept in the original group assignment, irrespective of whether or not the individual received the allocated exposure (Hennekens and Buring, 1987). The advantage of the intention to treat approach in the context of the Birthplace cohort study is that it preserves the baseline comparability achieved by the rigorous prospective study design. As such, the analyses provide estimates that illustrate what happens in real world practice, by taking into account the women who do not adhere to treatment. For example, for the three non-obstetric unit settings, the percentage of women that was transferred ranged between 21 and 26 percent. Transfer rates were much higher for nulliparous women ranging from 36 to 45 percent, than for multiparous women, which ranged between 9 and 13

percent (Birthplace in England Collaborative Group, 2011). Nevertheless, Hewitt, Torgerson and Miles (2006) point out that when there is no compliance with the treatment, the estimates obtained will be diluted by the data from individuals who do not receive the intervention to which they were initially allocated. In this case, the same authors also stress that an intention to treat approach will answer a different research question; whether the offer of treatment to the intervention population is effective. The substantive findings of this case study indicate that, overall, when clinical effectiveness is defined as normal birth, offering women the choice to plan birth in different places is very cost-effective in the short-term, with planned birth in a freestanding midwifery unit likely being the most cost-effective option. This is also the case for women with no complications at the start of care in labour, as well as for both nulliparous and multiparous low risk women.

This particular case study has some caveats, which should be acknowledged. First, certain limitations arise from the cohort study design, the costing methodology and the perspective of the evaluation undertaken. These have already been explained thoroughly in the reports by Hollowell et al. (2011) and Schroeder et al. (2011). In addition, as also stated in the same chapter, the data in the obstetric unit group are not representative of the population from which they are drawn because of the differential probability of selection of each unit and the duration of data collection in that unit. A typical solution to this problem is to use weights to restore the profile of the sample to that of the population, an approach that was also exemplified in the analysis by Schroeder and colleagues (2011). Accommodating sampling weights in methods relying on the propensity score is a topic currently subject to debate in the

causal inference literature, with analysts typically tending to ignore weights for differential probabilities of selection into the sample (Bryson, Dorsett and Purdon, 2002). This was also the case here, in order to facilitate a straightforward comparison of the results obtained across methods. As such, the estimates presented probably only apply to the specific sample on which the analysis was undertaken and cannot be generalised to the population from which the sample was drawn. Non-random samples may also involve clustering and stratification adjustments. Although the Birthplace data require such adjustments in order to take into account the stratification used in the random sampling of the obstetric units, as well as the clustering of women and babies into units (Hollowell et al., 2011), again for comparative purposes the analyses presented here did not take into account inefficiencies arising from these features of the cohort study design. It is also worth pointing out that all the econometric methods adjusting for selection bias in this case study rely on the unconfoundedness assumption and although sensitivity analysis attempted to explore the likely impact of confounding arising from unobserved sources of bias, ultimately, unconfoundedness is an assumption which cannot be fully tested. Consequently, the possibility that the estimates obtained are not free from selection bias on the unobservables cannot be ruled out. Finally, given the methodological focus of the thesis, the results presented were generated based on a single case study which relate only to the 'normal birth' clinical outcome without exploring other clinical outcomes included the cohort study, such as the primary clinical outcome 'adverse perinatal outcomes' which was the main focus of the studies by Hollowell and colleagues (2011) and Schroeder et al. (2011).

CONCLUSIONS

Health economics has emerged over the past forty years as a major branch of economics that has considerably influenced the way in which the organisation and delivery of health care is undertaken (Wagstaff and Culyer, 2012). Several countries now require well-documented evidence that health care interventions represent good value for money before they approve their use in the health care system (Sorenson, Drummond and Kanavos, 2008). Economic evaluation has evolved into a robust framework that renders the estimation process of the costs and effectiveness of health care interventions a positive research question subject to transparent scientific analysis (Rovithis, 2009). In recent years, the use of microdata for the purpose of evaluating cost-effectiveness has been at the forefront of applied research (Sculpher et al., 2006). The imposed focus on trial-based studies has also sparked an interest in the development of statistical methods that deal with a range of issues relevant to the analysis of such data (O'Hagan and Stevens, 2001; Hoch, Briggs and Willan, 2002; Willan, Briggs and Hoch, 2004; Nixon and Thomson, 2005; Quinn, 2005). In contrast, methodological advances in economic evaluations using observational microdata have been relatively modest (Sekhon and Grieve, 2012). This is a challenging research area, not only because treatment effects from such analyses are prone to bias arising from the non-random manner that individuals are selected into treatment, but also because

the cost-effectiveness evaluative framework requires consideration of a broader range of issues compared to other observational studies.

The contributions of this thesis comprise the extension of knowledge within both the substantive area of organisation of maternity services, as well as methods for cost-effectiveness analysis. The thesis achieves the former by using cutting-edge econometric methodology to adjust key results of the economic evaluation undertaken by Schroeder et al. (2011), strengthening in this way the evidence-base literature on the cost-effectiveness of planned place of birth. On the methodological side, the thesis presents a new structured template that can be used for reviewing and critically appraising economic evaluations employing observational microdata. In addition, the empirical part extends and explores the use of the propensity score in seemingly unrelated regression cost-effectiveness modeling, while also demonstrating innovative matching methods such as entropy balancing and coarsened exact matching, which are used in the context of cost-effectiveness for the first time. Emphasis is placed on doubly robust and bias-corrected solutions relying on parametric bivariate modeling that allow the joint uncertainty in the incremental cost-effectiveness estimates to be recognised. Finally, the thesis exemplifies a first-ever comprehensive empirical comparison of traditional regression and matching methods, which complements other recent studies in the area that compare methods relying on the assumption of no unmeasured confounding such as regression analysis, propensity score matching, as well as genetic matching (Kreif et al., 2012a;b). A more detailed synopsis of the contributions of the thesis to knowledge is given below.

The thesis began by offering an overview of the conceptual literature on economic evaluation and counterfactual analysis, which was then followed by a conceptual review of the econometric methods that can be used to measure average treatment effects when observational microdata are available. This informed the development of a structured template, which incorporated a checklist of items that was subsequently used to review and critically appraise in a thorough manner the application of these methods in the relevant economic evaluation literature. In addition, the thesis discussed and contrasted how different analytic approaches solve the evaluation problem in the presence of selection bias, while also provided a comprehensive account of the issues surrounding their use in cost-effectiveness analysis specifically. Evidently, analytic approaches operate under an array of assumptions, with different estimators being better suited in recovering distinctive treatment effect parameters, under different research questions and data regimes. These considerations will inevitably involve trade-offs between imposed assumptions, bias, consistent and efficient estimation, as well as interpretability of the results obtained. In the context of cost-effectiveness analysis, the review revealed that currently the application of econometric methods is accompanied by a number of important limitations. These as already stated earlier include, among other, lack of good quality evidence regarding the comparative performance of different estimators; inadequate assessment of the sensitivity of their results to violations of fundamental assumptions or variations to crucial estimator parameters; failure to combine the cost and effectiveness outcomes in a summary measure; and no consideration of stochastic uncertainty for the purpose of evaluating cost-effectiveness.

The empirical part of the thesis exploited data from the Birthplace national prospective cohort study to evaluate the short-term cost-effectiveness of alternative planned places of birth in England. The analyses undertaken also attempted to address some of the methodological limitations identified by the literature review. A comparison of regression methods in the context of cost-effectiveness using the net benefit framework was first carried out. Two broad approaches adjusting for selection bias were exemplified and contrasted; net benefit regression and seemingly unrelated regression. The former directly employs an amalgamation of cost, effectiveness and decision-maker's threshold value in a linear regression model adjusting for potential confounders (Hoch, Briggs and Willan, 2002). The latter estimates cost and effectiveness jointly, preserving in this way the correlation between them, while also allowing the use of different confounders for each outcome (Zellner, 1962). Both methods allow for quantification of the stochastic uncertainty surrounding the resulting incremental net benefit estimate. Nevertheless, the seemingly unrelated regression approach can exploit more information in the modeling process and under certain conditions it has the potential of producing a more efficient (in terms of precision) two-dimensional summary measure (Willan, Briggs and Hoch, 2004). A central feature of the analysis was the use of propensity score for the purpose of evaluating cost-effectiveness in a regression context. The analysis extended the methodology exemplified in previous studies (Mitra and Indurkhya, 2005), by combining propensity score adjustment and seemingly unrelated regression. A comparison of this approach with the net benefit regression framework was also undertaken, while as sensitivity analysis an effort was made to exploit regression methods based on the instrumental variable concept, which can adjust for unobserved

bias. However, the data available in the cohort study did not contain enough information that could potentially be used to construct relevant and valid instruments. Instead, further analyses were undertaken to assess the validity of the unconfoundedness assumption by testing the robustness of the results obtained using both Rosenbaum's bounds, as well as a range of alternative regression specifications.

The substantive results appear to confirm the fact that when effectiveness is defined as 'normal birth', offering the option to plan birth in a non-obstetric setting is cost-effective in the short-term. This finding is important in terms of policy implications since, in principle, it provides further evidence in support of giving women at low risk of complications before the onset of labour a broad choice of planning birth. Nevertheless, the results should be interpreted in the context of the various assumptions postulated by the analytical methods employed, as well as in light of the design and the wider findings of the cohort study. For example, regression and matching methods assume that all potential confounders have been observed and cannot address unobserved confounding. Indeed, current subject matter knowledge indicates that Birthplace did not collect a handful of important confounders. Consequently, the associations observed in this case study, although strong, may ultimately not reflect causal statements. On the other hand, the results of the extensive sensitivity analysis demonstrate that the baseline estimates are generally robust and more unobserved potential confounders do not always lead to more biased estimates. Of particular importance is also the fact that some heterogeneity in the baseline treatment effects is present. In this case study, although womens' treatment effect heterogeneity did not alter the main conclusion of the baseline analysis, it

nevertheless had a considerable impact on some of the estimates obtained. For example, when in the context of the exploratory subgroup analysis the study population was restricted to the potential treatment effect modifiers parity and complicating conditions at the start of care in labour, incremental costs between alternative planned places of birth presented stark differences compared to the baseline estimates.

The results of the methodological comparisons also indicate that although different regression approaches may yield nearly identical incremental net benefit estimates, choice of method may in fact influence their interpretation, with those produced from seemingly unrelated regression offering a more plausible interpretation when the effectiveness outcome is binary. In addition, a maximum likelihood implementation of seemingly unrelated regression has indeed the potential to yield more efficient (in terms of precision) estimates than the net benefit regression, although gains may be relatively modest. Nevertheless, the findings of the different comparisons undertaken here suggest that these efficiency gains will not always be guaranteed and may depend every time on the data at hand. Likewise, when high dimensionality is not an issue, the combination of the propensity score with seemingly unrelated regression may not offer any additional advantages over traditional covariate adjustment, such as improved efficiency or model fit, despite previous such claims in the literature (Indurkhya, Mitra and Schrag, 2006). This finding was consistent when the propensity score was employed only in one of the equations of the seemingly unrelated regression model. At this point, it is crucial to stress that the analyses exemplified placed particular emphasis on achieving robustness because the binary nature of the

effectiveness outcome rendered certain assumptions such as homoskedaticity and normality unrealistic. Indeed, the quest for more efficiency may prove elusive in settings where heteroskedasticity is of major concern. This is also because covariance estimators that are robust to heteroskedasticity typically achieve the additional robustness at the expense of efficiency (Woolridge, 2009).

An important shortcoming of regression methods is that their checking procedures do not involve examination of the overlap and covariate balance between groups before and after adjustment (Stuart, 2010). Although the use of the conditional probability of receiving treatment in the outcome model will typically reduce dimensionality, making it easier in this way for the analyst to find subjects that are similar in terms of propensity scores, there is no guarantee that the exposed and comparison groups will have sufficient balance in terms of individual characteristics (Sekhon and Grieve, 2012). When groups exhibit poor overlap and/or covariate imbalances, drawing causal inferences will involve model-based extrapolation beyond the support of data, which will often require strong external assumptions (Imbens and Woolridge, 2009). This issue was explored in detail using an extensive matching exercise, which introduced matching as a data pre-processing stage before inference, an approach that was recently exemplified in the causal inference literature by Ho et al. (2007). Such an approach, allows the analyst to consider problems of imbalance in individual covariates and assess overlap between groups in a direct and explicit manner (Stuart, 2010). A comprehensive comparison of different matching methods was undertaken, based predominantly on the propensity score (Imbens, 2004). The use of inverse probability weighting, which according to the literature review so far has never been

used in the context of the cost-effectiveness evaluative framework for the purpose of adjusting for selection bias, was exemplified in a variant of a version known as doubly robust estimation (Robins, Rotnitzky and Zhao, 1994). This approach weighs a regression model with the inverse propensity score, offering in this way additional protection against misspecification and bias in the estimates obtained (Wooldridge, 2007). The matching exercise also presented the application of other well-established procedures, including a number of variants of nearest neighbour matching, Mahalanobis and kernel matching, as well as two innovative solutions, namely coarsened exact matching and entropy balancing (Hainmueller, 2012; Iacus, King and Porro, 2012). A "bias-correction" stage using regression analysis was undertaken after these matching procedures, a solution that has been shown to adjust for remaining finite sample bias in the estimates, while potentially also making violations of functional form assumptions less consequential (Abadie and Imbens, 2011; Iacus, King and Porro, 2011). This parametric adjustment, which is similar to the idea of double robustness, was implemented in all cases using the seemingly unrelated regression solution.

In terms of cost-effectiveness, the bias-corrected estimates produced from different matching methods were in the majority of cases very similar. This finding was largely expected, since all estimators rely on the same identification conditions and are asymptotically equivalent (Imbens and Wooldridge, 2009). In addition, the alternative regression specifications in the context of sensitivity analysis demonstrated that omitting a potential confounder from the model could lead to similar levels of unobserved bias in the estimates of each of the regression method considered. Given

the fact that matching in the first place is essentially a weighted regression estimator relying on the same identification assumptions (Angrist and Pischke, 2009), it is unlikely that unmeasured confounding drives any differences in the estimates produced from regression and bias-corrected matching. It should be recognised however that the estimates obtained from the different regression and matching methods might not be in all cases directly comparable, as certain methods such as caliper matching can change the estimand. It is also important to note that the inference strategy employed here is in theory considered conservative, since it produces an estimate of the asymptotic variance of the treatment effect that ignores the fact that there is an error component associated to the estimation of the propensity score, and also to the ordering of the matching process itself (Moreno-Serra, 2007). In general, although there is little to choose between the results from the different matching procedures, certain approaches do not lose any observations, while at the same time they achieve the best covariate balance and the greatest reduction of bias in the estimates.

Indeed, the balance that different matching procedures achieved varied markedly. The results here suggest that in large samples, entropy balancing, inverse probability weighting and Mahalanobis matching, may be preferable over nearest neighbour matching methods. In addition, when entropy balancing and inverse probability weighting achieve similar balance in means, the former should be preferred because it can also balance higher order moments, with the analyst specifying beforehand the desired balance constraints (Hainmueller, 2012). For nearest neighbour matching, a maximum tolerated distance of around 0.2 standard deviations of the probit of the

propensity score appears to be a good starting point for analysts, although more scenarios should always be explored as sensitivity analysis. This finding appears to be in agreement with those of Austin (2009a) who reported that a caliper of 0.2 standard deviations of the logit of the propensity score removed 98 percent of the bias in the crude estimator, while producing confidence intervals with approximately the correct coverage rates. Nevertheless, in large samples attempts to enforce further common support through stricter caliper widths may not always translate to reductions in imbalance across all covariates or greater numbers of treated observations lost. The use of replacement in combination with a maximum tolerated distance should be preferred as it has the potential to provide consistent improvements on the balance achieved in individual covariates. Kernel matching may also be proven a better alternative than simple nearest neighbour matching. Interestingly, the results of this matching exercise indicate that despite the widespread view that bandwidth is the most important parameter of a kernel estimator (Imbens, 2004; Guo and Fraser, 2010), the choice of kernel function may also considerably influence the balance achieved in individual covariates. Thus, it is crucial for the analyst to choose a kernel function that is every time most appropriate for the data at hand. Choice of a particular kernel function will also affect the composition of the common support region, with the Gaussian kernel using all comparison individuals in constructing the counterfactual, as opposed to the other functions that restrict the neighbourhood set to those comparison individuals for which the absolute difference in the propensity score is smaller than the bandwidth (Smith and Todd, 2005). Consequently, different levels of bandwidth will typically result in different common support restrictions with different kernel functions. Finally, an important conclusion of the matching exercise

was that the theoretical advantages of novel approaches, such as that of coarsened exact matching, might not always materialise. In fact, balance even worsened in some cases, a finding however that either reflects inadequacy of the binning algorithm used for the coarsening, or limitations arising from the nature of the data. An exact matching exercise that was undertaken as a further sensitivity analysis, matched women using the grouping levels of each covariate that were decided by the Birthplace analytical team based on substantive knowledge. The results (not presented here) were nearly identical to those obtained from coarsened exact matching, indicating that the poor performance of the latter was potentially due to non-compliance, that is, the transfer of women because of complications at the start of care in labour (Rowe et al., 2012).

The findings from the matching exercise appear to have some resemblance to those reported from other similar comparisons undertaken in the interdisciplinary causal inference literature. For example, Dehejia and Wahba (1999) concluded that sampling with replacement has the potential to decrease bias, particularly when some individuals in the comparison group have extreme values. Morgan and Harding (2006) also found that nearest neighbour caliper matching with replacement and kernel matching are closely related and should be preferred to nearest neighbour matching without replacement. In contrast, Austin (2009a) reported that caliper matching without replacement achieved slightly better balance between exposed and non-exposed individuals compared with inverse probability weighting, which clearly was not the case here. Finally, Iacus, King and Porro (2011) have shown that coarsened exact matching outperforms both propensity score nearest neighbour

matching without replacement, as well as Mahalanobis matching, a finding which also contrasts the results reported in here.

As already explained in detail in previous chapters, the particular work presented in this thesis is subject to limitations. These include the fact that only one reviewer carried out the literature review and as such the classification of the information and the interpretation of the results may be subject to an element of subjectivity. Nevertheless, the use of the structured template ensured that the reviewing process was as rigorous and transparent as possible. In addition, the empirical investigation relied on a single case study, which although with respect to methods was extensive, for pragmatic reasons, it only evaluated certain aspects of planning birth in England by focusing solely on the 'normal birth' clinical outcome. The estimates obtained from the adjusted analyses are also not directly comparable to those of the unadjusted analysis carried out by Schroeder and colleagues (2011), since they were not weighted to reflect each unit's duration of participation, the sampling of obstetric units, as well as to take the clustered nature of the data into account. This was necessary in order to accommodate the methodological component of the thesis. Finally, certain limitations of the study by Schroeder et al. (2011) also remain relevant here. These include a short time horizon that allowed only the short-term cost-effectiveness to be evaluated; a narrow perspective; as well as the fact that calculations for statistical power were only based on those for the clinical outcome and not for the economic endpoints.

Overall, the thesis supports the view that observational microdata from large prospective cohort studies can successfully be exploited in the development of

reliable and informative cost-effectiveness analyses. Formulation of well-defined research questions, when combined with appropriate use of causal inference methodology, has the potential to enable robust assessment of treatment effects (Imbens and Wooldridge, 2009). Econometric methods are increasingly becoming more sophisticated, with the use of a variety of estimators already exemplified in the analysis of observational datasets (Jones and Rice, 2011). Nevertheless, these methods are data driven, being applicable only in situations where good quality microdata are available to support them. For example, the lack of repeated cross-sectional or longitudinal data in the empirical case study exemplified in this thesis, rendered impossible the use of quasi-experimental methods such as difference-in-differences. In addition, the difficulty in this particular empirical context to identify strong exclusion restrictions that could be used to carry out instrumental variable analysis reflects more generally the challenges of crafting plausible natural experiments for economic decision-making in the absence of large administrative datasets. Indeed, the use of such methods could have acted complementarily to regression and matching, either by providing an additional way of recovering the treatment effect, or acting as an alternative sensitivity analysis for indirectly assessing the validity of the unconfoundedness assumption. Ultimately, when selection is on the observables, the thesis advocates that for the purpose of evaluating cost-effectiveness, a novel approach to bias-corrected matching, combining entropy balancing with seemingly unrelated regression, has the potential to offer important advantages over traditional methods, both in terms of minimal (observed) bias, as well as consideration of the stochastic uncertainty surrounding the summary outcome measure. The net economic

benefit can be employed in a straightforward manner to exploit the strengths of rigorous econometric methodology in the context of cost-effectiveness analysis.

The thesis in its entirety establishes the rationale for future research on econometric methods that can be used for evaluating the cost-effectiveness of health care interventions using observational data. This could explore the generalisability of the findings presented here in other settings, for example when small sizes are used or overlap between groups is poor. The comparison of other novel methods including genetic matching with entropy balancing is also desirable. As such, it will be beneficial to undertake further comparative work, both in the context of additional case studies, as well as carefully designed Monte Carlo simulations that allow the further assessment of relative bias and efficiency under a range of hypothetical scenarios. In addition, of particular interest is the exploration of methods in the context of heterogeneous treatment effects, the combination of analytical approaches that can potentially offer improved covariate balance while also accounting for unobservables, and the extension of methods for the cost-effectiveness evaluation of non-binary interventions. The latter is an area of increasing methodological interest in econometrics (Imbens and Wooldridge, 2009) since at present it is not clear how to generalise the weights produced for more than two treatment groups (Blackwell et al., 2009). Last, but not least, the thesis motivates further progress on bivariate cost-effectiveness approaches, particularly those that can better take into account the nature of complex survey data, as well as the investigation of issues arising when feeding estimates from econometric methods in decision analytic models.

APPENDIX

Table A1 – Search strategy and results for Ovid databases

#	Searches	MEDLINE	EMBASE	Econlit
1	cost*.mp.	324,344	482,126	105,298
2	benefit*.mp.	324,043	409,579	47,616
3	effective*.mp.	870,965	1,113,105	30,734
4	cost-benefit*.mp.	52,765	57,632	6,131
5	cost-effective*.mp.	52,903	105,502	2,547
6	matching.mp.	31,429	42,836	6,894
7	stratification.mp.	18,861	25,160	10,347
8	regression*.mp.	315,709	369,473	28,658
9	propensity score*.mp.	2,004	2,674	515
10	instrumental variable*.mp.	443	533	2,521
11	difference in difference*.mp.	177	212	734
12	control function.mp.	370	414	89
13	discontinuity.mp.	3,081	3,624	616
14	3 or 2	1,126,765	1,441,303	75,207
15	14 and 1	130,829	207,619	26,148
16	5 or 4	84,586	149,867	8,461
17	16 or 15	130,829	207,619	26,148
18	13 or 11 or 10 or 12 or 6 or 8 or 9 or 7	366,616	437,436	48,102
19	17 and 18	4,122	5,916	953
20	limit 19 to (yr="1990 - 2010" and english)	3,767	5,394	890

Letter requesting additional studies from other experts

Dear _____

We are currently conducting a project to explore methods for the evaluation of cost-effectiveness using observational data. We are particularly interested in methods that address heterogeneity and selection bias in health economic evaluations arising from the use of individual patient level data. We are now in the process of finalising a literature review exploring the current methods used in the economic evaluation of medical technology and have already identified several published papers in this area. Below you can find a list of selected key references, which indicate the type of studies that we are looking for.

As an expert in the use of these methods, we would be very grateful if you could let us know of further published work, work in progress, or projects related to methods that account for heterogeneity and selection bias in full health economic evaluations, which we should include in our review. Copies of relevant papers (if not publicly available) will be greatly appreciated and duly referenced in our review.

Thank you in advance for your time and assistance.

Sincerely,

Dimitrios Rovithis, Stavros Petrou, Borislava Mihaylova

For Contact:

Dimitrios Rovithis
National Perinatal Epidemiology Unit
University of Oxford
Old Road Campus
Headington
Oxford OX3 7LF
Tel: +44 (0)1865 617 779
Fax: +44 (0)1865 289 701
dimitrios.rovithis@npeu.ox.ac.uk

Selected Bibliography:

Alegria, M., R. Frank, et al. (2005). "Managed care and systems cost-effectiveness: treatment for depression." Medical Care 43(12): 1225-1233.

Cutler, D. M. (2007). "The Lifetime Costs and Benefits of Medical Technology." Journal of Health Economics 26(6): 1081-1100.

De Ridder, A. and D. De Graeve (2009). "Comparing the cost effectiveness of risperidone and olanzapine in the treatment of schizophrenia using the net-benefit regression approach." PharmacoEconomics 27(1): 69-80.

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

Gilmer, T. P., S. Roze, et al. (2007). "Cost-effectiveness of diabetes case management for low-income populations." Health Services Research 42(5): 1943-1959.

Grieve, R., J. S. Sekhon, et al. (2008). "Evaluating health care programs by combining cost with quality of life measures: A case study comparing capitation and fee for service." Health Services Research 43(4): 1204-1222.

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.

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.

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.

Sekhon, Jasjeet S., and Richard Grieve. 2008. “A New Non-parametric Matching Method for Bias Adjust- ment with Applications to Economic Evaluations.” http://sekhon.berkeley.edu/papers/GeneticMatching_SekhonGrieve.pdf.

Table A2 – Structured template used for the review

General information	
Bibliographic information:	
Country:	
Funding:	☐ Industry ☐ Non-industry ☐ Not stated
Type of Study:	☐ CEA ☐ CUA ☐ CBA
Summary Measure:	☐ Net-Benefit ☐ ICER ☐ Costs and outcomes separately
Disease(s):	
Intervention(s):	
Category of Intervention(s): *(check all that apply)*	☐ Surgical ☐ Diagnostic ☐ Medical ☐ Preventative ☐ Rehabilitation ☐ Public health policy
Outcome(s):	
Name(s) of Dataset(s):	
Design / Sample Size/ Type of Data:	
Analytical approaches	
Method(s) for handling selection bias: *(check all that apply)*	☐ Regression analysis ☐ Matching ☐ Stratification ☐ Propensity scores ☐ Instrumental variables ☐ Difference-in-differences ☐ Control function ☐ Regression discontinuity
Adjustment performed on: *(check all that apply)*	☐ Costs ☐ Effects ☐ Costs and Effects ☐ Net-Benefit ☐ Not stated

Estimation method(s):	
Treatment effect (T.E):	
Software used:	☐ Reported (specify): ☐ Not stated
Handling of Uncertainty:	
Comparisons	

Conclusions concerning the methods used for handling selection bias:

Reviewer's appraisal and comments

Was any justification provided for the method used?

☐ No ☐ Yes (specify):

Was any justification provided for the specification used?

☐ No ☐ Yes (specify):

Has the analysis reported the use of any alternative specifications?

☐ No ☐ Yes (specify):

Have appropriate tests been undertaken?

☐ No ☐ Yes (specify):

Table A3 – General information of the reviewed studies

#	Study	Source	Country	Funding	Type	Summary Measure	Disease(s)	Outcome(s)	Interventions	Category	Design / Data	Sample Size
1	Akazawa et al (2008)	Health Services Research	USA	Industry, Non-industry	CEA	ICER	Chronic Obstructive Pulmonary Disease	Severe exacerbation avoided	Inhaled corticosteroids (ICS) treatment	Medical	Retrospective using a claims database	10,271
2	Alegria et al. (2005)	Medical Care	Puerto Rico	Non-industry	CEA	None	Depression care	Percent of respondents effectively treated	Managed care	Public Health Policy	Retrospective Before-After study using survey data	3,504 (wave 1), 3,263 (wave 2), 2,928 (wave 3)
3	Barnett and Swindle (1997)	Health Services Research	USA	Non-industry	CEA	ICER	Substance abuse disorders	Readmission rates	Inpatient substance abuse treatment programmes in terms of intended length of stay, programme size, staffing level, or history of prior treatment **(M)**	Medical	Retrospective using survey data, administrative records	38,683 patients in 98 programs
4	Blanchette et al. (2008)	American J of Ger Pha-rmacother	USA	Industry	CEA	None	Exacerbations associated with COPD	Risk reduction in COPD-related exacerbations	Fluticasone propionate salmeterol (FSC); ipratropium (IPR)	Preventative	Retrospective using administrative records	1,051 (952 in IPR and 99 in FSC)
5	Cakir et al. (2006)	European Spine Journal	Germany	Not stated	CEA	None	Blood loss in posterior spinal instru-mentation	Haemodilu-tion and various other	Harmonic scalpel; electrocauterisation	Preventative	Retrospective	100 (50 per group)
6	Castelli et al. (2007)	Statistics in Medicine	France	Not Stated	CEA	Net Benefit	Colorectal cancer	Life Years	Follow up strategies for curative resection of colorectal cancer	Preventative	Retrospective using a registry database	240 (225 for costs)

#	Study	Source	Country	Funding	Type	Summary Measure	Disease(s)	Outcome(s)	Interventions	Category	Design / Data	Sample Size
7	Chen et al. (2000)	Inquiry	USA	Non-industry	CEA	ICER	Five diagnosis-related groups	% functional improvement of individual patient	Post-acute care in different settings **(M)**	Rehabilitation	Retrospective using inter-views, hospital records and administrative data	2,137
8	Coleman et al. (2006)	Clinical Therapeutics	USA	Not stated	CEA	ICER	ST-segment elevation myocardial infarction	Combined incidence of major adverse cardiac end points	Facilitated PCI; Primary PCI	Surgical	Prospective using data from a laboratory database	538 / 254 matched (127 per group)
9	Coyte et al. (2000)	Journal of Health Economics	USA	Non-industry	CEA	ICER	Joint replacement surgery	Acute care readmission rates	Alternative discharge strategies after joint replacement surgery **(M)**	Rehabilitation	Retrospective using administrative records	29,131
10	Cutler (2007)	Journal of Health Economics	USA	Non-industry	CEA	ICER	Myocardial Infarction	Life-Years	Revascularisation; admission to high volume hospital	Surgical	Retrospective using administrative records	124,950
11	De Natale et al. (2009)	Clinical Drug Investigation	UK	Industry	CEA	None	Ocular hypertension or glaucoma	Treatment failure	Travopost; combination of latanoprost and timolol	Medical	Retrospective using administrative records	815 (639 and 176)
12	De Ridder et al. (2009)	Pharmaco Economics	Belgium	Industry	CUA	Net Benefit	Schizophrenia	QALYs	Olanzapine; risperidon	Medical	Prospective follow up Survey	265 (136 and 129)

#	Study	Source	Country	Funding	Type	Summary Measure	Disease(s)	Outcome(s)	Interventions	Category	Design / Data	Sample Size
13	Dhainaut et al. (2007)	Critical Care	France	Non-industry	CUA	ICER	Severe sepsis	QALYs; Life-Years	Recombinant human activated protein C (rhAPC)	Medical	Prospective Before-After study using a variety of databases	840 (420 per group
14	Farias-Eisner et al. (2009)	Current Medical Research and Opinion	USA	Industry	CEA	None	Venous Thrombo-embolism	Venous Thrombo-embolism occurrence	Fondaparinux; enoxaparin	Preventative	Retrospective using administrative data	5,364 (2,682 per group)
15	Franks et al. (2005)	BMC Health services Research	USA	Not Stated	CUA	ICER	Uninsured elderly population aged 65 or above	Life-Years; QALYs	Medicare Supplemental health insurance; Medicare Part A and B	Public Health Policy	Retrospective using survey	Not reported
16	Givon et al. (1998)	Int J Tech Assessement Health care	Israel	Not Stated	CUA	ICER	Osteoarthritis of the hip joint	QALYs	Total Hip Arhtroplasty using 4 implants: cementless, cemented, hybrid, HA-coated **(M)**	Surgical	Retrospective using mailed questionnaires	363
17	Goeree et al (2009)	Int J Tech Assessement Health care	Canada	Non-industry	CUA	ICER	Coronary artery disease	QALYs; Revascularisa-tion avoided	Drug-eluting stents; bare metal stents	Surgical	Prospective using a patient registry database and other external sources	7502
18	Grieve et al. (2008)	Health Services Research	USA	Non-industry	CUA	ICER Net Benefit	Mental health care	QALYs	Direct capitation; indirect capitation; fee for service **(M)**	Public Health Policy	Retrospective using administrative records	522 (see also Table A5)

#	Study	Source	Country	Funding	Type	Summary Measure	Disease(s)	Outcome(s)	Interventions	Category	Design / Data	Sample Size
19	Grieve et al. (2000)	Int J Tech Assessment Health care	UK / Denmark	Industry, Non-industry	CEA	ICER	Stroke	Life-Years	Models of stroke care (London; Copenhagen)	Medical	Prospective observational study	625
20	Griffin et al. (2007)	British Medical Journal	UK	Non-industry	CUA	ICER	Angina pectoris	QALYs	Coronary artery bypass grafting; percutaneous management; medical management **(M)**	Surgical	Prospective using survey, hospital case records and questionnaires	1,720
21	Groeneveld et al. (2008)	Heart Rhythm	USA	Non-industry	CEA	None	Congestive heart failure	Hazard ratio for mortality	Implantable cardioverter defibrillator (ICD)	Surgical	Retrospective using administrative records	7,125
22	Heaton et al. (2006)	Journal of Managed Care Pharmacy	USA	Non-industry	CEA	None	Asthma	Emergency room visits; hospitalisa-tions; steroid bursts	Use of Leukotriene modifiers (LM)	Medical	Retrospective using administrative records	5,541 (1,290 and 4251 in each group)
23	Indurkhya et al. (2006)	Statistics in Medicine	USA	Non-industry	CEA	Net Benefit	Muscle-invasive bladder cancer	Survival (days)	Cystectomy	Surgical	Retrospective from registry & administrative records	2,133 (1,295 and 838 in each group).
24	Kariv et al. (2007)	Dis Colon Rectum	USA	Not stated	CEA	None	Ulcerative colitis or familial polyposis	Disease-specific endpoints	Fast track (FT); control (CTL) post-operative management	Surgical	Prospective case-control study	194 (97 per group - 83 for costs)

#	Study	Source	Country	Funding	Type	Summary Measure	Disease(s)	Outcome(s)	Interventions	Category	Design / Data	Sample Size
25	Knapp et al. (2008)	Pharmaco economics	Various European	Industry	CUA	ICER	Schizophrenia	QALYs	Olanzapine; risperidone; quetiapine; amisulpride; clozapine; others **(M)**	Medical	Prospective cohort study	10,972 but less was used (unclear what)
26	Lairson et al. (2008)	Disease Management	USA	Not stated	CEA	None	Diabetes	HbA1c values, complications, hospital admissions	CareEnhance Clinical management software; Usual Care Diabetes Management	Public Health Policy	Retrospective using administrative records	870 (435 in each group)
27	Linden et al. (2005)	Dis Manage Health Outcomes	USA	None	CEA	None	Congestive Heart Failure	Emergency department visits; hospita-lisations	A disease management programme	Public Health Policy	Retrospective before-after study	188 (94 per group)
28	Manca, Austin (2008)	Working Paper	Canada	Non-industry	CEA	None	Post-Acute Myocardial Infarction (AMI)	Odds ratios for mortality	Percutaneous Transluminal Coronary Angioplasty; Coronary Artery Bypass Crafting Surgery	Surgical	Retrospective using administrative records	15,943
29	McClellan, Newhouse (1997)	Journal of Econometrics	USA	Non-industry	CEA	ICER	Acute myocardial infarction	Deaths avoided	Catheterisation	Surgical / Diagnostic	Retrospective using administrative records	819,563
30	Merito, Pezzoti (2006)	European Journal of Health Economics	Italy	Industry	CEA	ICER / Net benefit	Human acquired immune-deficiency virus	Disease progression or death avoided	Immediate highly active anti-retroviral therapies (at least three drugs or active components); deferred	Medical	Prospective observational study	1,962

#	Study	Source	Country	Funding	Type	Summary Measure	Disease(s)	Outcome(s)	Interventions	Category	Design / Data	Sample Size
31	Mihaylova et al. (2010)	Value in Health	Various European	Industry	CUA	ICER	Urinary incontinence	QALYs	Duloxetine; Duloxetine plus conservative; conservative; no treatment **(M)**	Medical	Prospective observational study	1,510
32	Mitra, Indurkhya (2005)	Health Economics	USA	Non-industry	CEA	Net Benefit	Muscle-invasive bladder cancer	Life Days	Cystectomy	Surgical	Retrospective from registry and administrative records	2,133
33	Mojtabai, Zivin (2003)	Health Services Research	USA	Non-industry	CEA	None	Substance disorders	Abstinent case; case of reduced use	Four treatment modalities for substance abuse **(M)**	Medical	CS using survey data	1,799
34	Polignano et al. (2008)	Surg Endosc	UK	Not stated	CEA	None	Liver surgery	Overall and liver-related morbidity, blood loss, Pringle mano-euvre, rese-ction margins	Laparoscopic; open liver resection	Surgical	Retrospective case-control study using hospital records	50 (25 per group)
35	Polsky et al. (2003)	Journal of Clinical Oncology	USA	Not stated	CUA	ICER	Breast cancer	QALYs	Breast conservation surgery with radiation (BCSRT); mastectomy. Also open; restricted regiment	Surgical	Retrospective using survey, administrative records	2,517

#	Study	Source	Country	Funding	Type	Summary Measure	Disease(s)	Outcome(s)	Interventions	Category	Design / Data	Sample Size
36	Polsky, Basu (2006)	Elgar Companion to Health Economics	USA	Not stated	CUA	ICER	Breast cancer	QALYs	Breast conservation surgery with radiation; mastectomy	Surgical	Retrospective using survey data and administrative records	Not reported
37	Sadhu et al. (2008)	Diabetes Care	USA	Non-industry	CEA	None	Hyperglycemia	Probability of dying	Intense; conventional insulin therapy	Medical	Retrospective Before-After design using a database and hospital accounting records	6,719 for main analyses. 5,787 for sensitivity analyses
38	Sekhon, Grieve (2009)	Working Paper	UK	Not stated	CUA	Net Benefit	Management of critically ill patients	QALYs	Pulmonary Artery Catheterisation (PAC)	Surgical	Retrospective using a critical care database	1,052 cases and 31,447 controls
39	Shih et al. (2007)	Pharmaco Economics	USA	None	CEA	Net Benefit	Depression	Avoidance of treatment failure	Paroxetine; sertraline; citalopram; escitalopram; fluoxetine after entry of generic paroxetine **(M)**	Public Health Policy	Retrospective Before-After design from a claims database	5,629 post-entry and 1901 pre-entry period patients
40	Shireman, Braman (2002)	Archives of Pediatrics & Adolescent Medicine	USA	Non-industry	CEA	None	Respiratory Syncytial Virus (RSV)	RSV hospitalisa-tions and their length of stay	RSV immune globulin and palivizumab	Medical	Retrospective using administrative records	1,506 children from which 137 were treated with further 137 controls selected

#	Study	Source	Country	Funding	Type	Summary Measure	Disease(s)	Outcome(s)	Interventions	Category	Design / Data	Sample Size
41	Soegaard et al. (2007)	European Spine Journal	Denmark	Not Stated	CEA	ICER Net Benefit	Chronic low back pain	Change in functional disability, change in degree of leg and back pain	Lumbar spinal fusion: Non-instrumented; instrumented; circumferential fusion **(M)**	Surgical	Prospective observational cohort study	136
42	Weiss et al. (2002)	The American J of Medicine	USA	Non-industry	CEA	ICER	Ventricular arrhythmias	Life-Years	Implantable cardioverter defibrillator	Surgical	Retrospective using administrative records	125,892 patients; 7,612 matched pairs identified
43	Windmeijer et al. (2006)	Int J Tech Assessment Health care	UK	Industry	CUA	ICER	Schizophrenia	QALYs	Two hypothetic antipsychotic treatments for schizophrenia	Medical	Prospective cohort study	10,972

Table A4 – Analytical approaches employed in the reviewed studies

#	Study	Method(s)	Parameters	Estimation	T. E.	Software	Uncertainty	Comparisons	Author(s) Conclusions
1	Akazawa et al. (2008)	Regression Analysis	Costs and Consequences	Longitudinal individual-level fixed effects linear regression for both costs and outcomes. The model for costs allowed the effect of treatment to vary with time through an interaction.	Not Stated	STATA 9.2	Standard Errors for costs and effects, CIs through bootstrap (1000 replications) / CEAC reported	With RCT and modelling C/E studies. Outcomes of these studies were different.	Bootstrapping the models accounts for the correlation between the numerator and the denominator. Fixed effects model takes into account unobserved time-invariant bias. Results might still be prone to time-varying unobserved bias.
2	Alegria et al. (2005)	Difference-in-Difference	Costs and Consequences	Difference-in-Difference; naive, with linear regression, and with matching by quintiles of the propensity score	Not stated	Not stated	Only p-values for effects on rates of effective treatment reported	Naïve DiD, DiD with covariate adjustment, DiD with PSM. Authors qualitatively reported that conclusion of their study consistent with other work	Baseline differences in treatment effectiveness between managed and non-managed care regions are considerable, and the methods may not have effectively contended with differences in unobserved variables.
3	Barnett and Swindle (1997)	Regression Analysis	Costs and Consequences	Random-intercept regression models were used to investigate the impacts, on the cost (linear) and effectiveness outcomes (logistic) of patient and programme characteristics	Not stated	Not stated	Sensitivity analysis using an alternative specification	Some consideration of previous effectiveness/cost literature but no actual comparison was made	None provided

#	Study	Method(s)	Parameters	Estimation	T. E.	Software	Uncertainty	Comparisons	Author(s) Conclusions
4	Blanchette et al. (2008)	Regression Propensity Score Analysis	Costs and Effects	Effects using Cox proportional hazards regression models. Costs using generalized linear models and gamma distribution as well as a log-link function to adjust for differences in baseline characteristics. Propensity score matching using logistic regression and then Mahalanobis matching with caliper.	Not Stated	SAS 9.1	CIs for hazard ratios and cost differences. For the latter the bootstrap method was used with 1000 replications	Qualitative and quantitative for outcomes and costs with other observational studies as well as trials.	A potential selection bias may also have been introduced by limiting the sample to only those patients who did not start another treatment within the first 60 days after initial treatment, which may account for sicker, less-stable patients in the sample.
5	Cakir et al. (2006)	Matching	Costs and Effects	The two groups were matched in a blinded manner with respect to several factors	Not Stated	SPSS 9	P values < 0.05	Mentioned qualitatively other studies' conclusions	The use of matching and independent observers ensured that the effect detected was mostly due to the treatment
6	Castelli et al. (2007)	Regression Analysis	Costs and Effects	Semi-Markov model with least-squares regression for a number of time intervals for costs and a hazard function using a Weibull distribution for transitions between model states	Not Stated	Not stated	Bootstrap procedure to evaluate INB distribution and CIs. See paper for more.	Comparison with Willan censoring-adjusted regression modelling	Costs and health outcomes can be linked in the model. Moreover, by using this method, cost data for health states are modelled and are therefore more homogeneous. Consequently, more reliable modelling is expected.

#	Study	Method(s)	Parameters	Estimation	T. E.	Software	Uncertainty	Comparisons	Author(s) Conclusions
7	Chen et al. (2000)	Instrumental Variables	Costs and Effects	Multinomial logit equation for the first stage and OLS for the second.	Not Stated	Not stated	CIs for ICER using Taylor's approximation method.	None reported	Consistency in findings suggests that the IV method adjust adequately for selection bias. A randomised controlled trial would be desirable to confirm the results obtained.
8	Coleman et al. (2006)	Propensity Score Analysis	Costs and Effects	Unspecified propensity score model for treatment assignment and 1:1 nearest neighbour matching.	Not stated	SPSS 11	CIs for ICER through non-parametric bootstrapping using 25000 replications with replacement	None reported	The use of propensity score matching minimizes biases for the end points evaluated. However, propensity score matching can only link patients on observable covariates, allowing unobservable covariates to potentially bias overall study conclusions.
9	Coyte et al. (2000)	Propensity Score Analysis	Costs and Effects	Multiple regression analysis for some costs. Logistic regression with two-way interaction terms employed to evaluate propensity scores. Stratification by propensity scores followed. Individual pairwise comparisons for multiple treatments.	Not stated	SAS 6.11	None reported	Authors stated that their results complement a recent national study.	While several alternative analyses were conducted to control for potential bias in the assignment of patients to various discharge destinations, the possibility that the adjustments were deficient in some respects could not be ruled out.

#	Study	Method(s)	Parameters	Estimation	T. E.	Software	Uncertainty	Comparisons	Author(s) Conclusions
10	Cutler (2007)	Regression analysis Instrumental Variables	Costs and Effects	Two separate models for spending and Mortality. OLS regression and two stage least squares regression for IV analysis	ATE / LATE	Not stated	None reported	Between OLS and IV estimates. Quantitative with other studies some using the same dataset. More details in Table A5.	Criteria for choice of instrument only partially testable. In the absence of strong assumptions one cannot necessarily attribute the estimated cost-effectiveness ratio as a causal statement.
11	De Natale et al. (2009)	Propensity Score Analysis Regression Analysis	Costs and Effects	Effects: logistic regression for propensity score. Propensity score quartiles included in a Cox model for an adjusted estimate of treatment effect for each drug group. Linear regressions for costs.	Not stated	SAS 9.1	P values for hazard ratio and cost difference.	Findings contrary to those reported by a randomised study, which however used second-line treatments.	Despite adjustments, results may have been confounded, at least partially, by disease severity.
12	De Ridder et al. (2009)	Regression Analysis	Net-Benefit	Linear net-benefit regressions (one with interactions).	Not stated	STATA 9	One-way Sensitivity Analysis / CEAC	Qualitative mentioning that studies provide conflicting evidence, with most not making adjustments and concerning only costs	Several patient characteristics influence the incremental net benefit of the drugs. Selection bias in terms of endogeneity could not be assessed.

#	Study	Method(s)	Parameters	Estimation	T. E.	Software	Uncertainty	Comparisons	Author(s) Conclusions
13	Dhainaut et al. (2007)	Propensity Score Analysis	Costs and Effects	Logistic regression for propensity score; 1:1 matching using the SAS 'match' macro. Linear regression model for hospital costs. Additional assumptions for life expectancy and quality of life.	Not stated	SAS	CIs through non-parametric bootstrap, with 10,000 samples of mean effectiveness, mean cost and ICER / CEAC	Quantitative with other cost-effectiveness studies. Discrepancies in results noted due to the use of trial effectiveness data and increased hospital costs.	The main limitation of the propensity score is that deals only with observed biases. Forty-six variables from case record forms ensured that the probability that a confounding factor was left out is quite low. Observed differences with regard to rhAPC cost-effectiveness were thus not related to the characteristics of the patients.
14	Farias-Eisner et al. (2009)	Propensity Score Analysis	Costs and Effects	Logistic regression for the propensity score. 1:1 greedy matching on 12 digits of the propensity score	Not Stated	SAS 9.1	None	Qualitative for costs and effectiveness with other clinical and observational studies	Acknowledgement of limitations arising from claims data: non-randomisation, missing data, improper data entry, and inability to establish causality and control for certain confounders.
15	Franks et al. (2005)	Regression Analysis	Costs and Effects	Generalised Linear Regression for expenditures using a gamma distribution and a log link function. Linear regression for HRQL. Markov Decision Analytic model.	Not Stated	SUDAAN 8.0.1, STATA 8.2, DATA 4.0	CI for ICER using Monte Carlo simulations. Univariate sensitivity analysis.	None reported	As any individual study employing observational data, this study does not adequately address the problems of endogeneity/confounding or establish causality.

#	Study	Method(s)	Parameters	Estimation	T. E.	Software	Uncertainty	Comparisons	Author(s) Conclusions
16	Givon et al. (1998)	Regression Analysis	Costs and Effects	Multiple regressions for continuous dependent variables including costs and QALYs. Details not reported.	Not Stated	SAS	CIs for QALYs and ICERs.	Results similar when different adjustments were carried out. Unadjusted results were not presented	Multiple regressions analysis controlling for all possible biases demonstrated one cementless implant as superior to all others.
17	Goeree et al. (2009)	Propensity Score Analysis	Some resource use, Effects	Logistic (logit) regression for propensity score. 1:1 nearest neighbour matching using a caliper width of less than 0.2 times the standard deviation of the propensity score. External resource use, cost and utility data used. Decision Analysis then followed.	Not stated	Not stated	Deterministic sensitivity analysis. Probabilistic using conventional stochastic distributions. Monte Carlo simulation. CEAC.	Mentioned that other C/E studies exist in the area but detailed comparison of results deemed inappropriate because different methodological approaches were used. Methodological assumptions might account for most differences.	The propensity score process identified a large well-matched cohort. Unmeasured confounders however may still affect the results of the study.
18	Grieve et al. (2008)	Matching	Costs and Effects	Two-stage approach employed. Similar areas across the three payment modes were selected. Genetic matching algorithm with covariate adjustment employed to improve comparability between groups.	Not Stated	Not Stated	CIs using non-parametric bias corrected bootstrap for incremental costs and QALYs. CEAC from bootstrap replicates.	Qualitative with previous cost-minimisation studies some of which use the same dataset and parametric methods to estimate costs	The application of this method achieves excellent covariate balance. The results are not sensitive to parametric assumptions in usual parametric regression models.

#	Study	Method(s)	Parameters	Estimation	T. E.	Software	Uncertainty	Comparisons	Author(s) Conclusions
19	Grieve et al. (2000)	Regression Analysis	Costs and Effects	Linear regression for costs and Cox regression model for survival both adjusting for case-mix.	Not stated	Not stated	Univariate sensitivity analysis	None reported	The authors concluded that the observational nature of the study meant that unmeasured case-mix differences between the centres could explain some of the residual differences in cost and consequences.
20	Griffin et al. (2007)	Regression Analysis Matching	Costs and Effects	Participants split into three groups based on rated clinical appropriateness. Regressions with interaction terms. OLS regression of life years. Seemingly Unrelated Regression for costs and effects.	Not stated	Not stated	Univariate sensitivity analysis, scenario analysis, CEAC	Clinical appropriateness, costs, mortality benefit and QoL, with RCTs	Authors acknowledge the risk of confounding in their study and stated that they sought to address this both by design and analysis.
21	Groeneveld et al. (2008)	Propensity Score Analysis	Costs and Effects	Logistic regression for the propensity score. Matching followed within 0.25 times the standard deviation of the propensity score and a minimum Mahalanobis distance calculated from key covariates. Cox proportional-hazards survival model for mortality. Median costs compared.	Not stated	SAS 9.1	Sensitivity Analysis	Unadjusted costs and mortality with adjusted. Mortality compared with that of other studies (trials). Also some quantitative comparison with other studies for costs and expected cost-effectiveness.	A strong point of the study was the use of propensity score matching, Propensity score models cannot adjust for inadequately measured or unmeasured covariates. It is possible that unmeasured factors were the actual cause of the mortality benefit and not the ICDs themselves. The method of selecting controls was biased, by design, toward inclusion of patients who were "healthier" than typical device recipients.

#	Study	Method(s)	Parameters	Estimation	T. E.	Software	Uncertainty	Comparisons	Author(s) Conclusions
22	Heaton et al. (2006)	Propensity Score Analysis	Costs and Effects	Logistic regression for the propensity score. Logistic regression for outcomes using the propensity score and covariates that were not balanced within quintiles of propensity score.	Not Stated	Not stated	CIs for outcomes and SDs for costs	Qualitative with other cost-effectiveness studies for key elements of study, plus method dealing with selection bias from other studies	Authors note limitations from claims data such as upcoding for reimbursement purposes or disease classification. Also, acknowledgement that propensity score analysis has been shown to be a valid method to reduce selection bias, it can only control for known variables, not unknown variables.
23	Indurkhya et al. (2006)	Propensity Score Analysis Regression Analysis	Net Benefit	Proportional hazards model for survival. Logistic regression for propensity score. Unadjusted Net Benefit regression using inverse probability weighting. Net Benefit regression with covariate and with/without propensity score adjustment.	Not stated	Not stated	SEs for propensity score means and NMB estimates / CEAC	Unadjusted, covariate adjusted and propensity score adjusted NMB	For large values of λ there are significant differences in NMB estimates obtained using unadjusted, covariate adjusted, and propensity score adjusted regressions. If significant imbalance in the covariate information across the treatment groups making propensity score adjustments as opposed to covariate adjustments is recommended.
24	Kariv et al. (2006)	Matching	Costs and Consequences	Matching with respect to a number of factors	Not stated	Not stated	P values (<0.05)	Qualitative for LOS and readmission rates mainly with observational studies	None provided with respect to matching.

#	Study	Method(s)	Parameters	Estimation	T. E.	Software	Uncertainty	Comparisons	Author(s) Conclusions
25	Knapp et al. (2008)	Regression Analysis	Costs and Effects	Separate fixed effects regression for the three study periods and results combined over duration of study (Epoch analysis). Linear OLS for EQ-5D and Poisson regression specified as an exponential function for costs	Not stated	Not stated	SEs and CIs for incremental treatment effects. Bootstrapping with replacement (200 replications) for ICER. CEACs.	Comparison with two RCTs and numerous other CEA studies.	The models did not explicitly consider correlation of unobservables over time and correlation between costs and effects. However, the use of bootstrap methods for inference takes into account the complex correlation structure between costs and consequences.
26	Linden et al. (2005)	Propensity Score Analysis	Costs and Effects	Logistic regression for the propensity score. Matching based on the nearest propensity score. Also stratification into 5 quintiles	Not stated	Not stated	Standard Errors for cost and health outcome means. P values.	Comparison of propensity score stratification and matching. Results from stratification support those of matching	Propensity scores only adjust for observed bias. However, study results are relatively insensitive and would require high levels of bias to alter the conclusions. Thus treatment effects are not a function of hidden bias. Stratification can remove more than 90% of initial bias.
27	Lairson et al. (2008)	Difference-in-Difference	Costs and Effects	Matching and then difference-in-difference two-way fixed effect linear regression that takes into account time	Not Stated	Not Stated	SEs for regression estimates.	Qualitative with a systematic reviews, a meta-analyses and trials for costs and health outcomes	The natural experiment with patient matching, but without patient choice, addresses the important problem of selection bias. Use of time series data and fixed effects multiple regression allowed for correction for time trends between the groups and for unmeasured differences between the individuals in the two groups.

#	Study	Method(s)	Parameters	Estimation	T. E.	Software	Uncertainty	Comparisons	Author(s) Conclusions
28	Manca, Austin (2008)	Propensity score analysis followed by (1) stratification in 5 strata based on propensity score, (2) nearest neighbour 1:1 matching within a calliper of 0.25 standard deviations of the propensity score, or (3) linear regression analysis including the propensity score in cost and effect models. Analysis employs Gamma distribution for costs and Bernulli for consequences.	Costs and Consequences	Logistic regression for propensity score	ATE	STATA 9, WinBUGS 1.4.2	Credibility intervals for differential costs and odd ratio. Correlation between costs and effects preserved as the logit of the binary outcome was conditioned on costs. For matching, the correlated models were applied to the propensity score matched cohort.	Unadjusted and propensity score based regression-adjusted, matched and stratified estimates	All four approaches led to the same conclusion. However, the estimates obtained after adjustment were considerably different than those from the unadjusted analysis. Acknowledgement of limitations of propensity score analysis based on administrative data and the selection on observables assumption.
29	McClellan, Newhouse (1997)	Difference-in-Difference	Costs and Effects	Least square methods with fixed effects, heteroskedasticity-consistent instrumental variable techniques with a weighted average estimate across the difference-in-difference comparisons in the data. Weights determined by estimated variance	ATE	Not stated	Standard Errors for incremental costs and effects / Scenario analysis adjusted for lead time.	Least squares estimates of average treatment effects; difference-in-difference with instrumental variables; difference-in-difference with instrumental variables and lead time adjustment. Some quantitative comparison with a previous study.	The panel instrumental variable estimation relied on minimal parametric assumptions and allowed for detailed analysis of the implications of partial failures of the strong identification conditions required for consistent difference-in-difference estimation.

#	Study	Method(s)	Parameters	Estimation	T. E.	Software	Uncertainty	Comparisons	Author(s) Conclusions
30	Merito, Pezzoti (2006)	Propensity Score Analysis Separate logit models for propensity score. Some variables were transformed by taking, respectively, the square root and the log base 10 to correct the skewed distributions of these variables. Stratification based on propensity scores in 4-5 strata. For consequences: Cox proportional hazards models stratified by propensity score blocks. Costs were computed as weighted sums of the differences between sample mean annual costs by treatment status within each propensity score block, with weights equal to the proportion of observations falling in each block.	Costs and Consequences Net Benefit	Analysis within CD4 counts subgroups.	ATE	Not stated	Standard Errors adjusted for clustering for propensity scores. CIs for hazard ratios and differences in mean costs. CIs for ICER from 10000 bootstrapped samples with bounds identified by the percentile. Also CEAC.	Various models for costs and effects unadjusted for lead-time bias, adjusted only for lead-time or adjusted for lead-time and for all baseline covariates. Clinical outcomes with randomised trial. Stressed that initiation of HAART in other studies sometimes was taken into account as well but also naïve analyses	Effort was made in the analysis to take into account all three mechanisms operating in a person who defers HAART in an observational setting, with selection bias potentially being one of those. Propensity score analysis eliminated imbalances.
31	Mihaylova et al. (2010)	Regression Propensity Score Analysis	Costs and Effects	Seemingly Unrelated (linear) Regression for costs and consequences. Propensity score matching based on nearest neighbour, Kernel and stratification.	Not stated	STATA	Standard Errors for incremental costs and effects. Probability of cost-effectiveness for willingness to pay of £20,000 per QALY.	Seemingly Unrelated Regression with different propensity score matching methods. Partial qualitative comparison with trials in the field.	Multivariate linear regression framework is more limited in its abilities to control for confounding. Propensity score analysis is likely more appropriate for the estimation of cost-effectiveness. Results from the two approaches were very close.

#	Study	Method(s)	Parameters	Estimation	T. E.	Software	Uncertainty	Comparisons	Author(s) Conclusions
32	Mitra, Indurkhya (2005)	Propensity Score Analysis	Net Benefit	Propensity Score for each patient via logistic regression predicting treatment assignment from a large number of covariates. Linear regression model employing the score as a covariate	Not stated	Not stated	CEAC. Simulation studies to assess sensitivity of results to dropped covariates. Sensitivity Analysis for different willingness to pay values	Unadjusted, covariate adjusted and propensity score adjusted net benefit regression models	Balance was achieved for all covariates after adjustment. Regardless of the presence of unobserved covariates propensity score adjustment estimates are less biased and more accurate with smaller standard errors. Propensity score adjustments are more sensitive to the assumption of strong ignorability for lower values of willingness to pay.
33	Mojtabai, Zivin (2003)	Propensity Score Analysis	Effects	Logistic regression for propensity scores for separate treatment comparisons. Stratification based on the propensity score followed. Effectiveness of the 4 modalities compared through logistic regression.	Not Stated	STATA 7	CIs for ICER through bootstrapping with 1000 replications and bias correction. Extreme scenario analysis.	Qualitative with other cost-effectiveness studies. Results in line with those of other studies.	While stratification according to propensity scores controls for the effect of observed confounders, it does not necessarily control for the effect of unobserved variables.
34	Polignano et al. (2008)	Matching	Costs and Effects	Groups were matched for age, sex, operation, magnitude of resection and for tumour location and size	Not stated	SPSS 12.0	None reported	None reported	Authors acknowledge that their results may somewhat depend on social and other local circumstances and they advocate further similar studies in different settings to confirm their findings. Nevertheless, they argue that matching, staged introduction of various laparoscopic liver resections and authors' increasing confidence and skills prevented any active selection bias.

#	Study	Method(s)	Parameters	Estimation	T. E.	Software	Uncertainty	Comparisons	Author(s) Conclusions
35	Polsky et al. (2003)	Regression Propensity Score Analysis	Costs and Effects	OLS regression for costs and consequences. Logistic regression for the propensity score and stratification in 4 groups. OLS regression in each group. Results were averaged across groups.	Not stated	Not stated	CIs for costs and effects. CIs for ICER using non-parametric bootstrap. Sensitivity analysis.	Unadjusted, regression adjusted, and propensity score adjusted estimates. Survival derived from clinical trials and observational study evidence on quality of life. Comparisons with other studies not directly relevant because of different time frames.	The negligible change in between the OLS-adjusted result and the propensity score result suggests there is little heterogeneity in treatment effects. Unobserved bias however may still exist. Instrumental variables analysis was employed but OLS was ultimately preferred.
36	Polsky, Basu (2006)	Regression Analysis Propensity Score Analysis Instrumental Variables	Costs and Effects	Costs and consequences using (1) linear regression, (2) propensity score (using logistic regression) stratification, followed by least-squares regression with covariate adjustment within stratas and averaging results across stratas. (3) Instrumental variables estimation (no details).	ATE	Not stated	CIs for costs, effects and ICERs. Bootstrapping for ICER	Quantitative comparison of unadjusted and adjusted results using different methods.	There is considerable selection bias in the observational data that diminishes as the selection correction methods are applied. Results using regression and propensity score analysis were similar but there were large differences with the instrumental variable approach. Either hidden bias is very important or the instruments used were weak.

#	Study	Method(s)	Parameters	Estimation	T. E.	Software	Uncertainty	Comparisons	Author(s) Conclusions
37	Sadhu et al. (2008)	Difference-in-Difference	Costs and consequences	Difference-in-difference regression with covariate adjustment. Linear regressions on log- transformed costs; the estimates were back-transformed to calculate intervention effects on costs on the original scales. For consequences logistic regressions were used to estimate mortality.	Not Stated	Not Stated	Sample mean of difference-in-difference estimate reported with the bias-corrected, 95% CI, from 1,000 bootstrap replicates with replacement. Outlier analysis	Quantitative comparison of LOS and costs with other studies.	Difference-in-difference study design relies on the assumption that the secular time trends affecting the intervention and comparison units are similar. In any event, this difference-in-difference assumption should be more valid than that of earlier pre-post study designs that did not take secular time trends into account at all.
38	Sekhon, Grieve (2009)	Matching Propensity Score Analysis	Costs and Effects	Logistic regression for propensity score, matching 1:1 (with replacement) based on the propensity score. Genetic matching algorithm using the same covariates for adjustment.	ATT	R	Confidence intervals for INB using non-parametric bootstrap conditional on the matched dataset for willingness to pay of 30,000 per QALY.	Genetic Matching achieved better balance for each covariate than propensity score matching. Matching without replacement gave same conclusions but worse covariate balance for both methods. Comparison with randomised controlled trial data.	Balance after matching of means between groups for each covariate as well as the distribution of each covariate is of primal importance. Genetic Matching can reduce but not eliminate selection bias as it improves the balance of observed characteristics when the treatment assignment mechanism is unknown the covariates have non-normal distributions and non-linear relationships with the outcome. Regression methods complementary to matching. Genetic Matching results robust after (semi) parametric models to matched data.

#	Study	Method(s)	Parameters	Estimation	T. E.	Software	Uncertainty	Comparisons	Author(s) Conclusions
39	Shih et al. (2007)	Regression Analysis	Net Benefit	Frequentist and Bayesian heteroskedasticity-robust net-benefit regressions using multiple comparators and covariate adjustment.	Not stated	WinBUGS	Robust standard errors, Markov chain Monte Carlo simulations, CEAC.	Adjusted and unadjusted results. Frequentist and Bayesian estimation.	There is potential for bias in the estimates of treatment effects because of endogeneity in treatment selection. The use of a polychotomous selection model to explore the issue of endogeneity in the frequentist framework found evidence of positive sample selection bias.
		Regression analysis with polychotomous sample selection was used Two-stage estimation procedure: multinomial logit model for factors associated with selection and a linear regression with the Mill's ratio in the net benefit regression. Time periods were taken into consideration in the analysis through interactions. Evaluation at three levels of willingness to pay values.							
40	Shireman, Braman (2002)	Propensity Score Analysis	Costs and Consequences	Logistic regression for the propensity score.	Not stated	Not stated	Confidence interval for odds of hospital admission, p-values for length of stay and costs differences.	Results concur with clinical trials for hospitalisations. Results also in line with most modelling studies. For the latter ranges provided.	Propensity score matching eliminated most of the differences. Authors acknowledge limitations of this approach with respect to unobserved bias.
		Stratification of treated cases in 5 groups based on the propensity score. 1:1 matching within groups followed. Logistic regression for probability of any RSV admission (controlling for the predicted propensity score). Multivariate regression for difference between the treated and untreated groups' RSV inpatient lengths of stay and costs, controlling for the predicted propensity score.							

#	Study	Method(s)	Parameters	Estimation	T. E.	Software	Uncertainty	Comparisons	Author(s) Conclusions
41	Soegaard et al. (2007)	Regression Analysis	Net Benefit	Net-benefit regression framework: Linear multiple regressions (ordinary least squares with bootstrapped confidence intervals for the coefficients (9199 replications) and different willingness to pay values: 2000, 4000, 8000, 16000.	Not Stated	STATA 8.2	Standard errors for significant determinants, Bootstrapped bias-corrected confidence intervals for costs, ICER (800 replications) CEAC*	Relevant literature was mentioned but not compared because of different methodologies. Some comparison of costs with a trial-based economic evaluation.	Despite the use of the net-benefit regression results are by definition biased. Further focus on the determinants for cost-effectiveness for the identification of subgroups. Patient characteristics that are modifiable at a relatively low expense may have greater influence on cost-effectiveness than the surgical technique itself.
42	Weiss et al. (2002)	Propensity score analysis Cumulative expenditures were then calculated and mortality at 1, 2, 3 years was estimated using logistic regression and 8-year Kaplan-Meier cumulative survival. ICER was calculated using the cumulative expenditures and mean cumulative survival in the two groups.	Costs and Consequences	Multivariate logistic regression for propensity score, 1:1 matching followed.	Not stated	SAS	Confidence Intervals for mortality, Standard deviations for costs	Unadjusted survival results, propensity score matching adjustment. Similar mortality with three trials. CE less favourable than that of another trial and more favourable with another	Some residual differences remained in the observed characteristics their small magnitude means that they are unlikely to be clinically significant. Administrative data lacked important clinical predictors of outcome. Nevertheless there was agreement with other studies suggests there is no selection bias

#	Study	Method(s)	Parameters	Estimation	T. E.	Software	Uncertainty	Comparisons	Author(s) Conclusions
43	Windmeijer et al. (2006)	Regression analysis	Costs and consequences	Separate regression models for different time periods; results combined over duration of study (Epoch analysis). Linear OLS regression for effects and Poisson regression with exponential mean function for costs	Not stated	Not stated	SEs and CIs for parameters in each Epoch. Bootstrapping using 200 samples with replacement on costs and effects / CEAC	Between different time periods (epochs)	Traditional methods of analysis are not adequate when it comes to assigning treatment effects to the drugs taken by patients when there is a tendency for them to switch their medication frequently. Epoch analysis addresses this issue and is flexible enough to incorporate current methods to address the modelling of skewed cost data, selection bias and sampling and decision-making uncertainty.

Notes: NMB = Net monetary benefit; ICER = Incremental cost-effectiveness ratio; CEAC = Cost-effectiveness acceptability curve; ATE = Average treatment effect,; DiD = Difference-in-differences

Table A5 - Reviewer's appraisal and comments

#	Study	Justification for		Alternative specifications	Tests	Comments
		method	specification			
1	Akazawa et al. (2008)	Individual fixed effects specifications for unobserved time-invariant bias.	Descriptive.	Longitudinal random effects model (not presented).	Hausman test for fixed vs. random effects (fixed effects judged appropriate).	Authors note that it is difficult to compare their results with those of other studies. They also note limitations with regards use of claims data particularly the use of proxy measures that can cause bias due to misclassification of the explanatory variables.
2	Alegria et al. (2005)	Difference-in-difference controls for baseline differences in regression analyses and exogenous changes over time. For potential imbalance in unobserved variables, propensity scores were used to match observations in the experimental and control regions on observables. The propensity score is the likelihood an observation came from an experimental region.	Descriptive.	Assessment of effectiveness using different definitions In specifications 1-6, the sample was split by diagnosis. In 2-4 larger numbers of covariates. In 5-7 propensity score matching was used.	None reported.	A systems cost-effectiveness framework was used. Difference-in-difference appropriate for analysis at an aggregate level. Lagged components to account for changes in number of providers or their practices over time were not included due to lack of data, but an interaction term between data wave and managed care was included. The mean balance of the covariates, the propensity score distribution and the type of matching performed was reported. Baseline comparability of the managed care and non-managed care cohorts was reported only with respect to treatment and its success and treatment costs.
3	Barnette, Swindle (1997)	Random-effects models treating the intercept as a random variable whose variation is explained by programme characteristics account for the correlation of patients within programmes.	Descriptive.	Cost and effectiveness models using a different survey-based definition of staffing intensity and cost.	None reported.	Patients were shown to be comparable in terms of the severity of illness index.

#	Study	Justification for		Alternative specifications	Tests	Comments
		method	specification			
4	Blanchette et al. (2008)	The use of propensity score matching was justified on the grounds of small sample size.	Descriptive.	Propensity score matching.	Wilcoxon rank sum tests (continuous variables) and χ^2 tests (categorical variables) for differences in baseline characteristics.	The results based on the regression and propensity score matching were similar.
5	Cakir et al. (2006)	Matching used to make groups comparable in important characteristics without knowledge of outcomes.	Variables used for matching were based on previous literature.	None reported.	Mann-Whitney for differences in continuous variables, Fisher's exact test for categorical (two-tailed).	Groups were mostly balanced after matching was performed.
6	Castelli et al. (2007)	Natural, flexible way of modelling clinical progression and cost accumulation.	Choice of covariates using a backward elimination approach.	Sub-group analysis for the incremental net benefit (not presented).	χ^2 test, Wald test for covariate selection, Goodness of fit for cost: BIAS, MSPE, MRSE and MAPE. Pearson for Markov.	Regression methods combined with decision analytic modelling can lead to more robust analysis but also incorporate additional assumptions. A feature of the semi-Markov model is that it explicitly considers the time spent in each state, in contrast to the Markov model, which has a single timescale, the time from entry into the study. This assumption is relevant in the setting of cost studies. Distribution of covariates in two arms not equal. Also normality assumed for costs.
7	Chen et al. (2000)	Functional outcomes and costs among patients of different types of PAC were not directly comparable due to possible selection bias.	Qualitative discussion of the covariates included in the equations.	Ordinary least squares regressions for costs and health outcomes on identified homogenous subgroup of patients.	Scheffe and χ^2 tests. Several specification tests were conducted to test the instrumental variable analysis assumptions.	Authors provided a comprehensive justification regarding the outcome measure used (instead of QALYs). Specification tests provided evidence on the validity of the instruments used. Another selection adjustment technique was used to verify the results and the authors stated that the findings were consistent. Authors stated that they addressed uncertainty for both costs and consequences but the approach used is superseded by more valid methods in the current literature. Authors defended the use of calculating confidence intervals instead of traditional sensitivity analysis. For multiple comparators, the authors used the coefficients estimated from the multinomial logit equation to adjust for selection effects in the ordinary least squares regression model for functional outcomes and costs.

#	Study	Justification for		Alternative specifications	Tests	Comments
		method	specification			
8	Coleman et al. (2006)	Propensity score matching to assure similarities between demographic and prior disease characteristics.	Descriptive.	None reported.	Categorical variables compared using χ^2 analysis or the Fisher exact test. Unpaired t-test to compare continuous data.	Based on a trial sample size calculation revealed that 54 patients in each group would be required to detect differences with an 80% of the power of the study. Post-match balance of means was reported. The size of the groups compared provides a low statistical power to detect significant differences in some of the outcomes.
9	Coyte et al. (2000)	Study addresses an important question which would be unethical to assess using a randomised controlled trial.	Descriptive sometimes backed up with literature references.	None reported.	Categorical variables compared using χ^2 analysis or the Fisher exact test. t-tests and ANOVA to compare continuous data.	To estimate the treatment costs and outcomes for the entire patient population, weighted sums of the stratum-specific results were calculated, using standard methods for stratified sampling. Multiple treatments were taken into account using a propensity score for different pairs. Authors claim that this allows different propensity score models for different comparisons. Nevertheless, results obtained may refer to different sub-populations.
10	Cutler (2007)	Instrumental variable analysis more appropriate than ordinary least squares for selection bias from unobserved sources.	Description of covariates included and relevant equations provided. Choice of covariates based on previous literature.	Models for the impact of revascularisation as sensitivity analysis of the basic instrumental variable results. Logarithmic specifications gave similar results.	None reported.	Comprehensive discussion about choice of instrument with evidence on whether it is appropriate and valid was based on looking at how observable risk factors are related to differential distance. Comparison with other studies using the same instrument and very similar datasets yielded comparable results. Study's strength was the availability of 17 additional years of follow-up data hence analysing outcomes over a longer period of time.
11	De Natale et al. (2009)	The propensity score method was used to reduce bias in estimation of effects when covariates in the two groups were unbalanced.	Descriptive.	None reported.	Continuous: Student's t-test when normality or Wilcoxon test otherwise. Categorical: χ^2 or Fisher's exact test when sample was small. All two-sided.	Groups were not balanced in few respects. Propensity score quartiles were used in the regression of the effects but it is unclear whether bias was properly adjusted for. No attempt was made to adjust cost estimation for the unbalanced covariates between groups.

#	Study	Justification for		Alternative specifications	Tests	Comments
		method	specification			
12	De Ridder et al. (2009)	Examine marginal impact of covariates on incremental net benefit, identify important subgroups straightforward handling of uncertainty.	Descriptive for the second model.	Simple net-benefit, covariate adjustment, interaction effects for the impact of covariates on incremental net benefit, different willingness to pay values.	t-tests and χ^2 tests for differences between treatment groups	Groups exhibited some differences in patient characteristics, but authors note that these are unlikely to affect the final results. Authors attempted to use instrumental variable analysis but no suitable instruments were available. Non-significance of interaction effects potentially due to the small sample size. Authors noted that it was uunnecessary to calculate confidence intervals for the net-benefit regression framework because the results for all parameters in the model are significant. They also noted that selection is a more important issue for effects rather than costs because physicians care less about costs. Authors justified the use of EQ-5D to calculate QALYs by stating that a literature review suggests that it is sensitive in detecting changes in quality of life when considering patients with schizophrenia.
13	Dhainaut et al. (2007)	Incomparability of the groups in terms of resource use and hence of costs in the initial cohort.	Descriptive.	None reported.	Standardized differences in each baseline variable between the two groups.	Sample size was designed for cost comparisons. As a result, the study is underpowered to deal with effectiveness issues. Post-match balance was reported.
14	Farias-Eisner et al. (2009)	None provided.	Descriptive.	None reported.	Unadjusted costs and clinical outcomes compared with t and χ^2 tests respectively.	Post-match balance of demographic, disease and treatment characteristics between groups reported.
15	Franks et al. (2005)	Regression models were developed to adjust for the complex sample designs used in the data sources.	Descriptive.	None reported.	None reported.	Sample size was not reported. Authors further acknowledged that additional studies are needed using different datasets and approaches. Quasi-experimental designs, propensity scores, instrumental variables employing good instruments may yield less biased estimates.

#	Study	Justification for		Alternative specifications	Tests	Comments
		method	specification			
16	Givon et al. (1998)	Multiple regressions to control for all possible biases.	Descriptive.	None reported.	χ^2 or Fisher exact test for discrete variables. Spearman correlation coefficient for continuous variables. t-tests for differences.	Patients were comparable in their baseline characteristics but different in terms of ethnicity and indication. It is unclear how uncertainty in ICER was evaluated and whether there was any uncertainty in cost estimates. Authors acknowledged the potential issues arising from the number of patients not returning the questionnaire.
17	Goeree et al. (2009)	Propensity score matching because of the non-randomized nature of recruitment.	Determination of variables for propensity score matching was made through univariate analysis on the available explanatory variables.	None reported.	None reported.	Post-match balance of means and covariates was reported. The analysis depends extensively on data collected outside the study, in particular for the valuation of costs.
18	Grieve et al. (2008)	No parametric assumptions. Also, allows for adjustment in baseline differences between the groups right across the distribution.	Descriptive.	Two-part model to estimate incremental costs and a multiple linear regression model to estimate incremental effectiveness.	Non-parametric bootstrap Kolomogorov-Smirnov (KS) distributional tests.	Sample size consisted of 522 patients before matching (151, 176 and 195 in each group) and 453 patients after matching (151 in each group). The non-parametric KS test is more appropriate given the highly non-normal distribution of the cost data. Post-match covariate balance was reported. Genetic matching does not rely on parametric assumptions such as assuming that the baseline costs are normally distributed. It also allows for adjustments of baseline differences across the groups right across the distribution The approach was used to independently match two of the intervention groups to the third.
19	Grieve et al. (2000)	None provided.	Descriptive.	Separate Cox regression analysis compared survival between the two hospitals.	For interaction effects in the Cox regression model.	Cohort study with comparable centres and patients. Multiple imputation for missing resource use values. Barthel index also used and functional outcome between centres were compared using logistic regression adjusting for case-mix.

#	Study	Justification for		Alternative specifications	Tests	Comments
		method	specification			
20	Griffin et al. (2007)	Classification of patients based on clinical appropriateness for valid comparisons. SUR deals with the potential correlation between costs and consequences.	Descriptive.	None reported.	None reported.	A cohort study for which 90% of unselected consecutive patients were matched to an appropriate rating. Correlation between costs and effects was taken into account using Seemingly Unrelated Regression (SUR). Missing data were imputed using ordinary least squares for length of stay and resource use and chained equations for adjusted analysis and utilities. Imputed datasets allowed for retention of between imputation variance in estimating standard errors. Groups were comparable with respect to their characteristics.
21	Groeneveld et al. (2008)	PSM approximates pseudo-randomisation of treatment and controls. It is also a simple and transparent statistical design.	Descriptive.	Two different Cox proportional-hazards survival models.	Comparisons between median costs using Wilcoxon rank-sum non-parametric tests.	The initial Cox model included only ICD as a predictor of survival. A subsequent model included ICD receipt, the propensity score, and demographic/clinical characteristics that remained imperfectly balanced between groups across quintiles of propensity scores. Post-match balance of means and covariates was reported. The method of selecting controls was biased, by design, toward inclusion of patients who were "healthier" than typical device recipients. As such, survival in the control groups cannot be compared to survival in the pharmacologic arms of randomised clinical trials.
22	Heaton et al. (2006)	The use of propensity scores can reduce selection bias by 90%.	Descriptive.	None reported.	Mann-Whitney U for comparing costs distributions. t-tests for continuous variables and chi-square tests for categorical variables.	Because balance in propensity score quintiles was not achieved in the propensity model for inhaled corticosteroids and short-acting beta2-agonists, the final logistic regression model for health outcomes had 4 independent variables: inhaled corticosteroids, short-acting beta2-agonists, LM use, and the propensity probability.

#	Study	Justification for		Alternative specifications	Tests	Comments
		method	specification			
23	Indurkhya et al. (2006)	Traditional model-based covariate adjusted estimates are biased if the covariate distributions in treatment groups do not have substantial overlap.	Descriptive.	Logit, quintile, and continuous (actual value) form of the propensity score.	Two-way analysis of variance model, which included main effects for propensity score quintile to check balance in covariates after propensity score adjustment.	Propensity score mean balance and covariate distributions reported. Net monetary benefit estimates for λ values of 100, 500 and 1000. The inclusion of propensity score as a covariate in regression analysis adds advantage only in terms of more precision in the estimation. However, it is unlikely to reduce the potential for bias compared to direct covariate-adjusted analysis.
24	Kariv et al. (2006)	Case control pairs were carefully matched to ensure similarity of patient characteristics and overcome potential selection bias.	Descriptive.	None reported.	Pearson χ^2 and Fisher's exact tests for categorical data and t-test for unpaired data. Wilcoxon signed ranks and paired t-tests for paired data.	As defined by the matching criteria patients were similar in age and identical in gender, preoperative diagnosis, and surgical procedure performed.
25	Knapp et al. (2008)	Epoch analysis considers patients that switch treatment. Allows short-term and cumulative estimation of treatment effects.	Descriptive. It was also noted that different periods have different requirements.	Different specifications used for the 3 periods (Epochs).	Modified Park Test.	A large naturalistic study with the analysis based on longitudinal data that took in consideration the different periods of treatment over 12 months. Development of combined linear and nonlinear models for repeated observations is required as will provide more efficient estimates. An extension to regression analysis for longitudinal data with treatment switches. An assumption that treatment effects are short term is made.

#	Study	Justification for		Alternative specifications	Tests	Comments
		method	specification			
26	Lairson et al. (2008)	Adjusts for time-invariant patient and environmental characteristics that may be correlated with outcomes, group selection, and time-varying factors common to both groups.	Descriptive.	None reported.	Student's t-test for paired data to compare the two groups for continuously distributed variables, chi-square test for binomially coded variables.	Post-match balance was reported. Difference-in-difference assumption was tested indirectly by examining pre-intervention trends in outcomes for the two groups. In results, individual and quarterly fixed effects included in the regression were not reported.
27	Linden et al. (2005)	Can reduce selection bias and regression to the mean when randomisation is impractical.	Descriptive. Covariates chosen mainly because they were readily available in the data.	Both stratification and matching was used.	None reported.	Authors note that the propensity score technique for DM programme evaluation requires large samples especially when using subclassification, which was not the case in the study. Most subclasses had extremely small number of participants. This leads to great variability to the covariate distribution. Administrative data suffer from lack of accuracy and also had limited variables. Post-match balance of means and the propensity score distributions, were reported. Graphical analysis was also used.
28	Manca, Austin (2008)	Propensity score analysis addresses some of the limitations of matching, stratification and regression. Unbiased estimation subject to ignorability.	Descriptive.	None reported.	Balance was checked with t or Wilcoxon rank sum tests for continuous variables and χ^2 tests for dichotomous variables. Distribution of the propensity score reported before and after matching.	Propensity score methodology could control for observable confounders but not for unobservable confounders.

#	Study	Justification for		Alternative specifications	Tests	Comments
		method	specification			
29	McClellan, Newhouse (1997)	Detailed analysis of the implications of partial failures of the identification conditions required for consistent difference-in-difference estimation	Descriptive. Also minimal parametric assumptions.	Reduced form models, different fixed effects and interactions in models.	F-tests for the six hospital type-time interactions included as instrumental variables demonstrated that there is no bias from weakly correlated instrumental variables.	Costs and Effects are adjusted separately but under the same model and therefore correlation is preserved in mean estimate. It is unclear how the correlation might be taken forward to the uncertainty in cost-effectiveness ratio. The comparison between instrumental variables panel method and the least squares approach shows that bias do exist in the latter when estimating incremental costs and outcomes. No evidence that the instruments are not correlated to the unobserved heterogeneity in outcomes.
30	Merito, Pezzoti (2006)	Propensity scores were used to account for selection bias. The propensity score methodology is one of the techniques recently introduced to address the issue of confounding in observational studies.	Descriptive. Also regressors in the logistic model chosen based on a forward-stepwise procedure.	Various Cox proportional hazards models and OLS models for costs and consequences.	Goodness of fit of logit models by χ^2 and Hosmer-Lemeshow tests. Cox model tested using Schoenfeld residuals and graphic methods.	Tests of the balancing property for the observed covariates in the two groups were restricted to the region of common support for the propensity score. The balancing property was checked using standard statistical tests for the comparison of the difference in means between immediate and deferred patients within each propensity score stratum for continuous covariates, and of the difference in the odds ratios for categorical variables.
31	Mihaylova et al. (2010)	Propensity scores more appropriate than regression. No suitable instruments for instrumental variable analysis. A degree of robustness can be achieved by considering results based on different methods jointly for the purpose of their interpretation.	Descriptive based on clinical opinion. Also, a stepwise backward elimination algorithm was used to identify significant covariates.	None reported.	None reported.	Post-match balance for means and covariates and post-match distribution of covariates were not reported. Correlation between costs and effects was preserved in regression adjustment using seemingly unrelated regression and in propensity score analysis. A limitation in the propensity score analysis in terms of separate adjustments for each for each separate treatment comparison rather than comparison of all treatment options simultaneously is noted.

#	Study	Justification for		Alternative specifications	Tests	Comments
		method	specification			
32	Mitra, Indurkhya (2005)	A new general linear model framework to estimate measures of cost–effectiveness and to demonstrate the advantages of using propensity score adjustment in assessing the cost–effectiveness of competing non-randomised treatments.	Based on the severity of non-cancer medical illness, using comorbidity indexes.	Linear net benefit model, linear net benefit with covariate adjustment, propensity score adjusted linear net benefit model.	Two-way analysis of variance model to check balance of each covariate.	Cost distributions in both groups were highly skewed with long tails; normality assumption for the net monetary benefit might not be appropriate. Authors note that propensity scores help make the treatment groups comparable with respect to important baseline characteristics. This in turn allows one to obtain more precise estimates of the net monetary benefit. The general linear model framework is useful in conducting subgroup net monetary benefit analysis by introducing a dummy variable for the subgroups and noting the estimate of the corresponding coefficient. Furthermore, this method provides estimates that are best linear unbiased estimates (BLUE) because they are the ordinary least squares solution to the normal regression equation.
33	Mojtabai, Zivin (2003)	Propensity score analysis was used to account for selection bias.	The socio-demographic and clinical variables that had shown significant variation across modalities were included.	None reported.	F-test and χ^2 test for continuous and categorical data comparison.	Mean balance of covariates in strata following calculation of propensity scores was reported. The cost-effectiveness analysis does not seem to be based on incremental costs and consequences but rather on average costs and consequences and their ratios.
34	Polignano et al. (2008)	None provided.	Descriptive.	None reported.	Student's t-test, χ^2, Fisher exact test.	The matched groups were homogenous in terms of age, sex, coexisted morbidity, type of resection and prevalence of liver cirrhosis. The groups were matched for magnitude of resection and for tumour location and size. After selection of the case-matched controls, the intention-to-treat principle was applied. Authors acknowledge influence of social factors on length of hospital stay.

#	Study	Justification for		Alternative specifications	Tests	Comments
		method	specification			
35	Polsky et al. (2003)	Propensity scores control for probability of treatment receipt.	Descriptive. Covariates theoretically predictive of the outcome.	None reported.	Group differences were checked with t-tests for continuous variables and χ^2 tests for dichotomous variables	Power calculations were not reported. Authors imputed costs based on survival by using a repeated-measures analysis of variance regression of interval costs estimated among patients who were alive during the interval in which the independent variables were treatment group, interval, interaction between interval and treatment group, and a standard set of explanatory variables. Also, they adjusted for the fact that patients who are no longer observed may not survive by multiplying imputed costs in the interval by the patient's predicted survival in that interval.
36	Polsky, Basu (2006)	The aim was to compare the performance of the methods when adjusting for selection bias.	Descriptive.	None reported.	None reported.	For instrument justification authors referred to another study. Unclear how confidence intervals reflecting uncertainty were calculated. Based on a prior publication it seems that the uncertainty was addressed using the non-parametric bootstrap approach. Very limited information is provided regarding the application of the instrumental variable approach.
37	Sadhu et al. (2008)	Difference-in-difference deals with secular time trends in hospital length of stay, costs and mortality.	Descriptive. Linear time trend that allows for secular trends in costs and length of stay. Interaction between time period and type of intervention for intervention effect.	Several alternative regression specifications including random effects models (results not provided).	χ^2 and Wilcoxon tests for differences in demographic and clinical characteristics. Graphs of the pre-existing time trends to test time trend assumption.	The specifications yielded findings consistent with the final specification, but because of concerns about over fitting and interpretability of the results, the most parsimonious specification was ultimately chosen. Because of the skewed distributions of the cost and length of stay measures, outcomes were log-transformed in linear regressions, and the estimates were retransformed to calculate intervention effects on costs and length of stay measured on the original scales.

#	Study	Justification for		Alternative specifications	Tests	Comments
		method	specification			
38	Sekhon, Grieve (2009)	Genetic Matching does not depend on the propensity score.	Descriptive. Covariate adjustment for propensity score and genetic matching based on literature recommendation.	Attempted to improve balance with interaction and higher order terms in the propensity score model.	QQ-plots and KS tests for continuous variables and t-tests for categorical variables.	Sample size also consisted of 31,447 potential controls. Similar populations and same methods to measure costs and outcomes, same exclusion criteria for randomised controlled trial-matching comparisons. Simulated non-randomised data were generated using data from a randomised controlled trial. In the simulated study costs were estimated using generalized linear model assuming a Gamma distribution and a log link. QALYs using a two-part model: a logistic regression and a generalized linear model with a Gamma distribution and an identity link. Costs and effects were fixed and treatment assignment was varied 1000 times determined each time by a propensity score estimated using logistic regression. This score does not capture the complexity of the true propensity score.
39	Shih et al. (2007)	A polychotomous selection model explored the issue of endogeneity in the frequentist framework. In the Bayesian approach the issue of sample selection bias was not examined as the methods are currently under development.	Descriptive. Also, for the multinomial model to be identifiable, variables in the first and second stage regressions can overlap but not be identical.	Different frequentist and Bayesian net benefit regressions.	Groups were not balanced in terms of socioeconomic characteristics. A fully interacted regression was used to test the overlap assumption. For heteroskedasticity of unknown form a Breusch-Pagan test was used. Group-wise heteroskedasticity was assessed by testing the equality of variance of the error term between patients in different groups. A Hausman test was used to check the independence of irrelevant alternative property for the multinomial logit model.	
40	Shireman, Braman (2002)	Propensity score analysis identifies a matched control group with similar risk factors and is a method adjusting for selection bias in observational research.	Descriptive.	None reported.	None reported.	No power calculations to determine the sample size were reported. Mean values of covariates for the two groups were presented but no tests to assess the comparability of the two groups were reported.

#	Study	Justification for		Alternative specifications	Tests	Comments
		method	specification			
41	Soegaard et al. (2007)	Regression analysis to investigate possible determinants for cost-effectiveness.	Model provisions by studying residuals vs. fitted values, residuals vs. possible determinants, normality of residuals.	None reported.	Bivariate correlation test of Kendall's tau-b for paired observations of costs and effects. Comparison of surgical groups using Kruskal-Wallis' test. Pair wise correlations and scatter diagrams for interactions.	Authors acknowledged problems of statistical power and noted that traditional power calculations for comparative analysis of cost-effectiveness are insufficient. Number of replications for boostrapping was calculated by means of Andrews and Buchinsky's method. Imputation was conducted to replace missing values in the Questionnaire. Horizontal (intra-patient) means of non-missing values within individual areas of functional disability were calculated and used for imputation. Non-response in 2-year disability was imputed by means of a regression approach. CEAC by means of a non-parametric method described by Lothgren and Zethraeus. Groups were balanced in terms of patient characteristics except for age. Very poor correlations were found between treatment costs and each of the four factors of the effect measure.
42	Weiss et al. (2002)	Propensity score matching to address selection bias.	Descriptive. All variables were retained in the propensity score model, regardless of the level of statistical significance.	Subgroup analysis of 1269 pairs in the middle tertile of the propensity score.	C-statistic.	Power calculations were not reported. Groups after matching were similar. Comparisons with other studies may be invalid because of different follow up.
43	Windmeijer et al. (2006)	A methodological framework allowing the treatment effects to be estimated in a longitudinal observational study where some patients have switched their treatment while accounting for selection bias.	Descriptive. Also, to allow for flexible treatment effects over time, separate coefficients for the different epochs were estimated.	Three different epochs.	Modified Park test.	An extension to regression analysis for longitudinal data with treatment switches. An assumption that treatment effects are short term is made. To control for the fact that the patients with repeated observations for the first epoch may be inherently different from those patients who do not switch treatment a switching/repeated observation binary indicator is fitted in the models. The epoch analysis is also flexible enough to allow for a reliable representation of uncertainty in sampling using nonparametric bootstrap resampling and uncertainty in the decision rule by means of the cost-effectiveness acceptability curve.

Table A6 – Key data sources identified from the reviewed studies

#	Study	Acronym	Name	Type	Format
1	Akazawa et al. (2008)	IHCIS	National Managed Care Benchmark Database, Integrated Healthcare Information Services	Longitudinal	Administrative database
2	Alegria et al. (2005)		Not reported	Repeated cross-sections	Survey
3	Barnette, Swindle (1997)	Not reported	Veteran Affairs (VA) Patient Treatment File, VA Cost Distribution Report, VA Computerized Accounting for Local Management.	Longitudinal (linked)	Administrative databases
4	Blanchette et al. (2008)	IHCIS	Integrated Healthcare Information Services	Longitudinal	Administrative database
5	Cakir et al. (2006)		Not reported	Short-term	Prospective observational matched-sample study
6	Castelli et al. (2007)		Not reported	Longitudinal	Registry database
7	Chen et al. (2000)	MADRS	Medicare Automated Data Retrieval System	Longitudinal	Administrative database and cohort study
8	Coleman et al. (2006)		Not reported	Longitudinal	Prospective cohort study
9	Coyte et al. (2000)	CIHI OHCAS	Canadian Institute for Health Information, Ontario Home Care Administrative System	Longitudinal (linked)	Administrative databases
10	Cutler (2007)		Not reported	Short-term	Administrative databases
11	De Natale et al. (2009)	GPRD	UK General Practitioner Research Database	Longitudinal	Administrative database
12	De Ridder et al. (2009)		Not reported	Longitudinal	Prospective observational Survey
13	Dhainaut et al. (2007)	CUB-Rea Not reported	College of Intensive Care Database Users Programme de Médicalisation des Systèmes d'Information	Follow-up	Before-after observational study
14	Farias-Eisner et al. (2009)	Not reported	Premier's Perspective Comparative Database	Short term follow-up	Administrative database
15	Franks et al. (2005)	NHIS - MEPS	National Health Interview Survey National Death Index Medical Expenditure Panel Survey	Not reported	Administrative databases
16	Givon et al. (1998)		Not reported	Longitudinal	Cohort study

#	Study	Acronym	Name	Type	Format
17	Goeree et al (2009)	ICES CARDI ACCESS	Institute of Clinical and Evaluative Sciences Cardiac Care Network Registry	Longitudinal (follow-up)	Not reported Registry database
18	Grieve et al. (2008)		Not reported		Administrative databases
19	Grieve et al. (2000)	Not reported	South London Stroke Register Hvidovre Hospital Stroke Database	Not reported	Medical Records
20	Griffin et al. (2007)	ACRE Not reported	The appropriateness of coronary revascularisation cohort UK Office of National Statistics	Not reported	
21	Groeneveld et al. (2008)	Not reported	Medicare Annual Denominator File Social Security Death Master File	Not reported	Administrative databases Not reported
22	Heaton et al. (2006)	Not reported	Ohio Medicaid Database	Not reported	Administrative database
23	Indurkhya et al. (2006)	SEER MEDPAR NCH SAF	Surveillance Epidemiology and End Results Medicare Provider Analysis and Review File National Claims Histories Standard Analytic Files	Not reported	Registry Administrative databases
24	Kariv et al. (2006)		Not reported		Institutionally maintained database
25	Knapp et al. (2008)	SOHO TFR2 MIMS - MIDAS PICAS	Schizophrenia Outpatient Health Outcomes Study Trust Financial Returns Monthly Index of Medical Specialties Chemist and Druggist Supplement IMS Health MIDAS database UK Pharmaceutical Industry Costing Analysis System database	SOHO: Longitudinal	Cohort Study Databases detailed description of which was not provided
26	Lairson et al. (2008)	Not reported	Not reported Centers for Medicare and Medicaid Services	Time-series Not reported	Clinical database Administrative database
27	Manca, Austin (2008)	OMID CIHI OHIP ODB RPDB	Ontario Myocardial Infarction Database Canadian Institute for Health Information Ontario Health Insurance Plan Ontario Drug Benefit Ontario Registered Persons Database	Not reported	Administrative databases

#	Study	Acronym	Name	Type	Format
28	McClellan, Newhouse (1997)		Not reported	Longitudinal	Administrative database
29	Merito, Pezzoti (2006)	ICONA -	Italian Cohort Naive Antiretrovirals Study Italian National Pharmaceutical Formulary	Not reported	Cohort Study Not reported
30	Mihaylova et al. (2010)	SUIT	Stress Urinary Incontinence Treatment Study	Not reported	Cohort Study
31	Mitra, Indurkhya (2005)	SEER MEDPAR	Surveillance Epidemiology and End Results Medicare Provider Analysis and Review File	Not reported	Registry Administrative databases
32	Mojtabai, Zivin (2003)	SROS	Services Research Outcomes Study	Not reported	Cohort Study
33	Polignano et al. (2008)	Not reported	Scottish Health Service Costs Book	Not reported	
34	Polsky et al. (2003)	OPTIONS - - -	Outcomes and Preferences for Treatment in Older Women Nationwide Survey United States Census The Area Resource File Centers for Medicare and Medicaid Services national claims database	Not reported	Not reported Administrative database
35	Polsky, Basu (2006)	Not reported	CMS Medicare Claims	Not reported	Administrative database
36	Sadhu et al. (2008)		Not reported		
37	Sekhon, Grieve (2009)	ICNARC CMP	Intensive Care National Audit Research Centre Case Mix Program database	Not reported	Administrative database
38	Shih et al. (2007)	Not reported	Medicare MarketScan® Database	Not reported	Administrative database
39	Shireman, Braman (2002)	Not reported	Kansas Medicaid Drug Utilization Review Program	Not reported	Administrative database
40	Soegaard et al. (2007)	Not reported	National Patient Registry, National Health Service National Health Insurance Service Registry, National Health Service Register of Prescribed Medication, Danish Medicines Agency Social Science Research Register, Statistics Denmark	Not reported	

#	Study	Acronym	Name	Type	Format
42	Weiss et al. (2002)	Not reported	Health Care Financing Administration Medicare Provider Analysis and Review inpatient hospitalization file Medicare Beneficiary Health Insurance Skeletonized Eligibility Write-off file	Longitudinal	Administrative databases
43	Windmeijer et al. (2006)	SOHO	Schizophrenia Outpatient Health Outcomes Study	Longitudinal	Cohort Study

Table A7 - Probit model results for Home vs. Obstetric Unit

Variable	Coefficient	Standard error	z-Stat
Constant	-0.331	0.026	-12.66
Parity: 0 previous pregnancy (reference)	-	-	-
Parity: 1 previous pregnancy	0.511	0.017	30.49
Parity: 2 previous pregnancies	0.792	0.022	36.35
Parity: 3+ previous pregnancies	0.757	0.027	27.93
37 weeks gestation	-0.319	0.045	-7.15
38 weeks gestation	-0.059	0.026	-2.28
39 weeks gestation	-0.001	0.019	-0.05
40 weeks gestation (reference)	-	-	-
41 weeks gestation	-0.106	0.019	-5.66
42-44 weeks gestation	-0.193	0.050	-3.87
Married (reference)	-	-	-
Single/unsupported partner	-0.387	0.030	-12.85
BMI not recorded	-0.009	0.020	-0.44
BMI 10-18	-0.137	0.049	-2.78
BMI 19-24 (reference)	-	-	-
BMI 25-29	-0.138	0.018	-7.54
BMI 30-35	-0.271	0.027	-10.09
White British (reference)	-	-	-
Indian/Bangladeshi	-1.215	0.067	-18.21
Pakistani	-1.382	0.088	-15.65
Black Caribbean	-0.360	0.073	-4.97
Black African	-0.979	0.061	-16.08
Mixed	0.038	0.056	0.68
Other	-0.580	0.048	-12.03
Maternal age <20	-0.696	0.045	-15.47
Maternal age 20-24	-0.335	0.023	-14.33
Maternal age 25-29 (reference)	-	-	-
Maternal age 30-34	0.182	0.019	9.69
Maternal age 35-39	0.250	0.022	11.56
Maternal age 40-60	0.079	0.041	1.93
IMD 0.37-8.31 (reference)	-	-	-
IMD 8.32-13.73	0.101	0.024	4.14
IMD 13.74-21.21	0.039	0.024	1.63
IMD 21.22-34.41	0.123	0.024	5.23
IMD 34.42-85.46	0.115	0.023	4.92
Fluent in English (reference)	-	-	-
Some English	-1.042	0.064	-16.38
No English	-1.162	0.129	-9.00
N	35034		
Pseudo R^2	0.147		
Log likelihood	-20622.8		

Table A8 - Probit model results for Freestanding Midwifery Unit vs. Obstetric Unit

Variable	Coefficient	Standard error	z-Stat
Constant	-0.399	0.027	-14.15
Parity: 0 previous pregnancy (reference)	-	-	-
Parity: 1 previous pregnancy	0.176	0.018	10.08
Parity: 2 previous pregnancies	0.281	0.025	11.21
Parity: 3+ previous pregnancies	0.120	0.034	3.59
37 weeks gestation	-0.175	0.044	-3.94
38 weeks gestation	-0.098	0.028	-3.53
39 weeks gestation	-0.010	0.020	-0.50
40 weeks gestation (reference)	-	-	-
41 weeks gestation	-0.049	0.020	-2.52
42-44 weeks gestation	-0.599	0.062	-9.67
Married (reference)	-	-	-
Single/unsupported partner	-0.334	0.029	-11.51
BMI not recorded	-0.126	0.021	-5.91
BMI 10-18	-0.198	0.050	-3.98
BMI 19-24 (reference)	-	-	-
BMI 25-29	-0.077	0.019	-4.02
BMI 30-35	-0.183	0.028	-6.59
White British (reference)	-	-	-
Indian/Bangladeshi	-0.317	0.047	-6.78
Pakistani	-0.302	0.055	-5.47
Black Caribbean	-0.553	0.089	-6.23
Black African	-0.731	0.062	-11.72
Mixed	-0.206	0.066	-3.13
Other	-0.265	0.044	-6.02
Maternal age <20	0.041	0.035	1.18
Maternal age 20-24	0.000	0.022	0.02
Maternal age 25-29 (reference)	-	-	-
Maternal age 30-34	-0.005	0.020	-0.26
Maternal age 35-39	-0.047	0.025	-1.88
Maternal age 40-60	-0.223	0.051	-4.40
IMD 0.37-8.31 (reference)	-	-	-
IMD 8.32-13.73	0.317	0.026	12.30
IMD 13.74-21.21	0.263	0.025	10.56
IMD 21.22-34.41	0.198	0.025	7.99
IMD 34.42-85.46	0.110	0.025	4.51
Fluent in English (reference)	-	-	-
Some English	-0.354	0.043	-8.23
No English	-0.598	0.083	-7.24
N	29818		
Pseudo R^2	0.040		
Log likelihood	-18827		

Table A9 – Probit model results for Alongside Midwifery Unit vs. Obstetric Unit

Variable	Coefficient	Standard error	z-Stat
Constant	0.051	0.023	2.23
Parity: 0 previous pregnancy (reference)	-	-	-
Parity: 1 previous pregnancy	0.119	0.016	7.62
Parity: 2 previous pregnancies	0.107	0.023	4.63
Parity: 3+ previous pregnancies	0.169	0.032	-5.34
37 weeks gestation	0.193	0.039	-4.90
38 weeks gestation	-0.104	0.024	-4.25
39 weeks gestation	-.0164	0.018	-0.93
40 weeks gestation (reference)	-	-	-
41 weeks gestation	-0.105	0.018	-5.91
42-44 weeks gestation	-0.545	0.053	-10.30
Married (reference)	-	-	-
Single/unsupported partner	-0.187	0.024	-7.71
BMI not recorded	-0.071	0.019	-3.75
BMI 10-18	-0.118	0.042	-2.83
BMI 19-24 (reference)	-	-	-
BMI 25-29	-0.090	0.017	-5.26
BMI 30-35	-0.221	0.025	-8.75
White British (reference)	-	-	-
Indian/Bangladeshi	-0.067	0.036	-1.85
Pakistani	-0.025	0.040	-0.61
Black Caribbean	-0.090	0.061	-1.47
Black African	-0.078	0.039	-1.98
Mixed	0.049	0.053	0.93
Other	0.103	0.033	3.17
Maternal age <20	-0.052	0.031	-1.69
Maternal age 20-24	-0.026	0.020	-1.33
Maternal age 25-29 (reference)	-	-	-
Maternal age 30-34	0.023	0.018	1.25
Maternal age 35-39	-0.001	0.023	-0.38
Maternal age 40-60	-0.265	0.049	-5.44
IMD 0.37-8.31 (reference)	-	-	-
IMD 8.32-13.73	-0.118	0.023	-5.12
IMD 13.74-21.21	-0.168	0.022	-7.58
IMD 21.22-34.41	-0.044	0.021	-2.09
IMD 34.42-85.46	0.018	0.020	0.90
Fluent in English (reference)	-	-	-
Some English	0.071	0.030	2.37
No English	-0.181	0.053	-3.40
N	34878		
Pseudo R^2	0.0134		
Log likelihood	-23739.9		

Table A10 – Regression analysis estimates

		Home vs. Obstetric Unit		
	NBR + PS	SUR + PS	SUR + PS@C	SUR + PS@E
Cost	N/A	-319	-319	-325
CI	N/A	-349 – -289	-349 -289	-354 – 297
Effect	N/A	0.18	0.18	0.18
CI	N/A	0.15 – 0.20	0.16 – 0.20	0.16 – 0.20
NMB	3,858	3,859	3,883	3,878
CI	3,398 – 4,319	3,410 – 4,307	3,428 – 4,337	3419 – 4337
		Freestanding Midwifery Unit vs. Obstetric Unit		
Cost	N/A	-117	-117	-118
CI	N/A	-151 – -83	-151 – -83	-150 – -86
Effect	N/A	0.26	0.26	0.26
CI	N/A	0.25 – 0.28	0.25 – 0.28	0.25 – 0.28
NMB	5,353	5,353	5,351	5,355
CI	5,005 – 5,701	4,965 – 5,741	5,011 – 5,693	5010 – 5701
		Alongside Midwifery Unit vs. Obstetric Unit		
Cost	N/A	-121	-121	-121
CI	N/A	-151 – -91	-150 – -91	-149 – -93
Effect	N/A	0.21	0.21	0.21
CI	N/A	0.19 – 0.22	0.19 – 0.22	0.19 – 0.22
NMB	4,212	4,212	4,210	4,213
CI	3,876 – 4,548	3,869 – 4,555	3,882 – 4,539	3,878 – 4,547

Notes: Table presents incremental estimates and associated 95% confidence intervals. NBR: net benefit regression; SUR: seemingly unrelated regression; PS: propensity score; PS@C: propensity score in cost equation; PS@E: propensity score in effectiveness equation; N/A: not applicable; NMB: net monetary benefit (lamba=£20,000); CI: confidence interval.

Table A11 – Doubly robust / bias-corrected matching estimates for Home vs. OU

	Coarsened Exact	Entropy Balancing	Inverse Probability	Nearest Neighbour
Cost	-325	-323	-324	-323
CI	-351 – -300	-347 – -300	-347 – -300	-351 – -296
Effect	0.19	0.17	0.17	0.18
CI	0.17 – 0.21	0.15 – 0.19	0.15 – 0.19	0.16 – 0.20
NMB	4,114	3,751	3,747	3,872
CI	3,644 – 4,584	3,348 – 4,154	3,344 – 4,151	3,433 – 4,311

	Nearest Neighbour (R)	Ratio	Caliper	Radius
Cost	-361	-339	-321	-327
CI	-394 – -329	-367 – -311	-352 – -291	-350 – -303
Effect	0.22	0.18	0.19	0.17
CI	0.19 – 0.25	0.16 – 0.21	0.16 – 0.21	0.15 – 0.19
NMB	4,786	3,944	4,025	3,777
CI	4,205 – 5,369	3,440 – 4,448	3,545 – 4,505	3,373 – 4,181

	Kernel (T)	Kernel (G)	Mahalanobis	M + PS
Cost	-326	-327	-387	-272
CI	-350 – -303	-351 – -304	-418 – -355	-304 – -240
Effect	0.17	0.17	0.21	0.18
CI	0.15 – 0.19	0.15 – 0.19	0.18 – 0.24	0.15 – 0.21
NMB	3,780	3,783	4,571	3,826
CI	3,376 – 4,184	3,380 – 4,187	4,017 – 5,127	3,249 – 4,403

Notes: Table presents incremental estimates and associated 95% confidence intervals. PS: propensity score; (R): replacement; (T): Tricube; (G): Gaussian; M: Mahalanobis; NMB: net monetary benefit; CI: confidence interval.

Table A12 – Doubly robust / bias-corrected matching estimates for FMU vs. OU

	Coarsened Exact	Entropy Balancing	Inverse Probability	Nearest Neighbour
Cost	-95	-119	-119	-112
CI	-127 – -62	-148 – -90	-148 – -90	-146 – -77
Effect	0.27	0.26	0.26	0.26
CI	0.25 – 0.29	0.24 – 0.27	0.24 – 0.27	0.24 – 0.28
NMB	5,390	5,280	5,280	5,329
CI	4,982 – 5,799	4,947 – 5,613	4,947 – 5,614	4,933 – 5,725

	Nearest Neighbour (R)	Ratio	Caliper	Radius
Cost	-103	-115	-112	-119
CI	-145 – -60	-151 – -78	-146 – -77	-148 – -90
Effect	0.28	0.26	0.26	0.26
CI	0.26 – 0.31	0.24 – 0.28	0.24 – 0.28	0.24 – 0.27
NMB	5,710	5,311	5,329	5,280
CI	5,197 – 6,222	4,885 – 5,736	4,933 – 5,725	4,947 – 5,614

	Kernel (T)	Kernel (G)	Mahalanobis	M + PS
Cost	-121	-122	-36	-77
CI	-150 – -92	-151 – -92	-76 – 4	-119 – -36
Effect	0.26	0.26	0.27	0.21
CI	0.24 – 0.28	0.24 – 0.28	0.24 – 0.30	0.19 – 0.23
NMB	5,299	5,312	5,453	4,246
CI	4,965 – 5,633	4,980 – 5,645	4,913 – 5,993	3,797 – 4,694

Notes: Table presents incremental estimates and associated 95% confidence intervals. PS: propensity score; (R): replacement; (T): Tricube; (G): Gaussian; M: Mahalanobis; NMB: net monetary benefit; CI: confidence interval.

Table A13 – Doubly robust / bias-corrected matching estimates for AMU vs. OU

	Coarsened Exact	Entropy Balancing	Inverse Probability	Nearest Neighbour
Cost	-113	-119	-119	-117
CI	-141 – -85	-146 – -92	-145 – -92	-145 – -89
Effect	0.21	0.20	0.20	0.21
CI	0.19 – 0.23	0.19 – 0.22	0.19 – 0.22	0.19 – 0.22
NMB	4,330	4,194	4,192	4,214
CI	3,966 – 4,693	3,868 – 4,521	3,865 – 4,519	3,874 – 4,554

	Nearest Neighbour (R)	Ratio	Caliper	Radius
Cost	-118	-121	-117	-119
CI	-154 – -82	-153 – -90	-145 – -89	-146 – -92
Effect	0.21	0.21	0.20	0.20
CI	0.19 – 0.23	0.19 – 0.23	0.19 – 0.22	0.19 – 0.22
NMB	4,232	4,247	4,214	4,192
CI	3,801 – 4,662	3,856 – 4,639	3,874 – 4,554	3,866 – 4,518

	Kernel (T)	Kernel (G)	Mahalanobis	M + PS
Cost	-118	-118	-124	-102
CI	-145 – -92	-145 – -91	-159 – -88	-137 – -68
Effect	0.20	0.20	0.22	0.22
CI	0.19 – 0.22	0.19 – 0.22	0.19 – 0.24	0.20 – 0.25
NMB	4,183	4,161	4,462	4,561
CI	3,858 – 4,508	3,837 – 4,484	4,003 – 4,920	4,083 – 5,039

Notes: Table presents incremental estimates and associated 95% confidence intervals. PS: propensity score; (R): replacement; (T): Tricube; (G): Gaussian; M: Mahalanobis; NMB: net monetary benefit; CI: confidence interval.

Table A14 – Sensitivity Analysis (Heckit models)

	INMB	lambda	z	P>|z|
Home vs. Obstetric unit	6205	-1424	-1.13	0.259
Freestanding Midwifery Unit vs. Obstetric Unit	5088	170	0.13	0.896
Alongside Midwifery Unit vs. Obstetric Unit	2510	1063	0.53	0.596

Notes:

INMB = incremental net monetary benefit (threshold value: 20,000)
Lambda = the inverse Mill's ratio

Table A15 – Sensitivity Analysis (Rosenbaum's bounds)

Home versus Obstetric Unit

Gamma	Q_mh+	Q_mh-	p_mh+	p_mh-
1	41.6169	41.6169	0	0
2	18.1858	67.0836	0	0
3	4.97911	83.492	3.2e-07	0
4	4.28853	96.095	9.0e-06	0
5	11.5277	106.567	0	0
6	17.4996	115.666	0	0
7	22.6145	123.797	0	0
8	27.1119	131.209	0	0
9	31.1436	138.061	0	0
10	34.8115	144.463	0	0

Freestanding Midwifery Unit versus Obstetric Unit

Gamma	Q_mh+	Q_mh-	p_mh+	p_mh-
1	37.6039	37.6039	0	0
2	15.0816	61.8986	0	0
3	2.31445	77.4491	.010322	0
4	6.67323	89.3376	1.3e-11	0
5	13.712	99.1798	0	0
6	19.5284	107.705	0	0
7	24.5167	115.305	0	0
8	28.9072	122.217	0	0
9	32.8462	128.595	0	0
10	36.4322	134.545	0	0

Table A15 – Sensitivity Analysis (Rosenbaum's bounds) (continued)

Alongside Midwifery Unit versus Obstetric Unit

Gamma	Q_mh+	Q_mh-	p_mh+	p_mh-
1	34.574	34.574	0	0
2	6.02122	64.5655	8.7e-10	0
3	10.4631	83.3211	0	0
4	22.2642	97.4223	0	0
5	31.5472	108.94	0	0
6	39.2617	118.804	0	0
7	45.9057	127.512	0	0
8	51.7722	135.366	0	0
9	57.0484	142.559	0	0
10	61.8609	149.226	0	0

Notes:

Mantel-Haenszel bounds for effectiveness endpoint 'normal birth'

Gamma : odds of differential assignment due to unobserved factors
Q_mh+ : Mantel-Haenszel statistic (assumption: overestimation of treatment effect)
Q_mh- : Mantel-Haenszel statistic (assumption: underestimation of treatment effect)
p_mh+ : significance level (assumption: overestimation of treatment effect)
p_mh- : significance level (assumption: underestimation of treatment effect)

```
// Stata code for Rosenbaum's sensitivity analysis

// Sort data by this variable to ensure random ordering

sort u

// Calling psmatch2

// Perform 1-to-1 caliper matching with no replacement (common support)

psmatch2 ppob c1cat2 c1cat3 c1cat4 b1cat1 gestcats1 gestcats2 gestcats3
                gestcats5 gestcats6 b4 b5cat1 b5cat3 b5cat4 b5cat5 b2regcat2
                b2regcat3 b2regcat4 b2regcat5 b2regcat6 b2regcat7 b1cat2
                b1cat4 b1cat5 b1cat6 b6quint1 b6quint2 b6quint3 b6quint4
                b3cat2 b3cat3, outcome(normal) caliper(0.053) common
                noreplacement descending

// Calling mbounds

// Calculate Mantel and Haenszel bounds using a gamma range 1-10

mhbounds normal, gamma(1 (1) 10)
```

Table A16 – Sensitivity analysis (Alternative specifications for Home vs. OU)

	Covariate adjustment			
Ceiling ratio	BMI	Parity	Maternal Age / BMI	BMI / IMD / Maternal Age
£20,000	3,873 (231)	5,566 (224)	3,633 (228)	3,620 (229)
£50,000	9,191 (571)	13,084 (554)	8,630 (564)	8,603 (567)
£100,000	18,054 (1,139)	25,614 (1,102)	16,957 (1,124)	16,909 (1,131)
	Propensity score adjustment			
Ceiling ratio	BMI	Parity	Maternal Age / BMI	BMI / IMD / Maternal Age
£20,000	3,873 (234)	5,561 (225)	3,642 (230)	3,629 (231)
£50,000	9,197 (578)	13,073 (555)	8,656 (569)	8,629 (572)
£100,000	18,071 (1,152)	25,594 (1,105)	17,012 (1,134)	16,962 (1,140)

Notes: Table presents incremental net monetary benefit estimates and associated robust standard errors in parentheses. BMI: body mass index; IMD: index of multiple deprivation.

Table A17 – Sensitivity analysis (Alternative specifications for FMU vs. OU)

	Covariate adjustment (SE)			
Ceiling ratio	BMI	Parity	Maternal Age / BMI	BMI / IMD / Maternal Age
£20,000	5,361 (172)	5,888 (174)	5,390 (172)	5,362 (174)
£50,000	13,222 (422)	14,434 (425)	13,290 (422)	13,231 (425)
£100,000	26,324 (838)	28,678 (844)	26,457 (839)	26,347 (845)
	Propensity score adjustment (SE)			
Ceiling ratio	BMI	Parity	Maternal Age / BMI	BMI / IMD / Maternal Age
£20,000	5,362 (177)	5,888 (175)	5,391 (177)	5,361 (178)
£50,000	13,227 (431)	14,435 (426)	13,293 (431)	13,229 (434)
£100,000	26,335 (856)	28,679 (846)	26,463 (856)	26,343 (861)

Notes: Table presents incremental net monetary benefit estimates and associated robust standard errors in parentheses. BMI: body mass index; IMD: index of multiple deprivation.

Table A18 – Sensitivity analysis (Alternative specifications for AMU vs. OU)

	Covariate adjustment (SE)			
Ceiling ratio	BMI	Parity	Maternal Age / BMI	BMI / IMD / Maternal Age
£20,000	4,217 (167)	4,410 (171)	4,208 (167)	4,228 (167)
£50,000	10,358 (409)	10,800 (418)	10,337 (410)	10,380 (409)
£100,000	20,593 (814)	21,451 (830)	20,551 (815)	20,633 (814)
	Propensity score adjustment (SE)			
Ceiling ratio	BMI	Parity	Maternal Age / BMI	BMI / IMD / Maternal Age
£20,000	4,218 (171)	4,411 (171)	4,209 (171)	4,228 (171)
£50,000	10,361 (420)	10,803 (419)	10,337 (419)	10,380 (419)
£100,000	20,599 (834)	21,455 (832)	20,552 (833)	20,633 (832)

Notes: Table presents incremental net monetary benefit estimates and associated robust standard errors in parentheses. BMI: body mass index; IMD: index of multiple deprivation.

Table A19 – Sensitivity analysis (Alternative specifications for SUR)

	Confounders (Home vs. OU)			
Equation	BMI	Parity	Maternal Age / BMI	BMI / IMD / Maternal Age
Cost	-328 (14.47)	-554 (14.64)	-302 (14.44)	-298 (14.43)
NMB	3,882 (231)	4,846 (234)	3,774 (231)	3,754 (231)
Effect	0.18 (0.01)	0.25 (0.01)	0.17 (0.01)	0.17 (0.01)
NMB	3,870 (231)	5,356 (223)	3,653 (228)	3,644 (229)

	Confounders (FMU vs. OU)			
Equation	BMI	Parity	Maternal Age / BMI	BMI / IMD / Maternal Age
Cost	-120 (16.27)	-191 (17.24)	-123 (16.33)	-116 (16.24)
NMB	5,358 (173)	5,652 (173)	5,373 (173)	5,341 (173)
Effect	0.26 (0.01)	0.29 (0.01)	0.26 (0.01)	0.26 (0.01)
NMB	5,359 (172)	5,825 (173)	5,385 (172)	5,364 (174)

Table A19 – Sensitivity analysis (Alternative specifications for SUR) (continued)

	Confounders (AMU vs. OU)			
Equation	BMI	Parity	Maternal Age / BMI	BMI / IMD / Maternal Age
Cost	-123 (14.20)	-150 (14.98)	-123 (14.22)	-127 (14.24)
NMB	4,220 (166)	4,331 (168)	4,219 (166)	4,237 (166)
Effect	0.21 (0.01)	0.21 (0.01)	0.20 (0.01)	0.21 (0.01)
NMB	4,215 (166)	4,385 (169)	4,206 (167)	4,223 (166)

Notes: Table presents incremental estimates and associated robust standard errors in parentheses when potential confounders are omitted from one equation of the seemingly unrelated regression (SUR) model. BMI: body mass index; IMD: index of multiple deprivation; NMB: net monetary benefit for £20,000.

Table A20 – Sensitivity analysis (Bias-corrected matching estimates for Home vs. OU)

	0.1 SDprobit(PS)	0.6 SDprobit(PS)	0.005	0.01
Cost	-322	-324	-314	-326
CI	-353 – -290	-353 – -296	-345 – -284	-358 – -295
Effect	0.18	0.18	0.19	0.18
CI	0.15 – 0.20	0.16 – 0.20	0.16 – 0.21	0.16 – 0.21
NMB	3,895	3,887	4,057	3,957
CI	3,395– 4,396	3,434 – 4,340	3,561– 4,554	3,455– 4,459
	0.02	0.03	0.05	0.1
Cost	-325	-325	-318	-325
CI	-357 – -294	-356 – -294	-349 – -288	-354 – -295
Effect	0.18	0.19	0.18	0.18
CI	0.15 – 0.20	0.16 – 0.21	0.16 – 0.21	0.15 – 0.20
NMB	3,881	4,020	3,987	3,870
CI	3,377 – 4,384	3,528 – 4,511	3,506 – 4,468	3,403 – 4,337

Notes: Table presents incremental estimates and associated 95% confidence intervals for varying caliper widths. SD: standard deviation; PS: propensity score; NMB: net monetary benefit; CI: confidence interval.

Table A21 – Sensitivity analysis (Bias-corrected matching estimates for FMU vs. OU)

	0.1 SDprobit(PS)	0.6 SDprobit(PS)	0.005	0.01
Cost	-112	-112	-112	-112
CI	-146 – -77	-146 – -77	-147 – -77	-146 – -77
Effect	0.26	0.26	0.26	0.26
CI	0.24 – 0.28	0.24 – 0.28	0.24 – 0.28	0.24 – 0.28
NMB	5,329	5,329	5,357	5,334
CI	4,933 – 5,725	4,933 – 5,725	4,958 – 5,756	4,937 – 5,730
	0.02	0.03	0.05	0.1
Cost	-112	-112	-112	-112
CI	-146 – -77	-146 – -77	-146 – -77	-146 – -77
Effect	0.26	0.26	0.26	0.26
CI	0.24 – 0.28	0.24 – 0.28	0.24 – 0.28	0.24 – 0.28
NMB	5,329	5,329	5,329	5,329
CI	4,933 – 5,725	4,933 – 5,725	4,933 – 5,725	4,933 – 5,725

Notes: Table presents incremental estimates and associated 95% confidence intervals for varying caliper widths. SD: standard deviation; PS: propensity score; NMB: net monetary benefit; CI: confidence interval.

Table A22 – Sensitivity analysis (Bias-corrected matching estimates for AMU vs. OU)

	0.1 SDprobit(PS)	0.6 SDprobit(PS)	0.005	0.01
Cost	-118	-117	-116	-117
CI	-146 – -89	-145 – -89	-145 – -88	-145 – -89
Effect	0.21	0.21	0.21	0.21
CI	0.19 – 0.22	0.19 – 0.22	0.19 – 0.22	0.19 – 0.22
NMB	4,232	4,213	4,234	4,224
CI	3,890– 4,574	3,874 – 4,554	3,891 – 4,577	3,884 – 4,564
	0.02	0.03	0.05	0.1
Cost	-117	-117	-117	-119
CI	-145 – -89	-145 – -89	-145 – -89	-145 – -89
Effect	0.21	0.21	0.21	0.21
CI	0.19 – 0.22	0.19 – 0.22	0.19 – 0.22	0.19 – 0.22
NMB	4,214	4,214	4,214	4,214
CI	3,874 – 4,554	3,874 – 4,554	3,874 – 4,554	3,874 – 4,554

Notes: Table presents incremental estimates and associated 95% confidence intervals for varying caliper widths. SD: standard deviation; PS: propensity score; NMB: net monetary benefit; CI: confidence interval.

Table A23 – Sensitivity Analysis (Functional form)

	Cost	CI
Home vs. Obstetric unit	-315	-339 – -292
Freestanding Midwifery Unit vs. Obstetric Unit	-130	-158 – -102
Alongside Midwifery Unit vs. Obstetric Unit	-131	-155 – -106

Notes: Table presents incremental estimates and associated 95% confidence intervals from the generalised linear model, which used a gamma distribution and identity link function. CI: confidence interval.

	Effectiveness	CI
Home vs. Obstetric unit	0.21	0.20 – 0.22
Freestanding Midwifery Unit vs. Obstetric Unit	0.23	0.22 – 0.24
Alongside Midwifery Unit vs. Obstetric Unit	0.17	0.16 – 0.18

Notes: Table presents average marginal effects and associated 95% confidence intervals from the separate probit model. CI: confidence interval.

Table A24 – Summary of the balance achieved in Home vs. OU

Matching Method	N_0 Before	N_1 Before	Pseudo R^2 Before	Pseudo R^2 After	$p>\chi^2$ After	Mean Difference (%) Before	Mean Difference (%) After	Lost to Common Support (%) After
Coarsened Exact	18,847	16,187	0.147	0.000	1.000	14.5	2.9	10.10
Entropy Balancing	18,847	16,187	0.147	0.000	1.000	14.5	0.0	0
Inverse Probability Weighting	18,847	16,187	0.147	0.000	1.000	14.5	0.4	0
Nearest Neighbour	18,847	16,187	0.147	0.090	0.000	14.5	9.2	0
Nearest Neighbour (R)	18,847	16,187	0.147	0.002	0.000	14.5	1.1	0
Ratio	18,847	16,187	0.147	0.003	0.000	14.5	1.4	0
Caliper	18,847	16,187	0.147	0.020	0.000	14.5	3.6	24.32

Table A24 – Summary of the balance achieved in Home vs. OU (continued)

Matching Method	N_0 Before	N_1 Before	Pseudo R^2 Before	Pseudo R^2 After	$p>\chi^2$ After	Mean Difference (%) Before	Mean Difference (%) After	Lost to Common Support (%) After
Radius	18,847	16,187	0.147	0.001	0.067	14.5	1.0	0
Kernel (Tricube)	18,847	16,187	0.147	0.001	0.684	14.5	0.7	0
Kernel (Gaussian)	18,847	16,187	0.147	0.003	0.000	14.5	1.6	0
Mahalanobis	18,847	16,187	0.147	0.000	0.991	14.5	0.4	0
Mahalanobis + PS	18,847	16,187	0.147	0.000	0.999	14.5	0.4	0

Notes: (R) = Replacement; PS = propensity score; OU = Obstetric Unit

Table A25 – Summary of the balance achieved in FMU vs. OU

Matching Method	N_0 Before	N_1 Before	Pseudo R^2 Before	Pseudo R^2 After	$p>\chi^2$ After	Mean Difference (%) Before	Mean Difference (%) After	Lost to Common Support After (%)
Coarsened Exact	18,847	10,971	0.040	0.000	1.000	8.1	2.7	9.85
Entropy Balancing	18,847	10,971	0.040	0.000	1.000	8.1	0.1	0
Inverse Probability Weighting	18,847	10,971	0.040	0.000	1.000	8.1	0.2	0
Nearest Neighbour	18,847	10,971	0.040	0.001	0.466	8.1	0.9	0
Nearest Neighbour (R)	18,847	10,971	0.040	0.001	0.620	8.1	0.7	0
Ratio	18,847	10,971	0.040	0.002	0.009	8.1	1.0	0
Caliper	18,847	10,971	0.040	0.001	0.466	8.1	0.9	0

Table A25 – Summary of the balance achieved in FMU vs. OU (continued)

Matching Method	N_0 Before	N_1 Before	Pseudo R^2 Before	Pseudo R^2 After	$p>\chi^2$ After	Mean Difference (%) Before	Mean Difference (%) After	Lost to Common Support After (%)
Radius	18,847	10,971	0.040	0.000	1.000	8.1	0.4	0
Kernel (Tricube)	18,847	10,971	0.040	0.000	1.000	8.1	0.5	0
Kernel (Gaussian)	18,847	10,971	0.040	0.003	0.000	8.1	1.7	0
Mahalanobis	18,847	10,971	0.040	0.000	1.000	8.1	0.3	0
Mahalanobis + PS	18,847	10,971	0.040	0.000	1.000	8.1	0.3	0

Notes: (R) = Replacement; PS = propensity score; FMU = Freestanding Midwifery Unit; OU = Obstetric Unit

Table A26 – Summary of the balance achieved in AMU vs. OU

Matching Method	N_0 Before	N_1 Before	Pseudo R^2 Before	Pseudo R^2 After	$p>\chi^2$ After	Mean Difference (%) Before	Mean Difference (%) After	Lost to Common Support After (%)
Coarsened Exact	18847	16031	0.013	0.000	1.000	4.3	4.6	17.03
Entropy Balancing	18847	16031	0.013	0.000	1.000	4.3	0.1	0
Inverse Probability Weighting	18847	16031	0.013	0.000	1.000	4.3	0.1	0
Nearest Neighbour	18847	16031	0.013	0.001	0.001	4.3	1.4	0
Nearest Neighbour (R)	18847	16031	0.013	0.002	0.000	4.3	1.5	0
Ratio	18847	16031	0.013	0.004	0.000	4.3	1.9	0
Caliper	18847	16031	0.013	0.001	0.001	4.3	1.4	0

Table A26 – Summary of the balance achieved in AMU vs. OU (continued)

Matching Method	N_0 Before	N_1 Before	Pseudo R^2 Before	Pseudo R^2 After	$p>\chi^2$ After	Mean Difference (%) Before	Mean Difference (%) After	Lost to Common Support After (%)
Radius	18847	16031	0.013	0.000	1.000	4.3	0.2	0
Kernel Tricube	18847	16031	0.013	0.000	0.886	4.3	0.7	0
Kernel Gaussian	18847	16031	0.013	0.004	0.000	4.3	2.1	0
Mahalanobis	18847	16031	0.013	0.001	0.511	4.3	0.6	0
Mahalanobis + PS	18847	16031	0.013	0.001	0.519	4.3	0.6	0

Notes: (R) = Replacement; PS = propensity score; AMU = Alongside Midwifery Unit; OU = Obstetric Unit

Table A27 – Summary of the balance achieved in sensitivity analysis (Home vs. OU)

Caliper Width	N_0 Before	N_1 Before	Pseudo R^2 Before	Pseudo R^2 After	$p>\chi^2$ After	Mean Difference (%) Before	Mean Difference (%) After	Lost to Common Support (%) After
0.1 SD of Probit of PS	18,847	16,187	0.147	0.011	0.000	14.5	3.2	27.82
0.6 SD of Probit of PS	18,847	16,187	0.147	0.046	0.000	14.5	5.6	11.79
Fixed 0.005	18,847	16,187	0.147	0.007	0.000	14.5	2.9	29.68
Fixed 0.01	18,847	16,187	0.147	0.008	0.000	14.5	3.0	29.21
Fixed 0.02	18,847	16,187	0.147	0.009	0.000	14.5	3.0	27.99
Fixed 0.03	18,847	16,187	0.147	0.011	0.000	14.5	3.3	26.74
Fixed 0.05	18,847	16,187	0.147	0.019	0.000	14.5	3.8	24.70
Fixed 0.1	18,847	16,187	0.147	0.033	0.000	14.5	4.9	17.39

Notes: SD = standard deviation; PS = propensity score; OU = Obstetric Unit

Table A28 – Summary of the balance achieved in sensitivity analysis (FMU vs. OU)

Caliper Width	N_0 Before	N_1 Before	Pseudo R^2 Before	Pseudo R^2 After	$p>\chi^2$ After	Mean Difference (%) Before	Mean Difference (%) After	Lost to Common Support (%) After
0.1 SD of Probit of PS	18,847	10,971	0.040	0.001	0.466	8.1	0.9	0
0.6 SD of Probit of PS	18,847	10,971	0.040	0.001	0.466	8.1	0.9	0
Fixed 0.005	18,847	10,971	0.040	0.001	0.589	8.1	0.9	0.87
Fixed 0.01	18,847	10,971	0.040	0.001	0.457	8.1	0.9	0.16
Fixed 0.02	18,847	10,971	0.040	0.001	0.466	8.1	0.9	0
Fixed 0.03	18,847	10,971	0.040	0.001	0.466	8.1	0.9	0
Fixed 0.05	18,847	10,971	0.040	0.001	0.466	8.1	0.9	0
Fixed 0.1	18,847	10,971	0.040	0.001	0.466	8.1	0.9	0

Notes: SD = standard deviation; PS = propensity score; FMU = Freestanding Midwifery Unit; OU = Obstetric Unit

Table A29 – Summary of the balance achieved in sensitivity analysis (AMU vs. OU)

Caliper Width	N_0 Before	N_1 Before	Pseudo R^2 Before	Pseudo R^2 After	$p>\chi^2$ After	Mean Difference (%) Before	Mean Difference (%) After	Lost to Common Support (%) After
0.1 SD of Probit of PS	18,847	15,868	0.013	0.001	0.025	4.3	1.3	1.03
0.6 SD of Probit of PS	18,847	15,868	0.013	0.001	0.001	4.3	1.3	0
Fixed 0.005	18,847	15,868	0.013	0.001	0.430	4.3	1.0	2.40
Fixed 0.01	18,847	15,868	0.013	0.001	0.003	4.3	1.3	0.15
Fixed 0.02	18,847	15,868	0.013	0.001	0.001	4.3	1.4	0
Fixed 0.03	18,847	15,868	0.013	0.001	0.001	4.3	1.4	0
Fixed 0.05	18,847	15,868	0.013	0.001	0.001	4.3	1.4	0
Fixed 0.1	18,847	15,868	0.013	0.001	0.001	4.3	1.4	0

Notes: SD = standard deviation; PS = propensity score; AMU = Alongside Midwifery Unit; OU = Obstetric Unit

Figure A1 – Propensity score distributions for Home vs. Obstetric Unit

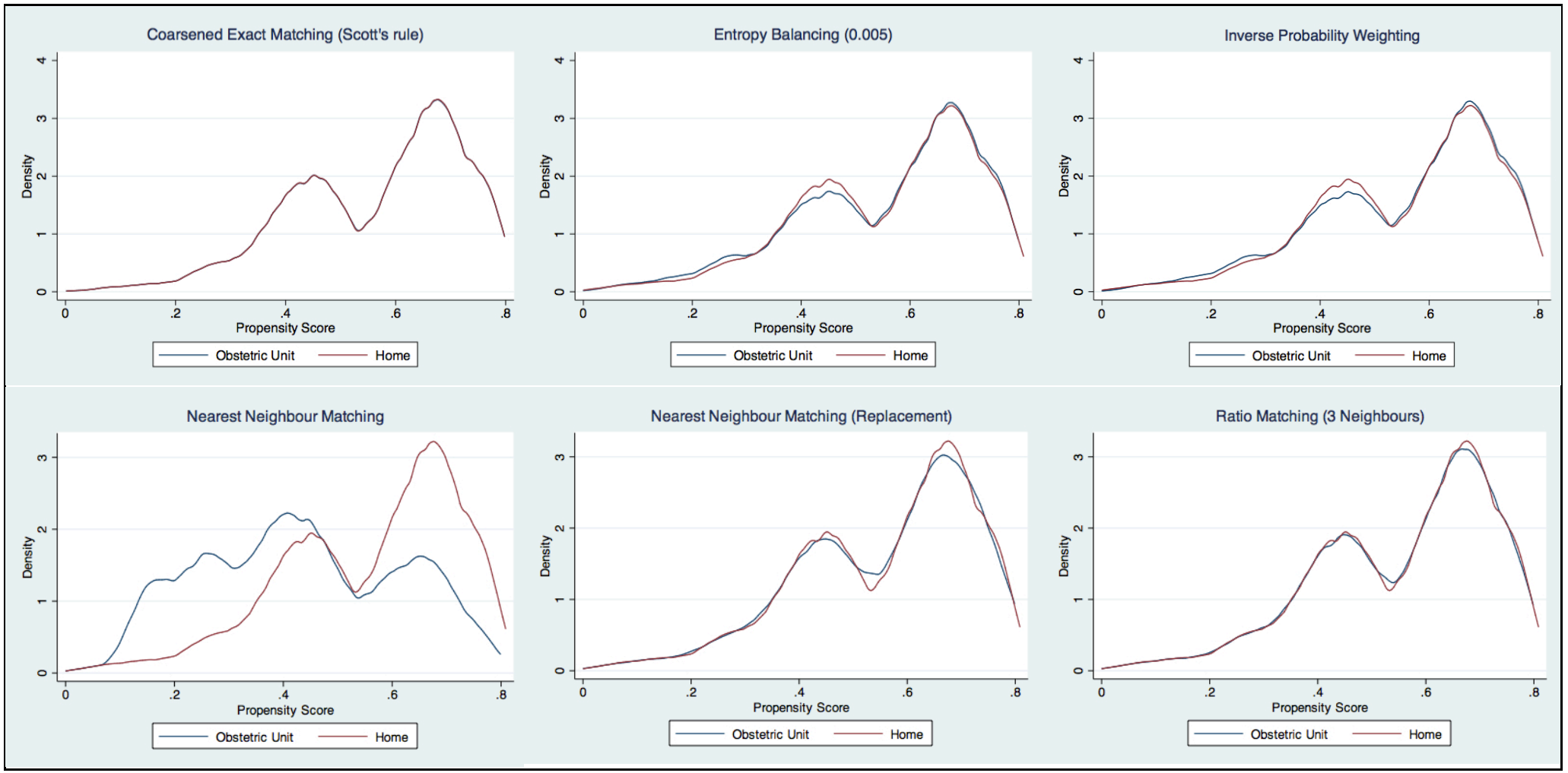


Figure A1 – Propensity score distributions for Home vs. Obstetric Unit (continued)

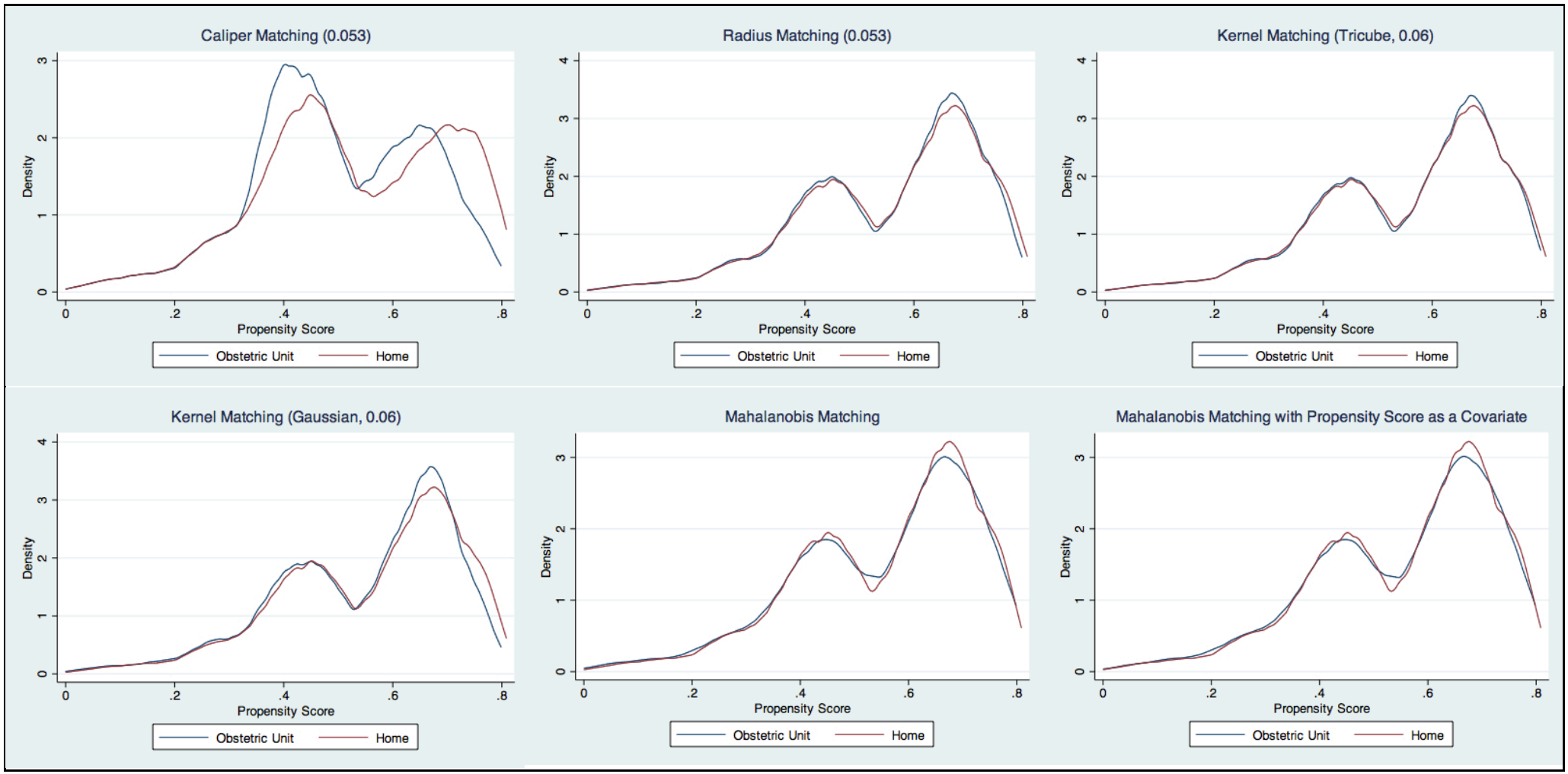


Figure A2 – Propensity score distributions for Freestanding Midwifery Unit vs. Obstetric Unit

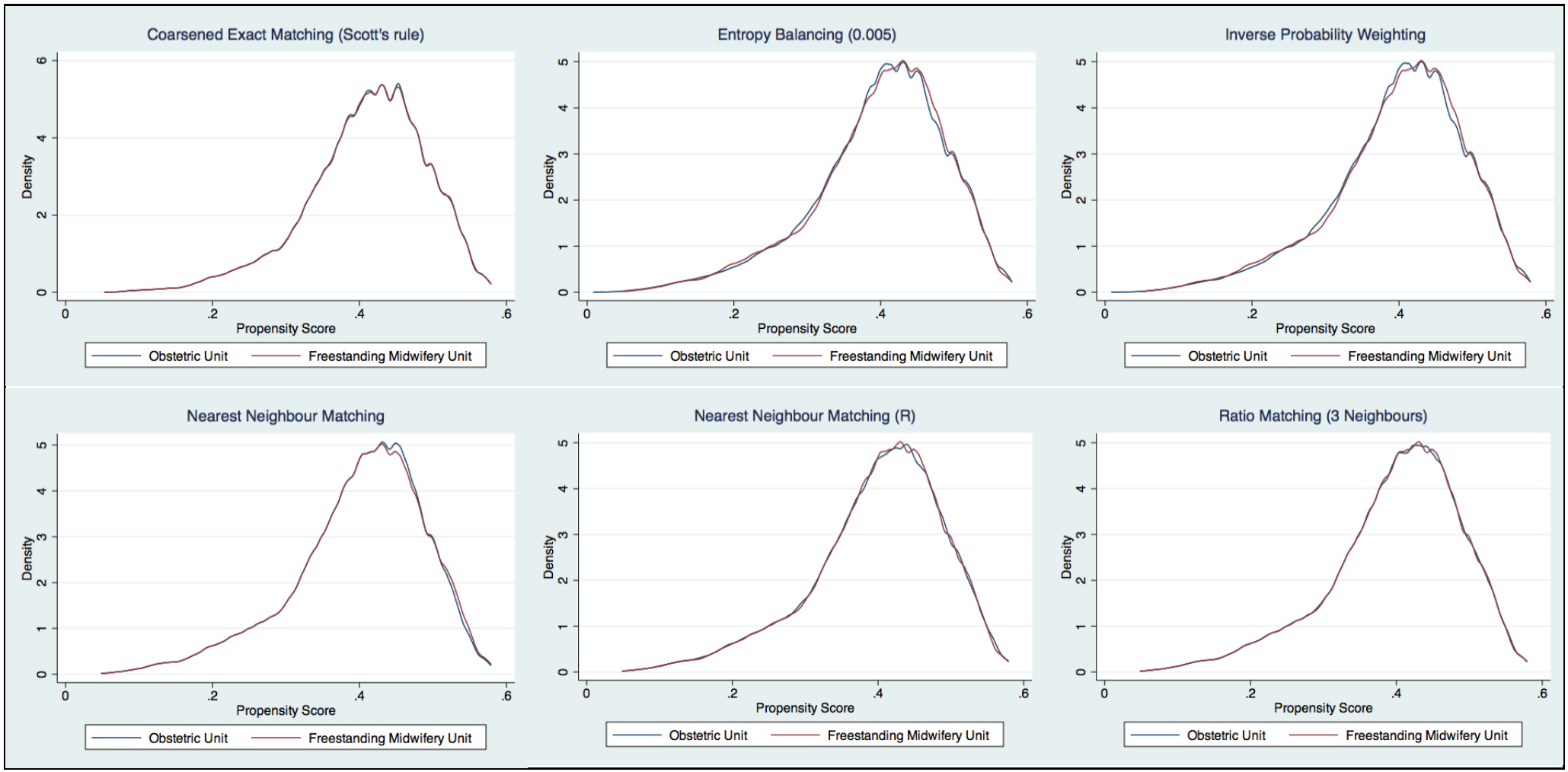


Figure A2 – Propensity score distributions for Freestanding Midwifery Unit vs. Obstetric Unit (continued)

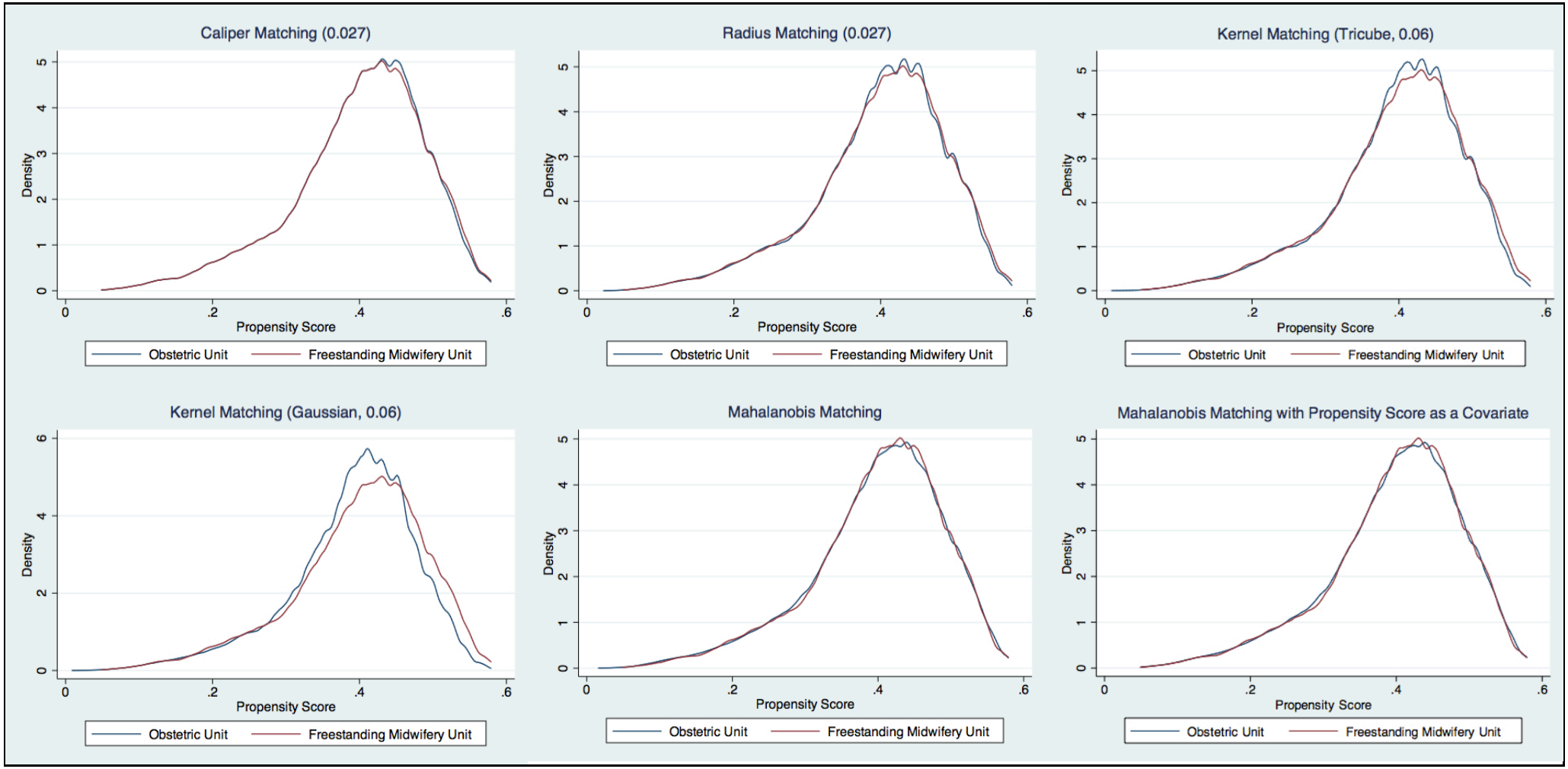


Figure A3 – Propensity score distributions for Alongside Midwifery Unit vs. Obstetric Unit

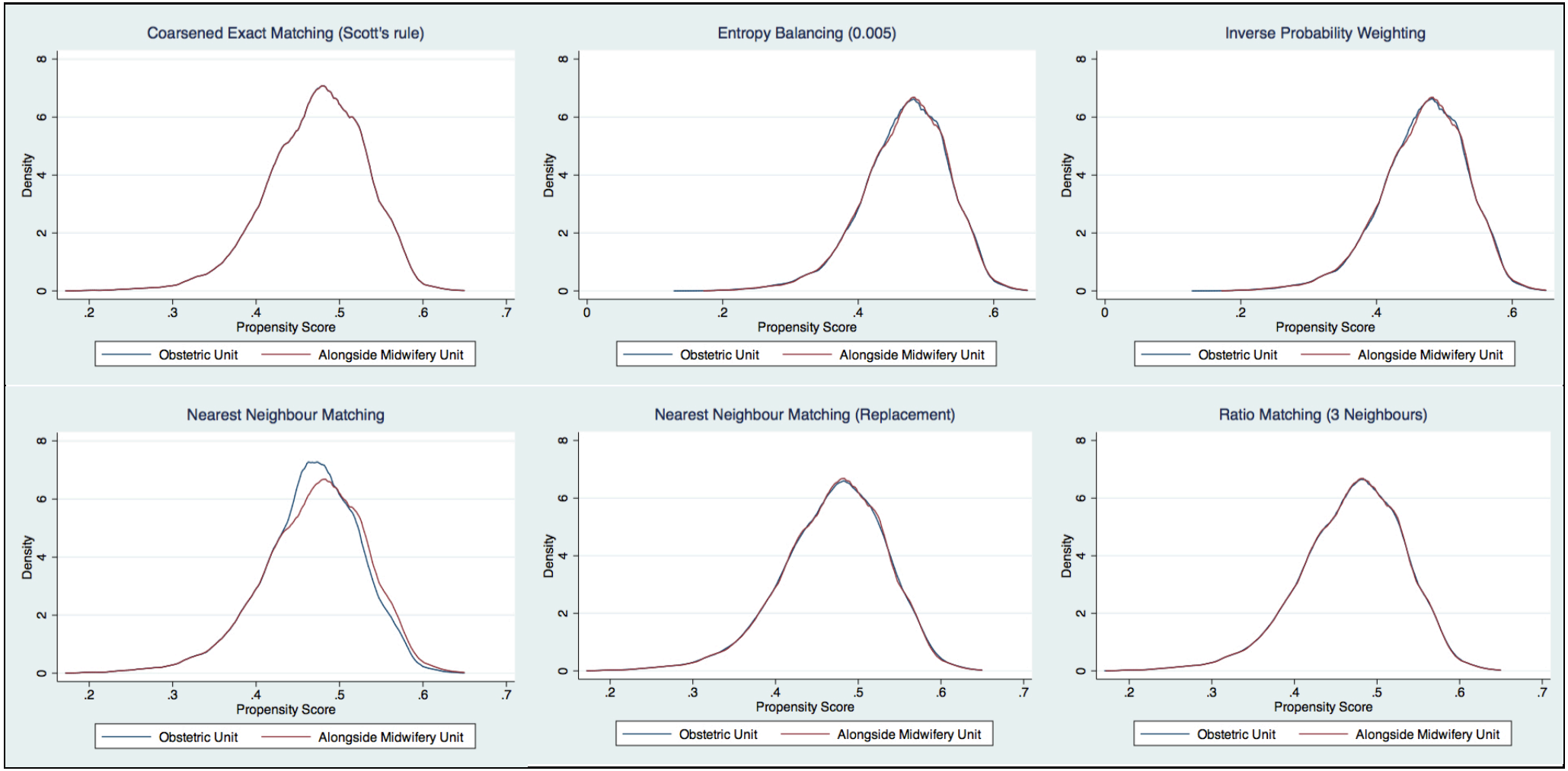


Figure A3 – Propensity score distributions for Alongside Midwifery Unit vs. Obstetric Unit (continued)

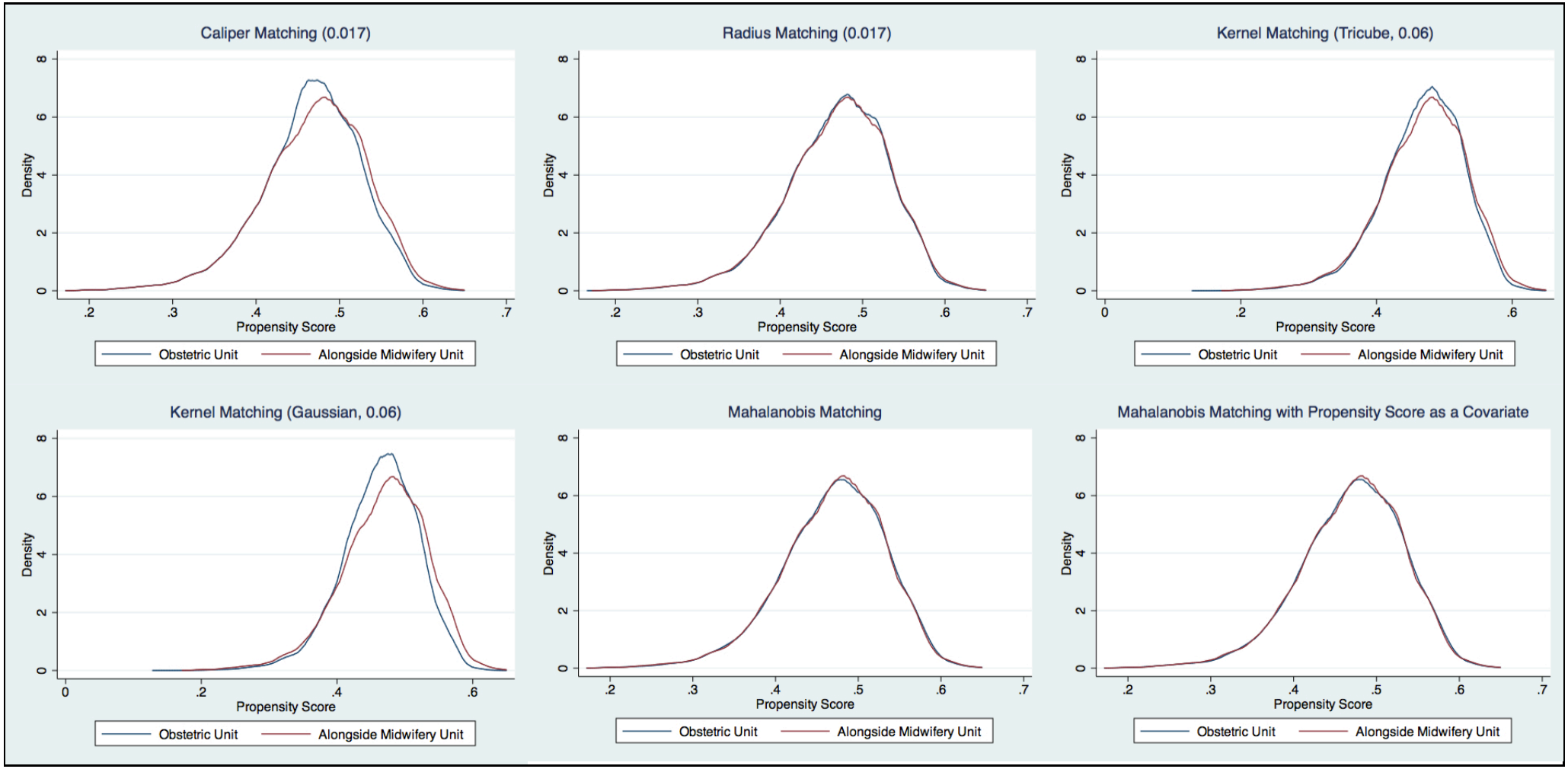


Figure A4 – Covariate balance for Home vs. Obstetric Unit

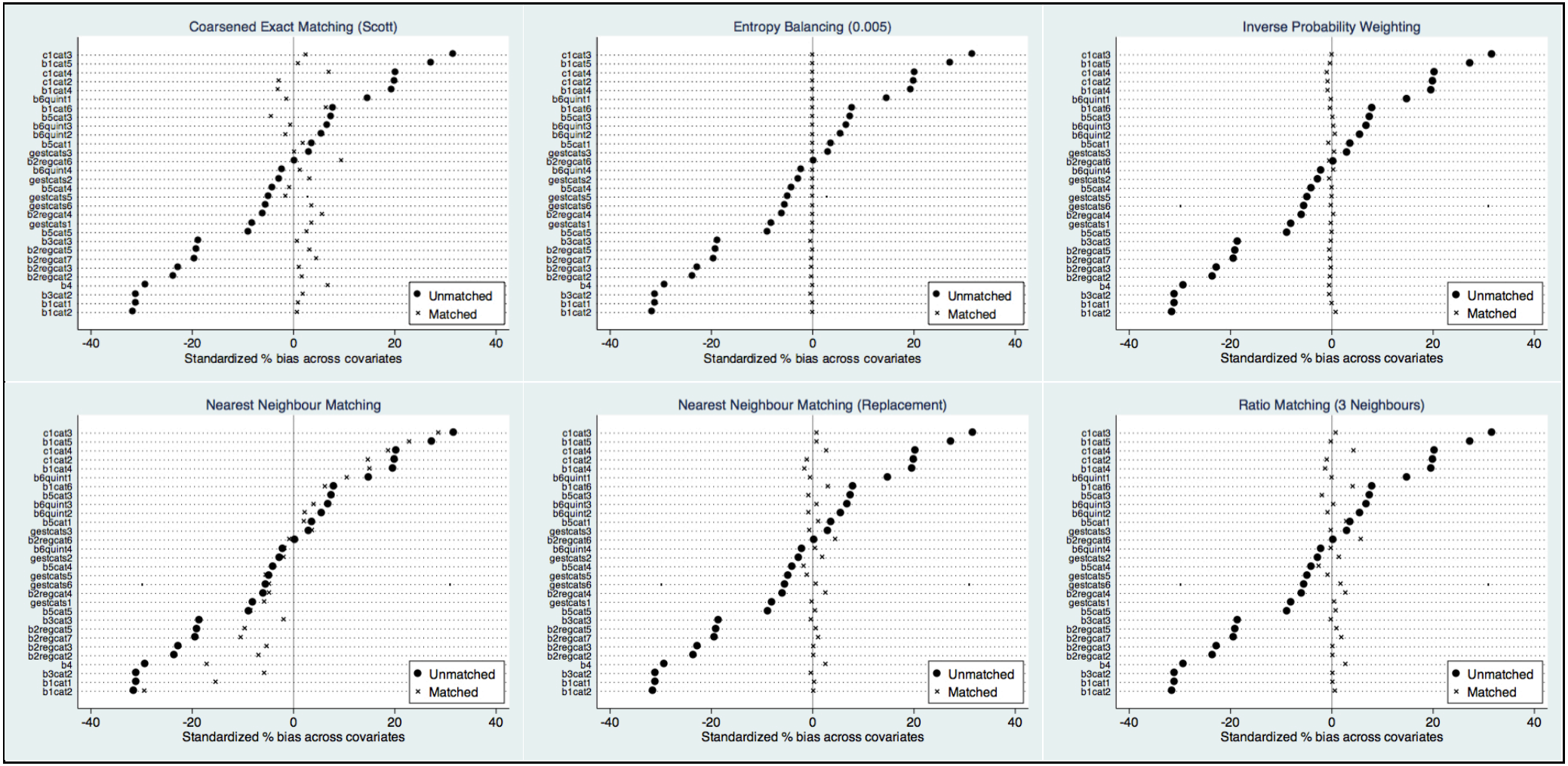


Figure A4 – Covariate balance for Home vs. Obstetric Unit (continued)

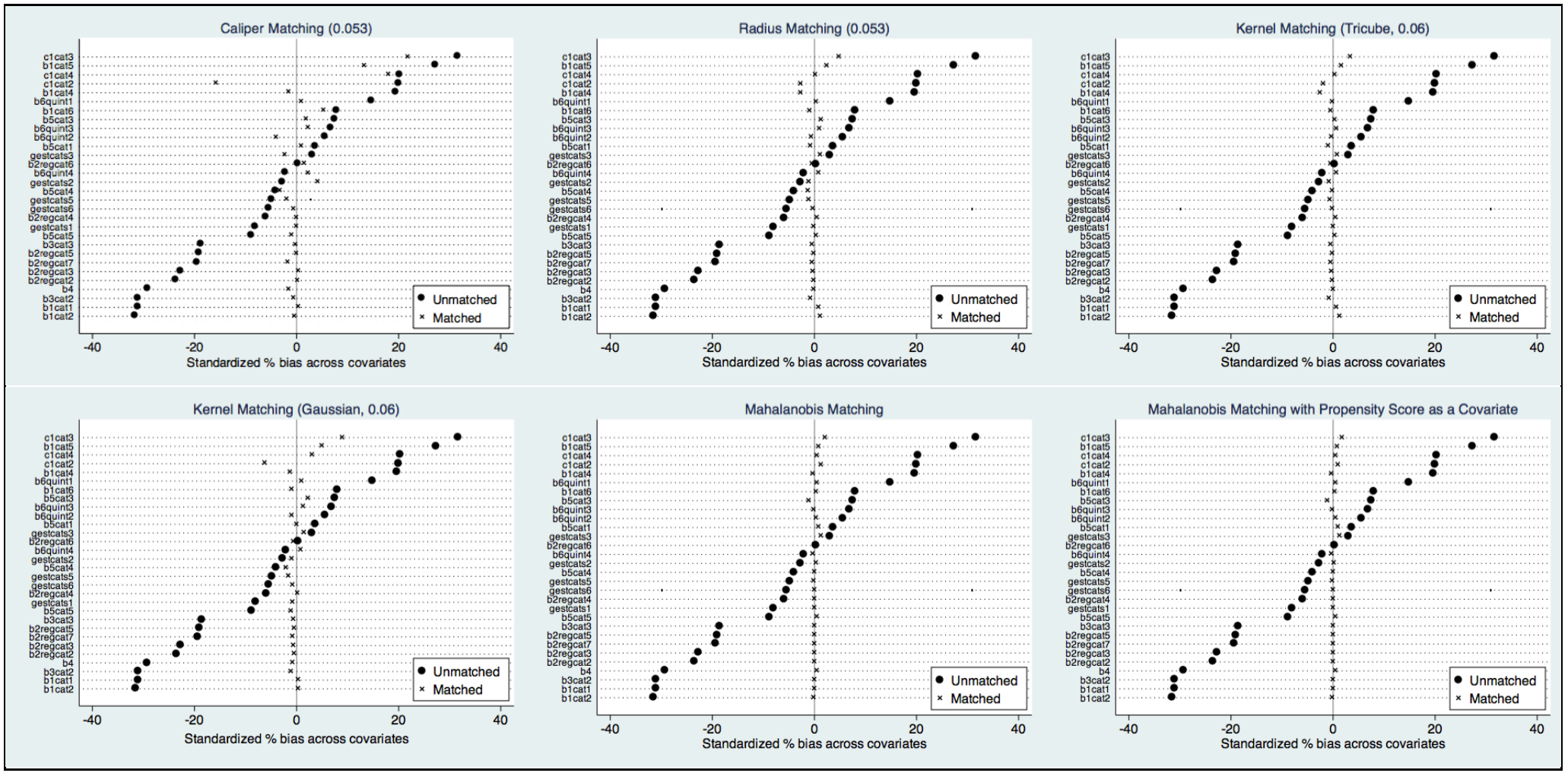


Figure A5 – Covariate balance for Freestanding Midwifery Unit vs. Obstetric Unit

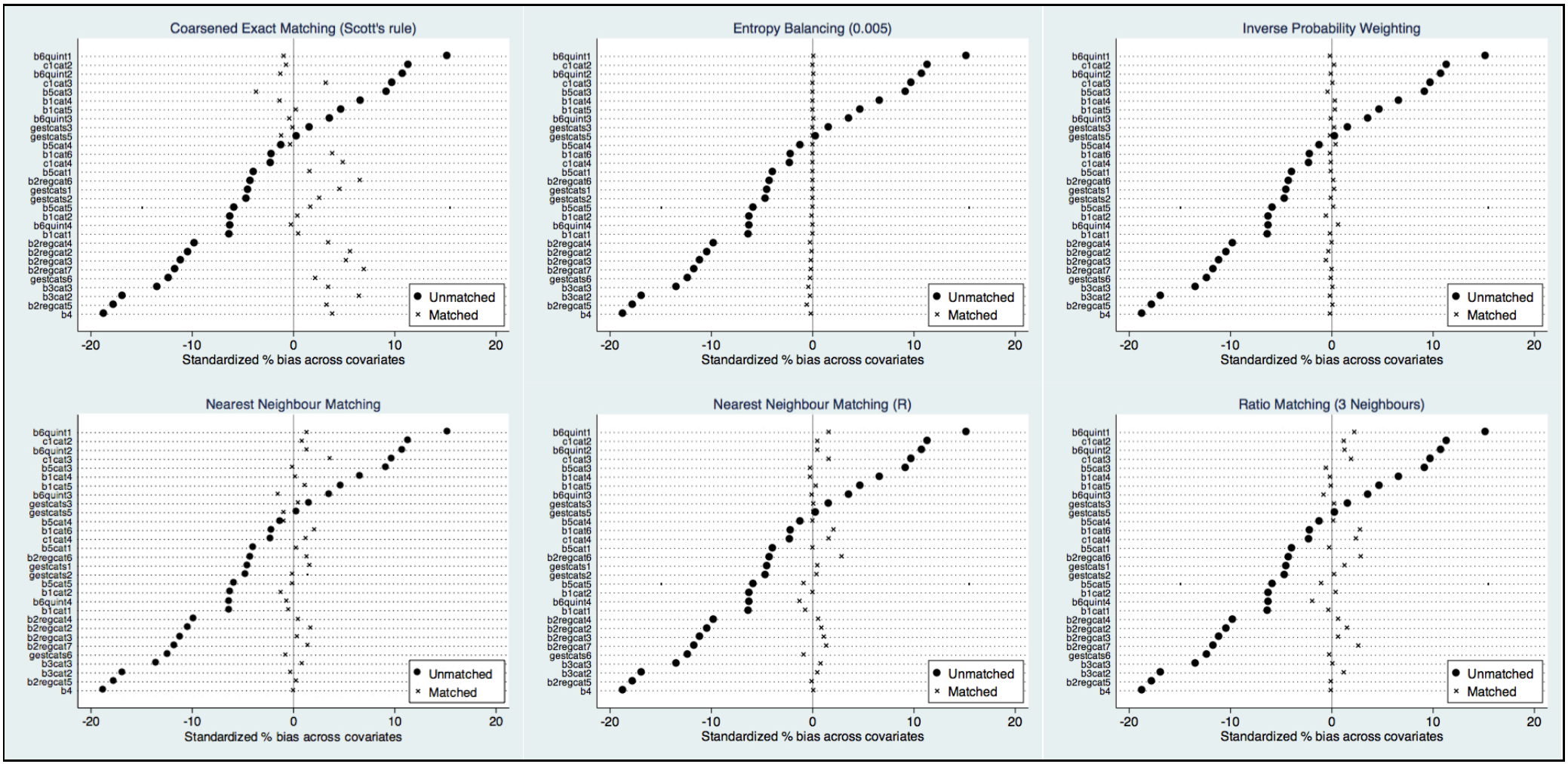


Figure A5 – Covariate balance for Freestanding Midwifery Unit vs. Obstetric Unit (continued)

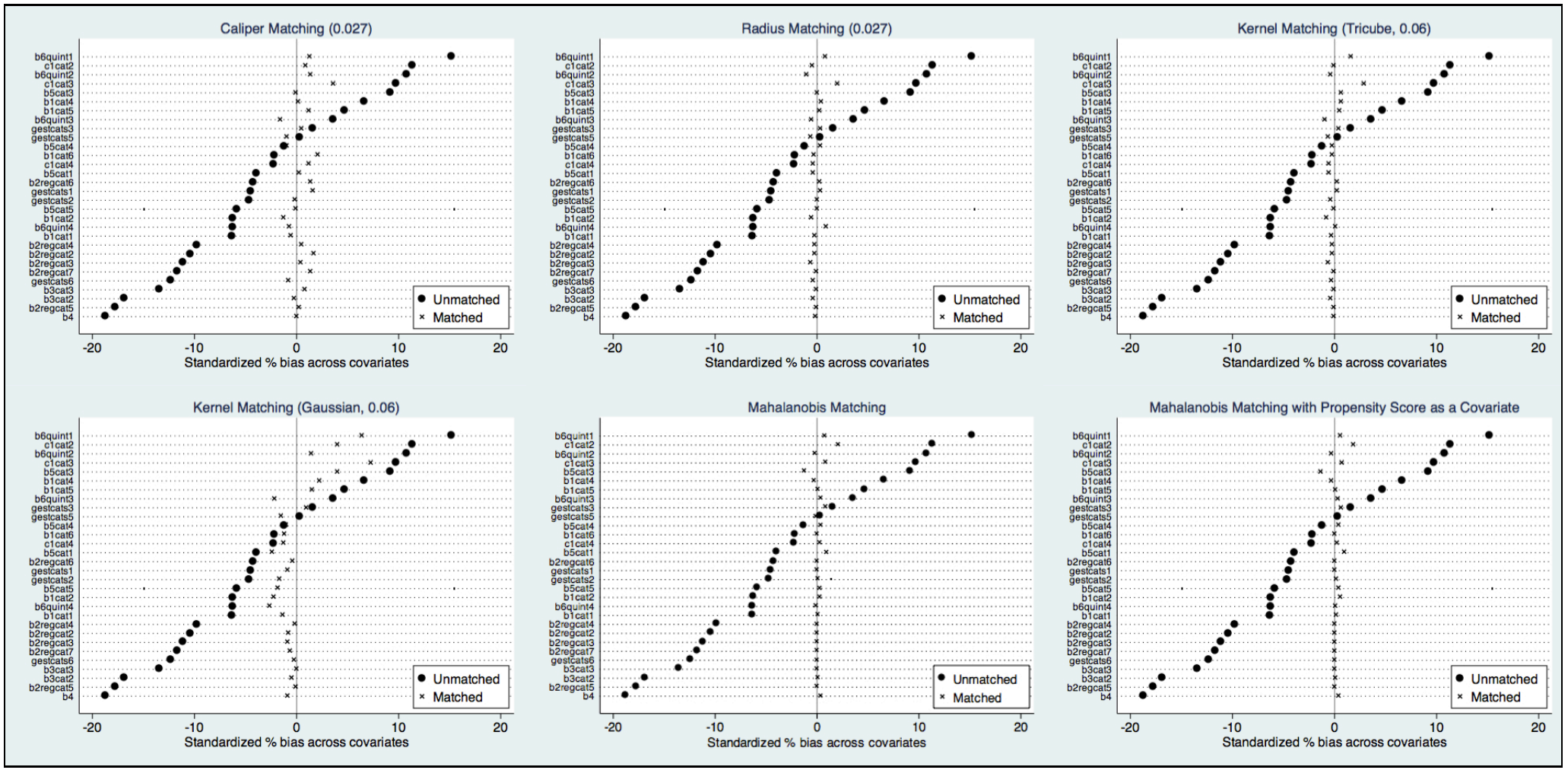


Figure A6 – Covariate balance for Alongside Midwifery Unit vs. Obstetric Unit

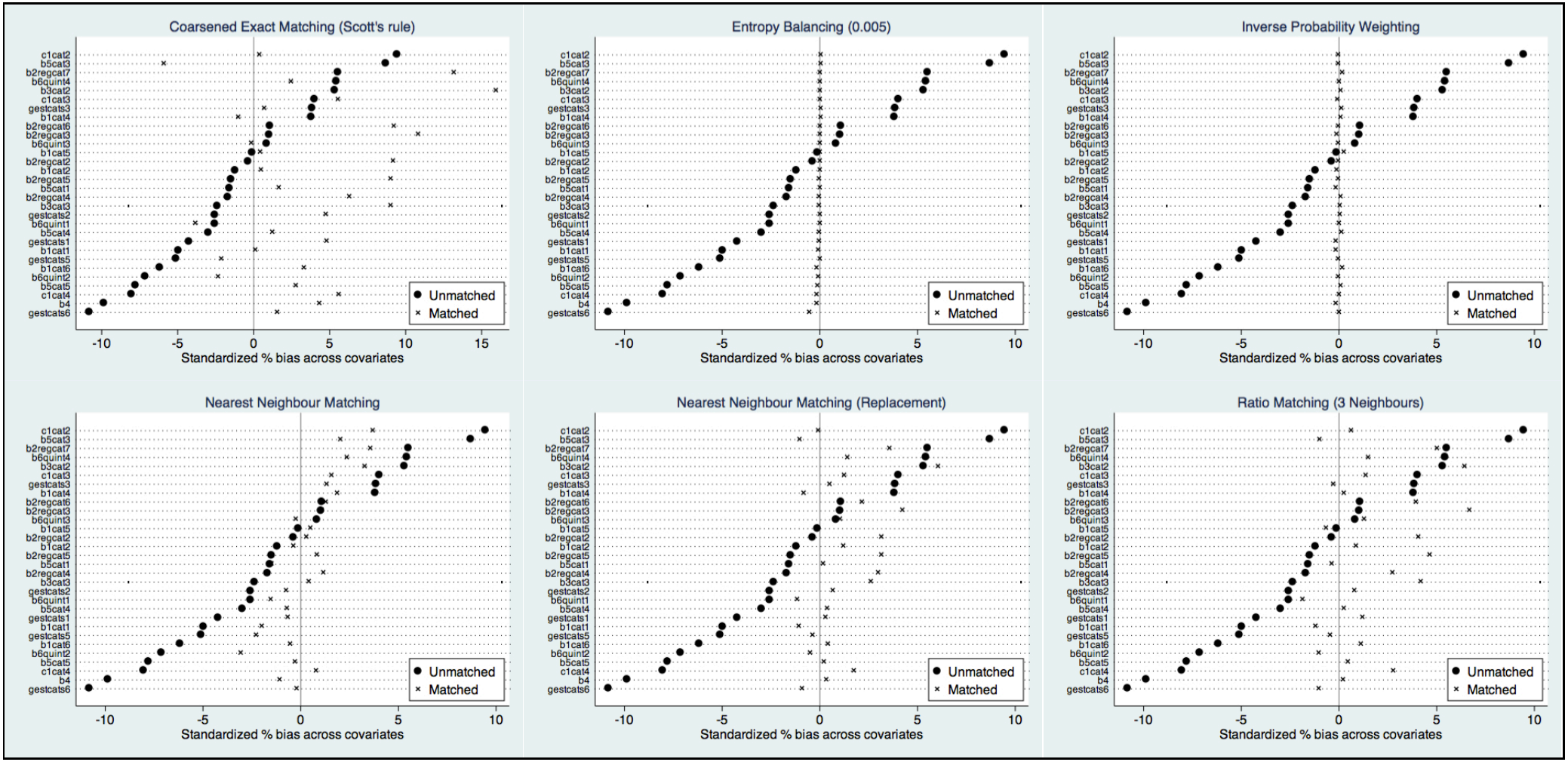


Figure A6 – Covariate balance for Alongside Midwifery Unit vs. Obstetric Unit (continued)

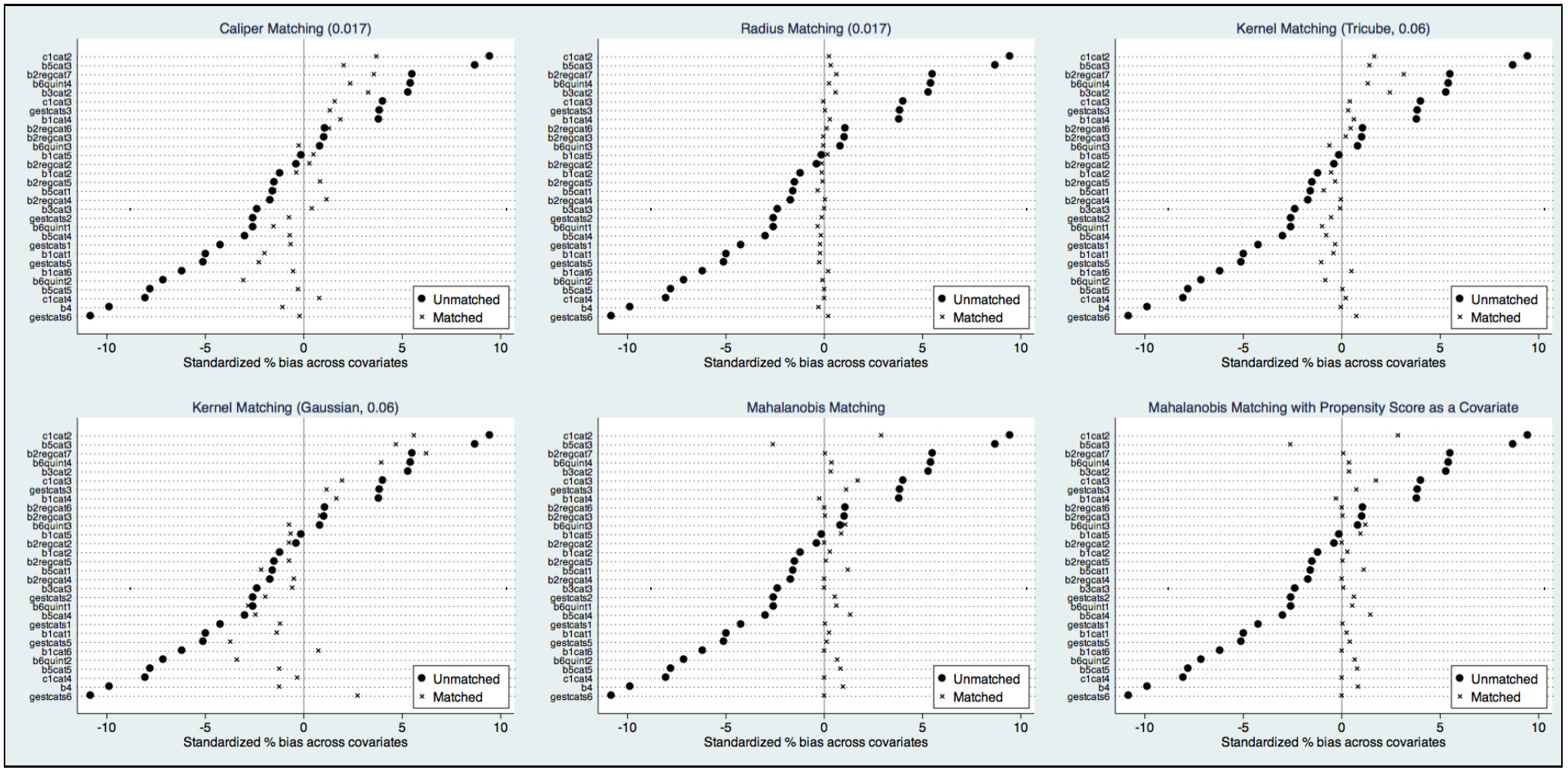


Figure A7 – Net benefit and 95% confidence limits for Home vs. Obstetric Unit

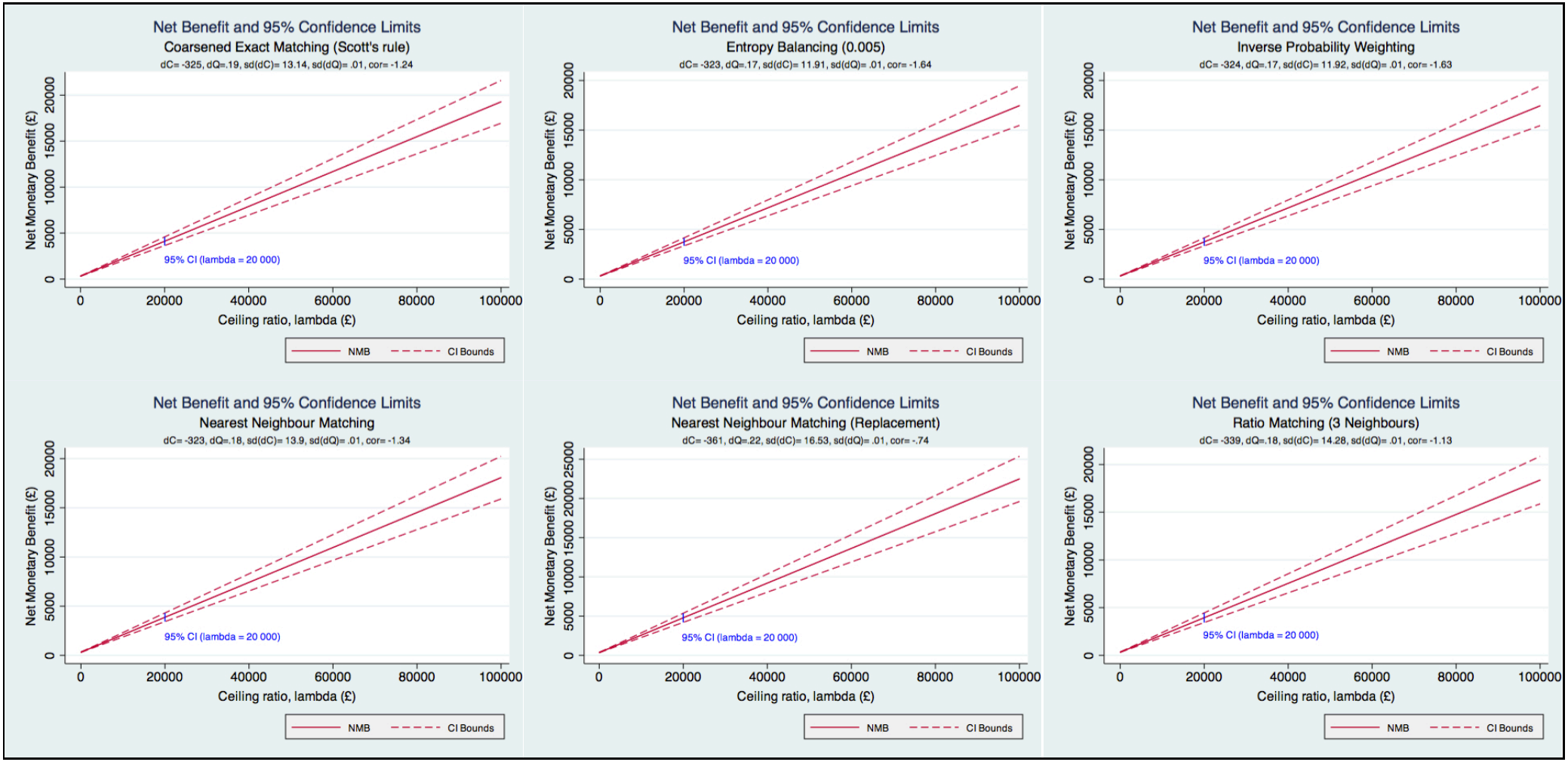


Figure A7 – Net benefit and 95% confidence limits for Home vs. Obstetric Unit (continued)

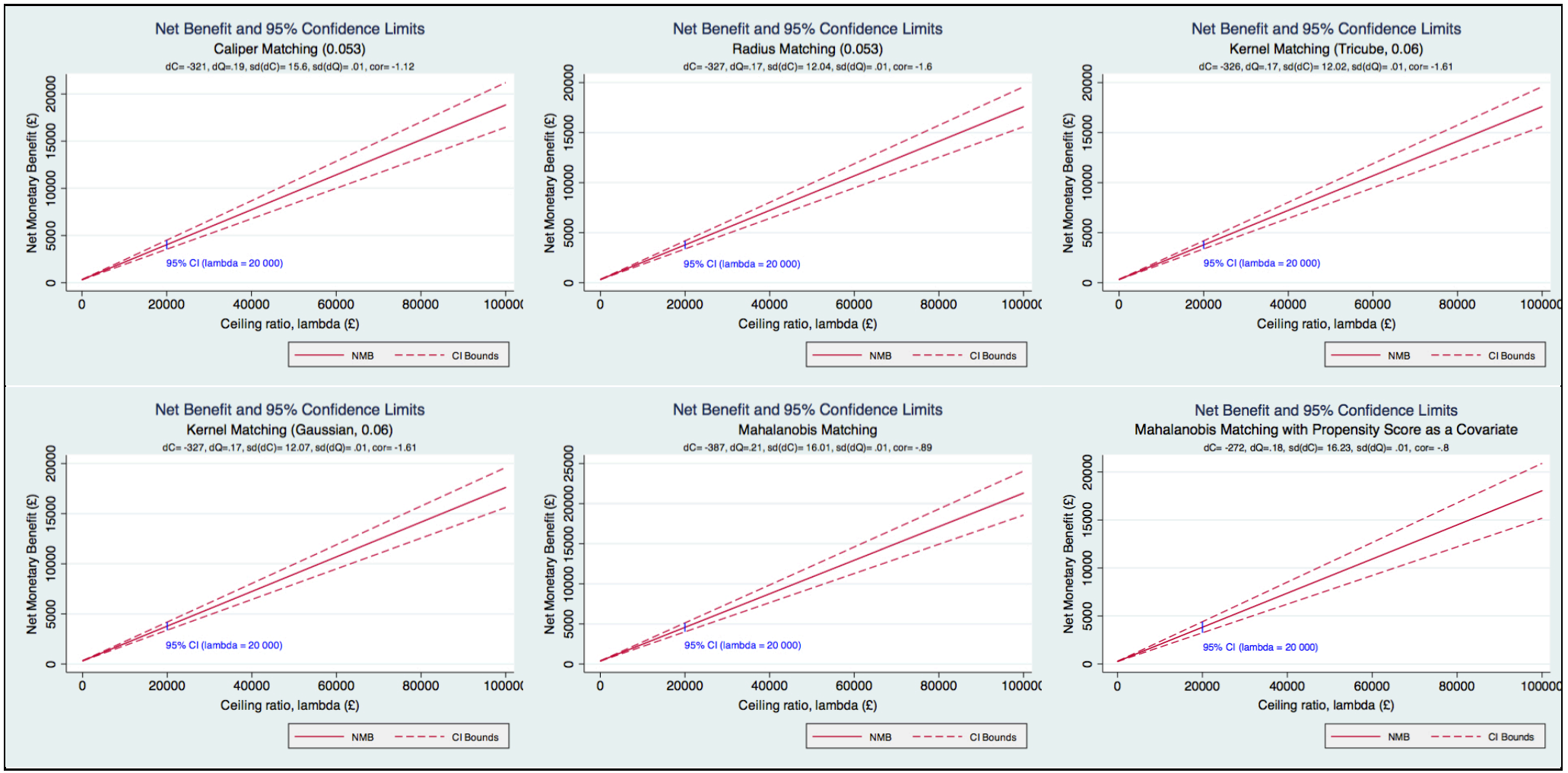


Figure A8 – Net benefit and 95% confidence limits for Freestanding Midwifery Unit vs. Obstetric Unit

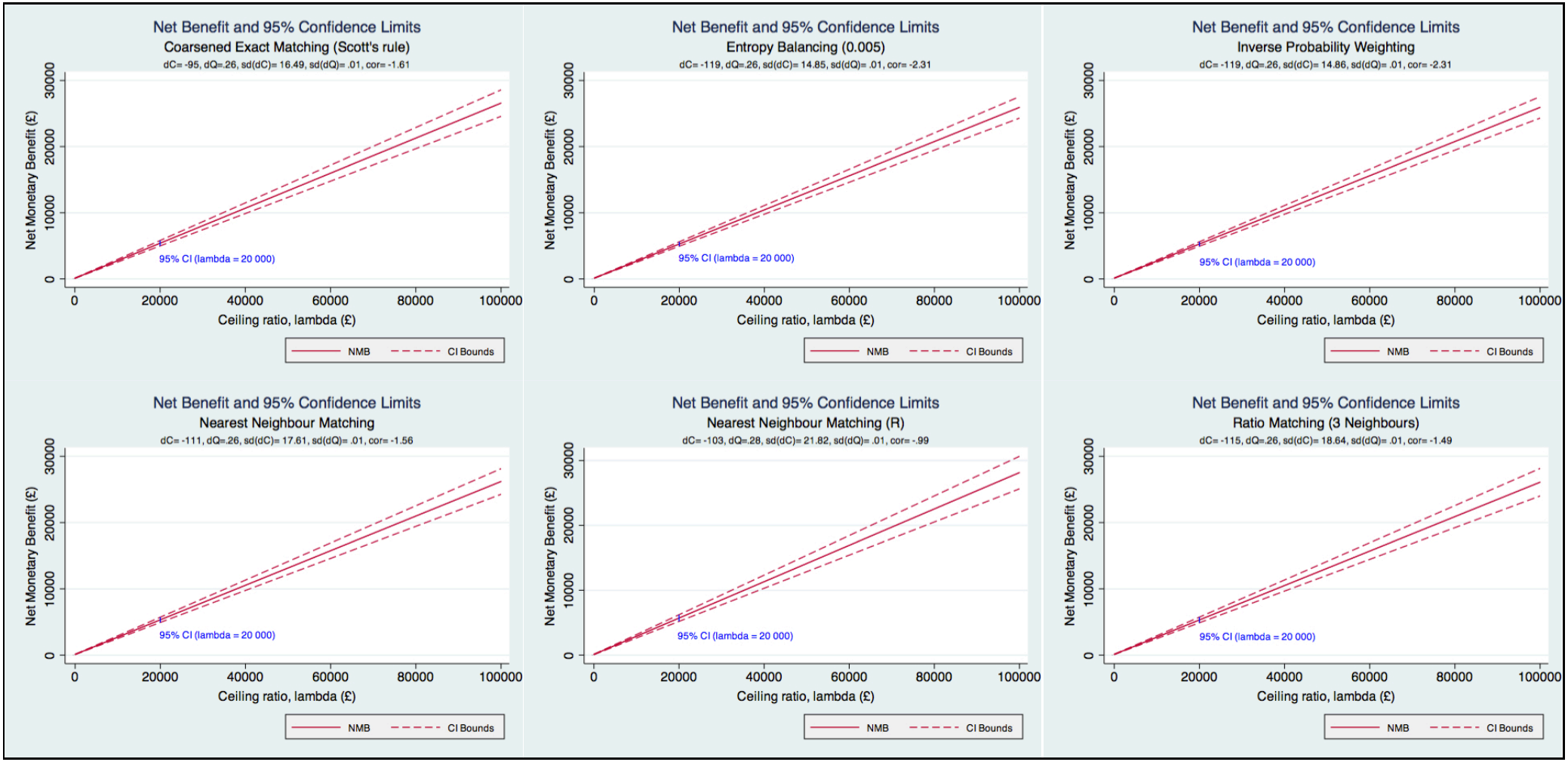


Figure A8 – Net benefit and 95% confidence limits for Freestanding Midwifery Unit vs. Obstetric Unit (continued)

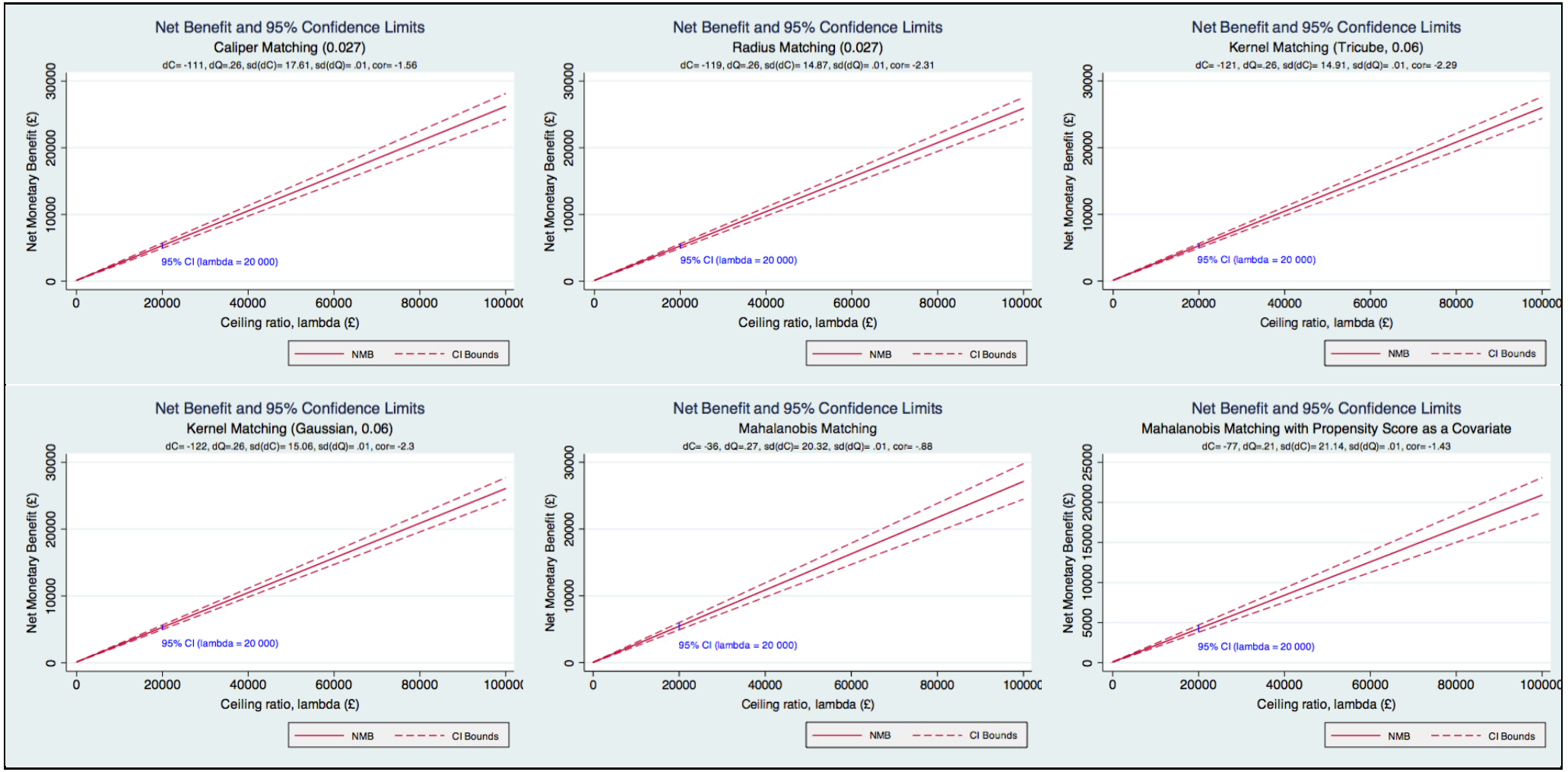


Figure A9 – Net benefit and 95% confidence limits for Alongside Midwifery Unit vs. Obstetric Unit

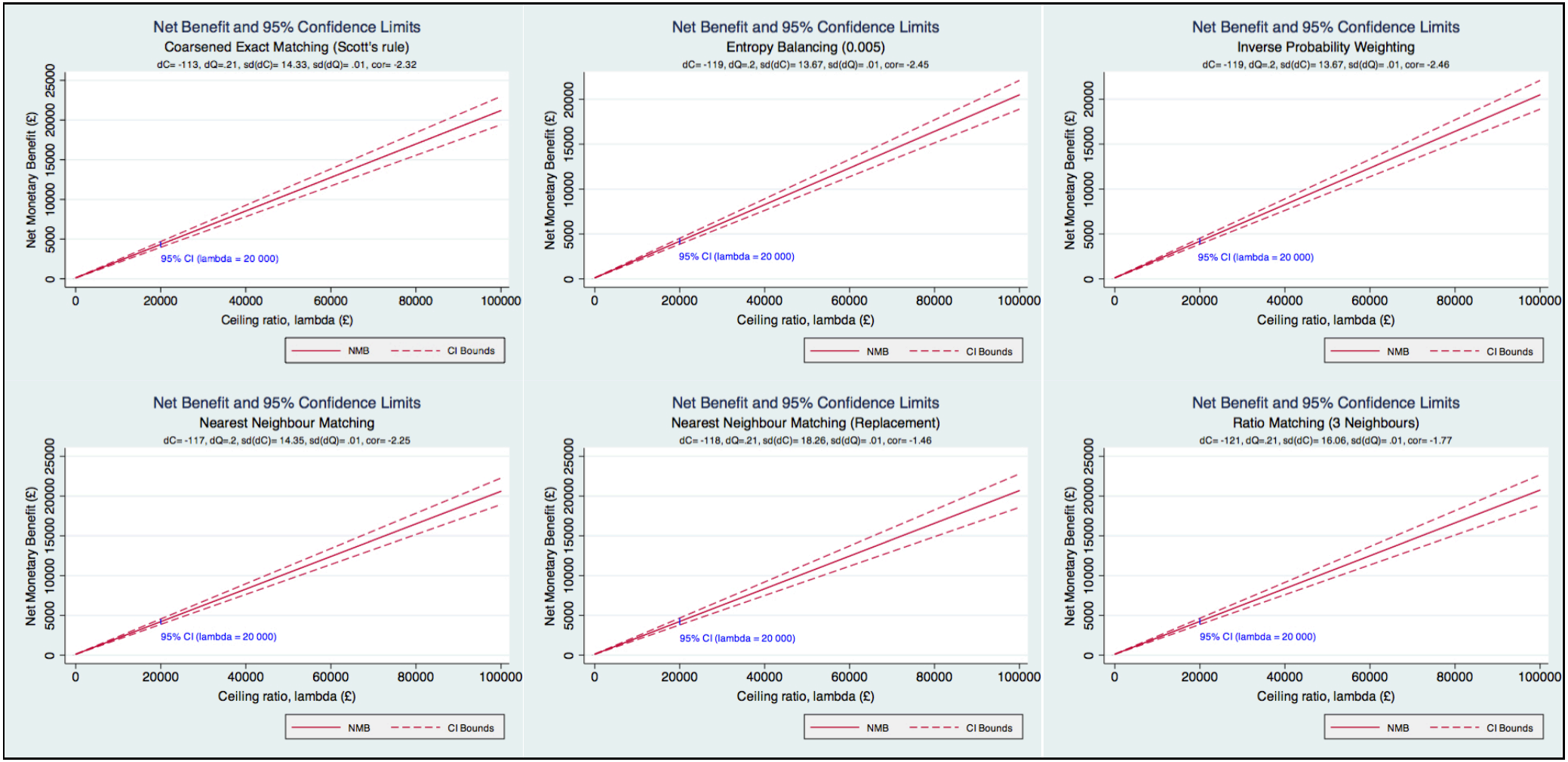


Figure A9 – Net benefit and 95% confidence limits for Alongside Midwifery Unit vs. Obstetric Unit (continued)

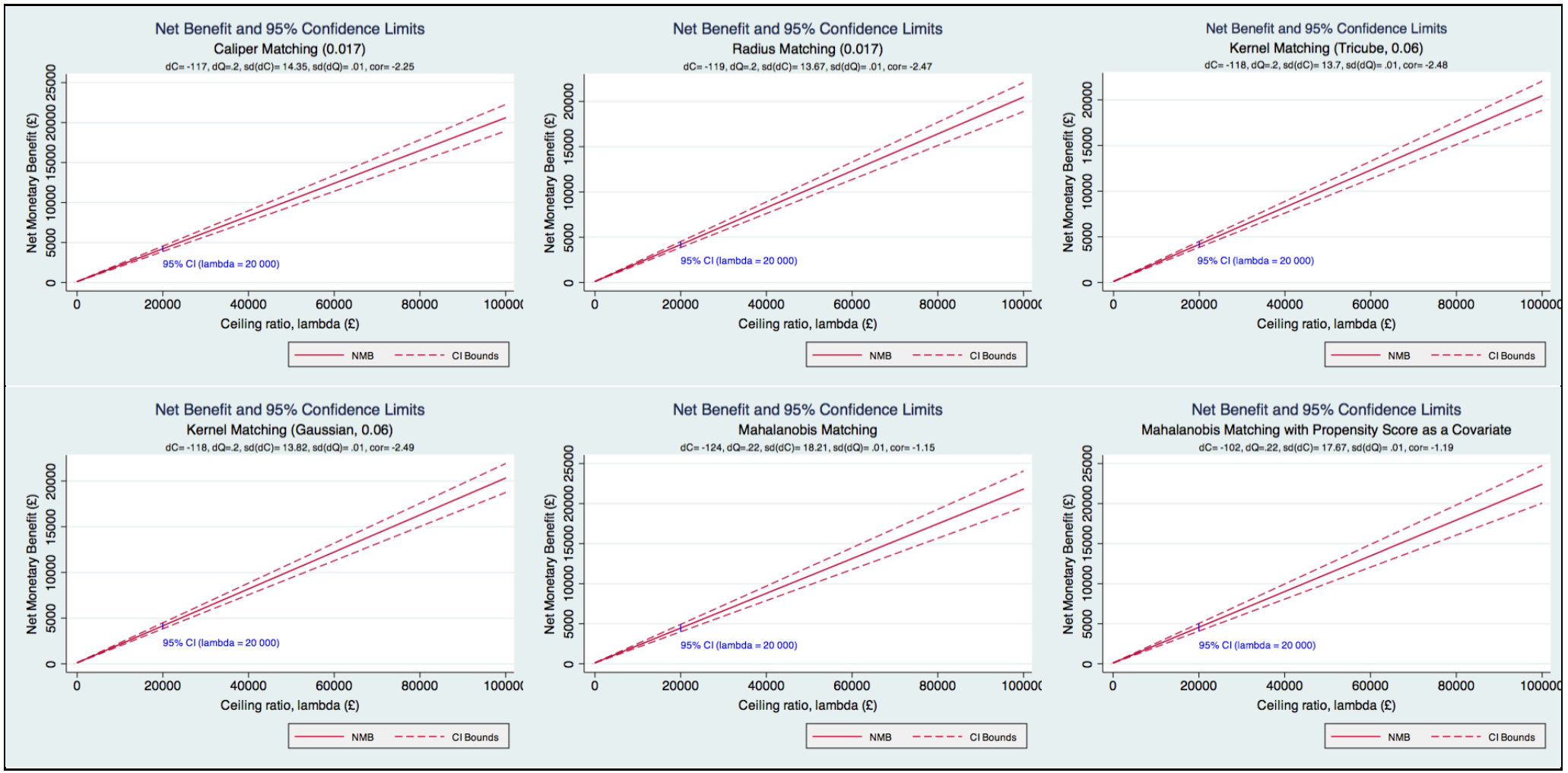


Figure A10 – Propensity score distributions for Home vs. OU

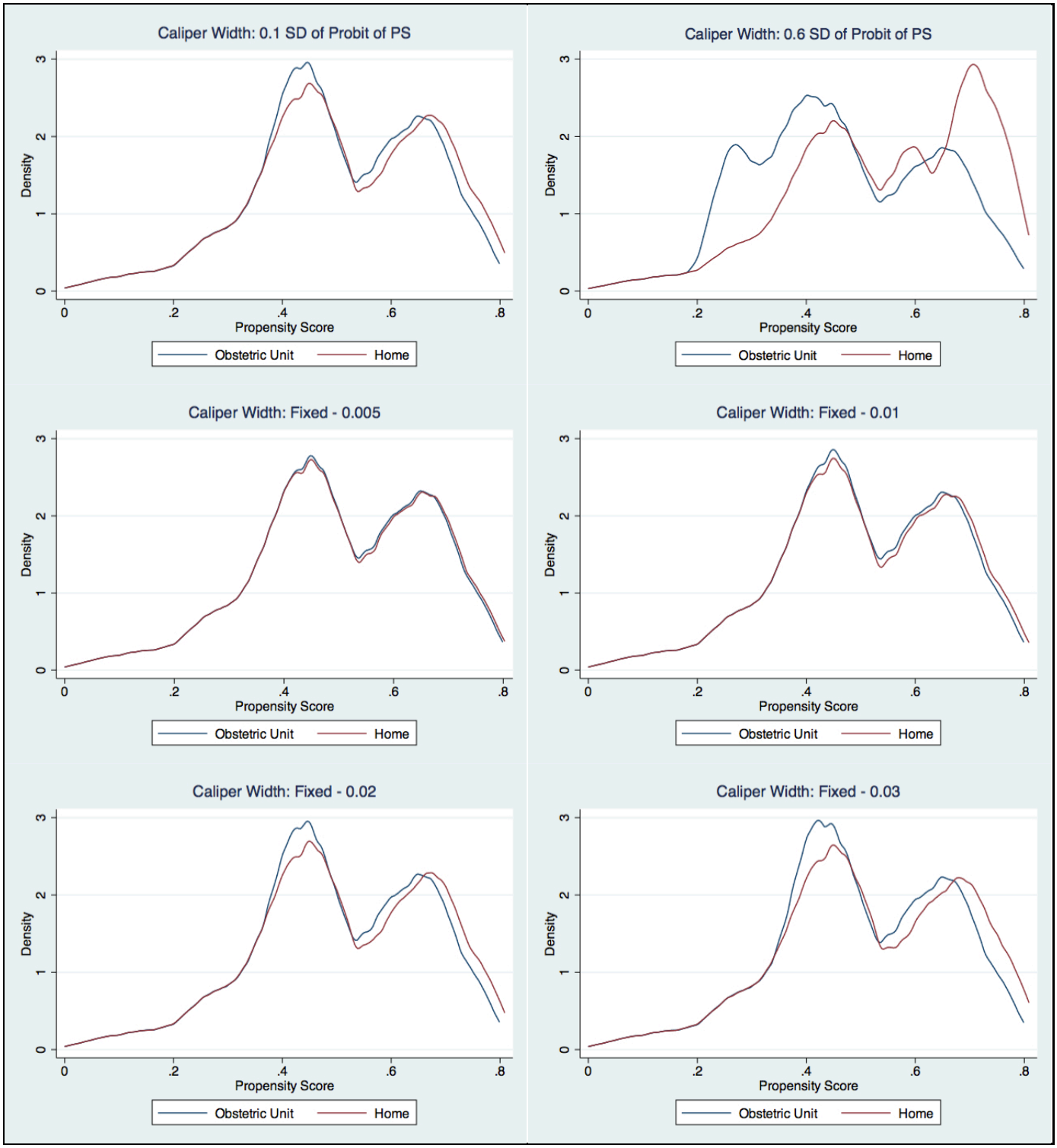


Figure A10 – Propensity score distributions for Home vs. OU (continued)

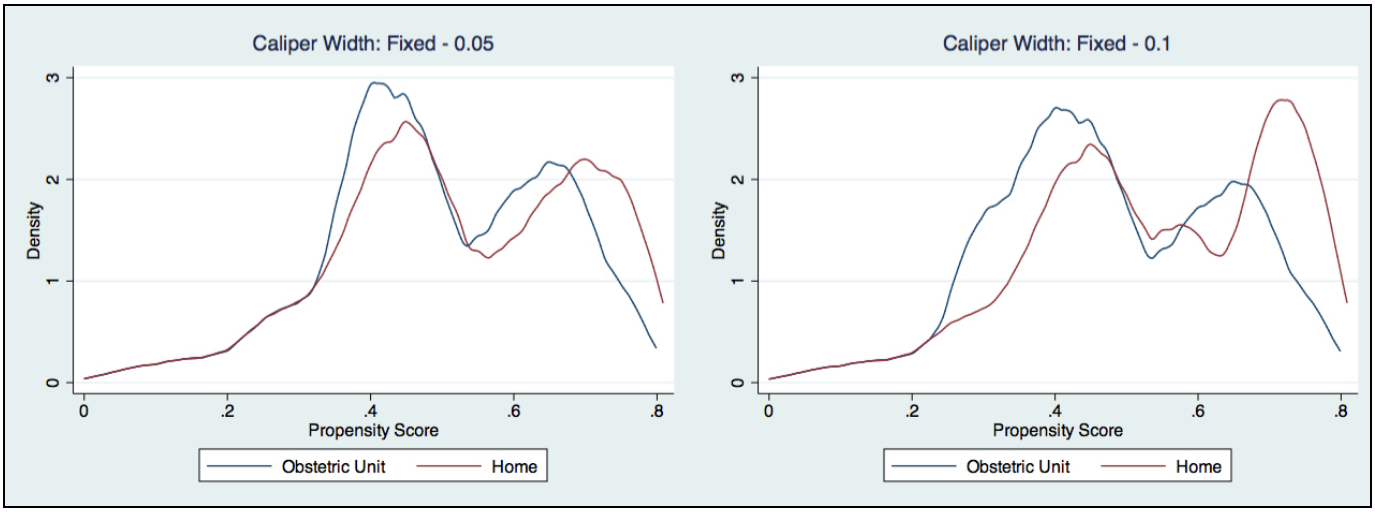


Figure A11 – Propensity score distributions for FMU vs. OU

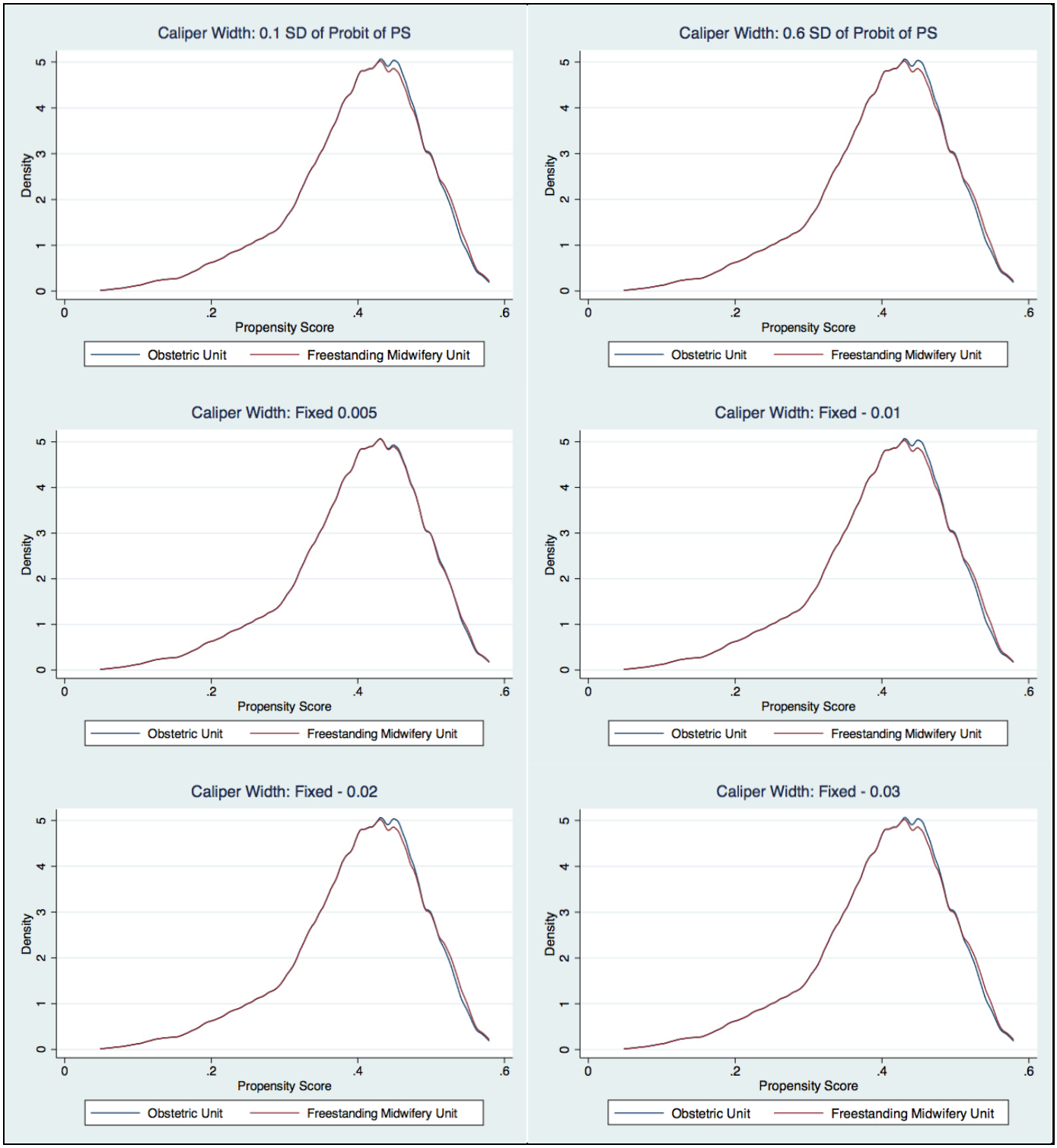


Figure A11 – Propensity score distributions for FMU vs. OU (continued)

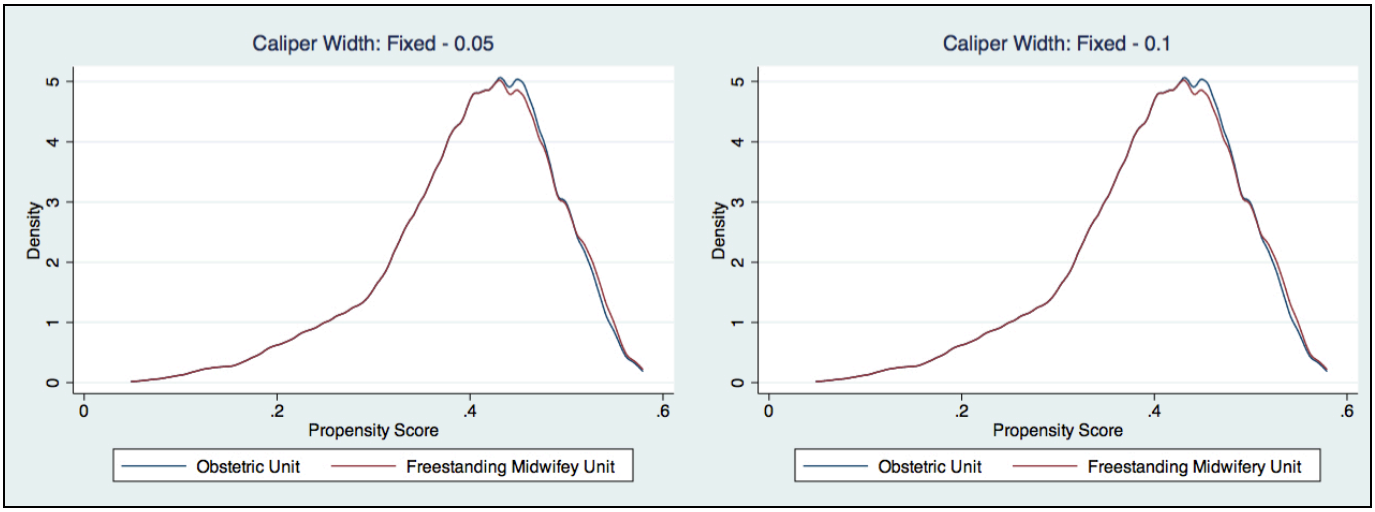


Figure A12 – Propensity score distributions for AMU vs. OU

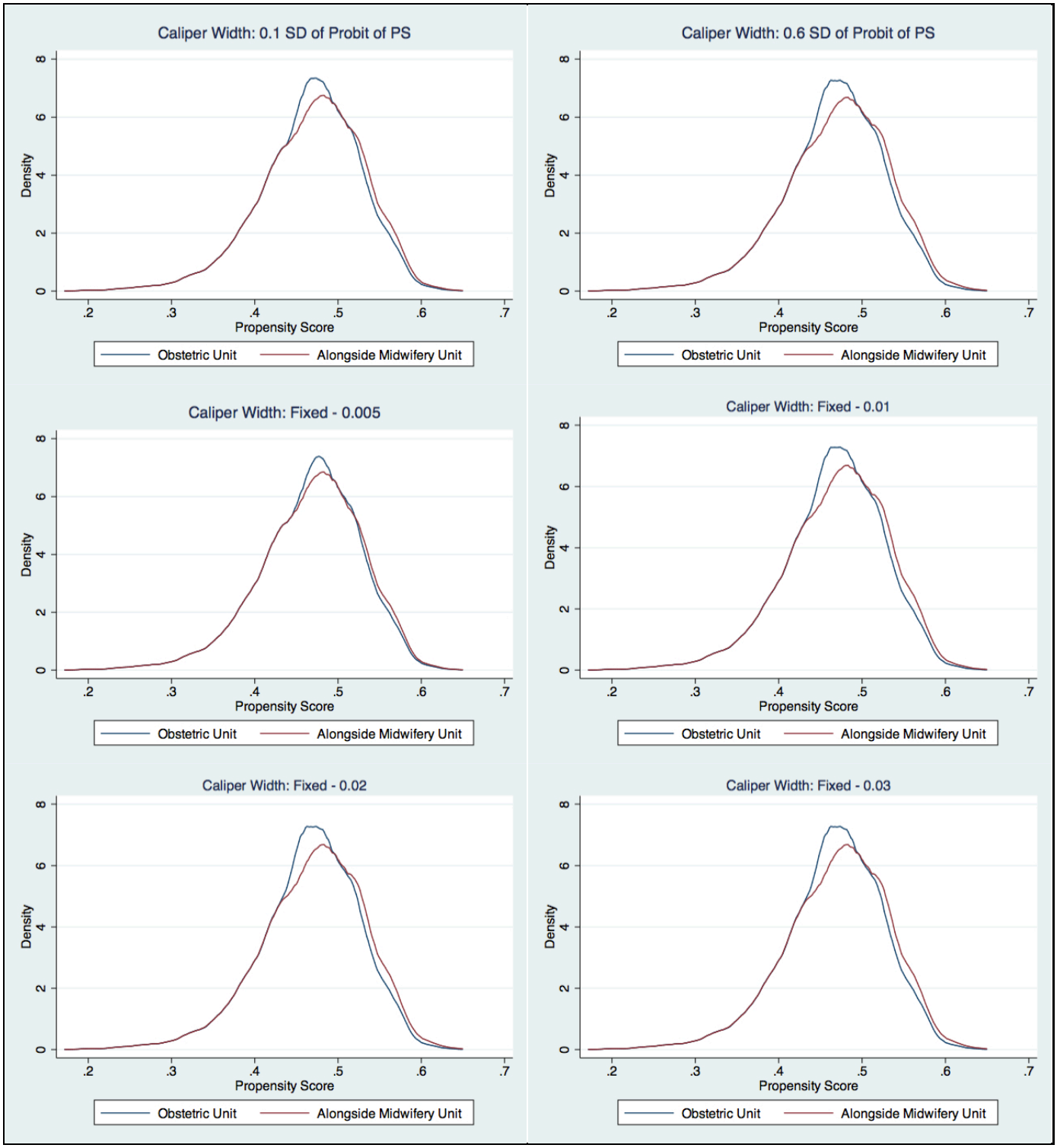


Figure A12 – Propensity score distributions for AMU vs. OU (continued)

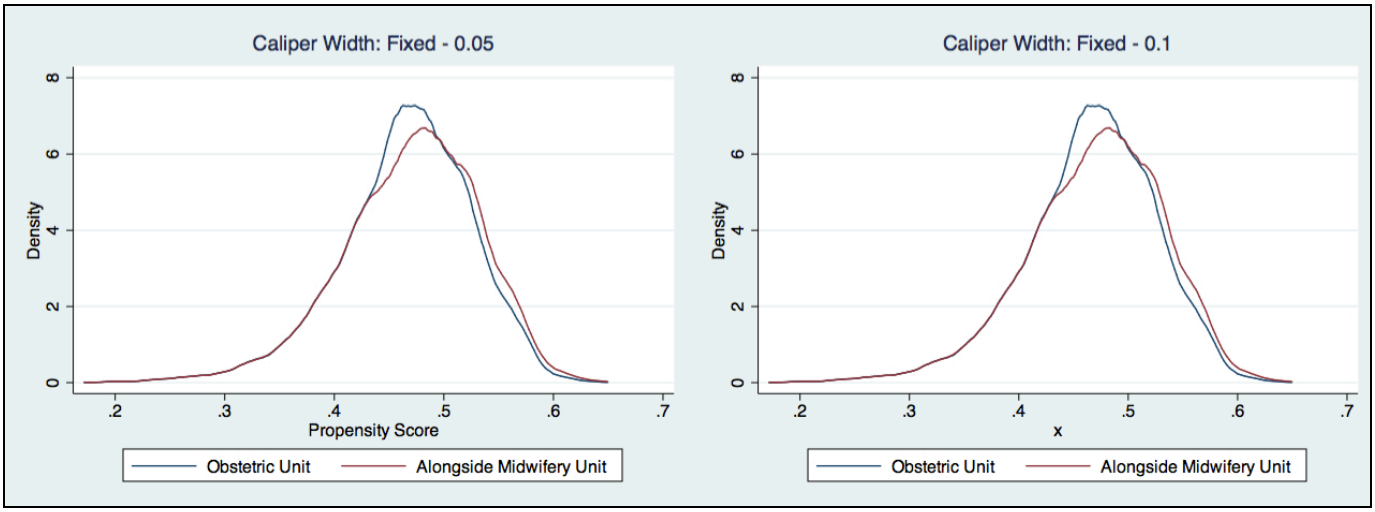


Figure A13 – Covariate balance for Home vs. OU

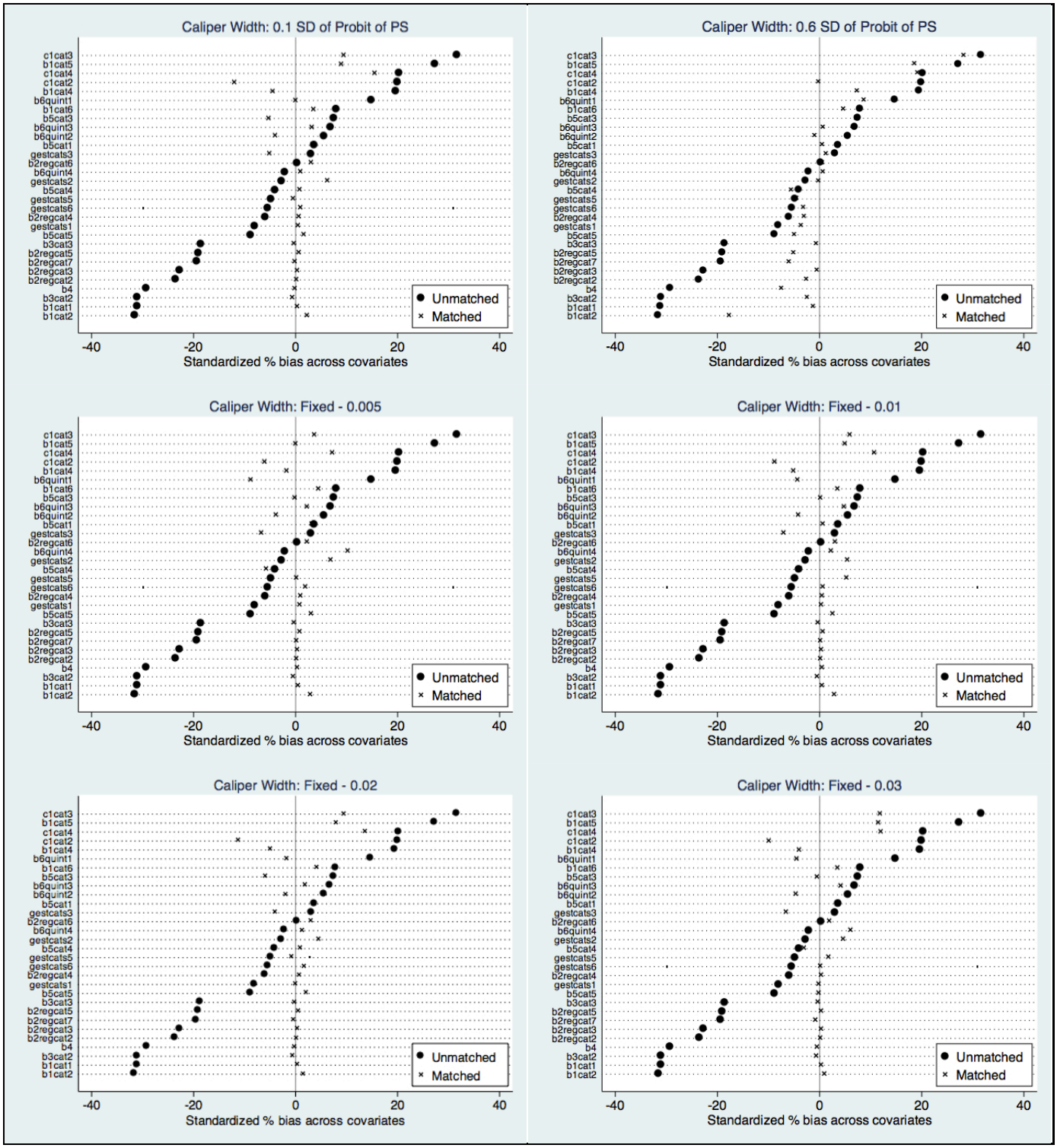


Figure A13 – Covariate balance for Home vs. OU (continued)

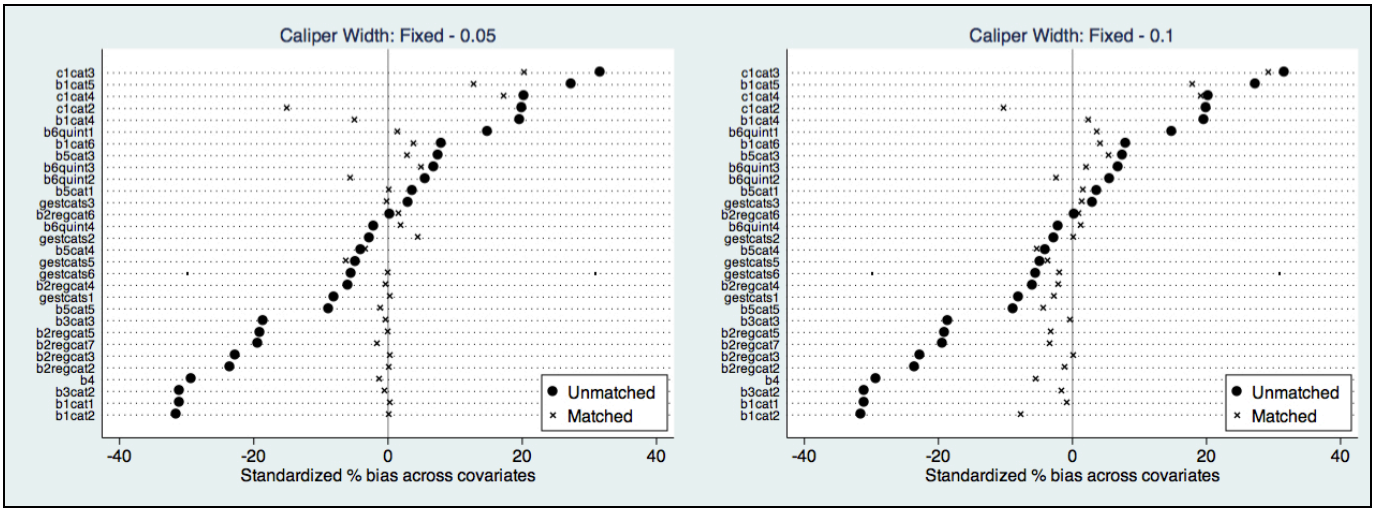


Figure A14 – Covariate balance for FMU vs. OU

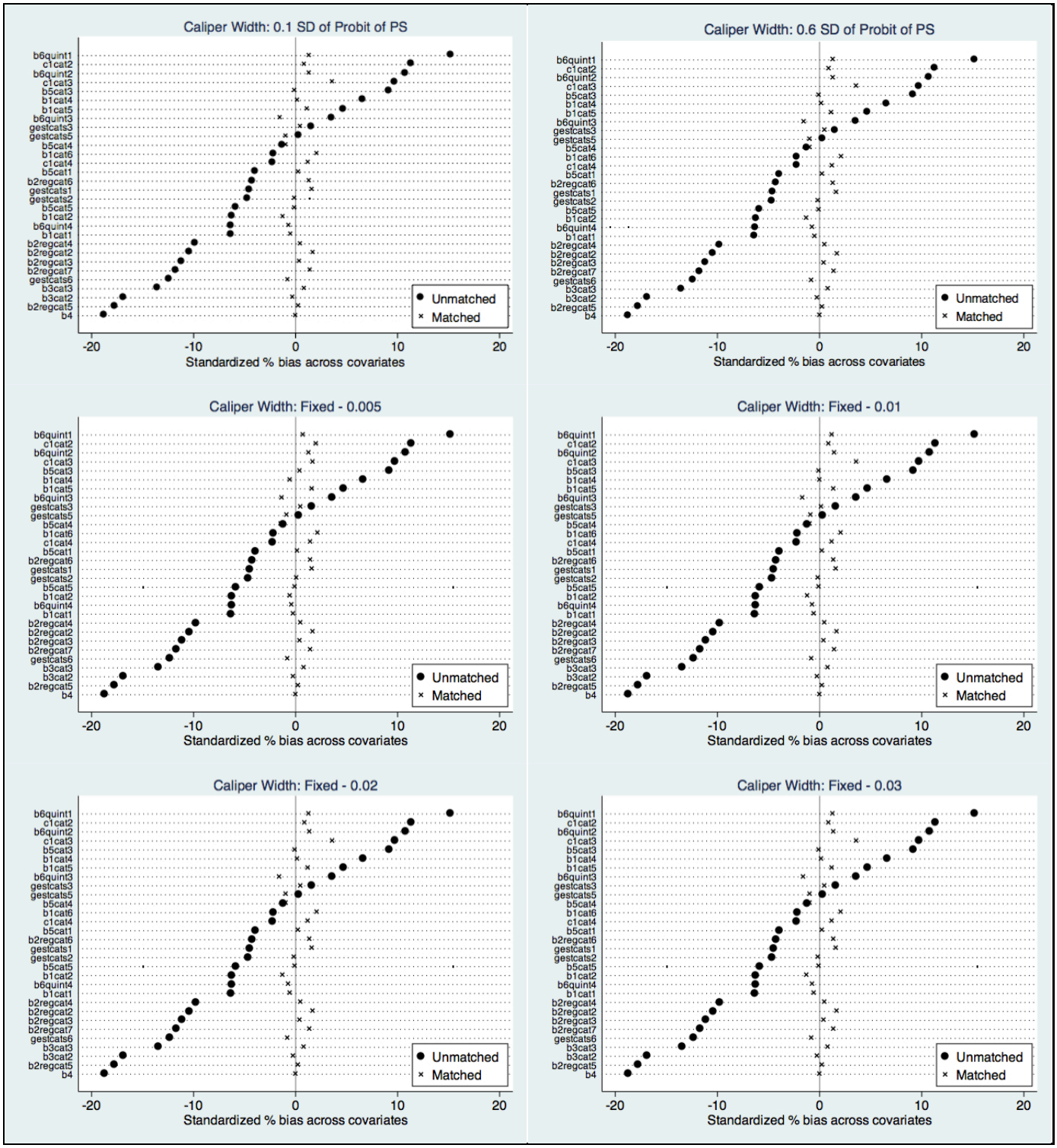


Figure A14 – Covariate balance for FMU vs. OU (continued)

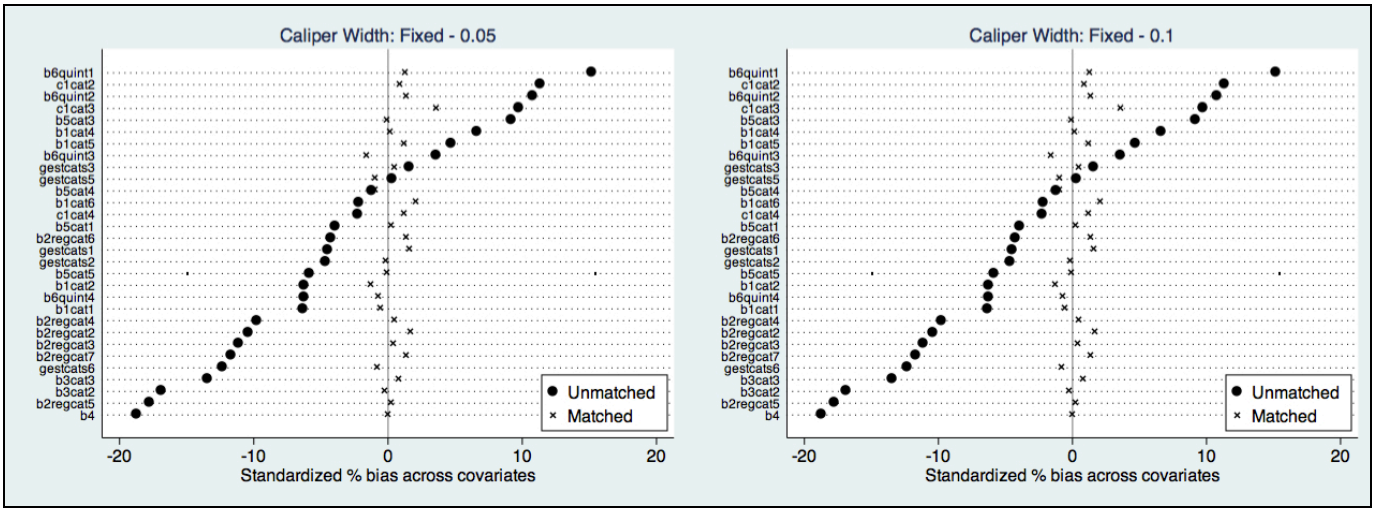


Figure A15 – Covariate balance for AMU vs. OU

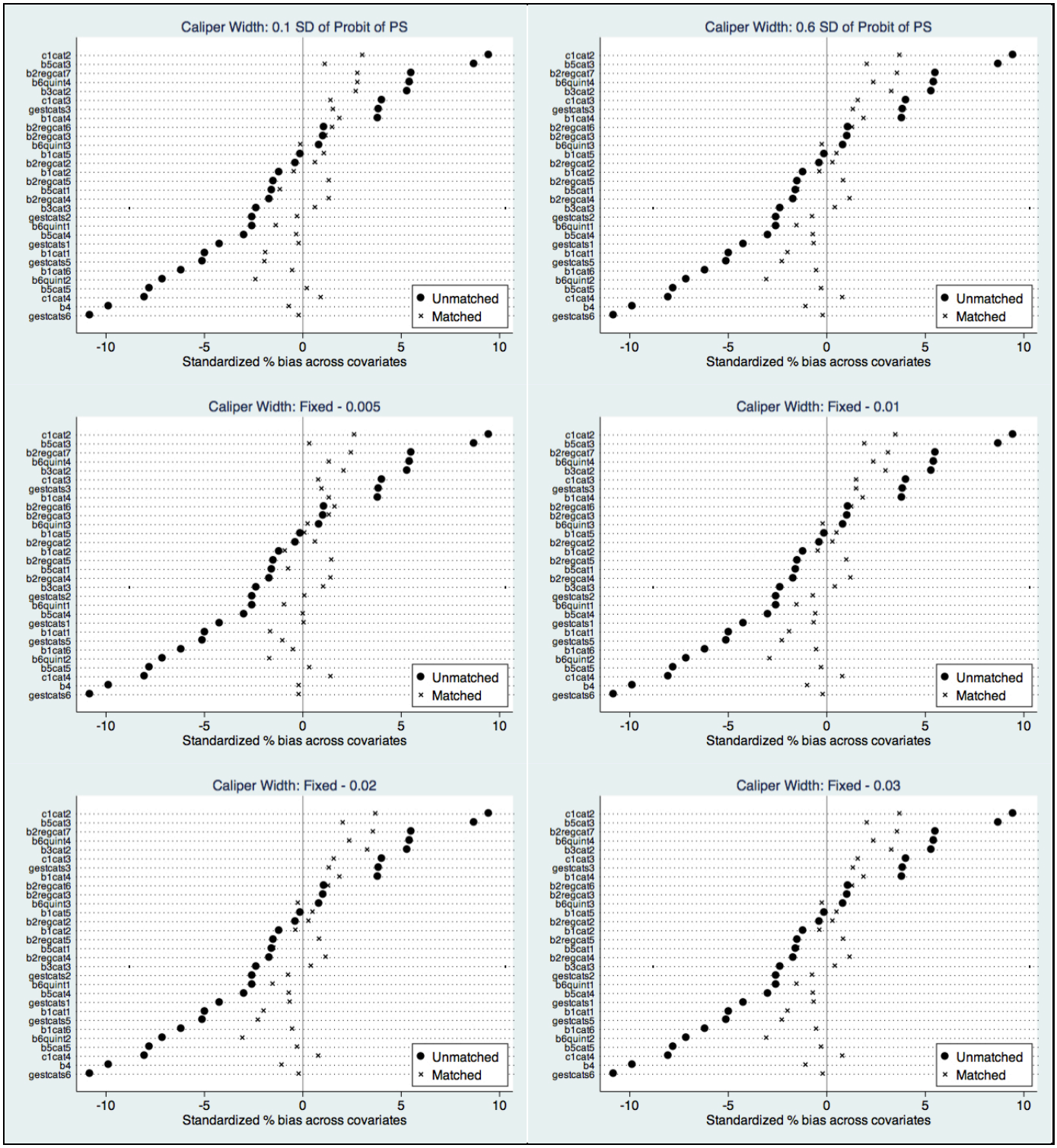


Figure A15 – Covariate balance for AMU vs. OU (continued)

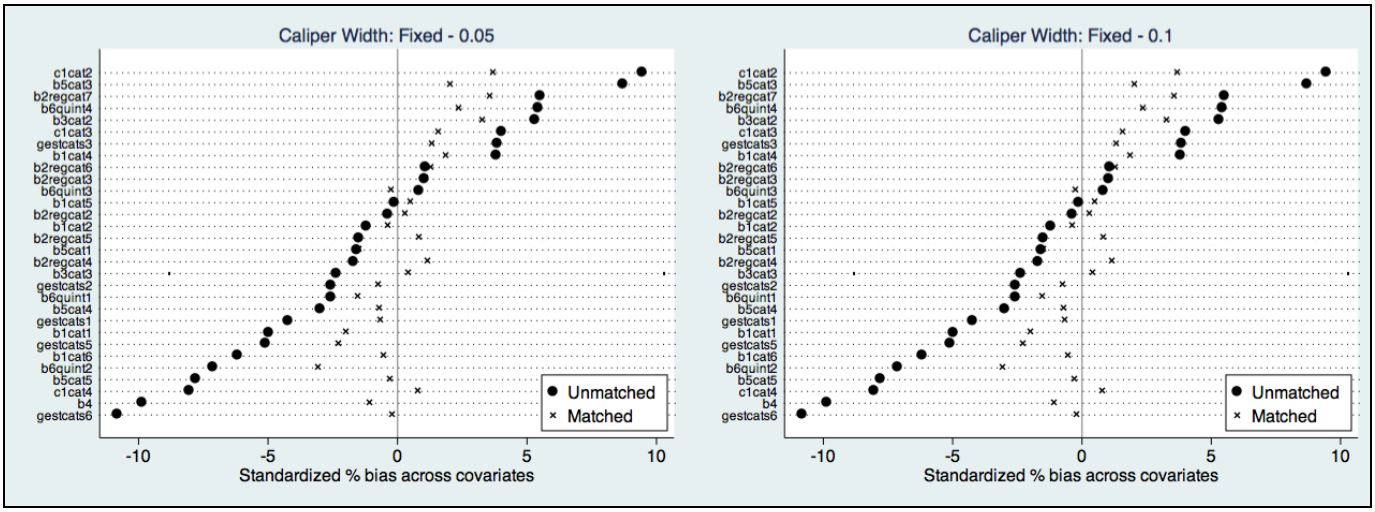


Figure A16 – Net benefit & 95% confidence limits - Home vs. OU

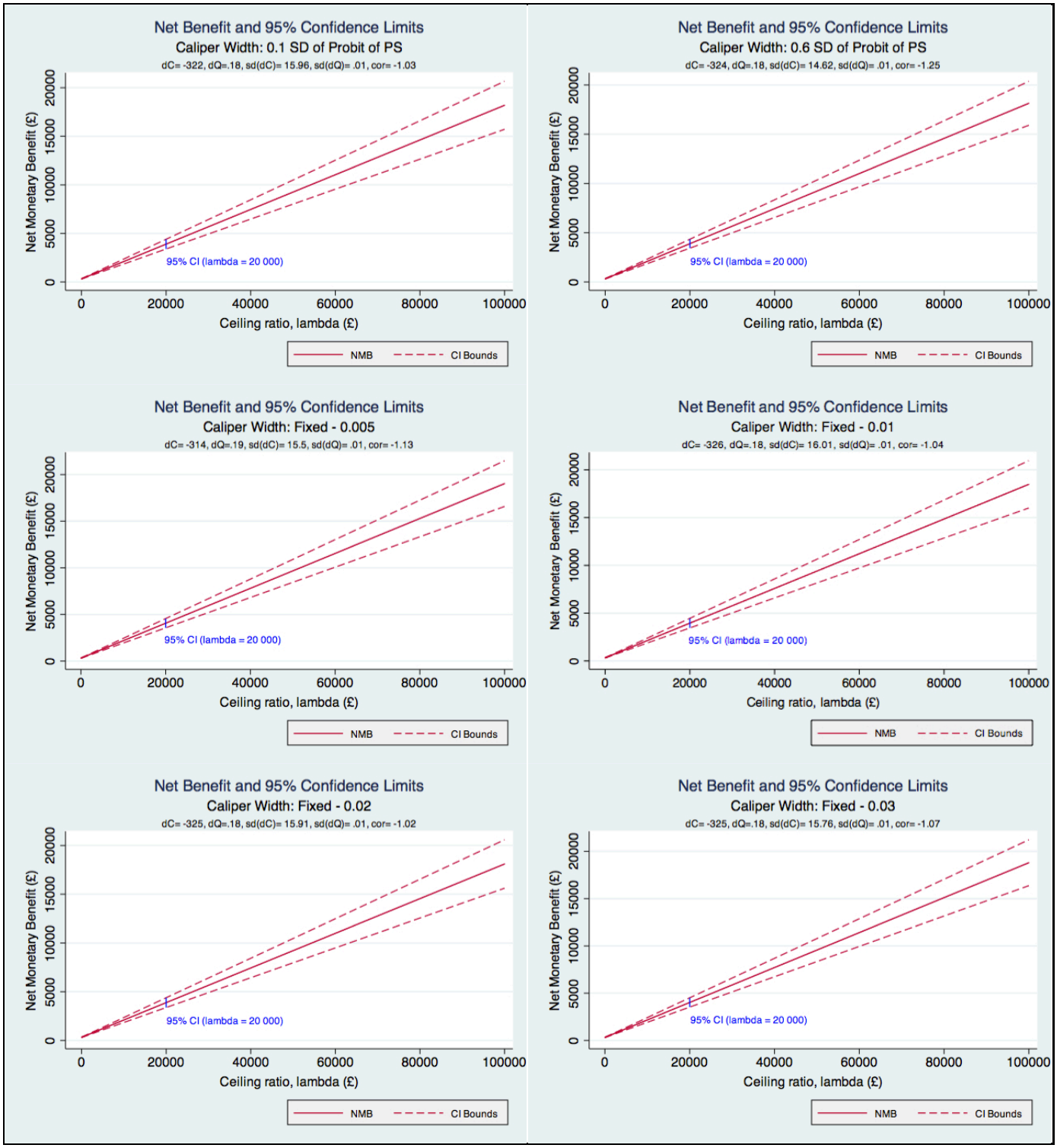


Figure A16 – Net benefit & 95% confidence limits - Home vs. OU (continued)

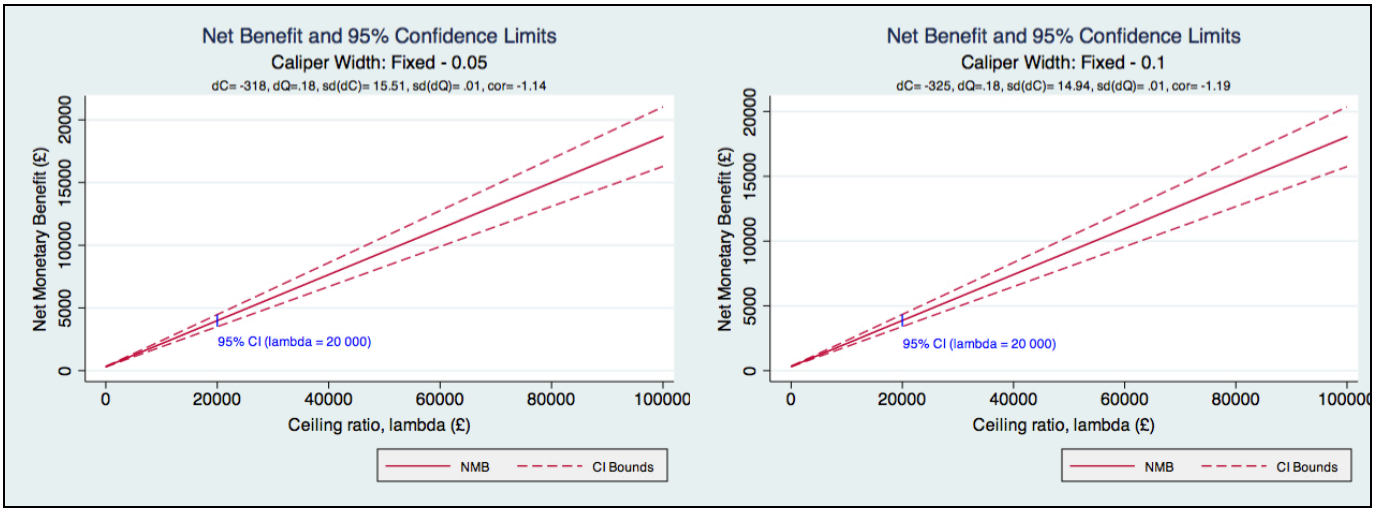


Figure A17 – Net benefit & 95% confidence limits - FMU vs. OU

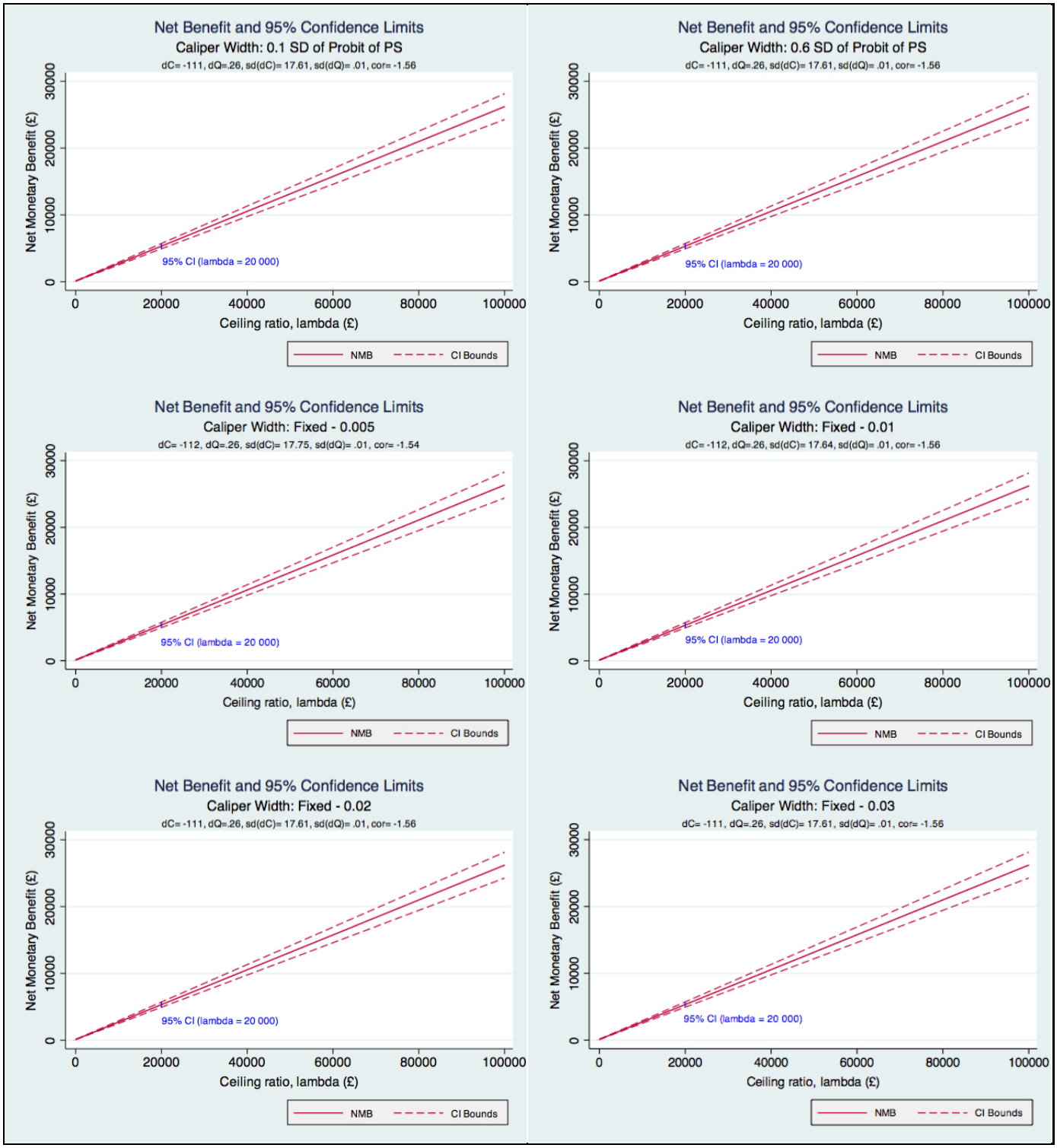


Figure A17 – Net benefit & 95% confidence limits - FMU vs. OU (continued)

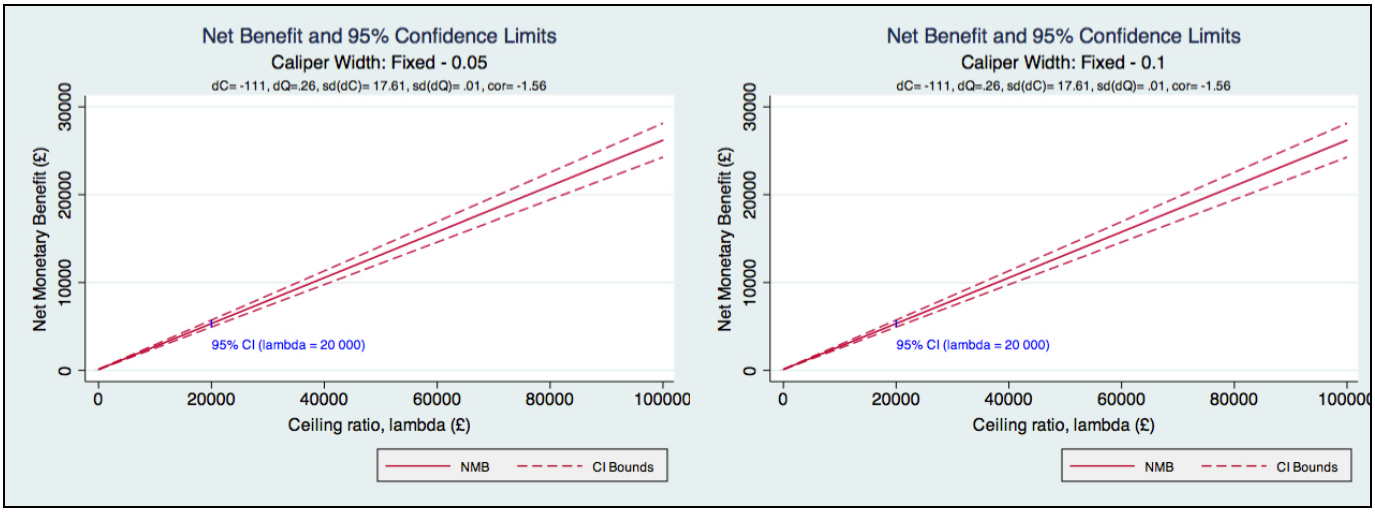


Figure A18 – Net benefit and 95% confidence limits for AMU vs. OU

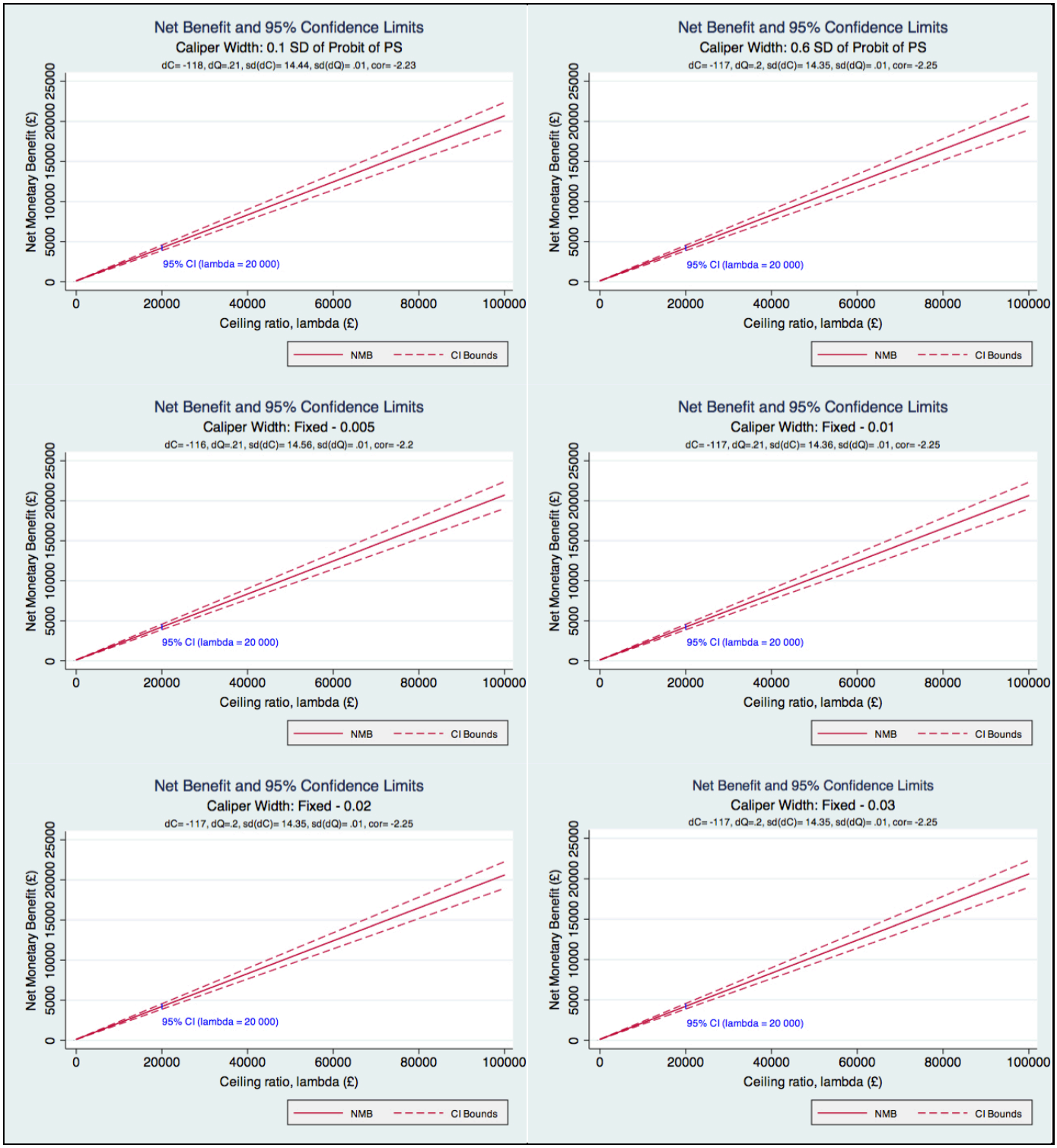


Figure A18 – Net benefit & 95% confidence limits - AMU vs. OU (continued)

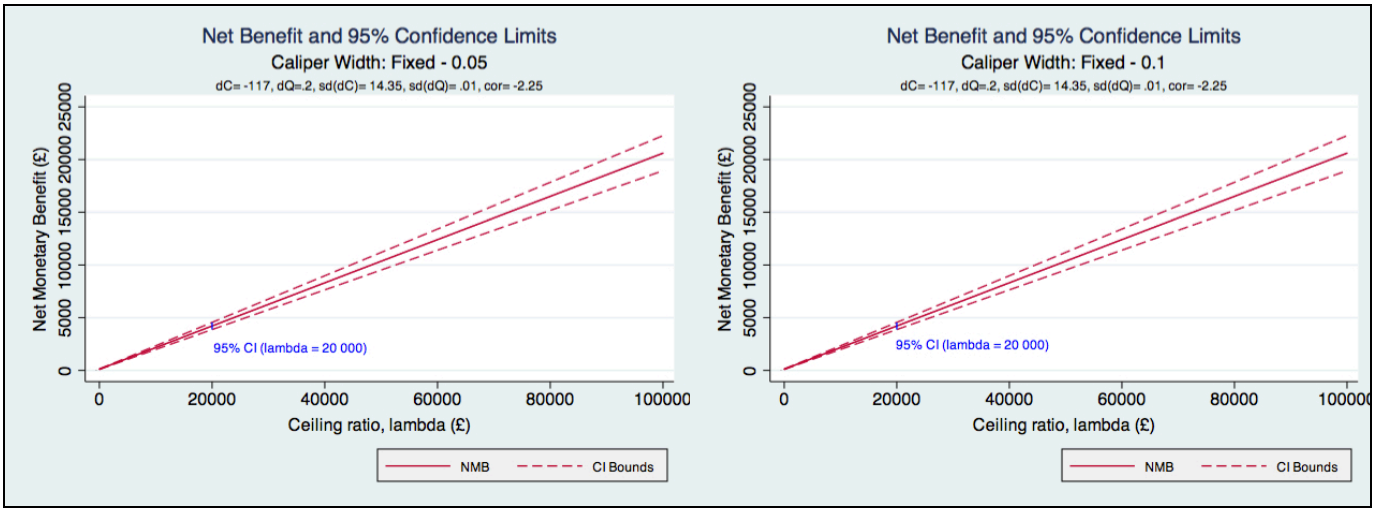
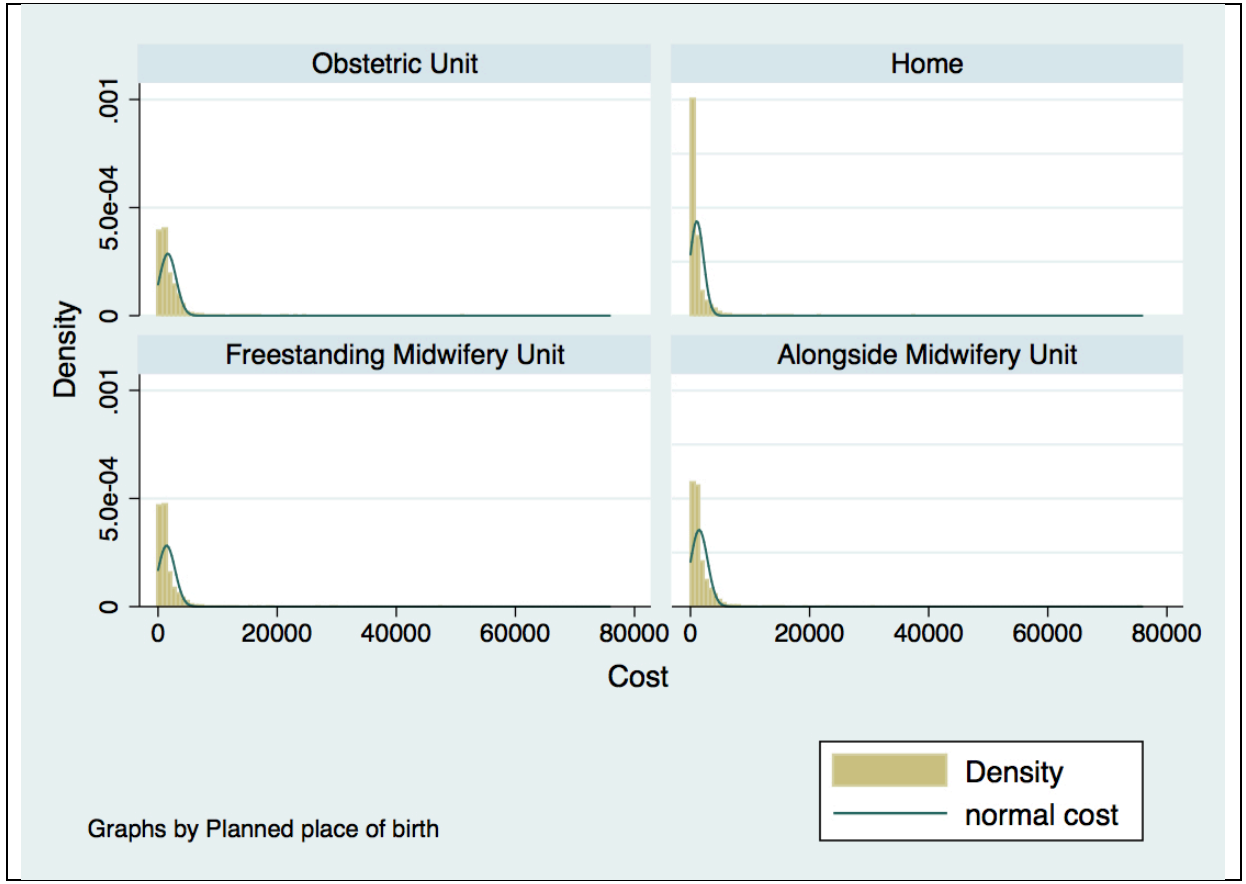


Figure A19 – Cost distribution



BIBLIOGRAPHY

Abadie, A., 2005. Semiparametric difference-in-differences estimators. *The Review of Economic Studies*, 72(1), 1-19.

Abadie, A. and Imbens, G., 2002. Simple and Bias-Corrected Matching Estimators for Average Treatment Effects, *NBER Technical Working Paper No. 283*.

Abadie, A. and Imbens, G., 2008. On the Failure of the Bootstrap for Matching Estimators. *Econometrica*, 76(6): 1537–57.

Abadie, A. and Imbens, G., 2011. Bias-corrected matching estimators for average treatment effects. *Journal of Business and Economic Statistics*, 29(1), 1-11.

Ahsan K.Z., Li, M. and Streatfield, P.K., 2007. Factors Affecting the Choice of Safe Delivery Practices for Pregnant Women in Bangladesh. [Abstract] 11th Annual Scientific Conference (ASCON), International Centre for Diarrhoeal Disease Research, Bangladesh. Abstract retrieved from: http://www.icddrb.org

Akazawa, M., Stearns, S. and Biddle, A., 2008. Assessing treatment effects of inhaled corticosteroids on medical expenses and exacerbations among COPD patients:

longitudinal analysis of managed care claims. *Health Services Research*, *43*(6), 2164-2182.

Alegría, M., Frank, R. and McGuire, T., 2005. Managed care and systems cost-effectiveness: treatment for depression. *Medical Care*, *43*(12), 1225.

Althauser, R. and Rubin, D., 1971. Measurement error and regression to the mean in matched samples. *Social Forces*, *50*(2), 206-214.

Angrist, J., 1998. Estimating the Labor Market Impact of Voluntary Military Service Using Social Security Data on Military Applicants. *Econometrica*, Vol. 66, 249-288.

Angrist, J., Imbens, G. and Rubin, D., 1996. Identification of causal effects using instrumental variables. *Journal of the American statistical Association*, *91*(434), 444-455.

Angrist, J. and Krueger, A., 2001. Instrumental variables and the search for identification: From supply and demand to natural experiments. Working Paper No. w8456. National Bureau of Economic Research.

Angrist, J. and Pischke, J., 2009. *Mostly harmless econometrics: An empiricist's companion*. New Jersey: Princeton University Press.

Asim, O. and Petrou, S., 2005. Valuing a QALY: review of current controversies. *Expert Review of Pharmacoeconomics and Outcomes Research*, *5*(6), 667-669.

Austin, P., 2008. A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003. *Statistics in medicine*, *27*(12), 2037-2049.

Austin, P., 2009a. Some Methods of Propensity-Score Matching had Superior Performance to Others: Results of an Empirical Investigation and Monte Carlo simulations. *Biometrical Journal*, *51*(1), 171-184.

Austin, P., 2009b. The relative ability of different propensity score methods to balance measured covariates between treated and untreated subjects in observational studies. *Medical Decision Making*, *29*(6), 661-677.

Bain, M., Chalmers, J. and Brewster, D., 1997. Routinely collected data in national and regional databases-an under-used resource. *Journal of Public Health*, *19*(4), 413-418.

Bang, H. and Robins, J., 2005. Doubly robust estimation in missing data and causal inference models. *Biometrics*, *61*(4), 962-973.

Barnett, P. and Swindle, R., 1997. Cost-effectiveness of inpatient substance abuse treatment. *Health Services Research*, *32*(5), 615.

Barnow, B., Cain, G. and Goldberger, A., 1980. Issues in the Analysis of Selectivity Bias. *Evaluation Studies Review Annual* Volume 5, 43.

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

Basu, A. and Manning, W., 2009. Issues for the next generation of health care cost analyses. *Medical Care*, *47*(7S1), S109-S114.

Basu, A., Polsky, D. and Manning, W., 2008. Use of propensity scores in non-linear response models: the case for health care expenditures. Working Paper No. w14086. National Bureau of Economic Research.

Benson, K. and Hartz, A., 2000. A comparison of observational studies and randomized, controlled trials. *New England Journal of Medicine*, *342*(25), 1878-1886.

Becker, S. and Caliendo, M., 2007. Sensitivity Analysis for Average Treatment Effects. *Stata Journal*, 7(1), 71-83.

Berk, A., 2004. *Regression Analysis: a constructive critique*. Thousand Oaks, CA:Sage.

Bhattacharya, J., Goldman, D. and McCaffrey, D., 2006. Estimating probit models with self-selected treatments. *Statistics in Medicine*, *25*(3), 389-413.

Baum, C., Schaffer, M. and Stillman, S., 2003. Instrumental variables and GMM: Estimation and testing. *Stata Journal*, *3*(1), 1-31.

Birnbaum, H., Cremieux, P., Greenberg, P., LeLorier, J., Ostrander, J. and Venditti, L., 1999. Using healthcare claims data for outcomes research and pharmacoeconomic analyses. *Pharmacoeconomics*, *16*(1), 1-8.

Birthplace in England Collaborative Group, 2011. Perinatal and maternal outcomes by planned place of birth for healthy women with low risk pregnancies: the Birthplace in England national prospective cohort study. *British Medical Journal*. 2011;343:d7400 doi:10.1136/bmj.d7400

Black, W., 1990. The CE Plane A Graphic Representation of Cost-Effectiveness. *Medical Decision Making*, *10*(3), 212-214.

Black, N., 1996. Why we need observational studies to evaluate the effectiveness of health care. *British Medical Journal*, *312*(7040), 1215.

Blackwell, M., Iacus, S., King, G. and Porro, G., 2009. CEM: Stata module to perform Coarsened Exact Matching. The *Stata Journal*, 9, Number 4, 524–546.

Blanchette, C., Akazawa, M., Dalal, A. and Simoni-Wastila, L., 2008. Risk of hospitalizations/emergency department visits and treatment costs associated with initial maintenance therapy using fluticasone propionate 500 microg/salmeterol 50 microg compared with ipratropium for chronic obstructive pulmonary disease in older adults. *The American Journal of Geriatric Pharmacotherapy*, *6*(3), 138.

Blundell, R. and Dias, M. C., 2009. Alternative approaches to evaluation in empirical microeconomics. *Journal of Human Resources*, *44*(3), 565-640.

Blundell, R. and Powell, J., 2001. Endogeneity in Nonparametric and Semiparametric Regression Models. Institute for Fiscal Studies, Department of Economics, UCL, CEMMAP Working Paper CWP09/01.

Blundell, R. and Powell, J., 2004. Endogeneity in semiparametric binary response models. *Review of Economic Studies*, *71*(3), 655-679.

Bound, J., Brown, C. and Mathiowetz, N., 2001. Measurement error in survey data. In: J. Heckman & E. Leamer ed., *Handbook of Econometrics*, edition 1, volume 5, chapter 59, 3705-3843

Briggs, A., 1999. Handling uncertainty in economic evaluation. *British Medical Journal*, *319*(7202), 120.

Briggs, A., 2000. Economic evaluation and clinical trials: size matters: The need for greater power in cost analyses poses an ethical dilemma. *British Medical Journal*, *321*(7273), 1362.

Briggs, A., 2001. A Bayesian approach to stochastic cost-effectiveness analysis. *International Journal of Technology Assessment in Health Care*, *17*(1), 69-82.

Briggs, A. and Fenn, P., 1998. Confidence intervals or surfaces? Uncertainty on the cost-effectiveness plane. *Health Economics*, *7*(8), 723-740.

Briggs, A. and Gray, A., 1998. Power and sample size calculations for stochastic cost-effectiveness analysis. *Medical Decision Making*, *18*(2), S81-S92.

Briggs, A., Claxton, K., and Sculpher, M., 2006. *Decision Modelling for Health Economic Evaluation*. Oxford: Oxford University Press.

Briggs, A., Nixon, R., Dixon, S. and Thompson, S., 2005. Parametric modelling of cost data: some simulation evidence. *Health Economics*, *14*(4), 421-428.

Briggs, A., Mooney, C. and Wonderling, D., 1999. Constructing confidence intervals for cost-effectiveness ratios: an evaluation of parametric and non-parametric techniques using Monte Carlo simulation. *Statistics in medicine*, *18*(23), 3245-3262.

Briggs, A., O'Brien, B. and Blackhouse, G., 2002. Thinking outside the box: recent advances in the analysis and presentation of uncertainty in cost-effectiveness studies. *Annual Review of Public Health*, *23*(1), 377-401.

Briggs, A., Wonderling, D. and Mooney, C., 1997. Pulling cost-effectiveness analysis up by its bootstraps: a non-parametric approach to confidence interval estimation. *Health Economics*, *6*(4), 327-340.

Brocklehurst, P., 2011. *Discussion on perinatal potential confounders.* [e-mail] (Personal communication, 6 December 2011).

Bryson, A., Dorsett, R. and Purdon, S., 2002. The use of propensity score matching in the evaluation of active labour market policies. Working Paper Number 4, Department for Work and Pensions.

Busse, R., 2001. Expenditure on health care in the EU: making projections for the future based on the past. *HEPAC Health Economics in Prevention and Care*, *2*(4), 158-161.

Buxton, M., Drummond, M., Van Hout, B., Prince, R., Sheldon, T., Szucs, T. and Vray, M., 1997. Modelling in ecomomic evaluation: An unavoidable fact of life. *Health economics*, *6*(3), 217-227.

Cakir, B., Ulmar, B., Schmidt, R., Kelsch, G., Geiger, P., Mehrkens, H., Puhl, W. and Richter, M., 2006. Efficacy and cost effectiveness of harmonic scalpel compared with electrocautery in posterior instrumentation of the spine. *European Spine Journal*, *15*(1), 48-54.

Caliendo, M. and Kopeinig, S., 2008. Some practical guidance for the implementation of propensity score matching. *Journal of Economic Surveys*, *22*(1), 31-72.

Cameron, A. and Trivedi, P., 2005. *Microeconometrics: methods and applications*. Cambridge: Cambridge University Press.

Campbell, R. and Macfarlane, A., 1986. Place of delivery: a review. *British Journal of Obstetrics and Gynaecology*, 93(7):675-83.

Campbell, M. and Torgerson, D., 1999. Bootstrapping: estimating confidence intervals for cost-effectiveness ratios. *Qjm*, *92*(3), 177-182.

Castelli, C., Combescure, C., Foucher, Y. and Daures, J., 2007. Cost-effectiveness analysis in colorectal cancer using a semi-Markov model. *Statistics in Medicine*, *26*(30), 5557-5571.

Cawley, J. and Meyerhoefer, C., 2011. The medical care costs of obesity: an instrumental variables approach. *Journal of Health Economics*. 31(1): 219–230.

Centre for Reviews and Dissemination, 2008. CRD's guidance for undertaking reviews in health care. University of York.

Chen, Q., Kane, R. and Finch, M., 2000. The cost effectiveness of post-acute care for elderly Medicare beneficiaries. *Inquiry*, 359-375.

Christensen, K. and Murray, J., 2007. What genome-wide association studies can do for medicine. *New England Journal of Medicine*, *356*(11), 1094-1097.

Claxton, K., 1999. The irrelevance of inference: a decision-making approach to the stochastic evaluation of health care technologies. *Journal of Health Economics*, *18*(3), 341.

Claxton, K., Martin, S., Soares, M., Rice, N., Spackman, E., Hinde, S., Devlin, N., Smith, P. and Sculpher, M., 2013. Methods for the Estimation of the NICE Cost-Effectiveness Threshold. CHE Research Paper 81. University of York.

Cochran, W., 1968. The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics*, 295-313.

Cochran, W. and Rubin, D., 1973. Controlling bias in observational studies: A review. *Sankhyā: The Indian Journal of Statistics, Series A*, 417-446.

Coleman, C., McKay, R., Boden, W., Mather, J. and White, C., 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, R. and MacMahon, S., 2001. Reliable assessment of the effects of treatment on mortality and major morbidity, I: clinical trials. *The Lancet*, *357*(9253), 373-380.

Concato, J., Shah, N. and Horwitz, R., 2000. Randomized, controlled trials, observational studies, and the hierarchy of research designs. *New England Journal of Medicine*, *342*(25), 1887-1892.

Cost-Effectiveness (CEA) Analysis Registry, n.d. Available at http://www.cearegistry.org

Coyte, P., Young, W. and Croxford, R., 2000. Costs and outcomes associated with alternative discharge strategies following joint replacement surgery: analysis of an observational study using a propensity score. *Journal of Health Economics*, *19*(6), 907-929.

Cox, D., 1958. *Planning of Experiments*. New York: Wiley.

Cox, J. and Koutroumanos, N., 2010. Comparing coding between interventional radiologists and hospital coding departments. *Clinical Audit*, 33.

Crowder, M., 1987. On linear and quadratic estimating functions. Biometrika, 74(3), 591-597.

Culyer, A., McCabe, C., Briggs, A., Claxton, K., Buxton, M., Akehurst, R. and Brazier, J., 2007. Searching for a threshold, not setting one: the role of the National Institute for Health and Clinical Excellence. *Journal of Health Services Research and Policy*, *12*(1), 56-58.

Cutler, D., 2007. The lifetime costs and benefits of medical technology. *Journal of Health Economics*, 26(6), 1081-1100.

Dawid, A., 2000. Causal inference without counterfactuals. *Journal of the American Statistical Association*, *95*(450), 407-424.

Dehejia, R. and Wahba, S., 1999. Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American Statistical Association*, *94*(448), 1053-1062.

Dehejia, R. and Wahba, S., 2002. Propensity score-matching methods for nonexperimental causal studies. *Review of Economics and Statistics*, *84*(1), 151-161.

Delgado-Rodríguez, M. and Llorca, J., 2004. Bias. *Journal of Epidemiology and Community Health*, *58*(8), 635-641.

de Jonge, A., van der Goes ,B.Y., Ravelli, A.C., Amelink-Verburg, M.P., Mol, B.W., Nijhuis, J.G., et al., 2009. Perinatal mortality and morbidity in a nationwide cohort of 529,688 low-risk planned home and hospital births. *BJOG*, 116(9):1177-84.

De Natale, R., Lafuma, A. and Berdeaux, G., 2009. Cost effectiveness of travoprost versus a fixed combination of latanoprost/timolol in patients with ocular hypertension or glaucoma: analysis based on the UK general practitioner research database. *Clinical Drug Investigation*, *29*(2), 111-120.

Department of Health, 2004. National Service Framework for Children, Young People and Maternity Services. Standard 11: Maternity Services. London.

Department of Health/Partnerships for Children Families and Maternity, 2007. Maternity Matters: Choice, access and continuity of care in a safe service. London.

De Ridder, A. and De Graeve, D., 2009. Comparing the cost effectiveness of risperidone and olanzapine in the treatment of schizophrenia using the net-benefit regression approach. *Pharmacoeconomics*, *27*(1), 69-80.

Dhainaut, J., Payet, S., Vallet, B., França, L., Annane, D., Bollaert PE., Le Tulzo, Y., Runge, I., Malledant, Y., Guidet, B., Le Lay, K., Launois R. and the PREMISS Study

Group, 2007. Cost-effectiveness of activated protein C in real-life clinical practice. *Critical Care*, *11*(5), R99.

Diamond, A. and Sekhon J.. 2005. Genetic Matching for Estimating Causal Effects: A General Multivariate Matching Method for Achieving Balance in Observational Studies." Working Paper. Available at http://sekhon.berkeley.edu/papers/GenMatch.pdf. [Accessed 12 March 2011].

DiNardo, J., 2008. Natural experiments and quasi-natural experiments. In Durlauf, S. and Blume, L. *The New Palgrave Dictionary of Economics* (Second ed.). Palgrave Macmillan.

Ding, W., Lehrer, S., Rosenquist, J. and Audrain-McGovern, J., 2009. The impact of poor health on academic performance: New evidence using genetic markers. *Journal of Health Economics*, *28*(3), 578-597.

Dixon, A., and Poteliakhoff, E., 2012. Back to the future: 10 years of European health reforms. *Health Economics, Policy and Law*, *7*(01), 1-10.

Drummond, M., 1998. Experimental versus observational data in the economic evaluation of pharmaceuticals. *Medical decision making*, *18*(2), S12-S18.

Drummond, M. and McGuire, A., 2001. *Economic evaluation in health care: merging theory with practice*. Oxford: Oxford University Press.

Drummond, M., Aguiar-Ibanez, R. and Nixon, J., 2006. Economic evaluation. *Singapore Medical Journal*, *47*(6), 456.

Drummond, M. and Sculpher, M., 2005. Common methodological flaws in economic evaluations. *Medical Care*, *43*(7), II-5.

Drummond, M., Sculpher, M., Torrance, G., O'Brien, B. and Stoddart, G., 2005. *Methods for the economic evaluation of health care programmes*. 3rd ed. Oxford: Oxford University Press.

Duflo, E., Glennerster, R. and Kremer, M., 2007. Using randomization in development economics research: A toolkit. In: *Handbook of Development Economics*, Volume *4*, eds. P. Schultz and J. Strauss, 3895-3962.

McLennan, D., Barnes, H., Noble, M., Davies, J., Garratt, E. and Dibben, C., 2011. *The English indices of deprivation 2010*. London, Department for Communities and Local Government.

Farias-Eisner, R., Horblyuk, R., Franklin, M., Lunacsek, O. and Happe, L., 2009. Economic and clinical evaluation of fondaparinux vs. enoxaparin for thromboprophylaxis following general surgery. *Current Medical Research and Opinion*, *25*(5), 1081-1087.

Fayers, P. and Hand, D., 1997. Generalisation from phase III clinical trials: survival, quality of life, and health economics. *The Lancet*, *350*(9083), 1025-1027.

Fenwick, E., Marshall, D., Levy, A. and Nichol, G., 2006. Using and interpreting cost-effectiveness acceptability curves: an example using data from a trial of management strategies for atrial fibrillation. *BMC Health Services Research*, *6*(1), 52.

Flury, B. and Riedwyl, H., 1986. Standard distance in univariate and multivariate analysis. *The American Statistician*, *40*(3), 249-251.

Folland, S., Goodman, A. and Stano, M., 2004. *The Economics of Health and Health Care*. New Jersey: Prentice Hall.

Fox-Rushby, J. and Cairns, J. eds, 2009. Economic Evaluation. London: Open University Press.

Frank, R., McGuire, T., Normand, S. and Goldman, H., 1999. The value of mental health care at the system level: the case of treating depression. *Health Affairs*, *18*(5), 71-88.

Franks, P., Muennig, P. and Gold, M., 2005. Is expanding Medicare coverage cost-effective?. *BMC Health Services Research*, *5*(1), 23.

Friedman, D., 2006. Assessing the potential of national strategies for electronic health records for population health monitoring and research. *Vital and health statistics. Series 2, Data evaluation and methods research*, (143), 1.

Fruehwirth, J., Navarro, S. and Takahashi, Y., 2011. How The Timing of Grade Retention Affects Outcomes: Identification and Estimation of Time-Varying Treatment Effects. *University of Western Ontario, CIBC Centre for Human Capital and Productivity Working Papers*.

Foster, V. and Young, A., 2012. The use of routinely collected patient data for research: A critical review. *Health:*, *16*(4), 448-463.

Gardner, M. and Altman, D., 1986. Confidence intervals rather than P values: estimation rather than hypothesis testing. *British Medical Journal, 292*(6522), 746.

Garen, J., 1984. The returns to schooling: A selectivity bias approach with a continuous choice variable. *Econometrica: Journal of the Econometric Society*, 1199-1218.

Givon, U., Ginsberg, G., Horoszowski, H. and Shemer, J., 1998. Cost-utility analysis of total hip arthroplasties: Technology assessment of surgical procedures by mailed questionnaires. *International Journal of Technology Assessment in Health Care*, *14*(04), 735-742.

Glazerman, S., Levy, D. and Myers, D., 2003. Nonexperimental versus experimental estimates of earnings impacts. *The Annals of the American Academy of Political and Social Science*, *589*(1), 63-93.

Glick, H., 2011a. Sample size and power for cost-effectiveness analysis (part 1). *Pharmacoeconomics*, *29*(3), 189-198.

Glick, H., 2011b. Sample size and power for cost-effectiveness analysis (part 2): the effect of maximum willingness to pay. *Pharmacoeconomics*, *29*(4), 287-296.

Glick, H., Doshi, J., Sonnad, S. and Polsky, D., 2007. *Economic evaluation in clinical trials*. Oxford: Oxford University Press.

Glynn, R., Schneeweiss, S. and Stürmer, T., 2006. Indications for propensity scores and review of their use in pharmacoepidemiology. *Basic & clinical pharmacology & toxicology*, *98*(3), 253-259.

Goeree, R., Bowen, J., Blackhouse, G., Lazzam, C., Cohen, E., Chiu, M., Hopkins, R., Tarride, J. and Tu, J., 2009. Economic evaluation of drug-eluting stents compared to bare metal stents using a large prospective study in Ontario. *International Journal of Technology Assessment in Health Care*, *25*(2), 196.

Gold, M., Siegel, J., Russell, L. and Weinstein, M. eds., 1996. *Cost-effectiveness in health and medicine*. Oxford University Press, USA.

Goodman, C. S., and Ahn, R., 1999. Methodological approaches of health technology assessment. *International journal of medical informatics*, *56*(1), 97-105.

Gray, A., Clarke, P., Wolstenholme, J. and Wordsworth, S., 2010. *Applied Methods of Cost-effectiveness Analysis in Healthcare*. Oxford: Oxford University Press.

Greenland, S., Robins, J. and Pearl, J., 1999. Confounding and collapsibility in causal inference. *Statistical Science*, 29-46.

Groeneveld, P., Farmer, S., Suh, J., Matta, M. and Yang, F., 2008. Outcomes and costs of implantable cardioverter-defibrillators for primary prevention of sudden cardiac death among the elderly. *Heart Rhythm*, *5*(5), 646-653.

Grieve, R., Porsdal, V., Hutton, J. and Wolfe, C., 2000. A comparison of the cost-effectiveness of stroke care provided in London and Copenhagen. *International Journal of Technology Assessment in Health Care*, *16*(2), 684-695.

Grieve, R., Sekhon, J., Hu, T. and Bloom, J., 2008. Evaluating health care programs by combining cost with quality of life measures: a case study comparing capitation and fee for service. *Health Services Research*, *43*(4), 1204-1222.

Griffin, S., Barber, J., Manca, A., Sculpher, M., Thompson, S., Buxton, M. and Hemingway, H., 2007. Cost effectiveness of clinically appropriate decisions on alternative treatments for angina pectoris: prospective observational study. *British Medical Journal*, *334*(7594), 624.

Gu, X. and Rosenbaum, P., 1993. Comparison of multivariate matching methods: Structures, distances, and algorithms. *Journal of Computational and Graphical Statistics*, *2*(4), 405-420.

Guinness, L. and Wiseman, V., 2011. *Introduction to health economics*. 2nd ed. London: Open University Press.

Gujarati, D., 2003. *Basic econometrics*. 4th *Mc Graw-Hills International edition*.

Guo, S., and Fraser, M., 2010. Propensity score analysis. *Statistical methods and applications*. SAGE Publications Inc.

Gyte, G., Dodwell, M., Newburn, M., Sandall, J., Macfarlane, A., Bewley, S., 2010. Safety of planned home births. Findings of meta-analysis cannot be relied on. *British Medical Journal*, 341:c4033.

Hahn, J., Todd, P. and Van der Klaauw, W., 2001. Identification and estimation of treatment effects with a regression-discontinuity design. *Econometrica*, *69*(1), 201-209.

Hall, J., 2003. Free standing maternity units in England. In: M. Kirkham, editor. *Birth centres*. Cheshire: Books for Midwives Press.

Hainmueller, J., 2012. Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. *Political Analysis*, *20*(1), 25-46.

Hainmueller, J. and Xu, Y., 2011. EBALANCE: Stata module to perform Entropy reweighting to create balanced samples. *Statistical Software Components*. Available at: http://ideas.repec.org/c/boc/bocode/s457326.html [Accessed 12 November 2011]

Hausman, J., 1978. Specification tests in econometrics. *Econometrica: Journal of the Econometric Society*, 1251-1271.

Hausman, J. and McFadden, D., 1984. Specification tests for the multinomial logit model. *Econometrica: Journal of the Econometric Society*, 1219-1240.

Heaton, P., Guo, J., Hornung, R., Johnston, J., Jang, R., Moomaw, C. and Cluxton, R., 2006. Analysis of the effectiveness and cost benefit of leukotriene modifiers in adults with asthma in the Ohio Medicaid population. *Journal of Managed Care Pharmacy. 12*(1), 33.

Heckman, J., 1979. Sample selection bias as a specification error. *Econometrica: Journal of the Econometric Society*, 153-161.

Heckman, J., 1992. Haavelmo and the birth of modern econometrics: A review of the history of econometric ideas by Mary Morgan. *Journal of Economic Literature*, 876-886.

Heckman, J., 1997. Instrumental variables: A study of implicit behavioral assumptions used in making program evaluations. *Journal of Human Resources*, 441-462.

Heckman, J., 2000. Causal parameters and policy analysis in economics: A twentieth century retrospective. *The Quarterly Journal of Economics*, *115*(1), 45-97.

Heckman, J., 2008. Econometric causality. *International Statistical Review*, *76*(1), 1-27.

Heckman, J., 2010. Building Bridges Between Structural and Program Evaluation Approaches to Evaluating Policy. *Journal of Economic Literature*, *48*(2), 356.

Heckman, J., Ichimura, H. and Todd, P., 1997. Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme. *The Review of Economic Studies*, *64*(4), 605-654.

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

Heckman, J. and Smith, J., 1995. Assessing the case for social experiments. *The Journal of Economic Perspectives*, *9*(2), 85-110.

Heckman, J., Smith, J. and Clements, N., 1997. Making the most out of programme evaluations and social experiments: Accounting for heterogeneity in programme impacts. *The Review of Economic Studies*, *64*(4), 487-535.

Heckman, J., Ichimura, H., Smith, J. and Todd, P., 1998. Characterizing Selection Bias Using Experimental Data. *Econometrica*, *66*(5), 1017-1098.

Heckman, J., LaLonde, R. and Smith, J., 1999. The Economics and Econometrics of Active Labor Market Programs. In: *Handbook of Labor Economics*, Volume 3A, pp. 1865-2097. Amsterdam: North-Holland.

Heckman, J. and Vytlacil, E., 2007. Econometric Evaluation of Social Programs, Part I: Causal Models, Structural Models and Econometric Policy Evaluation. In *Handbook of Econometrics*, Volume 6B, eds. J. Heckman and E. Leamer, 4779–4874. Amsterdam and Oxford: Elsevier, North-Holland.

Heitjan, D., 2000. Fieller's method and net health benefits. *Health Economics*, 9(4), 327-335.

HEED: Health Economic Evaluations Database, n.d. Available at http://heed.wiley.com/ohe/autolog.asp

Henderson, J. and Petrou, S., 2008. Economic implications of home births and birth centers: a structured review. - *Birth*, 35(2):136-46.

Hennekens, C. and Buring, J., 1987. *Epidemiology in Medicine.* Lippincott Williams and Wilkins.

Hernán, M., Alonso, A., Logan, R., Grodstein, F., Michels, K., Willett, W., Manson, J. and Robins, J., 2008. Observational studies analyzed like randomized experiments: an application to postmenopausal hormone therapy and coronary heart disease. *Epidemiology*, *19*(6), 766-779.

Hernán, M., Hernandez-Diaz, S. and Robins, J., 2004. A structural approach to selection bias. *Epidemiology*, *15*(5), 615-625.

Hernán, M. and Robins, J., 2006. Instruments for causal inference: an epidemiologist's dream? *Epidemiology*, *17*(4), 360-372.

Hernán M. and Robins J., 2013. *Causal Inference*. Chapman & Hall/CRC. Available at:http://www.hsph.harvard.edu/faculty/miguel-hernan/causal-inference-book/ [Accessed 23 July 2012].

Herron, M. and Wand, J., 2007. Assessing partisan bias in voting technology: The case of the 2004 New Hampshire recount. *Electoral Studies*, *26*(2), 247-261.

Hewitt, C., Torgerson, D. and Miles, J., 2006. Is there another way to take account of noncompliance in randomized controlled trials? Canadian Medical Association Journal, 75(4): 347.

Ho, D., Imai, K., King, G. and Stuart, E., 2007. Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis*, *15*(3), 199-236.

Hoch, J., 2009. Improving efficiency and value in palliative care with net benefit regression: an introduction to a simple method for cost-effectiveness analysis with person-level data. *Journal of Pain and Symptom Management*, *38*(1), 54.

Hoch, J. and Dewa, C., 2008. A clinician's guide to correct cost-effectiveness analysis: Think incremental not average. *Canadian Journal of Psychiatry*, *53*(4), 267-274.

Hoch, J., Briggs, A. and Willan, A., 2002. Something old, something new, something borrowed, something blue: a framework for the marriage of health econometrics and cost-effectiveness analysis. *Health Economics*, *11*(5), 415-430.

Hodnett, E.D., Downe, S., Walsh, D., Weston, J., 2010. Alternative versus conventional institutional settings for birth. Cochrane Database Of Systematic Reviews, 9:CD000012.

Hogan, J. and Lancaster, T., 2004. Instrumental variables and inverse probability weighting for causal inference from longitudinal observational studies. *Statistical Methods in Medical Research*, *13*(1), 17-48.

Holland, P., 1986. Statistics and causal inference. *Journal of the American Statistical Association*, *81*(396), 945-960.

Hollowell J., 2011. Birthplace programme overview: background, component studies and summary of findings. Birthplace in England research programme. Final report part 1: NIHR Service Delivery and Organisation programme.

Hollowell J., Puddicombe D., Rowe R., Linsell L., Hardy P., Stewart M., Redshaw, M., Newburn M., McCourt, C., Sandall, J., Macfarlane, A., Silverton, L. and Peter Brocklehurst on behalf of the Birthplace in England Collaborative Group, 2011. The Birthplace national prospective cohort study: perinatal and maternal outcomes by planned place of birth.Birthplace in England research programme. Final report part 4: NIHR Service Delivery and Organisation programme.

Hotz, V., Goerge, R., Balzekas, J. and Margolin, F., 1998. Administrative data for policy-relevant research: Assessment of current utility and recommendations for development. *Report of the Advisory Panel on Research Uses of Administrative Data of the Northwestern University/University of Chicago Joint Center for Poverty Research*.

Huber, P., 1967. The behavior of maximum likelihood estimates under nonstandard conditions. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*. Berkeley, CA: University of California Press, vol. 1, 221–233.

Huber, P. and Ronchetti, E., 2009. *Robust Statistics*, Wiley: New York.

Hutchings, H., Cheung, W., Williams, J., Cohen, D., Longo, M. and Russell, I., 2005. Can electronic routine data act as a surrogate for patient-assessed outcome measures? *International Journal of Technology Assessment in Health Care*, *21*(1), 138-143.

Iacus, S., King, G. and Porro, G., 2009. cem: Software for coarsened exact matching. Journal of Statistical Software 30:9.

Iacus, S., King, G. and Porro, G., 2011. Multivariate matching methods that are monotonic imbalance bounding. *Journal of the American Statistical Association*, *106*(493), 345-361.

Iacus, S., King, G. and Porro, G., 2012. Causal inference without balance checking: Coarsened exact matching. *Political Analysis*, *20*(1), 1-24.

Imai, K., King, G. and Stuart, E., 2008. Misunderstandings between experimentalists and observationalists about causal inference. *Journal of the Royal Statistical Society: Series A (statistics in society)*, *171*(2), 481-502.

Imbens, G. and Angrist, J., 1994. Identification and estimation of local average treatment effects. *Econometrica: Journal of the Econometric Society*, 467-475.

Imbens, G., 2000. The role of the propensity score in estimating dose-response functions. *Biometrika*, 87, 706-710

Imbens, G. and Lemieux, T., 2008. Regression discontinuity designs: A guide to practice. *Journal of Econometrics*, *142*(2), 615-635.

Imbens, G., 2004. Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and statistics*, *86*(1), 4-29.

Imbens, G. and Rubin, D., 2008. Rubin Causal Model. In: Durlauf, S. and Blume, L. eds. *The New Palgrave Dictionary of Economics*. Macmillan Publishers Ltd.

Imbens, G. and Wooldridge, J., 2009. Recent Developments in the Econometrics of Program Evaluation. *Journal of Economic Literature*, 5-86.

Indurkhya, A., Mitra, N. and Schrag, D., 2006. Using propensity scores to estimate the cost–effectiveness of medical therapies. *Statistics in Medicine*, *25*(9), 1561-1576.

Ioannidis, J., 2005. Why most published research findings are false. *PLoS medicine*, *2*(8), e124.

Ioannidis, J., Haidich, A., Pappa, M., Pantazis, N., Kokori, S. I., Tektonidou, M. Contopoulos-Ioannidis, D. and Lau, J, 2001. Comparison of evidence of treatment effects in randomized and nonrandomized studies. *JAMA: the journal of the American Medical Association*, *286*(7), 821-830.

Janssen, P.A., Saxell, L., Page, L.A., Klein, M.C., Liston, R.M., Lee, S.K., 2009. Outcomes of planned home birth with registered midwife versus planned hospital birth with midwife or physician. Canadian Medical Association Journal. 181(6-7):377-83.

Jones, A., 2000. Health econometrics. In: Culyer A. and Newhouse, J. eds. *Handbook of Health Economics*, Elsevier, Volume 1, Part A, 265-344.

Jones, A., 2007a. *Applied econometrics for health economists: a practical guide*. Radcliffe Publishing.

Jones, A., 2007b. Identification of treatment effects in Health Economics. *Health Economics*, *16*(11), 1127-1131.

Jones, A., 2011. Panel Data Methods and Applications to Health Economics. In: *Palgrave Handbook of Econometrics*, Volume 2: Applied Econometrics, eds. T. Mills and K. Petterson. Palgrave Macmillan.

Jones A. and Rice N., 2011. Econometric evaluation of health policies. In Glied S. and Smith P., eds. *The Oxford Handbook of Health Economics*. Oxford: Oxford University Press.

Kariv, Y., Delaney, C., Senagore, A., Manilich, E., Hammel, J., Church, J., Ravas, J. and Fazio, V., 2007. Clinical outcomes and cost analysis of a "fast track" postoperative care pathway for ileal pouch-anal anastomosis. A case control study. *Diseases of the Colon and Rectum*, *50*(2), 137-146.

Kelly, M. P., McDaid, D., Ludbrook, A., and Powell, J., 2005. *Economic appraisal of public health interventions*. London: Health Development Agency.

Kennedy, P., 2008. *A guide to econometrics*. 6th ed. Boston: MIT press.

Knapp, M., Windmeijer, F., Brown, J., Kontodimas, S., Tzivelekis, S., Haro, J., Ratcliffe, M., Hong, J. and Novick, D., 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(4): 341-358.

Kreif, N., Grieve,R., Sadique, Z., 2013. Statistical methods in cost-effectiveness analyses that use observational data: a critical appraisal of current practice. *Health Economics,* 22(4):486-500.

Kreif, N., Grieve, R., Radice, R. and Sekhon, J., 2012a. Regression-adjusted matching and double-robust methods for estimating average treatment effects in health economic evaluation. Available at: http://www.lshtm.ac.uk/php/hsrp/reducing-selection-bias/output/publications.html [Accessed 12 June 2013]

Kreif, N., Grieve, R., Radice, R., Sadique, Z., Ramsahai, R., Sekhon, J., 2012b. Methods for estimating subgroup effects in cost-effectiveness analyses that use observational data. *Medical Decision Making*, 32(6):750-63.

Lairson, D., Yoon, S., Carter, P., Greisinger, A., Talluri, K., Aggarwal, M. and Wehmanen, O., 2008. Economic evaluation of an intensified disease management system for patients with type 2 diabetes. *Disease Management*, *11*(2), 79-94.

LaLonde, R., 1986. Evaluating the econometric evaluations of training programs with experimental data. *The American Economic Review*, 604-620.

Lawlor, D., Harbord, R., Sterne, J., Timpson, N. and Davey Smith, G., 2008. Mendelian randomization: using genes as instruments for making causal inferences in epidemiology. *Statistics in Medicine*, *27*(8), 1133-1163.

Lechner, M., 2001. Identification and estimation of causal effects of multiple treatments under the conditional independence assumption. *Econometric Evaluation of Labour Market Policies*, 43-58.

Lechner, M., 2002. Some practical issues in the evaluation of heterogeneous labour market programmes by matching methods. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *165*(1), 59-82.

Lee D. and Lemieux, T., 2010. Regression Discontinuity Designs in Economics. *Journal of Economic Literature*, vol. 48(2), pages 281-355.

Leuven E. and B. Sianesi. B., 2003. PSMATCH2: Stata module to perform full Mahalanobis and propensity score matching, common support graphing, and covariate imbalance testing. Available at: http://ideas.repec.org/c/boc/bocode/s432001.html. [Accessed 14 February 2011]

Linden, A., Adams, J. and Roberts, N., 2005. Using propensity scores to construct comparable control groups for disease management program evaluation. *Disease Management and Health Outcomes*, *13*(2), 107-115.

Linden, A. and Adams, J., 2012. Combining the regression discontinuity design and propensity score-based weighting to improve causal inference in program evaluation. *Journal of Evaluation in Clinical Practice*. 18(2),: 317–325.

Lindgren, H.E., Radestad, I.J., Christensson, K., Hildingsson, I.M., 2008. Outcome of planned home births compared to hospital births in Sweden between 1992 and 2004. A population-based register study. Acta Obstetricia et Gynecologica Scandinavica, 87(7):751-9.

Lunn, D., Thomas, A., Best, N. and Spiegelhalter, D., 2000. WinBUGS-a Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and computing*, *10*(4), 325-337.

Macfarlane A., Mugford M., Henderson J., Furtado A., Stevens J. and Dunn A., 2000. *Birth counts: statistics of pregnancy and childbirth*. Volume 2, Tables. 2nd edition. London: The Stationery Office.

MacMahon, S. and Collins, R., 2001. Reliable assessment of the effects of treatment on mortality and major morbidity, II: observational studies. *The Lancet*, 357(9254), 455-462.

Maddala, G., 1983. *Limited Dependent and Qualitative Variables in Econometrics*. Cambridge: Cambridge University Press.

Manca, A. and Austin, P., 2008. Using propensity score methods to analyse individual patient level cost effectiveness data from observational studies (No. 08/20). HEDG, c/o Department of Economics, University of York.

Maniadakis, N. and Gray, A., 2000. Health economics and orthopaedics. *Journal of Bone and Joint Surgery*, 82-B:2-8.

Manning, W., 1998. The logged dependent variable, heteroscedasticity, and the retransformation problem. *Journal of Health Economics*, *17*(3), 283-296.

Manski, C., 2007. *Identification for Prediction and Decision*. Cambridge: Harvard University Press.

Marsh, J., Hutton, J. and Binks, K., 2002. Removal of radiation dose response effects: an example of over-matching. *British Medical Journal. 325*(7359), 327-330.

McCarron, C., Pullenayegum, E., Marshall, D., Goeree, R. and Tarride, J., 2009. Handling uncertainty in economic evaluations of patient level data: A review of the use of Bayesian methods to inform health technology assessments. *International Journal of Technology Assessment in Health Care*, *25*(4), 546.

McClellan, M. and Newhouse, J., 1997. The marginal cost-effectiveness of medical technology: a panel instrumental-variables approach. *Journal of Econometrics*, *77*(1), 39-64.

McCourt, C., Rance, S., Rayment J., Sandall J., 2011. Birthplace qualitative organisational case studies: How maternity care systems may affect the provision of care in different birth settings. Birthplace in England research programme. Final report part 6. NIHR Service Delivery and Organisation programme; 2011.

McGuinness, D., Bennett, S. and Riley, E., 1997. Statistical analysis of highly skewed immune response data. *Journal of Immunological Methods*, 201:99–114.

McIntosh, E., Clarke, P., Frew, E.,and Louviere, J. eds, 2010. *Applied methods of cost-benefit analysis in health care*. Oxford: Oxford University Press.

Meenan, R., Goodman, M., Fishman, P., Hornbrook, M., O'Keeffe-Rosetti, M. and Bachman, D., 2002. Issues in pooling administrative data for economic evaluation. *The American Journal of Managed Care*, *8*(1), 45.

Merito, M. and Pezzotti, P., 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), 30-36.

Meyer, B., 1995. Natural and Quasi-Experiments in Economics. *Journal of Business and Economic Statistics*. 13:2, pp. 151– 61.

Miguel, E. and Kremer, M., 2004. Worms: Identifying impacts on education and health in the presence of treatment externalities. *Econometrica*, 72: 159-217.

Mihaylova, B., Briggs, A., Armitage, J., Parish, S., Gray, A. and Collins, R., 2006. Lifetime cost effectiveness of simvastatin in a range of risk groups and age groups derived from a randomised trial of 20,536 people. *British Medical Journal*, *333*(7579), 1145.

Mihaylova, B., Pitman, R., Tincello, D., Van Der Vaart, H., Tunn, R., Timlin, L., Quail, D., Johns, A. and Sculpher, M., 2010. Cost-effectiveness of duloxetine: the Stress Urinary Incontinence Treatment (SUIT) study. *Value in Health*, *13*(5), 565-572.

Mitra, N. and Indurkhya, A., 2005. A propensity score approach to estimating the cost–effectiveness of medical therapies from observational data. *Health Economics*,

14(8), 805-815.

Mojtabai, R. and Zivin J., 2003. Effectiveness and Cost-effectiveness of Four Treatment Modalities for Substance Disorders: A Propensity Score Analysis. *Health Services Research*, *38*(1p1), 233-259.

Moreno-Serra, R., 2007. *Matching estimators of average treatment effects: a review applied to the evaluation of health care programmes* (No. 07/02). HEDG, c/o Department of Economics, University of York.

Mori, R., Dougherty, M., Whittle, M., 2008. An estimation of intrapartum-related perinatal mortality rates for booked home births in England and Wales between 1994 and 2003. *BJOG*, 115(5):554-9.

Morgan, S. and Harding, D., 2006. Matching Estimators of Causal Effects Prospects and Pitfalls in Theory and Practice. *Sociological Methods and Research*, *35*(1), 3-60.

Mossialos, E., Dixon, A., Figueras, J. and Kutzin, S. eds., 2002. *Funding health care: Options for Europe*. Buckingham, UK: Open University Press.

Muennig, P., Franks, P. and Gold, M., 2005. The cost effectiveness of health insurance. *American Journal of Preventive Medicine*, *28*(1), 59-64.

Mullahy, J., 1998. Much ado about two: reconsidering retransformation and the two-part model in health econometrics. *Journal of Health Economics*, *17*(3), 247-281.

Musgrove, P. and Fox-Rushby, J., 2006. Cost-effectiveness analysis for priority setting. In: Jamison, D., Breman, J., Measham, A., Alleyne, G., Claeson, M., Evans, D. and

Musgrove, P. eds. *Disease control priorities in developing countries*. Oxford University Press, USA. Ch 15.

National Institute for Health and Clinical Excellence, 2007. *Intrapartum care care of healthy women and their babies during childbirth.* National Collaborating Centre for Women's and Children's Health.

National Institute for Health and Clinical Excellence, 2008. *Guide to the Methods of Technology Appraisal*. London: NICE.

Neyman, J., 1923. Statistical Problems in Agricultural Experiments. *Journal of the Royal Statistical Society II* (S2): 107–80.

NHS EED, n.d. Available at http://www.york.ac.uk/inst/crd/

Nixon, R. and Thompson, S., 2004. Parametric modelling of cost data in medical studies. *Statistics in Medicine*, *23*(8), 1311-1331.

Nixon, R. and Thompson, S., 2005. Methods for incorporating covariate adjustment, subgroup analysis and between-centre differences into cost-effectiveness evaluations. *Health Economics*, *14*(12), 1217-1229.

Nixon, R., Wonderling, D. and Grieve, R., 2010. Non-parametric methods for cost-effectiveness analysis: the central limit theorem and the bootstrap compared. *Health Economics*, *19*(3), 316-333.

Norton, E. and Han, E., 2008. Genetic information, obesity, and labor market outcomes. *Health Economics*, *17*(9), 1089-1104.

Nuijten, M., 1998. The selection of data sources for use in modelling studies. *Pharmacoeconomics*, *13*(3), 305-316.

O'Hagan, A. and Stevens, J., 2001b. A framework for cost-effectiveness analysis from clinical trial data. *Health Economics*, *10*(4), 303-315.

O'Hagan, A., Stevens, J. and Montmartin, J., 2001a. Bayesian cost-effectiveness analysis from clinical trial data. *Statistics in Medicine*, *20*(5), 733-753.

O'Sullivan, A., Thompson, D. and Drummond, M., 2005. Collection of health-economic data alongside clinical trials: is there a future for piggyback evaluations?. *Value in Health*, *8*(1), 67-79.

Olsen, O., 1997 Meta-analysis of the safety of home birth. *Birth*, 24(1):4-13.

Olsen, O. and Jewell M.D., 1998. Home versus hospital birth. *Cochrane Database of Systematic Reviews*, (3):CD000352.10.

OvidSP, n.d. Available at: http://ovidsp.ovid.com/

Papay, J., Willett, J. and Murnane, R., 2011. Extending the regression-discontinuity approach to multiple assignment variables. *Journal of Econometrics*, *161*(2), 203-207.

Petrou, S., and Gray, A., 2011. Economic evaluation using decision analytical modelling: design, conduct, analysis, and reporting. *British Medical Journal*, *342*.

Pocock, S. and Elbourne, D., 2000. Randomized trials or observational tribulations?. *New England Journal of Medicine*, 342(25), 1907-1909.

Polignano, F., Quyn, A., de Figueiredo, R., Henderson, N., Kulli, C. and Tait, I., 2008. Laparoscopic versus open liver segmentectomy: prospective, case-matched, intention-to-treat analysis of clinical outcomes and cost effectiveness. *Surgical Endoscopy*, *22*(12), 2564-2570.

Polsky, D. and Basu, A., 2006. Selection bias in observational data. In: Jones, A. ed. *The Elgar Companion to Health Economics*. Edware Elgar Publishing.

Polsky, D., Mandelblatt, J., Weeks, J., Venditti, L., Hwang, Y., Glick, H., Hadley, J. and Schulman, K., 2003. Economic evaluation of breast cancer treatment: considering the value of patient choice. *Journal of Clinical Oncology*, *21*(6), 1139-1146.

Prentice, R., Rossouw, J., Furberg, C., Johnson, S., Henderson, M., Cummings, S., Manson, J., Freedman, L., Oberman, A., Kuller, L. and Anderson, G., 1998. Design of the WHI Clinical Trial and Observational Study. *Control Clinical Trials*, 19:61-109.

Quinn, C., 2005. Generalisable regression methods for cost-effectiveness using copulas. *Health, Econometrics and Data Group Working Paper WP 05/13, University of York*.

Quinn, C., 2007. *Improving precision in cost-effectiveness analysis using copulas* (No. 07/23). HEDG, Department of Economics, University of York.

Radice, R., Ramsahai, R., Grieve, R., Kreif, N., Sadique, Z. and Sekhon, J.S., 2012. Evaluating treatment effectiveness in patient subgroups: a comparison of propensity score methods with an automated matching approach. *The International Journal of Biostatistics*, 8(1):25. doi: 10.1515/1557-4679.1382.

Raessler, S. and Rubin, D., 2005. Complications when using nonrandomized job training data to draw causal inferences. Proceedings of the International Statistical Institute.

Ramsey, J.B., 1969. Tests for specification errors in classical linear least-squares regression analysis. *Journal of the Royal Statistical Society*. Series B (Methodological), 31(2):350-371.

Ramsey S., Willke R., Briggs A., Brown R., Buxton M., Chawla A., Cook J, Glick H., Liljas B., Petitti D. and Reed S., Good Research Practices for Cost-Effectiveness Analysis Alongside Clinical Trials: The ISPOR RCT-CEA Task Force Report, *Value in Health*, Volume 8, Issue 5.

Reiss, P. and Wolak, F., 2007. Structural econometric modeling: Rationales and examples from industrial organization. *Handbook of Econometrics*, *6*, 4277-4415.

Robbins L., 1932. *An Essay on the Nature and Significance of Economic Science*, London: Macmillan.

Robins, J., 2002. Comment on "Covariance adjustment in randomised experiments and observational studies". *Statistical Science*, 17, 309-321.

Robins, J., Hernán, M. and Brumback, B., 2000. Marginal structural models and causal inference in epidemiology. *Epidemiology*, *11*(5), 550-560.

Robins, J., Rotnitzky, A. and Zhao, L., 1994. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, *89*(427), 846-866.

Rosenbaum, P., 2002. *Observational studies*. 2nd ed. Springer.

Rosenbaum, P. and Rubin, D., 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*(1), 41-55.

Rosenbaum, P. and Rubin, D., 1984. Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association*, 79(387), 516-524.

Rosenbaum, P., and Rubin, D., 1985a. The bias due to incomplete matching. *Biometrics*, 103-116.

Rosenbaum, P. and Rubin, D., 1985b. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician*, *39*(1), 33-38.

Rovithis, D., 2006. Health economic evaluation in Greece. *International journal of technology assessment in health care*, *22*(3), 388-395.

Rovithis, D., 2009. Health economic evaluation: in need of more analytical rigor or more practical relevance? *Expert Review of Pharmacoeconomics and Outcomes Research*, *9*(2), 107.

Rowe, R., Fitzpatrick, R., Hollowell, J. and Kurinczuk, J., 2012. Transfers of women planning birth in midwifery units: data from the Birthplace prospective cohort study. *BJOG: An International Journal of Obstetrics and Gynaecology*.

Roy, A., 1951. Some thoughts on the distribution of earnings. *Oxford economic papers*, 135-146.

Royston, P., 1992. Approximating the Shapiro-Wilk W-Test for non-normality. *Statistics and Computing*, 2:117-119.

Rubin, D., 1973. Matching to Remove Bias in Observational Studies. *Biometrics*, 29(1): 159–83.

Rubin, D., 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, *66*(5), 688.

Rubin, D., 1976. Inference and missing data. *Biometrika*, *63*(3), 581-592.

Rubin, D., 1977. Assignment to Treatment Group on the Basis of a Covariate. *Journal of Educational and Behavioral Statistics*, *2*(1), 1-26.

Rubin, D., 1980. Bias reduction using Mahalanobis-metric matching. *Biometrics*, 293-298.

Rubin, D., 1986. Statistics and causal inference: Comment: Which ifs have causal answers. *Journal of the American Statistical Association*, 81(396), 961-962.

Rubin, D., 1997. Estimating Causal Effects from Large Data Sets Using Propensity Scores. *Ann Intern Med.* 1997;127(8:2):757-763.

Rubin, D., 2006. *Matched Sampling for Causal Inference*. Cambridge: Cambridge University. Press.

Rubin, D., 2007. The design versus the analysis of observational studies for causal effects: Parallels with the design of randomized trials. *Statistics in Medicine*. 26 20–36.

Rubin, D. and Thomas, N., 2000. Combining propensity score matching with additional adjustments for prognostic covariates. *Journal of the American Statistical Association*, *95*(450), 573-585.

Sadhu, A., Ang, A., Ingram-Drake, L., Martinez, D., Hsueh, W. and Ettner, S., 2008. Economic Benefits of Intensive Insulin Therapy in Critically Ill Patients The Targeted Insulin Therapy to Improve Hospital Outcomes (TRIUMPH) Project. *Diabetes Care*, *31*(8), 1556-1561.

Schroeder, E., Petrou, S., Patel, N., Hollowell, J., Puddicombe, D., Redshaw, M. and Brocklehurst, P. on behalf of the Birthplace in England Collaborative Group, 2011. Birthplace cost-effectiveness analysis of planned place of birth: individual level analysis. Birthplace in England research programme. Final report part 5. NIHR Service Delivery and Organisation programme.

Schroeder, E., Petrou, S., Patel, N., Hollowell, J., Puddicombe, D., Redshaw, M. and Brocklehurst, P., 2012. Cost effectiveness of alternative planned places of birth in woman at low risk of complications: evidence from the Birthplace in England national prospective cohort study. *British Medical Journal*, *344*.

Sculpher, M., Claxton, K., Drummond, M. and McCabe, C., 2006. Whither trial-based economic evaluation for health care decision making?. *Health economics*, *15*(7), 677-687.

Sculpher M., Pang F., Manca A., Drummond M., Golder S., Urdahl H., Davies, L. and Eastwood, A., 2004. Generalisability in economic evaluation studies in healthcare: a review and case studies. *Health Technology Assessment*, 8(49).

Sekhon, J. Alternative balance metrics for bias reduction in matching methods for causal inference. Survey Research Center, University of California, Berkeley. Available at: http://sekhon.berkeley.edu/papers/SekhonBalanceMetrics.pdf [Accessed 21 February 2010].

Sekhon, J., 2009. Opiates for the matches: Matching methods for causal inference. *Annual Review of Political Science*, *12*, 487-508.

Sekhon, J. and Grieve, R., 2009. *A Nonparametric Matching Method for Covariate Adjustment with Application to Economic Evaluation*. [online] Available at: http://sekhon.berkeley.edu/papers/GeneticMatching_SekhonGrieve.pdf [Accessed 17 March 2010].

Sekhon, J. and Grieve, R., 2012. A matching method for improving covariate balance in cost-effectiveness analyses. Health Economics, 21(6): 695–714.

Sekhon, J. and Mebane, W., 1998. Genetic Optimization Using Derivatives: Theory and Application to Nonlinear Models. *Political Analysis* 7: 189–203

Sequist T. and Bates D., 2010. Developing information technology capacity for performance measurement. In: *Performance Measurement for Health System Improvement*. P. Smith, E. Mossialos, I. Papanicolas and S. Leatherman eds. Cambridge: Cambridge University Press.

Shah, B., Laupacis, A., Hux, J. and Austin, P., 2005. Propensity score methods gave similar results to traditional regression modeling in observational studies: a systematic review. *Journal of Clinical Epidemiology*, *58*(6), 550.

Shapiro, S. and Wilk, M.B., 1965. An analysis of variance test for normality. *Biometrika*;52:591–611.

Shi, L., Wu, E., Hodges, M., Yu, A. and Birnbaum, H., 2007. Retrospective economic and outcomes analyses using non-US databases: A review. *PharmacoEconomics*, *25*(7), 563-576.

Shih, Y., Bekele, N. and Xu, Y., 2007. Use of Bayesian net benefit regression model to examine the impact of generic drug entry on the cost effectiveness of selective serotonin reuptake inhibitors in elderly depressed patients. *Pharmacoeconomics*, *25*(10), 843-862.

Shireman, T. and Braman, K., 2002. Impact and cost-effectiveness of respiratory syncytial virus prophylaxis for Kansas medicaid's high-risk children. *Archives of Pediatrics and Adolescent Medicine*, *156*(12), 1251.

Shpitser, I., VanderWeele, T. and Robins, J., 2012. On the validity of covariate adjustment for estimating causal effects. *arXiv preprint arXiv:1203.3515*.

Sica, G., 2006. Bias in Research Studies. *Radiology*, *238*(3), 780-789.

Sianesi, B., 2004. An Evaluation of the Active Labour Market Programmes in Sweden. *The Review of Economics and Statistics*, 86(1), 133{155.

Smith, H., 1997. Matching with multiple controls to estimate treatment effects in observational studies. *Sociological Methodology* 27:325–353.

Smith, J., 2000. *A critical survey of empirical methods for evaluating active labor market policies*. Department of Economics, University of Western Ontario.

Smith, J. and Todd, P., 2001. Reconciling conflicting evidence on the performance of propensity-score matching methods. *The American Economic Review*, *91*(2), 112-118.

Smith, J. and Todd, P., 2005. Does matching overcome LaLonde's critique of nonexperimental estimators? *Journal of Econometrics*, *125*(1), 305-353.

Soegaard, R., Bünger, C., Christiansen, T. and Christensen, F., 2007. Determinants of cost-effectiveness in lumbar spinal fusion using the net benefit framework: a 2-year follow-up study among 695 patients. *European Spine Journal*, *16*(11), 1822-1831.

Sørensen, H. and Gillman, M., 1995. Matching in case-control studies. *British Medical Journal*, *310*(6975), 329.

Sorenson, C., Drummond, M. and Kanavos, P., 2008. *Ensuring value for money in health care: the role of health technology assessment in the European Union* (No. 11). WHO Regional Office Europe.

Soto, J., 2002. Health economic evaluations using decision analytic modeling. *International Journal of Technology Assessment in Health Care*, *18*(1), 94-111.

Spaepen, E., Demarteau, N., Van Belle, S. and Annemans, L., 2008. Health economic evaluation of treating anemia in cancer patients receiving chemotherapy: a study in Belgian hospitals. *The Oncologist*, *13*(5), 596-607.

Spielauer, M., 2007. Dynamic microsimulation of health care demand, health care finance and the economic impact of health behaviours: survey and review. *International Journal of Microsimulation*, *1*(1), 35-53.

Staiger, D. and Stock, J., 1997. Instrumental Variables Regression with Weak Instruments. *Econometrica: Journal of the Econometric Society*, 557-586.

StataCorp. 2009. *Stata Statistical Software: Release 11*. College Station, TX: StataCorp LP.

Stevens, W. and Normand, C., 2004. Optimisation versus certainty: understanding the issue of heterogeneity in economic evaluation. *Social Science and Medicine*, *58*(2), 315-320.

Stevens, A., Raftery, J. and Roderick, P., 2005. Can health technologies be assessed using routine data?. *International Journal of Technology Assessment in Health Care*, *21*(1), 96-103.

Stinnett, A. and Mullahy, J., 1998. Net health benefits a new framework for the analysis of uncertainty in cost-effectiveness analysis. *Medical Decision Making*, *18*(2), S68-S80.

Stuart, E., 2010. Matching methods for causal inference: A review and a look forward. *Statistical Science*, *25*(1), 1.

Stürmer, T., Joshi, M., Glynn, R., Avorn, J., Rothman, K. and Schneeweiss, S., 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. *Journal of Clinical Epidemiology*, *59*(5), 437.

Tannen, R., Weiner, M. and Xie, D., 2009. Use of primary care electronic medical record database in drug efficacy research on cardiovascular outcomes: comparison of database and randomised controlled trial findings. *British Medical Journal*. 338.

Terza, J., Basu, A. and Rathouz, P., 2008. Two-stage residual inclusion estimation: addressing endogeneity in health econometric modeling. *Journal of Health Economics*, *27*(3), 531.

Terza, J., Bradford, W. and Dismuke, C., 2007. The use of linear instrumental variables methods in health services research and health economics: a cautionary note. *Health Services Research*, *43*(3), 1102-1120.

Todd, P., 2008. Matching estimators. In: *The New Palgrave Dictionary of Economics*. 2nd edition. eds. S. Durlauf and L. Blume. Palgrave Macmillan.

Todd, P. and Wolpin, K., 2008. Ex ante evaluation of social programs. *Annales d'Economie et de Statistique*, 263-291.

Uysal, S., 2011. Doubly Robust IV Estimation of the Local Average Treatment Effects. Available at: http://www.ihs.ac.at/vienna/resources/Economics/Papers/Uysal_paper.pdf [Accessed 11 October 2012].

Uysal, S., 2012. Doubly Robust Estimation of Causal Effects with Multivalued Treatments. Available at: https://espe.conference-ervices.net/resources/321/2907/pdf/ESPE2012_0221_paper.pdf [Accessed 20 November 2012].

Van Hout, B., Al, M., Gordon, G. and Rutten, F., 1994. Costs, effects and c/e-ratios alongside a clinical trial. *Health economics*, *3*(5), 309-319.

Van Der Klaauw, W., 2005. [The Sensitivity of Economic Statistics to Coding Errors in Personal Identifiers]: Comment. *Journal of Business and Economic Statistics*, 154-157.

Varadhan R, Seeger J. Estimation and reporting of heterogeneity of treatment effects. In: Velentgas P, Dreyer NA, Nourjah P, et al., eds. Developing a Protocol for Observational Comparative Effectiveness Research: A User's Guide. AHRQ Publication No. 12(13)-EHC099. Rockville, MD: Agency for Healthcare Research and Quality; January 2013: Chapter 3, pp. 35-44.

Von Elm, E., Altman, D., Egger, M., Pocock, S., Gøtzsche, P. and Vandenbroucke, J., 2007. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. *Preventive Medicine*, *45*(4), 247-251.

Wagstaff, A. and Culyer, A., 2012. Four decades of health economics through a bibliometric lens. *Journal of Health Economics*, *31*(2), 406-439.

Walsh, D. and Downe, S.M., 2004. Outcomes of free-standing, midwife-led birth centers: a structured review. *Birth*, 31(3):222-9.

Wax, J.R., Lucas, F.L., Lamont, M., Pinette, M.G., Cartin, A., Blackstone, J., 2010. Maternal and newborn outcomes in planned home birth vs planned hospital births: a meta-analysis. *American Journal of Obstetrics and Gynecology*., 203(3):243 e1-8.

Wehby, G, Ohsfeldt, R. and Murray, J., 2008. 'Mendelian randomization'equals instrumental variable analysis with genetic instruments. *Statistics in Medicine*, *27*(15), 2745-2749.

Wedderburn, R., 1974, Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method. *Biometrika*. 61, 439-447.

Weeks, M., 1997. The multinomial probit model revisited: A discussion of parameter estimability, identification and specification testing. *Journal of Economic Surveys*, *11*(3), 297-320.

Weinstein, M., O'Brien, B., Hornberger, J., Jackson, J., Johannesson, M., McCabe, C. and Luce, B., 2003. Principles of good practice for decision analytic modeling in health-care evaluation: report of the ISPOR Task Force on Good Research Practices—Modeling Studies. *Value in Health*, *6*(1), 9-17.

Weinstein, M., Siegel, J., Gold, M., Kamlet, M. and Russell, L., 1996. Recommendations of the Panel on Cost-effectiveness in Health and Medicine. *JAMA: the journal of the American Medical Association*, *276*(15), 1253-1258.

Weiss, J., Saynina, O., McDonald, K., McClellan, M. and Hlatky, M., 2002. Effectiveness and cost-effectiveness of implantable cardioverter defibrillators in the treatment of ventricular arrhythmias among Medicare beneficiaries. *The American Journal of Medicine*, *112*(7), 519-527.

Wells, B., Nowacki, A., Chagin,, K. and Kattan, M., 2013. Strategies for Handling Missing Data in Electronic Health Record Derived Data. *eGEMs (Generating Evidence & Methods to improve patient outcomes)*, Vol. 1: Iss. 3, Article 7.

White, H., 1980. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica* 48: 817–830.

Wickramaratne, P., 1995. Sample size determination in epidemiologic studies. *Statistical Methods in Medical Research*, *4*(4), 311-337.

Willan, A. and Briggs, A., 2006. *Statistical analysis of cost-effectiveness data*. Wiley.

Willan, A. and O'Brien, B., 1996. Confidence intervals for cost-effectiveness ratios: An application of Fieller's theorem. *Health Economics*, *5*(4), 297-305.

Willan, A., Briggs, A. and Hoch, J., 2004. Regression methods for covariate adjustment and subgroup analysis for non-censored cost-effectiveness data. *Health economics*, *13*(5), 461-475.

Willan, Lin, Manca,. 2005. Regression methods for cost-effectiveness analysis with censored data. *Statistics in Medicine*;24:131-45.

Williams J., Cheung W., Cohen D., Hutchings H., Longo M. and Russell I.. Can randomised trials rely on existing electronic data? A feasibility study to explore the value of routine data in health technology assessment. *Health Technology Assessement*. 7 (26).

Windmeijer, F., Kontodimas, S., Knapp, M., Brown, J. and Haro, J., 2006. Methodological approach for assessing the cost-effectiveness of treatments using longitudinal observational data: the SOHO study. *International Journal of Technology Assessment in Health Care*, *22*(4), 460-468.

Wooldridge, J., 1997. On two stage least squares estimation of the average treatment effect in a random coefficient model. *Economics Letters*, *56*(2), 129-133.

Wooldridge, J., 2007. Inverse probability weighted estimation for general missing data problems. *Journal of Econometrics*, 141:1281–1301.

Wooldridge, J., 2009. *Introductory econometrics: A modern approach*. 4th ed. South-Western Pub.

Wooldridge, J., 2010. *Econometric Analysis of Cross Section and Panel Data*. 2nd ed. Cambridge, MA: MIT Press

Zellner, A., 1962. An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias. *Journal of the American Statistical Association*, *57*(298), 348-368.

Zohoori, N. and Savitz, D., 1997. Econometric approaches to epidemiologic data: relating endogeneity and unobserved heterogeneity to confounding. *Annals of Epidemiology*, *7*(4), 251.

Zethraeus, N. and Löthgren, M., 2000. On the equivalence of the net benefit and the Fieller's methods for statistical inference in cost-effectiveness analysis. *Stockholm School of Economics Working Paper Series in Economics and Finance*.

Zhao, Z., 2004. Using matching to estimate treatment effects: data requirements, matching metrics, and Monte Carlo evidence. *Review of Economics and Statistics*, *86*(1), 91-107.