STATISTICAL METHODS TO ADDRESS SELECTION BIAS IN ECONOMIC EVALUATIONS THAT USE PATIENT-LEVEL OBSERVATIONAL DATA

Noémi Kreif

London School of Hygiene and Tropical Medicine

Faculty of Public Health and Policy

Thesis submitted to the University of London for Doctor of Philosophy degree

I, Noémi Kreif confirm that the work presented in this thesis is my own.

Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Signature:_____ Date: _____

Abstract

This thesis compares statistical methods for addressing selection bias in cost-effectiveness analyses (CEA) that use observational data. The thesis has four objectives: (1) to critically appraise currently recommended statistical methods, (2) to consider alternative statistical methods for CEA, (3) to compare propensity score (PS) approaches and Genetic Matching (GM) for estimating subgroup-effects in CEA, and (4) to compare methods that combine regression with PS approaches, for CEA.

I developed a new checklist for critically appraising statistical methods for addressing selection bias in CEA, and applied it in a systematic review of published CEA. Most studies used regression or matching methods, and did not assess their underlying assumptions, such as the correct specification of the PS or the endpoint regression model.

I identified methods that can make less restrictive assumptions: GM, a multivariate matching method that can directly balance covariates, double-robust (DR) methods, regression-adjusted matching, and machine learning estimation of the PS and the endpoint regression. I compared these methods across a range of typical CEA circumstances, using simulations and case studies.

In the first case study, where cost-effectiveness estimates for subgroups were of interest, I found that the cost-effectiveness results differed according to the statistical approach. The accompanying simulation study found that GM was relatively robust to the misspecification of the PS, and provided the least biased and most precise estimates of cost-effectiveness for each subgroup.

The second simulation study considered DR methods and regression-adjusted matching for estimating overall cost-effectiveness and found that regression-adjusted matching was relatively robust to misspecification of the PS and the regression model. The third study extended these approaches with machine learning estimation of the PS and the endpoint regression, and found that bias due to misspecification could be further reduced.

This thesis concludes that those approaches that relax the assumption that the statistical model for addressing selection bias is correctly specified, can give more accurate and precise estimates of cost-effectiveness than previously recommended methods. Findings from this thesis can improve the quality of CEA that use patient-level observational data, to help future studies provide a sounder basis for policy making.

Table of contents

List of tables

List of figures

List of appendices

Acknowledgements

First and foremost, I would like to thank my supervisor Richard Grieve. I am beyond grateful for all the prompt and thoughtful help and constructive criticism, as well as the challenging discussions. Richard continues to be a great example in doing research: he showed me how to keep focused on the relevant questions while maintaining high scientific standards, to cooperate with fellow academics, and most importantly, to always keep a good sense of humour.

I have benefitted from a great advisory committee including John Cairns, James Carpenter and Rhian Daniel, who offered me guidance and thoughtful insights throughout my studies. Due to my wonderful co-authors and colleagues, PhD was not a lonely experience for me. I am grateful to Zia Sadique, Roland Ramsahai, Mark Pennington, Karla Diaz-Ordaz and Rosalba Radice, for all the stimulating discussions and encouraging words, especially Rosalba's advice to stay "quiet" - I always knew what she meant. While we have not (yet) co-authored a paper, special thanks go to Manuel Gomes for being a great friend and for constantly showing me the light at the end of the tunnel, or at least for telling to stop halfway, for lunch.

I have been delighted to have the collaboration of Jasjeet S. Sekhon and Susan Gruber, who introduced me to the fascinating world of causal inference. In particular, I thank Jas for providing me with inspiring discussions accompanied with great café latte; and Susan for being an amazing role model with her passion for science and elegant R functions.

Special thanks go to my friends for seeing me through the ups and downs. Some were within a bike ride, such as Réka and Alan, while others, like Hédi and Mireille, looked after me from a distance. Thanks to Matt for showing me Selected Ambient Works 85-92 by Aphex Twin. Funding for this PhD by the Economic and Social Research Council is greatly appreciated.

I would like to dedicate this thesis to my family: Anya, Apa, Zsuzska and Kinga. Instead of questioning my life choices, you have always given me loving and unconditional support. I am so lucky to have you.

Abbreviations

AIPTW - Augmented inverse probability of treatment weighting
ATE - Average treatment effect
ATT - Average treatment effect on the treated
BCM - Bias-corrected matching
CARTs - Classification and regression trees
CEA - Cost-effectiveness analysis
CI - Confidence interval
CEAC - Cost-effectiveness acceptability curve
C-TMLE - Collaborative targeted maximum likelihood estimation
DR - Double-robust
DrotAA – Drotrecogin alfa activated
EQQ plots – Empirical quantile-quantile plots
GLM - Generalised linear model
GM - Genetic Matching
HEED - Health Economic Evaluations Database
HPC - High performance computing
HTA - Health technology assessment
HRQoL - Health-related quality of life
ICNARC - The Intensive Care National Audit and Research Centre
ICU - Intensive care unit
INB - Incremental net monetary benefit
IPD - Individual patient data
IPT weights - Inverse probability of treatment weights
IPTW - Inverse probability of treatment weighting
IV- Instrumental variable
KS test - Kolmogorov-Smirnov test
MD - Mahalanobis distance
ML - Maximum likelihood
NEED - NHS Economic Evaluations Database
NICE - National Institute for Health and Clinical Excellence
NRS - Non-randomised studies
OLS - Ordinary least squares
PROMs - Patient reported outcome measures
PS - Propensity score
QALY – Quality-adjusted life year
RCT - Randomised controlled trial
RMSE - Root mean squared error
TMLE - Targeted maximum likelihood estimation
WLS - Weighted least squares
WTP - Willingness to pay

Chapter 1 - Introduction

1.1 Economic evaluation to inform health policy

An important objective of health care systems is the allocation of scarce resources in order to maximise health gain (Gray et al., 2010). Health economic evaluation can address this optimisation problem, by comparing alternative options in terms of their costs and consequences (Drummond et al., 2005). Health economic evaluation is increasingly used for centralised decision making worldwide (NICE, 2008, IQWIG, 2009, PBAC, 2008, CADTH, 2006).

Economic evaluation can rely on various sources of evidence, depending on the decision context. For example the Australian Pharmaceutical Benefits Advisory Committee requires the appraisal of new pharmaceuticals and vaccines (PBAC, 2008). As randomised controlled trials (RCTs) are generally mandated in the drug development process (Glick et al., 2001), RCT evidence for such evaluations is widely available. In other settings, for example in the NHS in England and Wales, economic evaluation is used for a wider range of technologies, including medical devices, surgical procedures, or the development of public health guidelines. Here, RCT evidence might be insufficient or lacking and has to be complemented with data from non-randomised studies (NRS) (NICE, 2008, NICE, 2009).

There would appear to be a consensus across methodological guidelines on several aspects of the study design of economic evaluations (Hjelmgren et al., 2001). These include the use of cost-effectiveness analysis (CEA), with health outcomes measured as quality adjusted life years (QALYs) (Dolan, 2000). A further general requirement is a time horizon that incorporates all relevant benefits and costs of the interventions under comparison (Kuntz and Weinstein, 2001). Decision analytical models provide a

framework for synthesising different sources of evidence (Caro et al., 2012), and analysts are encouraged to consider and report the uncertainty that surrounds any recommendation in regards to the relative cost-effectiveness of the alternatives under consideration (Briggs et al., 2012).

The incorporation of individual patient data (IPD) in decision models has been recommended by methodological guidelines (Briggs et al., 2006, Briggs et al., 2012, Philips et al., 2006). Using IPD can help studies fully account for parameter uncertainty, by estimating the standard errors and correlations between model parameters. IPD can also help analysts address heterogeneity, by allowing input parameters to differ for patient subgroups (Cooper et al., 2007, Koerkamp et al., 2010). With the availability of IPD, a range of further challenges can be addressed, such as censoring (Willan et al., 2005), missing data (Noble et al., 2012) and confounding (Thompson et al., 2010).

Methodological guidance on the analysis of IPD in CEA has been through significant development, mostly in the context of cost and effectiveness data from RCTs (Glick et al., 2001). While RCT data is generally the preferred evidence for deriving the effectiveness of an intervention (NICE, 2008), it has been recognised that CEA based on a single study rarely provides a sufficient basis for decision making (Sculpher et al., 2006). For example, a protocol driven, multinational phase III trial might not reflect real world treatment patterns and resource use in a particular country; an RCT might not measure the relevant endpoints or include all the relevant comparators, while the time horizon can be too short to capture long-term costs and health benefits (Briggs et al., 2006). Hence the synthesis of data from different sources is required, including RCTs, NRS, epidemiological databases or patient registries (NICE, 2008).

Observational studies often provide a relevant data source for input parameters in a decision model. In some cases, observational data provides the main source of evidence

for CEA (Polsky and Basu, 2006, Manca and Austin, 2008). With observational studies, a general methodological challenge is handling selection bias due confounding, i.e. differences in prognostic factors between treatment groups of interest (Fung et al., 2011, Pizer, 2009, Rubin, 2010, Tunis et al., 2010, Polsky and Basu, 2006). When IPD from observational studies are available, appropriate statistical methods can be used to reduce selection bias. Current methodological guidance on economic evaluation warns of potential biases from using estimates based on observational studies (Philips et al., 2006, NICE, 2008). Currently there is no detailed guidance on the appropriate statistical analysis of observational data for CEA (Kearns et al., 2012), which was raised as a priority in a recent review on priorities for methodological research in health technology assessments (HTA) used by NICE (Longworth et al., 2009).

1.2 Observational data in CEA

In this thesis I define observational data as data from studies that do not have random allocation to alternative treatments (Deeks, 2003), which includes cohort studies, case-control studies, surveys, registries, administrative records or census data. Observational data can be used to estimate a wide range of parameters in CEA (Drummond, 1998, Deeks, 2003), including clinical endpoints, prevalence of side effects, health-related quality of life (HRQoL) measures, and long-term costs. While some economic evaluations do not use decision models and may rely heavily on RCT data, even here they may still use observational data, for example to obtain unit costs and HRQoL tariffs (Glick et al., 2007). When evidence is synthesised in decision analytical models (Briggs et al., 2006, NICE, 2008), observational data can be used to inform model parameters by making use of published external information, for example for transition probabilities. A more flexible use of observational data may be feasible where patient-level observational data are available. Here, for example, it may be possible to calibrate

estimates of long-term outcomes for the patient characteristics of the treatment groups of interest, accounting for patient heterogeneity. If the patient-level observational data include information on the treatments of interest, incremental cost and effectiveness parameters can be calculated. This can be done either by complementing parameters derived from RCT data, or in some settings parameters may be estimated exclusively from observational data. Examples for incremental effectiveness parameters are relative risk of mortality or clinical events, HRQoL differences, or incremental QALYs. These parameters can be then used in a decision-analytical model when extrapolating RCT data, for example.

In many cases, particularly in evaluations of medical technologies other than pharmaceuticals (e.g. health services and public health interventions), there may be no RCT data available and both incremental cost and effectiveness parameters are calculated based on IPD from a single observational study (Polsky and Basu, 2006, Manca and Austin, 2008). Here, additional aggregate information (e.g. HRQoL) may or may not be used.

The focus of this thesis is on CEA that uses patient-level observational data for estimating incremental cost and effectiveness parameters. Unless specified otherwise, by referring to observational data, this thesis will refer to patient-level observational data. Contributions of the thesis to the more general use of observational data in CEA will be noted in section 5 of this chapter (conceptual framework), and in the discussion of the thesis (chapter 7).

1.3 Statistical methods in CEA that use patient-level observational data

Statistical methods for CEA that use IPD predominantly from RCTs have seen considerable development in the last decade (Willan and Briggs, 2006, Glick et al., 2007, Gray et al., 2010). It has been recommended that regression methods adjust for covariate imbalances between treatment groups and estimate subgroup-specific treatment effects (Hoch et al., 2002, Willan et al., 2004, Nixon and Thompson, 2005), while maintaining the correlation between cost and effectiveness endpoints. Statistical methods have been proposed to address further challenges, such as: hierarchical data in multicentre CEA (Grieve et al., 2007, Manca et al., 2007), missing data (Noble et al., 2012), non-compliance to randomised treatment (Hughes et al., 2001) and censoring (Willan et al., 2002, Willan et al., 2005). Guidelines (NICE, 2008, Philips et al., 2006) and quality assessment tools (Doshi et al., 2006, Gomes et al., 2011) emphasise the use of appropriate methods to analyse patient-level data, mostly in the context of RCT data.

Less attention has been given to methods development in CEA that use patient-level observational data. Here, the general concern is that the treatment groups under evaluation can be imbalanced in observed and unobserved characteristics, which might be prognostic for the cost and effectiveness endpoints. Unadjusted comparisons of cost and effectiveness outcomes are then prone to selection bias, which in epidemiology is referred to as bias due to confounding (Greenland et al., 1999), and in econometrics as bias due to endogeneity (Imbens and Wooldridge, 2009a).

Selection bias can be reduced if appropriate statistical methods are applied (Jones, 2007, Polsky and Basu, 2006, Pizer, 2009, Rubin, 2010). The estimation of treatment effects using observational data has been at the centre of methodological research in the last decade in the general causal inference literature, including the fields of statistics (Pearl,

2009, Rubin, 2006), econometrics (Imbens and Wooldridge, 2009a), the social sciences (Morgan and Winship, 2007) and medical statistics (Austin, 2008, Shah et al., 2005, Stuart, 2010), but has received relatively little attention in CEA (Polsky and Basu, 2006, Sekhon and Grieve, 2011, Manca and Austin, 2008, Mitra and Indurkhya, 2005).

There are specific complexities of IPD used for CEA that statistical methods must acknowledge. Firstly, correlations between parameters, such as the incremental costs and effectiveness need to be estimated (O'Hagan and Stevens, 2001). Secondly, the distributions of cost (Mihaylova et al., 2010, Basu et al., 2011, Manning et al., 2005) and effectiveness endpoints (Basu and Manca, 2011) are likely to be irregular (non normal), and relationships between covariates and endpoints can be nonlinear (Basu et al., 2011). Thirdly, decision makers may want estimates of cost-effectiveness for particular patient subgroups (Sculpher, 2008). These issues have been considered in methods proposed for statistical analysis of IPD, but mainly in the context of studies that use RCT data (Willan and Briggs, 2006, Glick et al., 2007).

The aim of this PhD is to help fill in these gaps of methodological literature in CEA by considering a range of statistical methods that can reduce selection bias when estimating parameters for CEA that use observational data.

1.4 Aims and objectives of the thesis

This thesis considers alternative statistical methods for addressing selection bias in CEA that use patient-level observational data. The thesis has four main objectives:

1. To develop and apply a new checklist for assessing the underlying assumptions made by statistical methods for addressing selection bias in CEA, that use patient-level observational data;

2. To consider which statistical methods from the general causal inference literature may be appropriate for addressing selection bias in CEA;
3. To compare the relative performance of propensity score (PS) approaches and Genetic Matching (GM), a multivariate matching method for estimating subgroup-effects in CEA;
4. To compare methods that combine regression with PS approaches for addressing selection bias when estimating incremental effectiveness and cost-effectiveness parameters.

1.5 Conceptual framework of the thesis

The four objectives of the PhD are strongly interlinked. The basis of this research is a conceptual literature review, which consists of a careful assessment of the methodological challenges that arise when addressing selection bias in CEA that use patient-level observational data (objective 1). Here I also review the general causal inference literature, to examine further promising methods for addressing selection bias in CEA (objective 2).

Regression and PS methods are currently proposed to address selection bias in CEA (Nixon and Thompson, 2005, Polsky and Basu, 2006, Mitra and Indurkhya, 2005). The crucial assumption behind these methods is that all confounders can be observed (unconfoundedness assumption), the covariate distributions of the treatment groups overlap and their underlying models such as endpoint regression and PS models are correctly specified. It is unlikely that all these assumptions are met in CEA. For example, health care costs and outcomes often have irregular distributions (Jones, 2010, Basu and Manca, 2011), with nonlinear relationships between the confounders and the endpoints (Basu et al., 2011), hence it can be challenging to correctly specify regression models. Policy makers are often interested in cost-effectiveness results for patient

subgroups (Sculpher, 2008), which can necessitate specifying regression models that can account for heterogeneous treatment effects, or PS models that can incorporate heterogeneous selection into treatment.

The first approach of investigating whether the assumptions behind the currently recommended statistical methods are plausible is to conduct a critical review of the applied CEA literature (objective 1, research paper 1). A checklist informed by the conceptual review can help assess whether the assumptions behind the statistical methods are appropriately assessed. The results of this review can highlight those methods that are under-utilised, or inappropriately used. These deficiencies in the applied literature motivated me to undertake further methodological work, to assess the relative performance of the methods under different circumstances faced in applied CEA, using case studies and simulation studies.

I use the conceptual review to identify alternative methods that are promising, because they have the potential to make less restrictive assumptions than standard regression and PS methods (objective 2). The following methods were identified to be promising: GM, a multivariate matching method that aims to balance individual covariates; and approaches that combine the PS and regression models, such as double-robust (DR) methods and regression-adjusted matching. I also consider machine learning approaches for estimating the PS and the endpoint regression, which can reduce misspecification and bias compared to using fixed parametric models.

The conceptual review provided hypotheses on how these methods perform in realistic CEA settings, however previous simulation evidence may not be directly applicable in a CEA setting. I undertake Monte Carlo simulation studies (objective 3 and 4; research papers 2, 3 and 4) that extend the current CEA methods literature, by assessing the selected methods under settings typical of CEA. These simulations are motivated by

CEA case studies, and aim to test hypotheses generated by the conceptual review, across different circumstances.

The first case study (research papers 2 and 3) highlights circumstances when cost-effectiveness for patient subgroups is of interest. The corresponding simulation study (research paper 2) compares methods for subgroup analysis (objective 3). Research papers 3 and 4 consider methods that combine the PS and endpoint regression models (objective 4). The second case study and corresponding simulation study (research paper 4) demonstrate settings where the correct specification of an effectiveness endpoint is challenging, and uses machine learning estimation techniques to reduce bias due to misspecification. The methods are also considered in the case studies, and the impact of different methods on the cost-effectiveness results and estimated treatment effects are reported.

This thesis considers CEA where IPD from RCTs is either unavailable, or insufficient to estimate either incremental costs or incremental effectiveness parameters, or both. The focus of the simulation studies (research papers 2, 3 and 4) is when IPD from a single observational study is used to calculate incremental effectiveness and cost-effectiveness parameters. The applied literature review (research paper 1) and the case studies however also consider settings where these incremental parameters are combined with aggregate data (research papers 2 and 3), and when input parameters for decision models need to be estimated using patient-level observational data (research paper 4). It is therefore expected that findings from the thesis will be applicable to a more general use of observational data in CEA.

1.6 Overall contribution of the thesis

I developed a new checklist for critical appraisal of statistical methods for addressing selection bias in CEA that use patient-level observational data (research paper 1). This checklist complements previous quality-assessment tools and methodological guidance (Drummond et al., 2005, Philips et al., 2006, Glick et al., 2007), which did not include specific criteria for the analysis of patient-level data from observational studies. Research paper 1 provides detailed guidance on how the underlying assumptions of the statistical methods can be assessed, and highlights how the choice of statistical approach can contribute to structural uncertainty in CEA (Bojke et al., 2009). In addition, prior to this work it was unknown whether applied CEA use appropriate statistical methods for addressing selection bias. The systematic review in research paper 1 addressed this gap, and found that CEA do not appropriately assess the main assumptions behind statistical methods. This checklist can raise awareness about these assumptions.

Research paper 2 compares GM, a multivariate matching method, with PS matching and inverse probability of treatment weighting (IPTW) for estimating subgroup effects in CEA. GM was previously demonstrated to reduce selection bias in CEA (Sekhon and Grieve, 2011), but not in the context of subgroup analysis. The paper found that GM was relatively robust to the misspecification of the PS, and provided the lowest bias and root mean squared error of the estimated incremental net benefit (INB) for each subgroup. This paper provides the first comparison of GM with IPTW in the general literature.

Research paper 3 considers methods that combine the PS with endpoint regression models for CEA. This paper considers DR methods and regression-adjusted matching for reducing selection bias for the first time in CEA. The paper found that regression-adjusted matching was the least biased method when both the endpoint regression

models and the PS model were misspecified (dual misspecification). This paper considers the performance of regression-adjusted matching under dual misspecification for the first time.

Research paper 4 extended the combined approaches presented in research paper 3, by considering recently proposed machine learning techniques for estimating the PS and the endpoint regression for estimating incremental effectiveness parameters. This paper extends the previous literature which recommended regression modelling of HRQoL endpoints (Basu and Manca, 2011). The paper also extends the general methodological literature by providing the first comparison of targeted maximum likelihood estimation (TMLE) and bias-corrected matching (BCM). Unlike previous papers on BCM, which used linear regression (Abadie and Imbens, 2011, Busso et al., 2011), this paper uses machine learning techniques for bias correction. The paper found that both TMLE and BCM could reduce bias due to misspecification, with machine-learning versus fixed parametric approaches.

Overall, this thesis compares the performance of statistical methods for addressing selection bias under realistic circumstances for CEA. The simulation studies provide new evidence on the relative robustness of methods when some of their underlying assumptions, such as correct model specification fail. The case studies help motivate the simulation studies, demonstrate the appropriate use of statistical methods proposed, and illustrate how structural uncertainty from the choice of method can be addressed, by reporting results across a range of methods. The research papers provide detailed guidance and software codes for the implementation of the methods. It is expected that the findings of this thesis can add to the methodological guidance for researchers conducting CEA that use patient-level observational data, and help future studies provide a sounder basis for policy making.

1.7 Structure of the thesis

The remaining chapters of the thesis are as follows. Chapter 2 first identifies the key challenges for statistical methods which aim to address selection bias in CEA that use observational data. This chapter then describes the assumptions behind previously recommended methods for addressing selection bias in CEA, informed by a conceptual review of the general causal inference literature. The chapter then reviews promising statistical methods from the causal inference literature, that have potential for addressing selection bias in CEA. Finally, chapter 2 identifies gaps in the methodological literature concerned with the relative performance of the methods in CEA.

Chapters 3 to 6 comprise of the four research papers, each prefaced with a brief preamble. Research paper 1 develops a critical appraisal tool to assess the statistical methods for addressing selection bias in CEA that use observational data, and applies this checklist in a systematic review of published studies. Motivated by a CEA of a pharmaceutical intervention for patients with severe sepsis, research paper 2 presents a simulation study that compares the relative performance of GM and PS methods in reporting cost-effectiveness for patient subgroups. Research paper 3 evaluates the relative performance of statistical methods that combine the PS and regression models, for the cost and effectiveness endpoint in CEA. Research paper 4 extends these combined methods by considering the recently proposed methods, TMLE and BCM, for estimating incremental effectiveness parameters. The simulation study presented in this paper was motivated by an evaluation of the effect of alternative hip prostheses on patients' HRQoL.

Chapter 7 provides an overview of the main findings and contributions of the thesis. The chapter then acknowledges the limitations of the thesis, and identifies potential

areas for future research. This chapter concludes by highlighting the implications of the findings of the thesis for applied researchers and policy makers.

1.8 Contribution of the candidate to the thesis

The work conducted in this thesis was linked to a research grant "Methods for reducing selection bias in cost-effectiveness analysis", funded by the Economic and Social Research Council (ESRC), and took a similar approach in using simulations and case studies to assess the relative merits of alternative methods for addressing selection bias in CEA. The focus of the ESRC project was to compare GM to PS matching. This thesis aimed to offer a more thorough comparison of alternative methods, and extended the comparison additional methods not included in the ESRC study, such as DR methods and regression-adjusted matching. It also looked at some of the methods in a new context, where cost-effectiveness for patient subgroups is of interest.

The research questions for research papers 1 and 2 were linked to the ESRC project and identified by the principal investigator, Richard Grieve. In the first study, the candidate carried out a conceptual review, and developed a checklist and accompanying methodological guidance for critical appraisal of CEA that uses observational data, in collaboration with her supervisor, Richard Grieve. The candidate applied this checklist in a systematic review of studies, and interpreted the findings. A further contributor to this paper was a research fellow linked to the project, Zia Sadique, who verified the exclusion criteria of the systematic review, and conducted a second review by independently appraising 50% of the studies.

For research paper 2, the candidate led the design of the simulation study, with Richard Grieve. The candidate wrote the simulation code, with help from post-doctoral researchers employed by the ESRC project, Roland Ramsahai and Rosalba Radice.

Roland Ramsahai helped the candidate run simulations on the LSHTM computational cluster. Zia Sadique led on the analysis of the motivating case study, with the candidate contributing to the analysis. The candidate led on the reporting and interpretation of the results of the case study and the simulation studies. For this paper, the candidate built on insights from another simulation study linked to the project, which aimed to compare GM, IPTW and PS matching for estimating subgroup-specific treatment effects on binary endpoints (Radice et al., 2012). For this study, aimed at a biostatistics audience, the candidate contributed to the design and the implementation of the simulation study and to the interpretation of the results, as well as to writing sections of the manuscript.

The candidate led on the conception of the research question for research paper 3 in collaboration with her supervisor, Richard Grieve and an external collaborator, Jasjeet S. Sekhon, while visiting the Center for Causal Inference at UC Berkeley (USA, CA). The candidate led on the design of the simulation study, with the collaboration of Richard Grieve, Rosalba Radice and Jasjeet S. Sekhon. Rosalba Radice contributed to the design of the simulation scenarios. The candidate wrote the code for the simulation study, with help from Rosalba Radice. The candidate conducted the statistical analysis for the motivating case study, and interpreted the results of the paper, with Rosalba Radice and Richard Grieve.

The candidate led the design of the research question for research paper 4, in collaboration with Richard Grieve and an external collaborator, Susan Gruber (Harvard School of Public Health). The candidate led on the design and implementation of the simulation scenarios, with help from Rosalba Radice, who also contributed to the implementation of the statistical methods in the motivating case study. The candidate led on the interpretation of the results, with Rosalba Radice, Susan Gruber, Jasjeet S. Sekhon and Richard Grieve.

For each of the research papers, the candidate wrote the first draft of the manuscripts. She managed each round of comments and suggestions from co-authors, in collaboration with Richard Grieve. All authors read and approved the final drafts of the research papers prior to journal submission and inclusion in this thesis. The remaining chapters of the thesis are the sole work of the candidate.

References

Abadie, A. & Imbens, G. W. 2011. Bias-Corrected Matching Estimators for Average Treatment Effects. *Journal of Business & Economic Statistics,* 29**,** 1-11.

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

Basu, A. & Manca, A. 2011. Regression Estimators for Generic Health-Related Quality of Life and Quality-Adjusted Life Years. *Med Decis Making,* 2011 Oct 18. [Epub ahead of print].

Basu, A., Polsky, D. & Manning, W. 2011. Estimating treatment effects on healthcare costs under exogeneity: is there a 'magic bullet'? *Health Services and Outcomes Research Methodology,* 11**,** 1-26.

Bojke, L., Claxton, K., Sculpher, M. & Palmer, S. 2009. Characterizing structural uncertainty in decision-analytic models: a review and application of methods. *Value in Health,* 12**,** 739-49.

Briggs, A., Sculpher, M. & Klaxton, K. (eds.) 2006. *Decision Modelling for Health Economic Evaluation*: Oxford University Press.

Briggs, A. H., Weinstein, M. C., Fenwick, E. A. L., Karnon, J., Sculpher, M. J. & Paltiel, A. D. 2012. Model Parameter Estimation and Uncertainty Analysis. *Medical Decision Making,* 32**,** 722-732.

Busso, M., DiNardo, J. & McCrary, J. 2011. New Evidence on the Finite Sample Properties of Propensity Score Reweighting and Matching Estimators. *Working paper.*

CADTH. 2006. *Guidelines for the Economic Evaluation of Health Tecnologies: Canada. 3rd Ed.* [Online]. Ottawa, Canada. Available: http://www.cadth.ca/media/pdf/186_EconomicGuidelines_e.pdf.

Caro, J. J., Briggs, A. H., Siebert, U. & Kuntz, K. M. 2012. Modeling Good Research Practices—Overview. *Medical Decision Making,* 32**,** 667-677.

Cooper, N. J., Sutton, A. J., Ades, A. E., Paisley, S. & Jones, D. R. 2007. Use of evidence in economic decision models: practical issues and methodological challenges. *Health Economics,* 16**,** 1277-1286.

Deeks, J. J. 2003. *Evaluating non-randomised intervention studies* [Online]. Tunbridge Wells: published by Gray Pub. on behalf of NCCHTA. Available: http://www.hta.ac.uk/execsumm/summ727.htm [Accessed 23/05/2009.

Dolan, P. 2000. The measurement of health-related quality of life for use in resource allocation decisions in health care. *Handbook of Health Economics.* Amsterdam: North-Holland.

Doshi, J. A., Glick, H. A. & Polsky, D. 2006. Analyses of Cost Data in Economic Evaluations Conducted Alongside Randomized Controlled Trials. *Value in Health,* 9**,** 334-340.

Drummond, M., Sculpher, M., Torrance, G., O'Brien, B. & Stoddart, G. 2005. *Methods for the Economic Evaluation of Health Care Programmes,* Oxford, Oxford University Press.

Drummond, M. F. 1998. Experimental versus observational data in the economic evaluation of pharmaceuticals. *Medical Decision Making,* 18.

Fung, V., Brand, R. J., Newhouse, J. P. & Hsu, J. 2011. Using Medicare Data for Comparative Effectiveness Research: Opportunities and Challenges. *Am J Manag Care,* 17**,** 489-496.

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

Glick, H., Polsky, D. & Schulman, K. 2001. Trial-based economic evaluations: an overview of desgin and analysis. *In:* DRUMMOND, M. & MCGUIRE, A. (eds.) *Economic evaluation in health care: Meging theory with practice.* Oxford: Oxford University Press.

Gomes, M., Grieve, R., Edmunds, J. & Nixon, R. 2011. Statistical methods for cost-effectiveness analyses that use data from cluster randomised trials: a systematic review and checklist for critical appraisal. *Medical Decision Making,* 32**,** 209-20.

Gray, A. M., Clarke, P. M., Wolstenholme, J. L. & Wordsworth, S. 2010. *Applied Methods of Cost-Effectiveness Analysis in Healthcare,* Oxford University Press.

Greenland, S., Pearl, J. & Robins, J. M. 1999. Confounding and Collapsibility in Causal Inference. *Statist. Sci.* , 14**,** 29-46.
Grieve, R., SG, T., Nixon, R. M. & Cairns, J. 2007. Multilevel models for estimating incremental net benefits in multinational studies. *Health Economics,* 16**,** 815–26.
Hjelmgren, J., Berggren, F. & Andersson, F. 2001. Health Economic Guidelines—Similarities, Differences and Some Implications. *Value in Health,* 4**,** 225-250.
Hoch, J. S., Briggs, A. H. & Willan, A. R. 2002. Something old, something new, something borrowed, something blue: a framework for the marriage of health econometrics and cost-effectiveness analysis. *Health Economics,* 11**,** 415-430.
Imbens, G. M. & Wooldridge, J. M. 2009a. Recent Developments in the Econometrics of Program Evaluation. *Journal of Economic Literature,* 47**,** 5–86.
IQWIG. 2009. *Methods for assessment of the relation of Benefits to Costs in the German Statutory Health Care System* [Online]. Cologne, Germany. Available: http://www.ispor.org/peguidelines/source/Germany_AssessmentoftheRelationofBenefitstoCosts_En.pdf.
Jones, A. M. 2007. Identification of treatment effects in Health Economics. *Health Economics,* 16**,** 1127-1131.
Jones, A. M. 2010. Models For Health Care. *HEDG Working Papers.* HEDG, c/o Department of Economics, University of York.
Kearns, B., Ara, R. & Wailoo, A. 2012. *A review of the use of statistical regression models to inform cost-effectiveness analyses withing the NICE Technology Appraisals Programme* [Online]. ScHARR, University of Sheffield. Available: http://www.nicedsu.org.uk/FINAL%20DSU%20Regressions%20report_09.10.12.pdf.
Koerkamp, G. B., Weinstein, M. C., Stijnen, T., Heijenbrok-Kal, M. H. & Hunink, M. G. M. 2010. Uncertainty and Patient Heterogeneity in Medical Decision Models. *Medical Decision Making,* 30**,** 194-205.
Kuntz, K. M. & Weinstein, M. C. 2001. Modelling in economic evaluation. *In:* DRUMMOND, M. & MCGUIRE, A. (eds.) *Economic Evaluation in Helath care. Merging theory with practice.* Oxford: Oxford University press.
Longworth, L., Bojke, L., Tosh, J. & Sculpher, M. 2009. MRC-NICE Scoping Project: Identifying the National Institute For Health And Clinical Excellence's Methodological Research Priorities and an Initial Set of Priorities. *In:* ECONOMICS, C. F. H. (ed.). University of York.
Manca, A. & Austin, P. C. 2008. *Using propensity score methods to analyse individual patient-level cost-effectiveness data from observational studies* [Online]. Available: http://www.york.ac.uk/res/herc/documents/wp/08_20.pdf.
Manca, A., Lambert, P., Sculpher, N. & Rice, N. 2007. Cost-effectiveness analysis using data from multinational trials: the use of Bayesian hierarchical modelling. *Medical Decision Making***,** 471–90.
Manning, W. G., Basu, A. & Mullahy, J. 2005. Generalized modeling approaches to risk adjustment of skewed outcomes data. *Journal of Health Economics,* 24**,** 465-488.
Mihaylova, B., Briggs, A., O'Hagan, A. & Thompson, S. 2010. Review of statistical methods for analysing healthcare resources and costs. *Health Economics,* DOI: 10.1002/hec.1653.
Mitra, N. & Indurkhya, A. 2005. A propensity score approach to estimating the cost-effectiveness of medical therapies from observational data. *Health Economics,* 14**,** 805-15.
Morgan, S. L. & Winship, C. 2007. *Counterfactuals and causal inference: methods and principles for social research,* New York, Cambridge University Press.
NICE. 2008. *Guide to the Methods of Technology Appraisal* [Online]. Available: http://www.nice.org.uk/media/B52/A7/TAMethodsGuideUpdatedJune2008.pdf [Accessed 24/10/2010.
NICE. 2009. *Methods for the development of NICE public health guidance (second edition)* [Online]. Available: http://www.nice.org.uk/media/2FB/53/PHMethodsManual110509.pdf.

Nixon, R. M. & Thompson, S. G. 2005. Methods for incorporating covariate adjustment, subgroup analysis and between-centre differences into cost-effectiveness evaluations. *Health Economics,* 14**,** 1217-1229.
Noble, S., Hollingworth, W. & Tilling, K. 2012. Missing data in trial-based cost-effectiveness analysis: the current state of play. *Health Economics,* 21**,** 187-200.
O'Hagan, A. & Stevens, J. 2001. A framework for cost-effectiveness analysis from clinical trial data. *Health Economics,* 10**,** 303-15.
PBAC. 2008. *Guidelines for preparing submissions to the Pharmaceutical Benefits Advisory Committee* [Online]. Camberra, Australia. Available: http://www.health.gov.au/internet/main/publishing.nsf/Content/D8EBFB77AC0E7552CA25717D000AE40B/$File/pbac_guidelines.pdf.
Pearl, J. 2009. *Causality : models, reasoning, and inference,* Cambridge, Cambridge University Press.
Philips, Z., Bojke, L., Sculpher, M., Claxton, K. & Golder, S. 2006. Good practice guidelines for decision-analytic modelling in health technology assessment: a review and consolidation of quality assessment. *Pharmacoeconomics,* 24**,** 355-71.
Pizer, S. 2009. An intuitive review of methods for observational studies of comparative effectiveness. *Health Services and Outcomes Research Methodology,* 9**,** 54-68.
Polsky, D. & Basu, A. 2006. Selection Bias in Observational Data. *The Elgar Companion to Health Economics.* Edward Elgar Publishing.
Radice, R., Grieve, R., Ramsahai, R., Kreif, N., Sadique, Z. & 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).
Rubin, D. B. 2006. *Matched sampling for causal effects,* Cambridge, Cambridge University Press.
Rubin, D. B. 2010. On the limitations of comparative effectiveness research. *Statistics in Medicine,* 29**,** 1991-1995.
Sculpher, M. 2008. Subgroups and Heterogeneity in Cost-Effectiveness Analysis. *Pharmacoeconomics,* 26**,** 799-806.
Sculpher, M. J., Claxton, K., Drummond, M., McCabe, C. & Mihaylova, B. 2006. Whither trial-based economic evaluation for health care decision making? *Health Economics,* 15**,** 677–687.
Sekhon, J. S. & Grieve, R. D. 2011. A Matching Method for Improving Covariate Balance in Cost-Effectiveness Analyses. *Health Economics,* doi: 10.1002/hec.1748.
Shah, B. R., Andreas, L., Janet, E. H. & Peter, C. A. 2005. Propensity score methods gave similar results to traditional regression modeling in observational studies: a systematic review. *Journal of Clinical Epidemiology,* 58**,** 550-559.
Stuart, E. A. 2010. Matching Methods for Causal Inference: A review and a look forward. *Statistical Science,* 25.
Thompson, S., Kaptoge, S., White, I., Wood, A., Perry, P., Danesh, J. & Collaboration, T. E. R. F. 2010. Statistical methods for the time-to-event analysis of individual participant data from multiple epidemiological studies. *International Journal of Epidemiology*.
Tunis, S. R., Benner, J. & McClellan, M. 2010. Comparative effectiveness research: Policy context, methods development and research infrastructure. *Statistics in Medicine,* 29**,** 1963-1976.
Willan, A. R. & Briggs, A. H. 2006. *Statistical Analysis of Cost-effectiveness Data,* John Wiley & Sons Ltd.
Willan, A. R., Briggs, A. H. & Hoch, J. S. 2004. Regression methods for covariate adjustment and subgroup analysis for non-censored cost-effectiveness data. *Health Economics,* 13**,** 461-475.
Willan, A. R., Lin, D. Y., Cook, R. J. & Chen, E. B. 2002. Using inverse-weighting in cost-effectiveness analysis with censored data. *Statistical Methods in Medical Research,* 11**,** 539-551.
Willan, A. R., Lin, D. Y. & Manca, A. 2005. Regression methods for cost-effectiveness analysis with censored data. *Statistics in Medicine,* 24**,** 131-145.

Chapter 2 - Conceptual review of statistical methods for addressing selection bias in CEA that use patient-level observational data

2.1 Introduction

Statistical methods for cost-effectiveness analyses (CEA) that use individual patient data (IPD) from randomised controlled trials (RCTs) have been through considerable development in the last decade. However, statistical methods developed for the analysis of RCT data might not be appropriate for observational data. In CEA that use data from NRS, the distribution of the baseline covariates can be highly imbalanced (Grieve et al., 2008). Under such circumstances, estimates can be sensitive to the specification of the regression models, for example the inclusion or exclusion of nonlinearities in the covariate-endpoint relationship (Ho et al., 2007). Instead of modelling the cost and effectiveness endpoint, a recommended approach in CEA is to try to achieve covariate balance using the propensity score (PS) (Manca and Austin, 2008, Mitra and Indurkhya, 2005). However, correctly specifying the unknown PS is also challenging (Dehejia and Wahba, 2002).

This chapter aims to consider several outstanding methodological concerns that face CEA that use observational data. First, what are the important underlying assumptions made by common statistical methods for addressing selection bias in CEA? Which statistical methods are most appropriate for addressing selection bias when estimates of cost-effectiveness are required for patient subgroups? Which methods from the general causal inference literature are potentially appropriate for addressing selection bias in CEA, and what are their relative merits under typical circumstances that arise in CEA?

The overall aim of this chapter is to identify gaps in the methods literature of CEA concerned with addressing selection bias, which the subsequent chapters of the thesis aim to address. The specific objectives of this chapter are:

1. To review methodological guidance on statistical methods for addressing selection bias in CEA that use patient-level observational data, and to describe challenges that arise when using these methods.
2. To describe the main underlying assumptions made by recommended statistical methods for addressing selection bias in CEA.
3. To identify statistical methods from the general causal inference literature that have the potential to reduce selection bias in a CEA.
4. To identify gaps in the methodological literature on the relative performance of alternative methods for addressing selection bias in CEA.

To address these objectives, I conducted a conceptual literature review consisting of two parts. First, I reviewed papers that provide methodological guidance for handling selection bias in CEA that use patient-level observational data. Second, I conducted a targeted review of the general causal inference literature, by reviewing seminal papers and their reference lists, and by consulting leading researchers in the field.

In the next section, I review the challenges statistical methods currently recommended for reducing selection bias in CEA need to address. Section 3 considers the main underlying assumptions statistical methods make in the context of CEA, based on methodological guidance from the general causal inference literature. Section 4 reviews further promising statistical methods and their appropriateness for CEA. Section 5 identifies gaps in the CEA methodological literature on the relative performance of the methods. The last section summarises the chapter and highlights areas for further research.

2.2 Statistical challenges in accounting for selection bias in CEA that use patient-level observational data

When RCT data is available for the cost and effectiveness endpoints, in general, randomisation ensures unbiased estimates of incremental cost and effects (Gray et al., 2010). When the CEA uses observational data, selection bias due to the lack of balance between treatment groups is a potential concern. This concern applies for a wide range of uses of observational data. These uses include settings when the CEA is conducted alongside a single observational study (Manca and Austin, 2008), and the parameters of interest are the incremental cost and incremental effectiveness, as well as when observational data is used to estimate parameters for a decision model such as hazard ratios or relative risks (Philips et al., 2006).

Polsky and Basu (2006) discuss the sources of selection bias in health economic evaluation. Overt bias is due to differences in observed provider or patient or characteristics, such as diseases severity, while hidden bias is due to unobserved characteristics, such as patient preferences. In a case study example, the authors demonstrate the use of regression and PS methods that can handle overt bias, and instrumental variables (IV) estimation that can potentially handle both observed and unobserved confounding. The authors note that in the context of economic evaluation, finding an appropriate IV is challenging: a good instrument is strongly related to treatment receipt, but unrelated to the endpoints. Moreover IV methods often estimate treatment effects for a narrower population than what is relevant for the original evaluation question (see more on IV estimation in section 3.2). My systematic review in research paper 1 finds that IV methods are rarely used in CEA.

Regression methods that can handle correlated costs and outcomes have been developed for covariate adjustment of RCT data, and are also recommended for addressing the potential selection bias that arises when using observational data (Nixon and Thompson, 2005). Even when using RCT data, the choice of statistical model - for example the error distribution chosen to model costs - has been shown to influence the estimated cost-effectiveness (Thompson and Nixon, 2005). When observational data is used for CEA, covariate distributions between treatment groups can be highly imbalanced (Grieve et al., 2008). This can make regression estimates sensitive to parametric model specification (Ho et al., 2007). In CEA, regression models may be specified for the cost and the effectiveness endpoints, which in the case of generalised linear models (GLMs), for example involves specifying the functional form relationship between the covariates and the endpoint, and the error distributions of the endpoint. The challenges of specifying regression models for resource use and cost data are widely recognised (Manning et al., 2005, Manning et al., 1987, Manning and Mullahy, 2001, Mihaylova et al., 2010, Basu and Rathouz, 2005, Jones, 2010). These are: nonlinear relationships between covariates and endpoints, irregular distributions and heavy tails of the endpoints. Similar challenges prevail when estimating treatment effects on health outcomes, such as health related quality of life (HRQoL) and quality-adjusted life years (QALYs), which often exhibit truncated supports with spikes at 0 or 1 (Basu and Manca, 2011). Recommended regression techniques to handle these endpoints include GLMs (Barber and Thompson, 2004a), two-part models (Buntin and Zaslavsky, 2004, Basu, 2011), or exponential conditional mean models (Manning et al., 2005). Specification tests recommended to rank competing regression models (Manning and Mullahy, 2001, Basu et al., 2004) can provide measures of model fit. However, failure

to reject a model in a specification test does not necessarily imply that the model is correctly specified (Horowitz, 2011).

Instead of specifying models for the endpoints, the PS can be used to create balance between observed characteristics of the treatment groups (Rosenbaum and Rubin, 1983). In CEA, the PS has been recommended for matching, stratification on the PS and as a covariate in either univariate net benefit regression models (Mitra and Indurkhya, 2005, Manca and Austin, 2008) or in bivariate regression models (Manca and Austin, 2008). Some of these methods, such as stratification and linear regression on the PS are outperformed by alternative PS methods, such as inverse probability of treatment weighting (IPTW) (Austin, 2009c, Lunceford and Davidian, 2004). These methods rely on the correct specification of the PS, which can be challenging (Westreich et al., 2010). For example, just as the investigator does not know the specification of the relationship between covariates and endpoints, they seldom know how covariates influence treatment receipt. A misspecified PS, due to, for example, omitting nonlinear relationships between covariates and treatment assignment (Basu et al., 2011), can lead to biased parameter estimates. To appropriately assess the specification of the PS, an assessment of balance on the full distribution of the covariates between the treatment groups must be made (Sekhon and Grieve, 2011), however this is rarely performed in practice (Austin, 2008).

These methods can be used to calculate cost-effectiveness parameters across all patients of interest. Decision makers are often also interested in cost-effectiveness for patient subgroups (Espinoza et al., 2011, Sculpher, 2008), and calculating overall estimates of cost-effectiveness could mask important heterogeneity between subgroups. Sources of heterogeneity can be characteristics of the patient (age, weight, gender, preferences, baseline outcomes etc.) or the health care provider (e.g. type of treatment centre)

(Sculpher, 2008). For example, in the CEA conducted by Mihaylova et al. (2005), cost-effectiveness of a statin therapy differed across subgroups defined by patients' baseline risk of a vascular event. In RCT-based CEA, a standard way of accounting for subgroup-level heterogeneity is to include covariate-by-treatment interactions in a regression model (Nixon and Thompson, 2005, Hoch et al., 2002, Willan et al., 2004). To provide unbiased cost-effectiveness estimates in an observational setting where subgroup effects are of interest is more challenging; the analyst needs to choose the correct model specification for the covariate-endpoint relationships, and the joint distribution of the endpoints. CEA may be sensitive to the choice of the model specification.

PS methods can potentially be used for subgroup-analysis, here they are required to balance baseline covariates between treatment and control groups within each subgroup. Just as relative costs and effectiveness can be heterogeneous across patients, in an observational setting it can also be expected that the mechanism of treatment assignment might systematically differ by patient subgroup. Balancing covariates at the subgroup level may prove particularly challenging: a PS approach has to recognise the differential treatment assignment mechanism, for example by estimating separate PS models for each subgroup.

The methodological guidance on decision models in CEA emphasises the need to appropriately represent various sources of decision uncertainty, such as parameter uncertainty, methodological uncertainty and structural uncertainty (Bojke et al., 2009). As Polsky and Basu (2006) illustrate, statistical methods that make different underlying assumptions can lead to substantially different cost-effectiveness results. This uncertainty, due to the choice of statistical method can be characterised as part of

structural uncertainty (Jackson et al., 2011), a relatively under-researched area (Gray et al., 2010). Here, detailed methodological guidance is currently lacking.

To conclude this section, the methodological literature in CEA typically considers regression and PS approaches for addressing selection bias. In an observational setting, high covariate imbalance can make estimates sensitive to the specification of the regression model. PS methods can only provide reliable estimates if good balance is achieved after PS adjustment. Each of these methods needs to acknowledge further challenges which arise when cost-effectiveness estimates for subgroups are of interest. CEA need to acknowledge the structural uncertainty due to the choice of statistical method used to address selection bias. The next section considers the main assumptions that underlie the methods currently recommended for CEA that use observational data.

2.3 Methodological guidance from the general causal inference literature

This review had two objectives: first, to inform the development of a new quality appraisal tool that focuses on the appropriate assessment of assumptions for statistical methods currently recommended for CEA (research paper 1), and second, to identify alternative statistical methods and consider their assumptions in the context of addressing selection bias in CEA (section 4).

I conducted a targeted review of the causal inference literature, including articles from the statistics, econometrics and epidemiology literature concerned with estimating the effects of exposures, treatments and policies, published between 1983 and 2011 (for example, Rosenbaum and Rubin, 1983, Imbens and Wooldridge, 2009a, Stuart, 2010, Morgan and Winship, 2007).

2.3.1 The potential outcomes framework

Following the causal inference literature on estimating treatment effects, I draw on the Neyman-Rubin potential outcomes framework (Imbens and Wooldridge, 2009a), and use its concepts and notation throughout.

Y_{ik} is the observed outcome (cost if $k = C$ and effectiveness if $k = E$), for individual[1] $i = 1, \ldots, n$, where n is the sample size. $t_i(0,1)$ is the indicator of the observed treatment[2]. The potential outcome of the individual that would be realised if she received the control is $Y_{ik}(0)$, and $Y_{ik}(1)$ if she received treatment. One of these outcomes - the actual treatment received - is observed, while the counterfactual is never observed.

The causal effect of the treatment is the difference between the potential outcomes for an individual[3]:

$$\tau_{ik} = Y_{ik}(1) - Y_{ik}(0).$$

The expected treatment effect across the whole population is the average treatment effect (ATE), given by

$$\tau_k = E(Y_{ik}(1) - Y_{ik}(0)).$$

Another relevant estimand is the average treatment effect on the treated (ATT), here the expectation is taken only for those who actually received the treatment. The individual

[1] For simplicity here I consider individuals, however treatment effects can also be calculated for different units such as practices and hospitals.

[2] Again, treatment is a generic term for any intervention or program. Here I consider a binary treatment (1= treated, 0 =control). The methods considered here can be generalised to several treatments.

[3] This review focuses on methods for analysing continuous outcomes (e.g. QALYs) and costs, where additive (incremental) effects are of interest. The potential outcomes framework also applies for different types of outcomes (e.g. count and binary data, or censored data such as survival time), where parameters such as odds ratios, relative risks or hazard ratios are of interest. These parameters are beyond the scope of this review.

level treatment effect (τ_{ik}) can be heterogeneous across the patient population, for example by subgroups, and due to this heterogeneity, the ATE and ATT will differ from each other, and across subgroups.

In the CEA context, a treatment effect of interest is the incremental net benefit, calculated as

$$INB = K * \tau_E - \tau_C,$$

where K is the willingness to pay.

In non-randomised settings individuals can be assigned to a treatment according to characteristics that are observed (x_i) and unobserved by the investigator. Under the assumption of "selection on observables", even if some of the unobserved characteristics influence treatment assignment, these factors are assumed not to influence the endpoint of interest, and not to be associated with unobserved factors influencing the endpoint. The mathematical counterpart of selection on observables is a combination of two assumptions: the unconfoundedness assumption and the overlap assumption. The combination of these two assumptions is also referred to as "strong ignorability" (Imbens and Wooldridge, 2009a).

2.3.2 The assumption of "no unobserved confounding"

The assumption of no unobserved confounding (also referred to as "unconfoundedness") states that, after controlling for a vector of observed covariates (x_i)[4], treatment assignment (t_i) is independent of both potential outcomes:

[4] Here I assume that the vector of observed confounders necessary for the unconfoundedness assumption to hold is the same for the cost and effectiveness endpoint. Research paper 2 presents a scenario where a

$$t_{i\_}||_{\_}\ (Y_{ik}(0), Y_{ik}(1))\ |x_i$$

Under this assumption, the allocation of two individuals who have similar observed characteristics but are in different treatment arms, can be thought of as effectively at random (Greenland et al., 1999). Approaches that use longitudinal data, such as panel data regression and "difference-in-differences", rely on a weaker form of this assumption; they assume that changes over time in unobserved confounders are conditionally independent of treatment (Imbens and Jeffrey, 2007).

The unconfoundedness assumption is not directly testable from the data (Imbens and Wooldridge, 2009a). Intuitively, it requires that the researcher has all relevant pre-treatment covariates at his or her disposal (Rubin, 2010). To consider this assumption, external evidence or expert opinion of the potential influence of observed and unobserved baseline covariates on treatment assignment, and endpoints need to be considered (Rubin, 2010). Causal diagrams can be useful for defining the structure of such relationships (Pearl, 2001). These considerations might be complemented with indirect statistical tests, so called "placebo tests" which can detect violations of the assumption (Imbens, 2004, Jones, 2007, Abadie et al., 2010). One possible implementation of these tests is to split those who did not receive the treatment into two control groups: one that was eligible for the treatment and one that was not, and estimate the "treatment effect" between these two groups (Imbens and Wooldridge, 2009a). After adjusting for the observed covariates, a nonzero estimated treatment effect between these groups might indicate that it is not plausible to assume

covariate which is a confounder for the QALY endpoint, the baseline QALY, is not associated with the cost endpoint.

unconfoundedness. However, a set of observed covariates “passing” the placebo test does not confirm that the unconfoundedness assumption is valid.

Another approach is to tests the sensitivity of the treatment effect estimates to the impact of potential unobserved confounders, for example using Rosenbaum’s method of sensitivity analysis (Rosenbaum, 2002). This method can characterise the strength of association between the potential unobserved confounder and treatment assignment, which is necessary to change the conclusions regarding the estimated treatment effect.

IV estimation is a recommended statistical approach that can account for both observed and unobserved confounding (Mullahy, 2011, Basu et al., 2007, Terza et al., 2008). IV methods assume that the instrument only influences endpoints through treatment, and is independent of the unobserved confounders (Hernán and Robins, 2006). This assumption is also untestable, however its plausibility can be assessed, for example, with causal diagrams informed by expert opinion and evidence from literature (Joffe and Mindell, 2006). Examples of proposed instruments in health economics are, the distance from hospital (Basu et al., 2007), or in Mendelian randomisation studies, the genotype (Didelez and Sheehan, 2007). In most CEAs, however, finding a valid instrument is challenging (Polsky and Basu, 2006). A further challenge of IV estimation is that, instead of ATEs, a feasible estimand is often the local average treatment effect, which can differ from the ATE when individual treatment effects are heterogeneous. Therefore, for CEA, statistical methods that assume unconfoundedness warrant consideration.

2.3.3 The overlap assumption

The overlap assumption (also known as the assumption of positivity, or the ”experimental treatment assignment” assumption) (Westreich and Cole, 2010) requires

that for any combination of covariate values, there is a nonzero probability of being assigned to each treatment arm. For a binary treatment, this requires that

$$0 < pr(t_i = 1 \,|x_i) < 1).$$

Intuitively this implies that no covariate or combination of covariates perfectly determine treatment assignment. The overlap assumption can be assessed using the data, for example by inspecting histograms or density plots (continuous covariates) and by reporting standardised differences (categorical or continuous covariates) (Imbens and Wooldridge, 2009a, Busso et al., 2011).

Poor overlap, or structural violations of the positivity assumption, are present when at certain combinations of covariates, it is impossible to observe both treated or control individuals (Westreich and Cole, 2010). In such cases, in order to identify the ATE for the whole population, statistical methods need to extrapolate. For example, if in certain age groups there are no treated individuals, treatment effects for these patients need to be extrapolated using information on younger patients. Regression methods make this extrapolation automatically, and can hide overlap problems (Crump et al., 2009, Ho et al., 2007, Westreich and Cole, 2010). If the regression model specification is incorrect, this can result in biased parameter estimates. A common remedy is to estimate the treatment effect only for the subsamples where there is good overlap (Crump et al., 2009). However, this approach of dropping treated and control individuals can lead to an estimated treatment effect for a population that differs from that of interest for the policy maker.

In finite samples, even when overlap is reasonable, "practical positivity violations" often arise: in certain covariate strata there might be small numbers or no individuals from each treatment group (Westreich and Cole, 2010). This can lead to the probability

of treatment assignment to be close to 0 or 1, which can be problematic for methods that use the inverse of the PS for addressing selection bias (Kang and Schafer, 2007). This problem can be especially severe in reduced sample sizes of patient subgroups.

This thesis follows the terminology of the econometrics literature (Imbens and Wooldridge, 2009a) and defines both structural and practical violations of the positivity assumption as poor overlap.

2.4 Currently recommended methods for accounting for selection bias in CEA

2.4.1 Regression methods

A commonly used regression approach for covariate adjustment for CEA is the net benefit regression framework (Hoch et al., 2002). Here, the treatment effect is estimated on the individual net benefit endpoint, for example by using ordinary least squares (OLS) regression. This approach accounts for the correlation between individual costs and effects. However, it assumes the same, linear functional form relationship between the covariates and the cost and effectiveness endpoints for a fixed value of WTP, which may not be plausible (Nixon and Thompson, 2005).

A more flexible approach is to model the cost and effectiveness endpoints using GLMs (Barber and Thompson, 2004b). The advantage of GLMs is their potential to address skewed endpoints and nonlinear response surfaces, where response surface describes the functional form relationship between the covariate and the endpoint (Rubin, 1979). GLMs can predict expected endpoints on the original scale of interest. Following Barber and Thompson (2004), GLMs for Y_{ik} can be written as

$$g_k(\mu_{i,k}) = \gamma_k t_i + x_i \beta_k; \quad Y_{ik} \sim F_k. \qquad (1)$$

Here $\mu_{i,k} = E(Y_{i,k})$ is the expectation of $Y_{i,k}$, g_k is the link function which describes the scale on which x_i are related to $Y_{i,k}$, γ_k and β_k are the regression coefficients, and F_k is an exponential family distribution. If treatment effects for subgroups are of interest, treatment-covariate interactions in the linear predictor can account for heterogeneous treatment effects (Nixon and Thompson, 2005). Parameters can be estimated via maximum likelihood (ML), quasi ML or Bayesian methods (Basu and Manca, 2011). Bivariate models (Nixon and Thompson, 2005) or non-parametric bootstrap (Davison and Hinkley, 1997) can be used to recognise the joint uncertainty in the estimates of the incremental costs and effectiveness, for example by estimating confidence intervals (CIs) around the INB, or cost-effectiveness acceptability curves (CEACs).

For unbiased estimation with GLMs, correct specification of the link function and linear predictor is needed. For efficient estimates, the correct specification of the error distribution is also necessary. With certain data typical of CEA, for example cost or HRQoL data with large spikes in their distributions, it is recommended that GLMs are extended, for example, by applying two-part models (Buntin and Zaslavsky, 2004, Basu, 2011). Further flexible approaches are semi-parametric methods such as extended estimating equations (Basu and Rathouz, 2005), or beta-type size distributions (Jones et al., 2011), which, however, are rarely used in practice (Mihaylova et al., 2010). A general concern is that, even a flexible parametric approach is not a substitute for finding the correct model specification (Manning et al., 2005). Non-parametric methods such as the quintile regression and the discrete conditional density estimator (Gilleskie and Mroz, 2004), or machine learning algorithms can be promising alternatives (Austin, 2012).

A general way to obtain ATEs, across a wide range of regression models, is with the method of recycled predictions (Basu and Rathouz, 2005), which is equivalent to the G-computation estimator of a point treatment (Robins, 1986, Imbens and Wooldridge, 2009a, Imbens, 2004). This method predicts the expected potential outcomes for each individual, under treated and control states:

$$\hat{\tau}_{k,reg} = \frac{1}{n}\sum_{i=1}^{n}\{\hat{\mu}_{i,k}(x_i, t_i = 1) - \hat{\mu}_{i,k}(x_i, t_i = 0)\}, \quad (2)$$

where $\hat{\mu}_{i,k}(.)$ is the predicted mean of $Y_{i,k}$ from the GLM in equation (1), given x_i, with t_i set to 1 and then to 0 for the whole sample. Standard errors around the estimated ATE can be obtained using the non-parametric bootstrap.

2.4.2 Propensity score matching

Matching methods aim to create treated and control groups with balanced covariate distributions. Ideally, treated and control units can be exactly matched on their observed covariate values. However, with high dimensional, continuous confounders, exact matching is not possible. Instead, the PS can be used as a balancing score. The PS is the conditional probability of treatment assignment given x_i (Rosenbaum and Rubin, 1983),

$$p_i = Pr(t_i = 1|x_i).$$

The estimated PS, $\hat{p}_i$, is generally obtained as a prediction from a logistic regression model (Westreich et al., 2010). The PS can create balance between treated and control comparison groups, and can be used in several ways: for matching, using the inverse of the PS for weighting the sample, stratifying on the PS, and including the PS as a covariate in regression (Rosenbaum and Rubin, 1983). Matching and weighting – selected for further investigation – have been shown to dominate stratification and

regression in terms of accuracy (Austin, 2009b, Lunceford and Davidian, 2004, Austin, 2009c, Austin et al., 2007).

In PS matching, matched treated and control comparison groups can be created, using $\hat{p}_i$ as a distance metric (Rosenbaum and Rubin, 1983). For each subject, the missing potential outcome, $Y_{i,k}(0)$ or $Y_{i,k}(1)$ is imputed, using the observed outcomes of the closest matches:

$$\hat{Y}_{i,k}(0, x_i) = \begin{cases} Y_{i,k} & \text{if } t_i = 0 \\ \frac{1}{M} \sum_{j \epsilon \zeta_M(i)} Y_{j,k} & \text{if } t_i = 1 \end{cases}$$

$$\hat{Y}_{i,k}(1, x_i) = \begin{cases} \frac{1}{M} \sum_{j \epsilon \zeta_M(i)} Y_{j,k} & \text{if } t_i = 0 \\ Y_{i,k} & \text{if } t_i = 1 \end{cases}$$

where $\zeta_M(i)$ is the set of *M* individuals matched to unit i. The matching estimator for the ATE is the mean of the estimated individual-level treatment effects:

$$\hat{\tau}_{k,match} = \frac{1}{n} \sum_{i=1}^{n} \left\{ \hat{Y}(1, x_i) - \hat{Y}(0, x_i) \right\} \qquad (3)$$

The true PS is a balancing score: conditional on the PS, treatment and control groups are expected to have the same distribution of observed baseline characteristics. Matching on a correctly specified PS can therefore eliminate selection bias (Rubin and Thomas, 1992). However, in an observational setting, the specification of the true PS is generally unknown. If the PS is misspecified, for example by incorrectly omitting higher order terms from the logistic model, or ignoring differences between subgroups in the treatment assignment mechanism, then PS matching can lead to biased estimates of treatment effects (Cole and Hernán, 2008).

An appropriate check of the PS specification is whether covariate distributions between the treatment groups are balanced in the matched sample (Stuart, 2010). Recommended ways of assessing balance include calculating the standardised mean differences of

covariates between matched treatment groups (Rosenbaum and Rubin, 1985, Austin, 2009a), defined as

$$d = \frac{\bar{x}_{treatment} - \bar{x}_{control}}{\sqrt{\frac{s^2 treatment + s^2 control}{2}}}, \qquad (4)$$

, where $\bar{x}$ and s^2 denote the covariate's weighted means and variances by treatment group. It is suggested (Stuart, 2010, Diamond and Sekhon, 2012) that the balance needs to be assessed not only on the means, but on the full distribution of the covariates, e.g. by comparing empirical quantile-quantile (EQQ) plots between the treatment groups (Diamond and Sekhon, 2012, Basu et al., 2008, Austin, 2008, Ho et al., 2007). If the balance between the treatment groups is poor after matching, the analyst should try to improve it by re-estimating the PS (Dehejia and Wahba, 2002, Imbens and Wooldridge, 2009b). If treatment effects for subgroups are of interest, this process of balance checks should be performed for each subgroup. This approach can, however, result in subjective decisions, and is rarely followed (Austin, 2008).

There is no consensus in the methodological literature on the estimation of variance for matching approaches (Hill, 2008, Hill and Reiter, 2005, Abadie and Imbens, 2006b, Austin, 2008, Stuart, 2010). Ideally, the estimated variance of the treatment effect should include the variance due to the estimation of the PS, and account for the dependences in the data created by the matching process (Hill and Reiter, 2005, Hill, 2008). A general suggestion is to estimate standard errors conditional on the matched data (Ho et al., 2007), for example using analytical standard errors from regression models applied on the matched data, or the non-parametric bootstrap. In CEA, the non-parametric bootstrap, by re-sampling individual cost and effectiveness pairs, can maintain the correlation between the incremental cost and effectiveness parameters (Sekhon and Grieve, 2011). While this approach does not account for the uncertainty

due to the estimated PS, under relatively general circumstances, it expected that using the estimated PS instead of the true PS provides conservative variance estimates (Stuart, 2010). Analytical variance formulas which can account for the matching process are subject to ongoing research (Abadie and Imbens, 2009, Abadie and Imbens, 2006a), and cannot be readily applied for the bivariate context of CEA.

2.4.3 Inverse probability of treatment weighting

A further recommended use of the estimated PS is inverse probability of treatment weighting (IPTW). IPTW can estimate ATEs, by reweighting the observed cost and effectiveness endpoints for treated and control samples. The IPT weight, w_i, is the inverse of the estimated probability of receiving the observed treatment:

$$w_i = \frac{t_i}{\hat{p}_i} + \frac{1-t_i}{1-\hat{p}_i} \qquad (5)$$

It is recommended that in practice the normalised IPTW estimator is implemented, where weights are divided with the sum of weights for the respective treatment group (Hirano and Imbens, 2001, Kang and Schafer, 2007a):

$$\hat{\tau}_{k,IPTW} \frac{\sum_{i=1}^{n} t_i w_i Y_{i,k}}{\sum_{i=1}^{n} t_i w_i} - \frac{\sum_{i=1}^{n} (1-t_i) w_i Y_{i,k}}{\sum_{i=1}^{n} (1-t_i) w_i}$$

In CEA, IPTW has been introduced for reducing selection bias in cost analysis (Basu et al., 2011) and for handling censored cost (Pullenayegum and Willan, 2011) and cost-effectiveness (Willan et al., 2002, Willan et al., 2005) data. IPTW has not previously been considered for addressing selection bias in CEA.

If the PS is correctly specified, IPTW can provide consistent estimates and reach semi-parametric efficiency (Hirano et al., 2003). This can be particularly attractive for subgroup analysis, where sample sizes at the subgroup level can be small. Covariate

balance can be assessed following IPTW according to standardised differences (Austin, 2009a), where the means and standard deviations in equation (4) are weighted by the IPT weights.

Misspecification of the PS, for example ignoring subgroup-specific treatment assignment, can cause IPTW to be biased, and it is expected to report more bias than PS matching. While for matching it is sufficient for the estimated PS to be a balancing score, for weighting, the PS needs to be the correct conditional probability of treatment assignment (Busso et al., 2009, Waernbaum, 2011).

Even when the PS model is correctly specified, poor overlap can result in PS values close to 0 and 1, and unstable IPT weights, which can lead to estimates of ATEs that are biased and inefficient (Kang and Schafer, 2007a, Lee et al., 2010, Busso et al., 2011). Such challenges are likely to arise in CEA, where covariate imbalance can be high (Grieve et al., 2008), possibly resulting in poor overlap.

After applying IPTW, uncertainty can be calculated using the sandwich estimator of a weighted regression of the endpoint on the treatment indicator [5] or the non-parametric bootstrap (Lunceford and Davidian, 2004), for example. Both approaches can maintain the correlation between the incremental cost and effectiveness parameters.

2.4.4 Structural uncertainty from the choice of statistical method to address selection bias

Each of the statistical approaches described above make assumptions that cannot be directly tested, for example the unconfoundedness assumption, or the assumption of

[5] This weighted regression estimator, often used in applied work, corresponds the normalised IPTW estimator defined earlier (Busso et al., 2009).

correct regression model specification. This implies that no one approach is ideal, and the choice of statistical method for estimating cost-effectiveness parameters from patient-level observational data can contribute to structural uncertainty (Bojke et al., 2009, Jackson et al., 2011). This type of structural uncertainty can be incorporated in CEA, by considering the impact of choosing alternative statistical methods on the estimated cost-effectiveness. In some cases structural uncertainty can be quantitatively incorporated in the analysis, for example using Bayesian model averaging for weighting regression models according to some measure of model adequacy (Jackson et al., 2011). Another approach, that can quantify the uncertainty due to possible violations of the unconfoundedness assumption, is calculating bounds around the estimated treatment effect, using Rosenbaum's method of sensitivity analysis (Rosenbaum, 2002). A more general approach is to repeat the analysis with alternative assumptions and carefully report and interpret their impact on the cost-effectiveness results. An example is provided by Polsky and Basu (2006) who report results obtained with regression, alongside with PS matching and IV methods. The critical appraisal tool, developed in research paper 1, provides guidance on addressing structural uncertainty from the choice of the statistical method to address selection bias in CEA.

2.4.5 Summary: underlying assumptions of statistical methods currently recommended to address selection bias in CEA

This section reviewed the main assumptions behind statistical methods currently proposed for addressing selection bias in CEA. These assumptions include:

1. Unconfoundedness.
2. When using IV estimation, the IV only influences the endpoint through the treatment.
3. Good overlap between covariate distributions of the treatment groups.

4. Correct specification of the endpoint regression model.
5. Correct specification of the PS model.

The critical appraisal tool, developed and applied in research paper 1, provides detailed guidance on how these assumptions can be assessed for CEA that use patient-level observational data, and gives guidance on how structural uncertainty from the choice of statistical method can be acknowledged.

In this conceptual review, I found that in CEA, some of these assumptions are unlikely to hold. For example, while IV methods can potentially handle selection bias due to both observed and unobserved confounding, finding an appropriate IV can be challenging for CEA (Polsky and Basu, 2006). While in health economics, genetic variation has been recently proposed as a type of valid instrument (Mullahy, 2011), its appropriateness for CEA has not been investigated. This thesis therefore considers methods that rely on the unconfoundedness assumption.

For these methods, including regression, PS matching and IPTW, in realistic circumstances of CEA, it is expected that three main challenges will prevail: misspecification of the endpoint regression model, misspecification of PS, and poor overlap. The previous sections reviewed the expected performance of these methods under realistic circumstances. This is summarised in Table 2.1.

Table 2.1 - The expected performance of currently recommended methods, under realistic circumstances in CEA

	Misspecification of endpoint regression model	Misspecification of PS	Poor overlap
Regression	Possible bias and inefficiency.	Not relevant.	Performance depends on correct specification of endpoint model. If misspecified, poor overlap magnifies bias.
PS matching	Not relevant.	Possible bias and inefficiency.	PS makes overlap explicitly testable. If weak overlap, treated observations might have to be dropped, alternatively bad quality matches, leading to bias and inefficiency.
IPTW	Not relevant.	Possible bias and inefficiency. More sensitive to misspecification than matching.	Unstable IPT weights, bias and inefficiency.

In the systematic review of research paper 1, I find that most applied CEA used regression and matching methods, including exact matching and PS matching. The review also revealed that the main underlying assumptions of these methods were not appropriately assessed. This motivates the consideration of alternative statistical methods for addressing selection bias in CEA, which rely on less restrictive assumptions. The following sections review these methods.

2.5 Statistical approaches identified in the general causal inference literature that have the potential to reduce selection bias in CEA

In this section, I consider further statistical methods that, based on my conceptual literature review, were deemed promising for addressing selection bias in CEA. Each of these methods has the potential to make more plausible underlying assumptions in the

CEA context, than the methods currently recommended for CEA. Genetic Matching relaxes the assumption of the correctly specified PS model, by aiming to directly maximise covariate balance, using machine learning. Double-robust methods and regression-adjusted matching exploit information from the endpoint regression and the PS models, and can be unbiased even if one of these models is misspecified. I also consider machine learning approaches for estimating the PS and the endpoints, which relax the assumption of knowing the correctly specified, fixed parametric models.

2.5.1 Genetic Matching

Genetic Matching (GM) is a multivariate matching approach whose explicit aim is to optimise covariate balance (Diamond and Sekhon, 2012, Ramsahai et al., 2011, Sekhon, 2011, Sekhon and Grieve, 2011). GM extends standard PS matching in two ways. First, instead of the manual process of modifying the PS and re-assessing covariate balance, GM harnesses an automated search algorithm that iteratively checks balance on observed confounders, and directs the search towards those matches that optimise balance (Diamond and Sekhon, 2012, Sekhon, 2011). Secondly, the GM algorithm can maximise covariate balance by matching on individual covariates as well as the PS. GM is a multivariate matching method that uses a generalised version of Mahalanobis distance (MD) metric. The MD between any two column vectors (here representing the covariate values of a treated and a control individual) is:

$$MD(x_i, x_j) = \left\{(x_i - x_j)^T S^{-1} (x_i - x_j)\right\}^{\frac{1}{2}}$$

, where *S* is the sample covariance matrix of x. The distance metric for GM contains an extra weight W, with dimensions of a $k \times x$ matrix with k free parameters in the diagonal, where *k* is the number of covariates:

$$GMD(x_i, x_j) = \left\{(x_i - x_j)^T (S^{-1/2})^T W S^{-1/2} (x_i - x_j)\right\}^{\frac{1}{2}}$$

The algorithm chooses the free parameters of the W matrix based on a loss function which minimises the imbalance between the covariate distributions of the matched treated and control groups. To measure imbalance, paired t-tests assess the equality of the means and Kolmogorov-Smirnov (KS) distributional tests assess covariate balance across the whole distribution. Using the optimal *W* matrix, matched pairs are selected, for example performing 1:1 matching, and the matching estimator can be obtained as described by equation (3).

As a default, the variables that receive the highest weight in the loss function are those that have the worse balance at each stage of the optimisation process. However, a researcher might want to ensure that good balance is achieved on a subset of covariates designated as "high priority". For example, prior clinical evidence might indicate that a covariate is particularly prognostic for the endpoint. GM can be tailored to prioritise achieving covariate balance on such high priority variables (Ramsahai et al., 2011), by modifying the loss function. The ability of GM to reduce bias hence relies on the correct specification of the loss function.

While it is an option to use the estimated PS in the matching process, GM does not rely on knowing the correct PS. Sekhon and Grieve (2011) report that when the PS is misspecified, GM can improve covariate balance, and reduce bias and variability in the cost-effectiveness estimates. For subgroup analysis, the GM algorithm can be modified to maximise balance for each subgroup.

A challenge GM shares with other multivariate matching methods is that matching on a high dimensional covariate vector can lead to finite sample bias and loss of precision (Abadie and Imbens, 2006a). This can be of particular concern when reporting cost-

effectiveness results by subgroup as sample sizes can be relatively small. GM has not been considered for subgroup analysis in CEA before, nor compared to alternatives other than PS matching.

2.5.2 Approaches that combine the PS with regression for the endpoints

Some commentators raise doubts about regression approaches, where model selection can be influenced by its consequences for the estimated treatment effects (Rubin, 2008, Rubin, 2007, Ho et al., 2007). In contrast, PS methods do not require information on the endpoint, and can be regarded as more objective. However, if some confounders are highly prognostic for the endpoint, even small imbalances that remain in these covariates after matching can translate into large biases in the estimated treatment effects. PS methods alone cannot formally take into account information on the confounder-endpoint relationship (Stuart, 2010). In CEA, where selection bias needs to be addressed for both the cost and the effectiveness endpoints, the PS might balance the most important confounders for one endpoint, but not for the other, leading to bias. It is recommended that PS and other matching methods form the design stage of an observational study, and are followed by regression modelling of the endpoint (Polsky and Basu, 2006, Rubin, 1973, Imbens and Wooldridge, 2009a, Rubin, 2007). The following sections describe two combined approaches: double-robust methods, and regression-adjusted matching.

Double-robust methods

Double-robust (DR) methods, proposed by Robins and colleagues (Robins et al., 1995, Bang and Robins, 2005, Robins et al., 2007) combine models for the PS and for the endpoint, with most implementations using the PS as IPT weights (Kang and Schafer, 2007b). The distinctive property of DR estimators is that they are consistent if either

(but not necessarily both) the PS or the endpoint regression model is correctly specified. If both components are correct, the DR estimator can be a semi-parametric efficient estimator (Robins et al., 2007).

In the context of CEA, DR estimators require specifying a model for both the cost and the effectiveness endpoints, as well as for the PS. If treatment effect estimates for subgroups are required, treatment-covariate interactions can be included in the regression models. DR methods have been proposed for addressing censoring (Pan and Zeng, 2011, Bang and Tsiatis, 2000), or selection bias in cost analyses (Basu et al., 2011), but have not been considered for addressing selection bias in CEA. Below I review two DR methods that are commonly used in the causal inference literature, augmented IPTW (AIPTW) and weighted regression. I also describe a recently proposed DR method, targeted maximum likelihood estimation (TMLE).

Commonly used DR methods

AIPTW (Robins et al., 1994, Basu et al., 2011) weights residuals from a regression model with the IPT weights. The AIPTW estimator is

$$\hat{\tau}_{k,AIPTW} \frac{\sum_{i=1}^{n} t_i w_i (Y_{i,k} - \hat{\mu}_{i,k}(x_i))}{\sum_{i=1}^{n} t_i w_i} - \frac{\sum_{i=1}^{n} (1 - t_i) w_i (Y_{i,k} - \hat{\mu}_{i,k}(x_i))}{\sum_{i=1}^{n} (1 - t_i) w_i} +$$

$$\frac{1}{n} \sum_{i=1}^{n} \{\hat{\mu}_{i,k}(x_i, t_i = 1) - \hat{\mu}_{i,k}(x_i, t_i = 0)\},$$

where $\hat{\mu}_{i,k}(.)$ is the predicted endpoint from a regression model, for example from a GLM defined in equation (1), and w_i is the IPT weight, defined in equation (5).

An alternative is the weighted regression estimator (Freedman and Berk, 2008, Kang and Schafer, 2007a), which weights the endpoint regression, for example a GLM, with w_i (Freedman and Berk, 2008, Kang and Schafer, 2007a). ATEs can be obtained using the method of recycled predictions,

$$\hat{\tau}_{k,wreg}\frac{1}{n}\sum_{i=1}^{n}\{\hat{\mu}_{i,k,wreg}(x_{i,},t_i=1)-\hat{\mu}_{i,k,wreg}(x_{i,},t_i=0)\},$$

where $\hat{\mu}_{k,wreg}(.)$ *is* the prediction from a weighted regression.

Due to the DR property, these methods are consistent even if one of the PS or the regression model are misspecified. Under circumstances of poor overlap, in theory, a correctly specified regression can help with extrapolation (Petersen et al., 2010), hence can reduce bias compared to using IPTW alone. DR methods can also increase efficiency, because using the predicted endpoint can stabilise weights (Glynn and Quinn, 2010).

In realistic settings, such as when there is poor overlap and the PS and the endpoint models are misspecified, DR methods can provide biased and inefficient estimates of ATEs (Kang and Schafer, 2007a, Porter et al., 2011, Freedman and Berk, 2008, Basu et al., 2011). An ongoing debate discusses the relative merits of different DR estimators under these circumstances (Porter et al., 2011, van der Laan and Gruber, 2010, Robins et al., 2007). It has been suggested that DR estimators should have a "boundedness property": they should respect the known bounds of the endpoint; so that the estimated parameter will always fall within the parameter space, i.e. the realistic range of values for that parameter (Robins et al., 2007, Rotnitzky et al., 2012). A recently proposed DR method , targeted maximum likelihood estimation, can have this boundedness property (Gruber and van der Laan, 2010a, Gruber and van der Laan, 2012a).

Targeted maximum likelihood estimation

While standard maximum likelihood estimation aims to find parameter values that maximise the likelihood function for the whole data, targeted maximum likelihood estimation (TMLE) is concerned with a particular feature of the distribution such as the

ATE (van der Laan and Rubin, 2006, Moore and van der Laan, 2009). Maximising a global likelihood function may not yield the least biased estimate of the target parameter, so TMLE is designed to target the initial estimate so as to reduce bias in the estimate of the parameter of interest. Performing TMLE involves two stages (Gruber and van der Laan, 2012b). In the first stage, an initial estimate of the conditional mean of Y_i , given t_i and x_i , is obtained. In practice, this involves using a regression approach to predict the expected potential outcomes conditional on the observed confounders, $\hat{\mu}_{i,k}(x_i, t_i = 0)$ and $\hat{\mu}_{i,k}(x_i, t_i = 1)$.

In the second stage, these initial estimates of the regression functions for the potential outcomes are fluctuated, exploiting the information in the treatment assignment mechanism, using $\hat{p}$, the estimated PS. For the ATE, the fluctuation corresponds to extending the parametric regression model for $Y_{i,k}$ with "clever covariates" (h), which are defined similarly to the IPT weights:

$$h_0(t, x) = \frac{1 - t_i}{1 - p_i}$$

$$h_1(t, x) = \frac{t_i}{p_i}$$

For continuous endpoints, it is recommended (Gruber and van der Laan, 2012a, Gruber and van der Laan, 2010a) that known bounds of the endpoint are exploited, by rescaling Y to between 0 and 1, to ensure that TMLE has the boundedness property. Then the fluctuation can be performed on the logistic scale: logistic regressions are fitted with the transformed endpoint on the left hand side, using the initial prediction as an offset, and the clever covariates as regressors. This regression can be interpreted as explaining the residual variability of the predicted endpoint, using information from the treatment assignment mechanism. The resulting targeted estimates of the expected potential

outcomes, $\hat{\mu}^1{}_{i,k}(x_i, t_i = 0)$ and $\hat{\mu}^1{}_{i,k}(x_i, t_i = 1)$ are used in the G-computation formula to obtain the TMLE estimator:

$$\hat{\tau}_{k,TMLE} = \frac{1}{n}\sum_{i=1}^{n}\left\{\hat{\mu}^1{}_{\mathrm{i,k}}(x_{i,}, t_i = 1) - \hat{\mu}^1{}_{\mathrm{i,k}}(x_{i,}, t_i = 0)\right\},$$

TMLE has the property of double-robustness, and if both components are correct, it reaches the semi-parametric efficiency bound (van der Laan, 2010). As any DR method, TMLE can be biased if both the endpoint and the PS model are misspecified, and can be sensitive to unstable IPT weights when overlap is poor (Porter et al., 2011). However, TMLE was demonstrated to report relatively low bias compared to other DR methods when machine learning techniques were used to obtain the initial estimate of the endpoint, and to estimate the PS (Porter et al., 2011).

TMLE has not been considered in CEA before. The method has particular appeal for estimating treatment effects on data where known bounds can be exploited, for example with common HRQoL endpoints, where distributions are bounded at small negative values and 1, or cost data that is typically bounded at 0. Standard errors can be estimated using the influence curve (van der Laan, 2010), but also with the non-parametric bootstrap. The latter method can be used for any DR method including AIPTW and weighted regression, and can maintain the correlation between the incremental cost and effectiveness parameters.

Regression-adjusted matching methods

The causal inference literature generally recommends that matching is followed by regression adjustment (Rubin, 1973, Rubin and Thomas, 2000, Abadie and Imbens, 2006a). The idea is similar to regression-adjustment in randomised trials: regression is used to "clean up" imbalances between treatment groups after matching (Stuart, 2010).

Here, I review two implementations: matching as non-parametric pre-processing (Ho et al., 2007) and bias-corrected matching (Abadie et al., 2004, Abadie and Imbens, 2011).

Matching as non-parametric pre-processing

Ho et al. (2007) proposes undertaking matching - for example PS matching or GM - as the first step of the analysis, and then use the frequency weights from the matching to weight endpoint regression models, for example GLMs. Using this approach, the regression-adjusted matching estimator (Hill and Reiter, 2005) of the ATEs for each endpoint can be obtained as:

$$\hat{\tau}_{k,reg-match} = \frac{1}{n}\sum_{i=1}^{n}\{\hat{\mu}_{i,k,reg_{match}}(x_{i,}, t_i = 1) - \hat{\mu}_{i,k,reg_{match}}(x_{i,}, t_i = 0)\},$$

where $\hat{\mu}_{i,k,reg_{match}}(.)$ are the predicted potential outcomes obtained from applying regression models on the matched data.

Regression-adjusted matching is expected to reduce finite sample bias and increase efficiency compared to matching alone (Hill and Reiter, 2005, Ho et al., 2007). Applying a regression model with treatment by covariate interactions can also facilitate hypothesis testing for differences between treatment effects among subgroups of interest. This approach can reduce the sensitivity of the estimated ATEs to the specification of the endpoint model (Ho et al., 2007). In CEA, regression-adjustment has been proposed after matching in CEA, in order to reduce conditional bias and to test the robustness of results obtained after matching (Sekhon and Grieve, 2011). Standard errors and confidence intervals of cost-effectiveness estimates can be obtained conditional on the matched data (Ho et al., 2007). No guidance on the implementation of this approach has been provided for CEA, nor has its performance been compared with alternative methods.

Bias-corrected matching

Bias-corrected matching (BCM) (Abadie et al., 2004, Abadie and Imbens, 2011) adjusts the imputed potential outcome with the difference in the predicted outcome that can be attributed to covariate imbalances between the matched pairs. The predicted potential outcomes are obtained using regression models of the endpoint on covariates, stratified by treatment assignment. The bias-corrected predictions of the potential outcomes are:

$$\hat{Y}(0,x_i) = \begin{cases} Y_{i,k} & \text{if } t_i = 0 \\ \frac{1}{M}\sum_{j\epsilon\zeta_M(i)} Y_{j,k} + \hat{\mu}_{i,k}(x_{i,}, t_i = 0) - \hat{\mu}_{i,k}(x_{j,}, t_i = 0) & \text{if } t_i = 1 \end{cases},$$

$$\hat{Y}(1,x_i) = \begin{cases} \frac{1}{M}\sum_{j\epsilon\zeta_M(i)} Y_{j,k} + \hat{\mu}_{i,k}(x_{i,}, t_i = 1) - \hat{\mu}_{i,k}(x_{j,}, t_i = 1) & \text{if } t_i = 0 \\ Y_{i,k} & \text{if } t_i = 1 \end{cases}.$$

For example, for an individual i who received control, the imputed potential outcome under treatment is the average outcome of its M closest matches from the treatment group (indexed by *j*), adjusted with the difference between the predicted outcomes under treatment, when covariate values are set to those of its own values, $\hat{\mu}_{i,k}(x_{i,}, t_i = 1)$ and the values of the match , $\hat{\mu}_{i,k}(x_{j,}, t_i = 1)$. The corresponding estimator, analogously to the PS matching estimator, is the mean of the estimated individual-level treatment effects:

$$\hat{\tau}_{k,BCM} = \frac{1}{n}\sum_{i=1}^{n}\{\hat{Y}(1,x_i) - \hat{Y}(0,x_i)\}$$

BCM is consistent if $\hat{\mu}_{i,k}(x,0)$ and $\hat{\mu}_{i,k}(x,1)$ are consistent estimators for the potential outcomes (Abadie and Imbens, 2011). It has been shown that for reducing bias, the correct specification of the regression model is not essential, for example, for moderate nonlinearities in the response surface, adjustment even with a linear model can remove most bias (Rubin, 1973, Rubin and Thomas, 2000, Abadie and Imbens, 2011, Busso et al., 2011). CEA typically has highly nonlinear response surfaces, and so adjustment

with linear models might be insufficient. One approach would be to consider non-parametric regression methods, such as series estimation which have been also recommended for bias-adjustment (Abadie and Imbens, 2011). However, such flexible approaches have not been considered in applied studies or simulations that implement BCM (Busso et al., 2011, Abadie and Imbens, 2011). In scenarios relevant to CEA, for example when overlap is poor, BCM has been shown to outperform reweighting approaches such as IPTW and DR methods (Busso et al., 2011). BCM has not been considered for CEA before.

2.5.3 Machine learning estimation for the PS and the endpoint regression

Each of the methods reviewed in the previous sections can use models for the costs and effectiveness endpoints, or a model for the PS. Single methods - for example, regression, or PS matching - can use these models to estimate ATEs, or they can be combined, for example in DR methods or in regression-adjusted matching. GM extends PS matching, by using a machine learning algorithm to directly maximise balance. In general, machine learning covers a wide range of classification and prediction algorithms (Westreich et al., 2010, Austin, 2012). Unlike approaches that assume a fixed statistical model, for example a GLM with a log link, machine learning aims to extract the relationship between the endpoint and covariates through a learning algorithm, without choosing one specific model *a priori* (Lee et al., 2010). These approaches can be used to estimate the PS (Westreich et al., 2010), and the endpoint regression functions (Austin, 2012), and are reviewed in the next section.

Machine learning estimation of the PS

A common approach to PS estimation is to use logistic regression models without interaction or higher order terms. Assumptions behind the logistic regression, for example the linearity of the relationship between covariates and the logit, are rarely assessed (Westreich et al., 2010). More flexible modelling approaches, such as series regression estimation (Hirano et al., 2003), and methods from the machine learning literature (Westreich et al., 2010), can increase the chances of correctly specifying the PS. Machine learning methods that can be used for PS modelling, include decision trees, neural networks, linear classifiers and boosting methods (Austin, 2012, Lee et al., 2010, van der Laan, 2007). While these algorithms mostly choose an estimated PS based on the predictive performance of a model for the binary indicator of treatment receipt, the ultimate test of the specification is whether the PS balances the distribution of potential confounders between the treatment groups (Stuart, 2010).

This thesis considers boosted classification and regression trees (CART), which can reduce bias in the estimated ATE when compared to using a misspecified logistic regression, and also compared to alternative machine learning approaches (Lee et al., 2010). The algorithm fits regression trees on random subsets of the data, and in each iteration, the data points which were incorrectly specified with the previous trees receive greater priority. The algorithm can be specified to stop when the best balance - measured as mean standardised differences or KS tests - is achieved (McCaffrey et al., 2004, Lee et al., 2010). As well as specifying a balance measure, tuning parameters, such as the number of iterations, depth of interactions or shrinkage parameters need to be chosen.

Machine learning estimation of the endpoint regression function

For estimating treatment effects, the expected potential outcomes need to be predicted under treated and control states, using regression functions of the endpoint. Here, recommended machine learning methods include bagged regression trees, random forests or boosted regression trees (Austin, 2012); and the "super learning" algorithm (van der Laan, 2007), considered for this thesis.

Super learning uses a collection of prediction algorithms pre-selected by the user, potentially exploiting subject-matter knowledge of the data-generating mechanism for the endpoint. For example, for CEA, if a multiplicative relationship between the covariates and the cost endpoint is likely, a GLM with log link can be included among the prediction algorithms. The super learner algorithm uses cross-validation to select a weighted combination of estimates reported by the prediction procedures (Polley and van der Laan, 2010). The selected combination is proposed to be asymptotically optimal (van der Laan and Dudoit., 2003): if the correct model is among the candidates, the algorithm is expected to select it. The predicted potential outcomes can then be used in a regression estimator (see equation 2) or in combined methods such as TMLE (Porter et al., 2011) or BCM.

There is little known about the performance of machine learning regression estimators when there is poor overlap between covariate distributions. It is expected that the increased flexibility of these methods can reduce model misspecification, as well as reduce bias from extrapolation with an incorrect model (Porter et al., 2011). A drawback of machine learning methods is computational time, especially if the non-parametric bootstrap is used for calculating standard errors for estimated treatment effects (Austin, 2012).

2.5.4 Summary: promising methods from the causal inference literature, for addressing selection bias in CEA

The objective of this section was to review promising statistical methods from the general causal inference literature which can be used in CEA to address selection bias, and to evaluate their appropriateness for CEA. Beyond the approaches recommended by the methods literature in CEA, such as regression, PS matching and IPTW, the following methods were deemed promising for CEA: GM, DR methods and regression-adjusted matching. I also considered machine learning methods for estimating the PS and the endpoint regression function. Each of these methods relies on the assumptions of unconfoundedness and good overlap.

This section investigated these methods in terms of the challenges identified for CEA, such as the specification of the endpoint regression models and the PS, and when there is poor overlap. I found that these methods can make less restrictive assumptions than previously proposed methods. Table 2.2 summarises the expected relative performance of these methods under realistic circumstances in CEA.

GM does not require a correctly specified PS, however the analyst does need to specify a loss function for the machine learning algorithm. This involves selecting those potential confounders that the algorithm is specified to balance. This choice of confounders also needs to be made when PS matching or IPTW is used for creating balance (Stuart, 2010). GM, similarly to PS matching, can result in poor quality matches, hence bias and high variability if overlap is weak. Combined methods can mitigate the disadvantages of either PS or regression methods: DR methods can be unbiased if either one of the PS or regression models is correct, but can be sensitive to unstable weights if overlap is poor. Regression-adjusted matching can decrease the sensitivity of estimates to the misspecification of the regression model, due to increased

balance; however can be less efficient than using regression. Machine learning estimation methods for the PS and the endpoints, while not requiring a correctly specified fixed parametric model, can be sensitive to subjective choices such as the choice of prediction algorithms and tuning parameters.

2.6 Identifying research gaps in the literature comparing alternative statistical methods for addressing selection bias in CEA

In the conceptual review, I identified a range of alternative statistical methods that are promising in addressing selection bias in CEA. While statistical theory offers guidance on how methods perform when their underlying assumptions fail, an important challenge for the applied researcher is to choose the least biased and most efficient estimator under realistic circumstances, such as not knowing the true statistical models that generate the endpoint and the treatment assignment.

The results of current comparative work in the methodology literature might not translate directly to CEA. Therefore new simulation studies, which incorporate important features of CEA, such as correlated cost and effectiveness endpoints and nonlinear covariate-endpoint relationships, are needed.

When cost-effectiveness for patient subgroups needs to be estimated using observational data, regression methods, recommended for subgroup analysis in CEA (Nixon and Thompson, 2005), can be sensitive to the specification of the regression model. Methods that use the PS to reduce selection bias are therefore of interest to estimate subgroup-effects. IPTW has particular appeal for subgroup analysis, where due to reduced sample sizes, precision can be a concern: if the PS is correctly specified, IPTW can provide more precise estimates of treatment effects than matching (Hirano et al., 2003).

Table 2.2 - The expected performance of proposed methods, under realistic circumstances in CEA

	Misspecification of the endpoint[1]	**Misspecification of the PS[1]**	**Poor overlap**
GM	Not relevant.	Does not matter if loss function correctly specified.	Treated observations might have to be dropped, or the quality of matches is bad, leading to bias and inefficiency.
DR methods	Consistent estimates if the PS is correct.	Consistent estimates if the endpoint regression model is correct.	A correctly specified regression can help with extrapolation (Petersen et al., 2010). In practice, with unstable weights, bias and inefficiency likely.
Regression-adjusted matching	Matching can reduce sensitivity to model misspecification.	Regression adjustment can correct for imbalance and bias due to misspecified PS.	Bias and inefficiency. However, bias due to model misspecification can be reduced by increased balance after matching.
Machine learning estimation of PS	Not relevant.	Reduces chance of misspecification. Can be sensitive to the choice of algorithm and tuning parameters.	Can provide remedy against unstable IPT weights (Lee et al., 2010).
Machine learning estimation of the endpoint	Reduces chance of misspecification. Can be sensitive to the choice of prediction algorithms.	Not relevant.	Reduced misspecification can reduce bias due to extrapolation (Porter et al., 2011).

Notes: [1] Misspecification is defined as functional form misspecification, for example using the incorrect link function or omitting higher order terms from the linear predictor.

However IPTW has not been considered in CEA for addressing selection bias. GM has been proposed for CEA (Grieve et al., 2008), and was compared to PS matching (Sekhon and Grieve, 2011). However GM has not been compared to IPTW in the general methodological literature before, and none of the PS or matching approaches have been considered for subgroup analysis in CEA.

While methodological guidance in CEA propose several ways of using the PS to create balance (Polsky and Basu, 2006, Mitra and Indurkhya, 2005, Manca and Austin, 2008), these studies did not consider the combination of the PS with regression models for the endpoint. Previous findings on the relative merits of alternative DR methods (Kang and Schafer, 2007a, Porter et al., 2011) may not translate to the CEA setting, where models for both costs (Jones, 2010) and health outcomes (Basu and Manca, 2011) may be required. While regression adjustment post matching has been proposed as a sensitivity analysis in the CEA literature (e.g. Sekhon and Grieve, 2011), the performance of this method under model misspecification has not been considered in a CEA setting before.

The evidence on the relative performance of regression-adjusted matching and DR methods is limited in the general causal inference literature. A recent working paper (Busso et al., 2011) compared BCM with IPTW and a DR method, across a range of scenarios, including poor overlap. However, BCM has not been compared to TMLE, which has been proposed to outperform alternative DR methods under circumstances of model misspecification and poor overlap (Porter et al., 2011, Gruber and van der Laan, 2010b).

Both TMLE and BCM can be coupled with machine learning estimation of the endpoint regression function and the PS. This can be a promising approach for estimating parameters for CEA, such as treatment effects on HRQoL data, where specifying the endpoint regression model can be challenging (Basu and Manca, 2011). There are no

previous studies that implement BCM with machine learning, and the simulation studies which consider TMLE with machine learning (Porter et al., 2011, Gruber and van der Laan, 2010b), did not consider typical circumstances of HRQoL data, such as spikes in the distribution of the endpoint.

To conclude, a number of gaps were identified in the methodological literature of CEA that considers statistical methods to address selection bias:

1. PS methods such as PS matching and IPTW have not been considered for estimating subgroup-effects in CEA.
2. GM has not been compared to IPTW before in the general methodological literature.
3. DR methods and regression-adjusted matching have not been considered in the context of CEA.
4. TMLE and regression-adjusted matching have not been compared in the methodological literature.
5. Machine learning estimation methods have not been considered for estimating parameters for CEA before.

The research papers included in this thesis (chapters 3 to 6) will aim to address these gaps, using simulations and case studies (see Table 2.3).

Table 2.3 - Summary of research papers to compare alternative statistical methods for addressing selection bias in CEA

	Context	Main comparators	Main challenges considered	Previous methodological papers extended	Gap addressed
Research paper 2	Subgroup analysis in CEA	IPTW, PS matching, GM	Misspecification of the PS; unstable IPT weights	Sekhon and Grieve, 2011	1,2
Research paper 3	CEA	Common DR methods, regression-adjusted PS matching	Misspecification of the PS, cost and effectiveness endpoints; unstable IPT weights	Basu et al., 2011 Kang and Schafer, 2007	3
Research paper 4	Estimating incremental effectiveness	TMLE, BCM, with PS and endpoint estimated using (1) fixed parametric, (2) machine learning methods	Misspecification of PS and HRQoL endpoint; poor overlap	Basu and Manca, 2011; Busso et al., 2011; Porter et al., 2011; Lee et al., 2010	4,5

2.7 Discussion

This chapter had four interlinked objectives. First, to review methodological guidance on the statistical methods for addressing selection bias in CEA that use patient-level observational data and to describe statistical challenges that arise when using these methods. Second, to describe the main underlying assumptions of statistical methods previously recommended for addressing selection bias in CEA. Third, to identify further promising statistical methods from the general causal inference. Fourth, to identify gaps

in the methodological literature on the relative performance of statistical methods for addressing selection bias in CEA.

CEA methodological guidance recommended that regression, PS methods and IV estimation are considered to address selection bias. IV methods can potentially reduce selection bias due to both observed and unobserved confounding, however they make further untestable assumptions that may be unrealistic in a CEA setting (Polsky and Basu, 2006). Hence methods that assume no unobserved confounding need to be considered for CEA. These methods make the following further assumptions:

1. Good overlap between the covariate distributions.
2. Correctly specifying regression models for cost and effectiveness endpoints.
3. Correctly specifying the PS model.

The conceptual review found that under realistic circumstances, these assumptions might not be plausible for methods currently recommended for CEA. For example, INB regression (Hoch et al., 2002) imposes a linear functional form on the relationships between the covariates and the net benefit endpoint. PS, when used for stratification or as a covariate in regression (Manca and Austin, 2008, Mitra and Indurkhya, 2005) has been shown to be dominated by alternative PS approaches such as IPTW and PS matching (Lunceford and Davidian, 2004). These methods are not further considered for this thesis.

Research paper 1 (chapter 3) uses findings from this conceptual review to give detailed guidance on how the plausibility of the underlying assumptions of statistical methods can be assessed in CEA. As this conceptual review highlighted, any statistical method relies on assumptions that cannot be directly tested from the data. As the checklist presented in research paper 1 highlights, the uncertainty due to the choice of statistical

approach needs to be acknowledged as part of a structural uncertainty in the CEA. My systematic review appraising published CEA (research paper 1) found that most studies did not appropriately assess the underlying assumptions their statistical methods made. Motivated by this finding, this conceptual review identified alternative statistical approaches from the causal inference literature that can potentially make less restrictive assumptions than previously proposed methods.

I found that the following alternative approaches held promise for addressing selection bias in CEA: GM, DR methods, regression-adjusted matching and machine learning estimation approaches for the PS and the endpoint regression. The relative performance of these methods is likely to depend on the specific circumstances of the CEA, such as the extent to which there is poor overlap or misspecification of the PS or the endpoints. There is limited evidence on the relative performance of these methods, and it is unknown how these methods would perform when compared to previously proposed methods for addressing selection bias, in typical CEA.

This thesis aims to address these gaps with three research papers. Research paper 2 compares PS matching, GM and IPTW, for estimating cost-effectiveness for patient subgroups. Research paper 3 considers the relative performance of DR methods and regression-adjusted matching, for CEA. Research paper 4 considers a recently proposed DR method, TMLE, and compares it to BCM, for estimating treatment effects on HRQoL data, when fixed parametric models and machine learning techniques are used for estimating the PS and the endpoint regression.

This review focused on methods that can address selection bias in CEA that use patient-level observational data to estimate incremental parameters of continuous endpoints, such as incremental costs or HRQoL. The methods reviewed here can be extended to other endpoints such as binary, count or time-to-event data, and different estimands

such as odds ratios (Radice et al., 2012, Moore and van der Laan, 2009) and hazard ratios (Thompson et al., 2010). The complexities of using these statistical methods for estimating alternative parameters are beyond the scope of this review.

This conceptual review focused on methods that can estimate the effect of a time constant, binary treatment. Extending the methods to categorical or continuous treatment is possible, for example using the generalised PS (Cole and Frangakis, 2009). IPTW and DR methods can also be extended to handle treatment and covariates that vary over time (Robins et al., 2000). Such extensions have relevance in CEA which need to handle cross-over between treatments, or treatment starting at different time points for patients. More generally, these methods can be used in CEA when developing input parameters for decision models, for example risk equations which need to account for time-varying confounding (Caro et al., 2012).

This review did not cover some further important statistical challenges in CEA which uses patient-level observational data. These include the appropriate analysis of missing data and censored endpoints such as survival times or costs. Some of the methods reviewed here, for example IPTW and DR methods have more general applicability to account for censoring (Willan et al., 2002, Willan et al., 2005, Bang and Tsiatis, 2000) and to estimate mean endpoints under missing data (Kang and Schafer, 2007a).

This review concludes that current methods recommended to address selection bias in CEA make assumptions that may not be plausible in practice. Further promising statistical methods are available from the general causal inference literature, but there is little evidence on their relative merits across settings typically observed in CEA. The subsequent chapters provide insights on the relative performance of statistical methods that aim to tackle selection bias in CEA that use patient-level observational data, to help address the gaps in the methodological literature identified in this chapter.

References

Abadie, A., Diamond, A. & Hainmueller, J. 2010. Synthetic Control Methods for Comparative Case Studies:Estimating the Effect of California's Tobacco Control Program. *Journal of the American Statistical Association***,** 493-505.

Abadie, A., Herr, J. L., Imbens, G. W. & Drukker, D. M. 2004. *NNMATCH: Stata module to compute nearest-neighbor bias-corrected estimators* [Online]. Boston College Department of Economics. Available: http://fmwww.bc.edu/repec/bocode/n/nnmatch.hlp.

Abadie, A. & Imbens, G. W. 2006a. Large Sample Properties of Matching Estimators for Average Treatment Effects. *Econometrica,* 74**,** 235-267.

Abadie, A. & Imbens, G. W. 2006b. On the Failure of the Bootstrap for Matching Estimators. *National Bureau of Economic Research Technical Working Paper Series,* No. 325.

Abadie, A. & Imbens, G. W. 2009. Matching on the Estimated Propensity Score. National Bureau of Economic Research, Inc.

Abadie, A. & Imbens, G. W. 2011. Bias-Corrected Matching Estimators for Average Treatment Effects. *Journal of Business & Economic Statistics,* 29**,** 1-11.

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

Austin, P. C. 2009a. Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. *Statistics in Medicine,* 28 3083-3107.

Austin, P. C. 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**,** 661-677.

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

Austin, P. C. 2012. Using Ensemble-Based Methods for Directly Estimating Causal Effects: An Investigation of Tree-Based G-Computation. *Multivariate Behavioral Research***,** 115-135.

Austin, P. C., Grootendorst, P. & Anderson, G. M. 2007. A comparison of the ability of different propensity score models to balance measured variables between treated and untreated subjects: a Monte Carlo study. *Statistics in Medicine,* 26**,** 734-753.

Bang, H. & Robins, J. M. 2005. Doubly Robust Estimation in Missing Data and Causal Inference Models. *Biometrics* 962–972.

Bang, H. & Tsiatis, A. A. 2000. Estimating medical costs with censored data. *Biometrica* 87**,** 329-343.

Barber, J. & Thompson, S. 2004a. Multiple regression of cost data: use of generalised linear models. *J Health Serv Res Policy,* 9**,** 197-204.

Barber, J. & Thompson, S. G. 2004b. Multiple regression of cost data: use of generalised linear models. *Journal of Health Services Research Policy,* 9**,** 197-204.

Basu, A. 2011. Economics of individualization in comparative effectiveness research and a basis for a patient-centered health care. *Journal of Health Economics,* 30**,** 549-559.

Basu, A., Heckman, J. J., Navarro-Lozano, S. & 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**,** 1133-1157.

Basu, A. & Manca, A. 2011. Regression Estimators for Generic Health-Related Quality of Life and Quality-Adjusted Life Years. *Med Decis Making,* 2011 Oct 18. [Epub ahead of print].

Basu, A., Manning, W. G. & Mullahy, J. 2004. Comparing alternative models: log vs Cox proportional hazard? *Health Economics,* 13**,** 749-765.

Basu, A., Polsky, D. & Manning, W. 2011. Estimating treatment effects on healthcare costs under exogeneity: is there a 'magic bullet'? *Health Services and Outcomes Research Methodology,* 11**,** 1-26.

Basu, A., Polsky, D. & Manning, W. G. 2008. Use of propensity scores in non-linear response models: The case for health care expenditures. HEDG, c/o Department of Economics, University of York.

Basu, A. & Rathouz, P. J. 2005. Estimating marginal and incremental effects on health outcomes using flexible link and variance function models. *Biostatistics,* 6**,** 93-109.

Bojke, L., Claxton, K., Sculpher, M. & Palmer, S. 2009. Characterizing structural uncertainty in decision-analytic models: a review and application of methods. *Value in Health,* 12**,** 739-49.

Buntin, M. B. & Zaslavsky, A. M. 2004. Too much ado about two-part models and transformation?: Comparing methods of modeling Medicare expenditures. *Journal of Health Economics,* 23**,** 525-542.

Busso, M., DiNardo, J. & McCrary, J. 2009. Finite Sample Properties of Semiparametric Estimators of Average Treatment Effects.

Busso, M., DiNardo, J. & McCrary, J. 2011. New Evidence on the Finite Sample Properties of Propensity Score Reweighting and Matching Estimators. *Working paper.*

Cole, S. R. & Hernán, M. A. 2008. Constructing inverse probability weights for marginal structural models. *Am J Epidemiol,* 168**,** 656-64.

Crump, R. K., Hotz, V. J., Imbens, G. W. & Mitnik, O. A. 2009. Dealing with limited overlap in estimation of average treatment effects. *Biometrika,* 96**,** 187-199.

Davison, A. & Hinkley, D. 1997. *Bootstrap Methods and Their Application,* New York, Cambridge University Press.

Dehejia, R. H. & Wahba, S. 2002. Propensity Score-Matching Methods For Nonexperimental Causal Studies. *The Review of Economics and Statistics,* 84**,** 151-161.

Diamond, A. & Sekhon, J. S. 2012. Genetic Matching for Estimating Causal Effects: A General Multivariate Matching Method for Achieving Balance in Observational Studies. *Review of Economic and Statistics,* Forthcoming.

Didelez, V. & Sheehan, N. 2007. Mendelian randomization as an instrumental variable approach to causal inference. *Statistical Methods in Medical Research,* 16**,** 309-330.

Espinoza, M. A., Manca, A., Claxton, K. & Sculpher, M. J. 2011. The value of identifying heterogeneity: a framework for subgroup cost-effectiveness analysis. *HESG Conference presentation.* York.

Freedman, D. & Berk, R. A. 2008. Weighting regression by propensity score. *Evaluation Review,* 32**,** 392-409.

Gilleskie, D. B. & Mroz, T. A. 2004. A flexible approach for estimating the effects of covariates on health expenditures. *Journal of Health Economics,* 23**,** 391-418.

Glynn, A. N. & Quinn, K. M. 2010. An Introduction to the Augmented Inverse Propensity Weighted Estimator. *Political Analysis,* 18**,** 36-56.

Gray, A. M., Clarke, P. M., Wolstenholme, J. L. & Wordsworth, S. 2010. *Applied Methods of Cost-Effectiveness Analysis in Healthcare,* Oxford University Press.

Greenland, S., Pearl, J. & Robins, J. M. 1999. Confounding and Collapsibility in Causal Inference. *Statist. Sci. ,* 14**,** 29-46.

Grieve, R., Sekhon, J. S., Hu, T.-w. & 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**,** 1204-1222.

Gruber, S. & van der Laan, M. 2010a. A targeted maximum likelihood estimator of a causal effect on a bounded continuous outcome. *Int J Biostat. ,* 6.

Gruber, S. & van der Laan, M. J. 2010b. An application of collaborative targeted maximum likelihood estimation in causal inference and genomics. *The International Journal of Biostatistics,* 6.

Gruber, S. & van der Laan, M. J. 2012a. Targeted minimum loss based estimation of causal effects on an outcome with known conditional bounds. *International Journal of Biostatistics,* in press.

Gruber, S. & van der Laan, M. J. 2012b. tmle: An R Package for Targeted Maximum Likelihood Estimation. *Journal of Statistical Software,* 51**,** 1-35.
Hernán, M. A. & Robins, J. M. 2006. Instruments for Causal Inference: An Epidemiologist's Dream? *Epidemiology,* 17**,** 360-372.
Hill, J. 2008. Discussion of research using propensity-score matching: Comments on 'A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003' by Peter Austin. *Statistics in Medicine,* 27 2055–2061.
Hill, J. & Reiter, J. P. 2005. Interval estimation for treatment effects using propensity score matching. *Statistics in Medicine,* 25.
Hirano, K. & Imbens, G. W. 2001. Estimation of Causal Effects using Propensity Score Weighting: An Application to Data on Right Heart Catheterization. *Health Services and Outcomes Research Methodology,* 2**,** 259-278.
Hirano, K., Imbens, G. W. & Ridder, G. 2003. Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score. *Econometrica,* 71**,** 1161-1189.
Ho, D. E., Imai, K., King, G. & Stuart, E. A. 2007. Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference *Political Analysis,* 15**,** 199-236.
Hoch, J. S., Briggs, A. H. & Willan, A. R. 2002. Something old, something new, something borrowed, something blue: a framework for the marriage of health econometrics and cost-effectiveness analysis. *Health Economics,* 11**,** 415-430.
Horowitz, J. L. 2011. Applied Nonparametric Instrumental Variables Estimation. *Econometrica,* 79**,** 347-394.
Imbens, G. & Jeffrey, M. W. 2007. Difference-in-Differences Estimation. *NBER Lecture Notes.*
Imbens, G. & Wooldridge, J. M. 2009b. New Developments in Econometrics. *Lecture Notes, CEMMAP, UCL.*
Imbens, G. M. & Wooldridge, J. M. 2009a. Recent Developments in the Econometrics of Program Evaluation. *Journal of Economic Literature,* 47**,** 5–86.
Imbens, G. W. 2004. Nonparametric Estimation of Average Treatment Effects Under Exogeneity: A Review. *The Review of Economics and Statistics,* 86**,** 4-29.
Jackson, C., Bojke, L., Thompson, S., Claxton, K. & Sharples, L. 2011. A Framework for Addressing Structural Uncertainty in Decision Models. *Medical Decision Making,* 31**,** 662–674.
Joffe, M. & Mindell, J. 2006. Complex Causal Process Diagrams for Analyzing the Health Impacts of Policy Interventions. *American Journal of Public Health,* 96**,** 473-479.
Jones, A., Lomas, J. & Rice, N. 2011. Applying Beta-type Size Distributions to Healthcare Cost Regressions. *HEDG Working Papers.* HEDG, c/o Department of Economics, University of York.
Jones, A. M. 2007. Identification of treatment effects in Health Economics. *Health Economics,* 16**,** 1127-1131.
Jones, A. M. 2010. Models For Health Care. *HEDG Working Papers.* HEDG, c/o Department of Economics, University of York.
Kang, J. D. Y. & Schafer, J. L. 2007a. Demystifying Double Robustness: A Comparison of Alternative Strategies for Estimating a Population Mean from Incomplete Data. *Statistical Science,* 22**,** 523-539.
Kang, J. D. Y. & Schafer, J. L. 2007b. Rejoinder: Demystifying Double Robustness: A Comparison of Alternative Strategies for Estimating a Population Mean from Incomplete Data. *Statistical Science,* 22**,** 574-580.
Lee, B. K., Lessler, J. & Stuart, E. A. 2010. Improving propensity score weighting using machine learning. *Statistics in Medicine,* 29**,** 337-346.
Lunceford, J. K. & Davidian, M. 2004. Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Statistics in Medicine,* 23**,** 2937-2960.
Manca, A. & Austin, P. C. 2008. *Using propensity score methods to analyse individual patient-level cost-effectiveness data from observational studies* [Online]. Available: http://www.york.ac.uk/res/herc/documents/wp/08_20.pdf.

Manning, W. G., Basu, A. & Mullahy, J. 2005. Generalized modeling approaches to risk adjustment of skewed outcomes data. *Journal of Health Economics,* 24**,** 465-488.
Manning, W. G., Duan, N. & Rogers, W. H. 1987. Monte Carlo evidence on the choice between sample selection and two-part models. *Journal of Econometrics,* 35**,** 59-82.
Manning, W. G. & Mullahy, J. 2001. Estimating log models: to transform or not to transform? *Journal of Health Economics,* 20**,** 461-494.
McCaffrey, D., Ridgeway, G. & Morral, A. 2004. Propensity Score Estimation with Boosted Regression for Evaluating Causal Effects in Observational Studies. *Psychol Methods,* 9**,** 403-425.
Mihaylova, B., Briggs, A., Armitage, J., Parish, S., Gray, A., Collins, R. & Group, H. P. S. C. 2005. Cost-effectiveness of simvastatin in people at different levels of vascular disease risk: economic analysis of a randomised trial in 20,536 individuals. *Lancet. ,* 365**,** 1779-85.
Mihaylova, B., Briggs, A., O'Hagan, A. & Thompson, S. 2010. Review of statistical methods for analysing healthcare resources and costs. *Health Economics,* DOI: 10.1002/hec.1653.
Mitra, N. & Indurkhya, A. 2005. A propensity score approach to estimating the cost-effectiveness of medical therapies from observational data. *Health Economics,* 14**,** 805-15.
Moore, K. L. & van der Laan, M. J. 2009. Covariate adjustment in randomized trials with binary outcomes: Targeted maximum likelihood estimation. *Statistics in Medicine,* 28**,** 39-64.
Morgan, S. L. & Winship, C. 2007. *Counterfactuals and causal inference: methods and principles for social research,* New York, Cambridge University Press.
Mullahy, J. 2011. Symposium on genetic data in health economics research. *Health Economics***,** n/a-n/a.
Nixon, R. M. & Thompson, S. G. 2005. Methods for incorporating covariate adjustment, subgroup analysis and between-centre differences into cost-effectiveness evaluations. *Health Economics,* 14**,** 1217-1229.
Pan, W. & Zeng, D. 2011. Estimating Mean Cost Using Auxiliary Covariates. *Biometrics,* 67**,** 996-1006.
Pearl, J. 2001. Causal Inference in the Health Sciences: A Conceptual Introduction. *Health Services and Outcomes Research Methodology,* 2**,** 189-220.
Petersen, M. L., Porter, K., Gruber, S., Wang, Y. & Laan., M. J. v. d. 2010. Diagnosing and Responding to Violations in the Positivity Assumption. *U.C. Berkeley Division of Biostatistics Working Paper Series*.
Philips, Z., Bojke, L., Sculpher, M., Claxton, K. & Golder, S. 2006. Good practice guidelines for decision-analytic modelling in health technology assessment: a review and consolidation of quality assessment. *Pharmacoeconomics,* 24**,** 355-71.
Polley, E. C. & van der Laan, M. J. 2010. Super Learner In Prediction. *U.C. Berkeley Division of Biostatistics Working Paper Series.*
Polsky, D. & Basu, A. 2006. Selection Bias in Observational Data. *The Elgar Companion to Health Economics.* Edward Elgar Publishing.
Porter, K. E., Gruber, S., Laan, M. J. v. d. & Sekhon, J. S. 2011. The Relative Performance of Targeted Maximum Likelihood Estimators. *The International Journal of Biostatistics,* 7.
Pullenayegum, E. M. & Willan, A. R. 2011. Marginal Models for Censored Longitudinal Cost Data: Appropriate Working Variance Matrices in Inverse-Probability-Weighted GEEs Can Improve Precision. *The International Journal of Biostatistics,* 7.
Radice, R., Grieve, R., Ramsahai, R., Kreif, N., Sadique, Z. & 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).
Ramsahai, R., Grieve, R. & Sekhon, J. S. 2011. Extending Iterative Matching Methods: An Approach to Improving Covariate Balance that Allows Prioritisation. *Health Services and Outcomes Research Methodology*.

Robins, J. 1986. A new approach to causal inference in mortality studies with a sustained exposure period – application to control of the healthy worker survivor effect. . *Mathematical Modelling.* , 7**,** 1393–1512.
Robins, J., Rotnitzky, A. & Zhao, L. P. 1994. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association,* 89**,** 846-866.
Robins, J., Sued, M., Lei-Gomez, Q. & Rotnitzky, A. 2007. Comment: Performance of Double-Robust Estimators When "Inverse Probability" Weights Are Highly Variable. *Statistical Science,* 22**,** 544-559.
Robins, J. M., Rotnitzky, A. & Zhao, L. P. 1995. Analysis of Semiparametric Regression Models for Repeated Outcomes in the Presence of Missing Data. *Journal of the American Statistical Association,* 90**,** 106-121.
Rosenbaum, P. R. 2002. *Observational studies,* New York ; London, Springer.
Rosenbaum , P. R. & Rubin, D. B. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika,* 70**,** 41-55.
Rosenbaum , P. R. & Rubin, D. B. 1985. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician,* 39.
Rotnitzky, A., Lei, Q., Sued, M. & Robins, J. M. 2012. Improved double-robust estimation in missing data and causal inference models. *Biometrika,* 99**,** 439-456.
Rubin, D. 2008. For Objective Causal Inference, Design Trumps Analysis. *The Annals of Applied Statistics,* 2**,** 808-840.
Rubin, D. & Thomas, N. 1992. Characterizing the Effect of Matching Using Linear Propensity Score Methods with Normal Distributions. *Biometrika,* 79.
Rubin, D. B. 1973. The Use of Matched Sampling and Regression Adjustment to Remove Bias in Observational Studies. *Biometrics,* 29.
Rubin, D. B. 1979. Using multivariate matched sampling and regression adjustment to control bias in observational studies. *Journal of the American Statistical Association***,** 318–328.
Rubin, D. B. 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. B. 2010. On the limitations of comparative effectiveness research. *Statistics in Medicine,* 29**,** 1991-1995.
Rubin, D. B. & Thomas, N. 2000. Combining propensity score matching with additional adjustments for prognostic covariates. *Journal of the American Statistical Association,* 95**,** 573–585.
Sculpher, M. 2008. Subgroups and Heterogeneity in Cost-Effectiveness Analysis. *Pharmacoeconomics,* 26**,** 799-806.
Sekhon, J. S. 2011. Matching: multivariate and propensity score matching with automated balance search. *Journal of Statistical Software*.
Sekhon, J. S. & Grieve, R. D. 2011. A Matching Method for Improving Covariate Balance in Cost-Effectiveness Analyses. *Health Economics,* doi: 10.1002/hec.1748.
Stuart, E. A. 2010. Matching Methods for Causal Inference: A review and a look forward. *Statistical Science,* 25.
Terza, J. V., Basu, A. & Rathouz, P. J. 2008. Two-Stage Residual Inclusion Estimation: Addressing Endogeneity in Health Econometric Modeling. *J Health Econ,* 27**,** 531–543.
Thompson, S., Kaptoge, S., White, I., Wood, A., Perry, P., Danesh, J. & Collaboration, T. E. R. F. 2010. Statistical methods for the time-to-event analysis of individual participant data from multiple epidemiological studies. *International Journal of Epidemiology*.
Thompson, S. G. & Nixon, R. 2005. How Sensitive Are Cost-Effectiveness Analyses to Choice of Parametric Distributions? *Medical Decision Making,* 25**,** 416-423.
van der Laan, M. J. 2010. Targeted Maximum Likelihood Based Causal Inference: Part I. *The International Journal of Biostatistics*.
van der Laan, M. J. & Dudoit, S. 2003. Unified cross-validation methodology for se- lection among estimators and a general cross-validated adaptive epsilon-net estimator: Finite

sample oracle inequalities and examples. *Technical report, Division of Biostatistics, University of California, Berkeley*.
van der Laan, M. J. & Gruber, S. 2010. Collaborative Double Robust Targeted Maximum Likelihood Estimation. *The International Journal of Biostatistics* 6.
van der Laan, M. J. & Rubin, D. 2006. Targeted maximum likelihood learning. *The International Journal of Biostatistics,* 2.
van der Laan, M. J. P., Eric C.; and Hubbard, Alan E. 2007. Super Learner *Statistical Applications in Genetics and Molecular Biology: ,* Vol. 6
Waernbaum, I. 2011. Model misspecification and robustness in causal inference: comparing matching with doubly robust estimation. *Stat Med,* DOI: 10.1002/sim.4496.
Westreich, D. & Cole, S. R. 2010. Invited Commentary: Positivity in Practice. *American Journal of Epidemiology*.
Westreich, D., Lessler, J. & Funk, M. 2010. Propensity score estimation: neural networks, support vector machines, decision trees (CART), and meta-classifiers as alternatives to logistic regression. *Journal of Clinical Epidemiology,* 63**,** 826-33.

Chapter 3 - Checklist for critical appraisal of statistical methods to address selection bias in CEA that use patient-level observational data

3.1 Preamble to research paper 1

Chapter 2 reviewed the current methodological guidance on statistical methods for addressing selection bias in CEA, and found that recommended methods make important underlying assumptions that may be implausible in typical CEA. General checklists and methodological guidelines (Drummond et al., 2005, Glick et al., 2007) do not include criteria for assessing the quality of statistical methods for CEA that use observational data. Research paper 1 aims to fill this gap in the methodological literature, by developing a new checklist to critically appraise statistical methods for addressing selection bias in CEA that use patient-level observational data.

The development of this checklist was informed by the conceptual review (chapter 2), and by insights from an expert panel. In order to help a reviewer judge whether a study meets the checklist criteria, and to help the analyst to appropriately apply statistical methods in CEA, this paper also presents detailed methodological guidance (Appendix 3.1).

The checklist is applied in a systematic review of the applied literature, which aims to identify which methods are frequently used in applied CEA, and evaluates whether the assumptions underlying these methods were appropriately assessed. Findings from the review will inform the choice of statistical methods and simulation scenarios in the subsequent empirical work (research papers 2, 3 and 4).

London School of Hygiene & Tropical Medicine
Keppel Street, London WC1E 7HT
www.lshtm.ac.uk

Registry
T: +44(0)20 7299 4646
F: +44(0)20 7299 4656
E: registry@lshtm.ac.uk

LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

COVER SHEET FOR EACH 'RESEARCH PAPER' INCLUDED IN A RESEARCH THESIS

Please be aware that one cover sheet must be completed for each 'Research Paper' included in a thesis.

1. **For a 'research paper' already published**

   1.1. Where was the work published? .....

   1.2. When was the work published? .....

   1.2.1. If the work was published prior to registration for your research degree, give a brief rationale for its inclusion

   .....

   .....

   .....

   1.3. Was the work subject to academic peer review? .....

   1.4. Have you retained the copyright for the work? **Yes / No**
   If yes, please attach evidence of retention.
   If no, or if the work is being included in its published format, please attach evidence of permission from copyright holder (publisher or other author) to include work

2. **For a 'research paper' prepared for publication but not yet published**

   2.1. Where is the work intended to be published? .....

   2.2. Please list the paper's authors in the intended authorship order

   .....

   2.3. Stage of publication – Not yet submitted / Submitted / Undergoing revision from peer reviewers' comments / In press

3. **For multi-authored work, give full details of your role in the research included in the paper and in the preparation of the paper. (Attach a further sheet if necessary)**

   .....

   .....

NAME IN FULL (Block Capitals) .....

STUDENT ID NO: .....

CANDIDATE'S SIGNATURE ..... Date .....

SUPERVISOR/SENIOR AUTHOR'S SIGNATURE (3 above) .....

Additional page for Question (3) on LSHTM cover sheet form:

The research question for this paper was linked to the ESRC project and identified by the principal investigator, RG. I carried out a conceptual review, developed a checklist and accompanying methodological guidance for critical appraisal of CEA that use observational data, in collaboration with RG. I applied this checklist in a systematic review of studies, and interpreted the findings. ZS verified the exclusion criteria of the systematic review and conducted a second review by independently appraising 50% of the included studies. I wrote the first draft of the manuscript and managed each round of comments and suggestions from RG and ZS. All authors read and approved the final draft prior to journal submission and inclusion in the dissertation.

3.2 Research paper 1- Statistical methods for cost-effectiveness analyses that use observational data: a critical appraisal tool and review of current practice

Noémi Kreif MSc[1], Richard Grieve PhD[1], Zia Sadique PhD[1]

[1] Department of Health Services Research & Policy, London School of Hygiene and Tropical Medicine, London, UK

Status: Published in Health Economics, 2012, DOI: 10.1002/hec.2806

Contributions: The research question for this paper was linked to the ESRC project and identified by the principal investigator, RG. The candidate carried out a conceptual review, developed a checklist and accompanying methodological guidance for critical appraisal of CEA that use observational data, in collaboration with RG. The candidate applied this checklist in a systematic review of studies, and interpreted the findings. ZS verified the exclusion criteria of the systematic review and conducted a second review by independently appraising 50% of the included studies. The candidate wrote the first draft of the manuscript and managed each round of comments and suggestions from RG and ZS. All authors read and approved the final draft prior to journal submission and inclusion in the dissertation.

The candidate_____ The supervisor_____

Abstract

Many cost-effectiveness analyses (CEAs) use data from observational studies. Statistical methods can only address selection bias if they make plausible assumptions. No quality assessment tool is available for appraising CEAs that use observational studies. We developed a new critical appraisal checklist to assess statistical methods for addressing selection bias in CEAs that use observational data.

The checklist criteria were informed by a conceptual review, and applied in a systematic review of economic evaluations. Criteria included whether the study assessed the "no unobserved confounding" assumption, overlap of baseline covariates between the treatment groups, and the specification of the regression models. The checklist also considered structural uncertainty from the choice of statistical approach.

We found 81 studies that met the inclusion criteria: studies tended to use regression (51%), matching on individual covariates (25%) or matching on the propensity score (22%). Most studies (77%) did not assess the "no observed confounding" assumption, and few studies (16%) fully considered structural uncertainty from the choice of statistical approach.

We conclude that published CEAs do not assess the main assumptions behind statistical methods for addressing selection bias. This checklist can raise awareness about the assumptions behind statistical methods for addressing selection bias and can complement existing method guidelines for CEAs.

Introduction

Methodological guidance for cost-effectiveness analyses (CEAs) emphasises the use of pragmatic randomised controlled trials (RCTs) (Willan and Briggs, 2006, Glick et al., 2007), but for many decision problems RCTs are unavailable or insufficient. Hence CEAs may use observational studies, for example to provide transition probabilities for decision models, or to estimate incremental costs or effectiveness. However, the non-random selection of patients into treatment can lead to selection bias (Jones and Rice, 2011). For CEAs where individual patient data (IPD) are available, statistical methods such as regression, matching or instrument variable (IV) estimation can address selection bias. For these methods to provide unbiased estimates, the underlying assumptions must be plausible. For example, regression and matching assume that there is no unobserved confounding (Greenland et al., 1999), regression that the endpoint model is correctly specified (Ho et al., 2007), and matching that baseline characteristics are balanced after matching (Stuart, 2010).

In CEA, the choice of method for addressing selection bias can lead to different conclusions. For example, after applying regression to adjust for baseline differences, a surgical intervention for breast cancer appeared cost-effective, whereas after IV estimation, the intervention was dominated (Polsky and Basu, 2006) (Table 3.1).

Table 3.1 - Incremental cost-effectiveness results according to statistical method for addressing selection bias: an illustrative example from a study comparing breast conserving surgery to mastectomy (Polsky and Basu, 2006).

	Covariate adjustment	Instrumental variables
Incremental cost [USD](95% CI)	14,199 (10,279 to 18,118)	50,997 (12,879 to 89,114)
Incremental QALY (95% CI)	0.12 (0.05 to 0.19)	-0.29 (- 0.095 to 0.38)
ICER (USD) (95% CI)	118,325 (70,040 to 250,000)	Dominated (150,200 to Dominated)

Abbreviations: USD- US Dollars; QALY, quality adjusted life year; CI, confidence interval.

Critical appraisal tools have been developed for CEAs (Drummond et al., 2005, Philips et al., 2006), but there is no tool for assessing the quality of CEAs that use observational data; it is currently unknown whether such studies adopt appropriate statistical methods. This paper introduces a new checklist for assessing the main assumptions made by statistical methods for addressing selection bias. We apply the checklist in a systematic review of published CEAs that use observational data.

Method

A critical appraisal checklist was developed for assessing whether CEAs used appropriate statistical methods for addressing selection bias, and to provide a tool for improving the quality of future studies. To inform development of the checklist, we undertook a conceptual review of the statistics, econometrics and epidemiology literatures (including work published between 1983 and 2011) to identify relevant statistical methods for addressing selection bias in CEAs (e.g. Rosenbaum and Rubin, 1983, Imbens and Wooldridge, 2009a, Stuart, 2010). The approaches judged most relevant were regression, matching on the propensity score, matching on individual covariates and IV methods (Polsky and Basu, 2006, Jones and Rice, 2011). To inform

the development of the checklist (Table 3.2) and accompanying guidance (Appendix 3.1), the conceptual review identified the main assumptions underlying each method. Provisional versions of the checklist were reviewed by a panel of health economists and statisticians. Three independent reviewers piloted the tool on 15 studies.

Assumption of no unobserved confounding

Regression and matching methods assume "*no unobserved confounding"*. Under this assumption, the allocation of two individuals, who have similar observed characteristics but are in different treatment arms, can be thought of as effectively at random (Greenland et al., 1999). So for example in the CEA described by Polsky and Basu (2006) (and Table 3.1), this assumption implies that after regression adjustment there are no differences in the distributions of unobserved confounders between treatment arms. Approaches that use longitudinal data such as panel data regression and "difference-in-differences" rely on a weaker form of this assumption; they assume that changes over time in unobserved confounders are conditionally independent of treatment (Imbens and Jeffrey, 2007).

The assumption of no unobserved confounding cannot be tested (Imbens and Wooldridge, 2009a). However, as ***Question 1a*** states, studies should assess whether this assumption is plausible (Table 3.2). To meet this criterion a study is required to draw on external evidence or expert opinion of the potential influence of observed and unobserved baseline covariates on treatment assignment and endpoints (Rubin, 2010). Causal diagrams can be useful for defining the structure of such relationships (Pearl, 2001).

Table 3.2 - Checklist for critically appraising statistical methods to address selection bias, in estimating incremental costs, effectiveness and cost-effectiveness

Q1a: Did the study assess the "no unobserved confounding" assumption?

a) Yes, for example with causal diagrams informed by external literature ☐
b) Partially, for example with external literature or expert opinion ☐
c) No ☐
d) Not applicable ☐

Q1b: Did the study assess the assumption that the IV was valid?

a) Yes, for example, with causal diagrams informed by external literature ☐
b) Partially, for example, with external literature or expert opinion ☐
c) No ☐
d) Not applicable ☐

Q2: Did the study assess whether the distributions of the baseline covariates overlapped between the treatment groups?

a) Yes, for example, with histograms and standardised differences ☐
b) Partially, for example, just with standardised differences ☐
c) No ☐
d) Not applicable ☐

Q3: Did the study assess the specification of the regression model for

	(i.) Health outcomes?	**(ii.) Costs?**
a) Yes, for example, with statistical tests or plots	☐	☐
b) Partially, with qualitative arguments	☐	☐
c) No	☐	☐
d) Not applicable	☐	☐

Q4: Was covariate balance assessed after applying a matching method?

a) Yes, with a measure that considered imbalance across different aspects of the distribution ☐
b) Partially, for example, by assessing mean differences ☐
c) No ☐
d) Not applicable ☐

Q5: Did the study consider structural uncertainty arising from the choice or specification of the statistical method for addressing selection bias?

a) Yes, the sensitivity of cost-effectiveness results to the choice of method was quantitatively assessed and interpreted ☐
b) Partially, for example, by additional statistical analysis but without interpetation, or commentary with no quantitave assessment ☐
c) No ☐

One way of handling both observed and unobserved confounding is with IV methods (Basu et al., 2007, Mullahy, 2011). IV estimation assumes that the instrument only influences endpoints through treatment and is independent of the unobserved confounders (Hernán and Robins, 2006). Although this assumption is also untestable, studies should again assess plausibility, for example with causal diagrams informed by expert opinion and evidence from the literature (Joffe and Mindell, 2006) (***Question 1b***). For instance in a Mendelian randomisation study where genotype is the proposed instrument, causal diagrams may help the assessment of whether genotype only influences the endpoint of interest through the treatment (Didelez and Sheehan, 2007).

Assumption of good overlap in the distribution of baseline covariates between treatment arms

Question 2 highlights that methods that assume no unobserved confounding also assume that there is *good overlap* in the distributions of baseline covariates between the treatment groups (Imbens and Wooldridge, 2009a). Good overlap implies there are no baseline characteristics which fully predict treatment status. With weak overlap, a regression model extrapolates beyond the observed covariate data, so unless the model is correctly specified, this will lead to biased estimates (Ho et al., 2007). Overlap can be assessed by inspecting histograms or density plots (continuous covariates), and by reporting standardised differences (categorical or continuous covariates) (Imbens and Wooldridge, 2009b), before statistical adjustment takes place. A remedy for weak overlap is to constrain the sample to the area of good overlap, but recognising that this alters the population of interest (Crump et al., 2009).

Assumption that the parametric regression model is correctly specified

Parametric regression models can only provide unbiased and efficient parameter estimates if they are correctly specified. For unbiased estimates, the model must state the true relationship between the covariates and the mean endpoint (Ho et al., 2007). For the model to provide the most precise estimates, the probability distribution of the endpoint or error terms must be correct. In CEAs, these two elements of correct model specification are challenging, especially as it is also important to recognise any correlation of costs with health outcomes. Flexible bivariate models have been proposed that allow for non-normal distributions, and can improve the precision of the estimates (Nixon and Thompson, 2005), but less attention has been given to specifying the correct relationship between the covariates and the mean endpoint (Thompson et al., 2006). While the true parametric model is always unknown, a study should assess the relative fit of alternative models, for example by using likelihood based model diagnostics or cross validation (Jones, 2010, Hill and Miller, 2010). ***Question 3*** of the checklist considers whether the study has evaluated the model specification appropriately.

Assumption that a matching method has balanced the matched samples

Matching aims to balance treatment and control groups, by creating matched samples with similar observed covariate distributions. The estimated propensity score (Pscore) is often used as a balancing score (Rosenbaum and Rubin, 1983). As ***Question 4*** emphasises, matching methods can only produce unbiased estimates if matching balances the distributions of baseline covariates. Hence the appropriate specification test is whether the matched samples are balanced (Stuart, 2010). Some balance diagnostics such as standardised differences use a comparison of means, but these are insufficient for capturing imbalances elsewhere in the covariate distributions (Sekhon and Grieve,

2011). In CEAs, covariates tend to be non-normal, making it important to assess balance on the full distributions of the covariates using graphical tools (e.g. quantile-quantile plots) and nonparametric tests (Stuart, 2010).

Structural uncertainty from the choice of statistical method for addressing selection bias

Each of the statistical approaches described makes untestable assumptions, which implies that no one approach is ideal. The choice of statistical method for estimating cost-effectiveness parameters from IPD can therefore contribute to structural uncertainty, both when the CEA takes data from a single study and if the estimates are used in a decision-model (Jackson et al., 2011). ***Question 5*** considers structural uncertainty in the context of addressing selection bias. These criteria assess whether the study has considered the impact of choosing alternative methods for addressing selection bias, for example by making alternative assumptions about unobserved confounders, or assuming different regression model specifications.

Even if, for example, the study has previously judged that a regression model is appropriate, it is still important to assess whether the results are sensitive to alternative approaches. The rationale is that even amongst regression models with similar fit, results can still be sensitive to the choice of model (Thompson and Nixon 2005). A recommended approach for characterising structural uncertainty is to repeat the analysis with alternative structural assumptions and carefully report and interpret the impact on results. One way to assess whether results are sensitive to the potential for unobserved confounders is to employ "Rosenbaum bounds" (Rosenbaum, 2002). Here, cost-effectiveness results can be reported according to alternative assumptions about the level of unobserved confounding (see Noah et al., 2011).

Systematic review of published cost-effectiveness analyses

A literature search identified published economic evaluations that used observational data. The databases searched were MEDLINE, EMBASE, NHS Economic Evaluations Database (NEED), and the Health Economic Evaluations Database (HEED) (Appendix 3.2). Inclusion and exclusion criteria were applied in title, abstract and full text reviews. Papers had to be published between 2000-2011, report a full economic evaluation (Drummond et al, 2005) that estimated at least one incremental parameter (e.g. incremental costs, incremental quality adjusted life years, or an incremental surrogate measure such as relative risks) using observational IPD. The studies had to employ a statistical method to address selection bias (Appendix 3.2).

Title and abstract screening were conducted by one reviewer (NK); a second reviewer (ZS) verified exclusion criteria on a random sample of 5% of excluded studies. There were no disagreements on the articles excluded. The selected studies were critically appraised by the first reviewer. The second reviewer independently appraised a random sample of 50% of these papers. The inter-rater reliability of the checklist was good (kappa > 0.95), disagreements were resolved by a third reviewer (RG). Pre-specified subgroup analyses were defined according to publication year (post 2006 versus 2006 or earlier), the observational study's design (prospective versus retrospective), and journal type (health economics or statistics versus other).

Results

The literature search yielded 4203 abstracts, 257 papers were selected for full text review with data extracted from 81 papers (Figure 3.1). The most common statistical method for addressing selection bias was regression (Table 3.3).

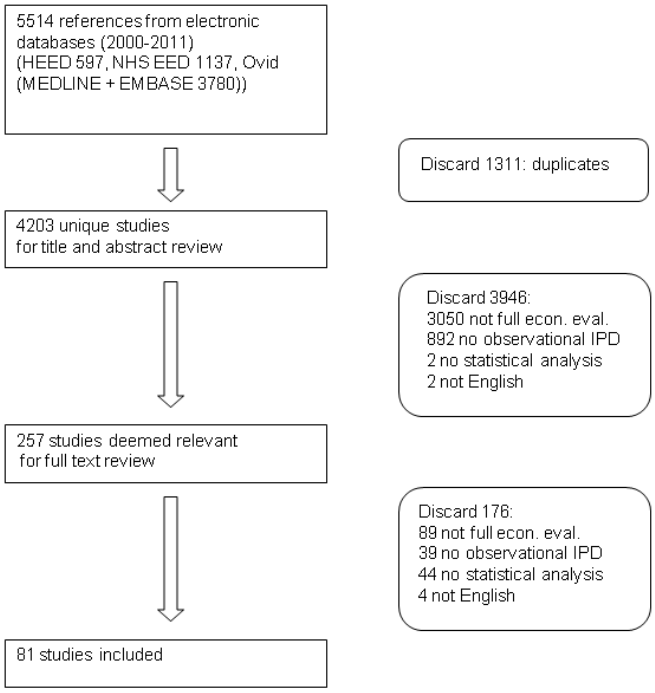


Figure 3.1. - Flow chart of studies included in the systematic review of published economic evaluations

Most studies (77%) did not assess the assumption of "no unobserved confounding" (Table 3.4). Both studies using IV methods only partially assessed the validity of the instrument. A small minority of regression-based studies fully assessed model specification with statistical tests; others gave a partial assessment by citing methodological work to justify the model choice. Around half of the matching studies assessed covariate balance by comparing means in the matched samples.

A minority of studies (16%) fully considered structural uncertainty. They reported and interpreted cost-effectiveness estimates across different methods (e.g. regression and matching); 35% partially assessed structural uncertainty by considering alternative model specifications, or by providing a "qualitative assessment" of the potential role of unobserved confounders.

Structural uncertainty was fully considered by a higher proportion of studies published in health economics and statistics journals (30% vs. 13%; p=0.091). For other items any differences between subgroups were relatively small.

Table 3.3 - Characteristics of studies included in the review (n=81)

Primary statistical method to address selection bias					
Regression	Matching on the Pscore	Matching on covariates	Instrumental variables		
41 (51%)	18 (22%)	20 (25%)	2 (2%)		
Year of publication					
2000-2005	2006-2011				
36 (44%)	45 (56%)				
Journal type					
Health economics	Statistics	Medical	Public health, health services.		Other
16 (20%)	1 (1%)	53 (65%)	9 (11%)		2 (3%)
Intervention type					
Health services	Disease management	Prevention, screening	Substance abuse treatment		Other
11 (14%)	54 (67%)	7 (9%)	6 (7%)		3 (4%)
Design of observational study					
Concurrent cohort	Before-after	Historic cohort	Cross-sectional	Case - control	Other
41 (51%)	2 (2%)	25 (31%)	2 (2%)	2 (2%)	9 (11%)
Observational data is used to estimate…			**Does the CEA use a decision model?**		
Incremental cost and effects or INB	Incremental effects [1]	Incremental cost [2]	Yes	No	
54 (67%)	26 (32%)	1 (1%)	12 (15%)	69 (85%)	

Notes: Pscore, propensity score; INB, incremental net benefit; IPD, individual patient data, CEA, cost-effectiveness analysis

[1] These studies used aggregate estimates of incremental costs, for example, from external literature.

[2] This study used an aggregate estimate of incremental effectiveness from external literature.

Table 3.4 - Results of applying the checklist to published CEAs (n=81)

Q1a. Was the "no unobserved confounding" assumption assessed?[1]	
Yes	Partially
1/79 (1%)	17/79 (22%)

Q2. Was the overlap of the covariate distributions between the treatment groups assessed?[1]	
Yes	Partially
2/79 (3%)	0/79 (0%)

Q3. For regression methods, was model specification assessed, for

Health outcomes?[2]		**Costs?[2]**	
Yes	Partially	Yes	Partially
8/41 (20%)	12/41 (29%)	3/23(13%)	9/23 (39%)

Q4. Was covariate balance assessed after applying a matching method?	
Yes	Partially
1/38 (3%)	20/38 (51%)

Q5. Was structural uncertainty from the choice of statistical method considered?	
Yes	Partially
13/81 (16%)	28/81 (35%)

Notes: [1] Two studies used IV estimation which does not rely on the assumption of no unobserved confounding. For these studies Q1a and Q2 do not apply. Therefore for Q1a and Q2 the denominator was 79. Results for Question 1b are given in the Results section.

[2] Regression adjustment was used in 41 papers for health outcomes, 23 for costs.

Discussion

This paper presents a critical appraisal tool for assessing and improving the way selection bias is addressed in CEAs that use observational data. The systematic review found that in the majority of published CEAs, the main assumptions underlying the statistical methods were not assessed. In particular, most studies assumed "no unobserved confounding" without any justification, raising concerns that the cost-effectiveness estimates were biased. To improve practice, studies could use external evidence, for example from previous clinical studies to carefully consider potential confounders. Synthesising this information in causal diagrams can help make the assumption of no unobserved confounding explicit, and help assess whether this assumption is credible (Joffe and Mindell, 2006).

We found that half the matching studies reported balance statistics, but only one study followed recent recommendations and assessed balance according to the full covariate distribution (Stuart, 2010). In CEAs, covariates tend to have irregular distributions, so using balance statistics that consider the full covariate distribution can be important in helping to address bias (Sekhon and Grieve, 2011).

A promising approach for CEAs would be to combine matching with regression; this can address residual imbalances post matching, and findings tend to be insensitive to model choice (Ho et al., 2007, Abadie and Imbens, 2011). Another alternative is to combine weighting on the inverse probability of treatment assignment with regression (Robins et al., 1994). Such estimators can have double-robust properties; if either the model for the endpoint or for the treatment assignment is correctly specified, estimates are unbiased and achieve semiparametric efficiency (Basu et al., 2008, Vansteelandt et al., 2011). Further work is required to test double-robust methods in CEA.

This paper has some limitations; as with any critical appraisal tool, study quality is judged according to the methods reported, and does not recognise that a study may have justified an assumption without reporting it. However, the goal is to encourage studies to adopt better methods and report them transparently. Studies which did not apply a statistical method for addressing selection bias were excluded from the review, so the findings do not apply to all CEAs that use observational data, nor to those that relied on aggregate estimates from previous studies. The checklist does not cover other aspects of statistical analysis, for example the handling of missing or censored data. The checklist is therefore intended to complement rather than substitute for current methods guides for CEA (Drummond et al., 2005, Philips et al., 2006, Glick et al., 2007).

Furthermore, as with any methodological guidance (Philips et al., 2006), our checklist cannot prescribe which specific statistical method should be used to address selection bias in CEAs. The checklist's main contribution is to raise awareness of the assumptions underlying alternative statistical methods.

There has been recent interest and investment in comparative effectiveness research, which tends to assess cost-effectiveness without using RCTs (Sox et al., 2010). The findings from our paper suggest that it is vital to scrutinise the assumptions behind the statistical methods that purport to address selection bias. Indeed critical appraisal may reveal that the study design is flawed in that key underlying assumptions, for example that there are no unobserved confounders, cannot be justified. These insights might encourage future studies with more rigorous designs such as prospective cohort studies that collect baseline data on rich set of measured covariates (Rubin, 2010) including plausible instruments, or RCTs.

In conclusion, CEAs that use observational data rarely assess the main assumptions behind statistical analyses for addressing selection bias. This checklist can raise

awareness about the major assumptions behind statistical methods for addressing selection bias and can complement existing method guidelines for CEAs.

Acknowledgments

We gratefully acknowledge John Cairns, Rhian Daniel, James Carpenter, Rosalba Radice (all LSHTM), Jasjeet S. Sekhon (UC Berkeley) and Roland Ramsahai (University of Cambridge) for reviewing provisional versions of the checklist. We also thank Carla Guerriero, Manuel Gomes and Mark Pennington (all LSHTM) for piloting the checklist. This work was funded by the Economic and Social Research Council (Grant no. RES-061-25-0343).

References

Abadie, A. & Imbens, G. W. 2011. Bias-Corrected Matching Estimators for Average Treatment Effects. *Journal of Business & Economic Statistics,* 29**,** 1-11.

Austin, P. C. 2009. Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. *Statistics in Medicine,* 28 3083-3107.

Barber, J. & Thompson, S. G. 2004. Multiple regression of cost data: use of generalised linear models. *Journal of Health Services Research Policy,* 9**,** 197-204.

Basu, A., Heckman, J. J., Navarro-Lozano, S. & 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**,** 1133-1157.

Basu, A., Manning, W. G. & Mullahy, J. 2004. Comparing alternative models: log vs Cox proportional hazard? *Health Economics,* 13**,** 749-765.

Basu, A., Polsky, D. & Manning, W. G. 2008. *Use of Propensity Scores in Non-Linear Response Models: The Case for Health Care Expenditures* [Online]. National Bureau of Economic Research, Inc. Available: http://ideas.repec.org/p/nbr/nberwo/14086.html [Accessed 05/10/2010.

Brookhart, M. A., Schneeweiss, S., Rothman, K. J., Glynn, R. J., Avorn, J. & Sturmer, T. 2006. Variable selection for propensity score models. *Am J Epidemiol,* 163**,** 1149-56.

Buntin, M. B. & Zaslavsky, A. M. 2004. Too much ado about two-part models and transformation?: Comparing methods of modeling Medicare expenditures. *Journal of Health Economics,* 23**,** 525-542.

Crump, R. K., Hotz, V. J., Imbens, G. W. & Mitnik, O. A. 2009. Dealing with limited overlap in estimation of average treatment effects. *Biometrika,* 96**,** 187-199.

Diamond, A. & Sekhon, J. S. 2012. Genetic Matching for Estimating Causal Effects: A General Multivariate Matching Method for Achieving Balance in Observational Studies. *Review of Economic and Statistics,* Forthcoming.

Didelez, V. & Sheehan, N. 2007. Mendelian randomization as an instrumental variable approach to causal inference. *Statistical Methods in Medical Research,* 16**,** 309-330.

Drummond, M., Sculpher, M., Torrance, G., O'Brien, B. & Stoddart, G. 2005. *Methods for the Economic Evaluation of Health Care Programmes,* Oxford, Oxford University Press.

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

Greenland, S., Pearl, J. & Robins, J. M. 1999. Confounding and Collapsibility in Causal Inference. *Statist. Sci. ,* 14**,** 29-46.

Hernán, M. A. & Robins, J. M. 2006. Instruments for Causal Inference: An Epidemiologist's Dream? *Epidemiology,* 17**,** 360-372.

Hill, S. C. & Miller, G. E. 2010. Health expenditure estimation and functional form: applications of the generalized gamma and extended estimating equations models. *Health Economics,* 19**,** 608-627.

Ho, D. E., Imai, K., King, G. & Stuart, E. A. 2007. Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference *Political Analysis,* 15**,** 199-236.

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

Hosmer, D. W., Lemeshow, S. & May, S. 2008. *Applied survival analysis : regression modeling of time-to-event data,* Hoboken, N.J., Wiley ; Chichester : John Wiley [distributor].

Imbens, G. & Jeffrey, M. W. 2007. Difference-in-Differences Estimation. *NBER Lecture Notes.*

Imbens, G. & Wooldridge, J. M. 2009b. New Developments in Econometrics. *Lecture Notes, CEMMAP, UCL.*

Imbens, G. M. & Wooldridge, J. M. 2009a. Recent Developments in the Econometrics of Program Evaluation. *Journal of Economic Literature,* 47**,** 5–86.

Jackson, C., Bojke, L., Thompson, S., Claxton, K. & Sharples, L. 2011. A Framework for Addressing Structural Uncertainty in Decision Models. *Medical Decision Making,* 31**,** 662–674.

Joffe, M. & Mindell, J. 2006. Complex Causal Process Diagrams for Analyzing the Health Impacts of Policy Interventions. *American Journal of Public Health,* 96**,** 473-479.

Jones, A. M. 2007. Identification of treatment effects in Health Economics. *Health Economics,* 16**,** 1127-1131.

Jones, A. M. 2010. Models For Health Care. *HEDG Working Papers.* HEDG, c/o Department of Economics, University of York.

Jones, A. M. & Rice, N. 2011. Econometric Evaluation of Health Policies. *In:* GLIED, S. & SMITH, P. (eds.) *The Oxford handbook of health economics.* Oxford: Oxford University Press.

Mullahy, J. 2011. Symposium on genetic data in health economics research. *Health Economics***,** n/a-n/a.

Nixon, R. M. & Thompson, S. G. 2005. Methods for incorporating covariate adjustment, subgroup analysis and between-centre differences into cost-effectiveness evaluations. *Health Economics,* 14**,** 1217-1229.

Noah, M. A., Peek, G. J., Finney, S. J., Griffiths, M. J., Harrison, D. A., Grieve, R., Sadique, M. Z., Sekhon, J. S., McAuley, D. F., Firmin, R. K., Harvey, C., Cordingley, J. J., Price, S., Vuylsteke, A., Jenkins, D. P., Noble, D. W., Bloomfield, R., Walsh, T. S., Perkins, G. D., Menon, D., Taylor, B. L. & Rowan, K. M. 2011. Referral to an Extracorporeal Membrane Oxygenation Center and Mortality Among Patients With Severe 2009 Influenza A(H1N1). *JAMA: The Journal of the American Medical Association,* 306**,** 1659-1668.

Pearl, J. 2001. Causal Inference in the Health Sciences: A Conceptual Introduction. *Health Services and Outcomes Research Methodology,* 2**,** 189-220.

Petersen, M. L., Porter, K., Gruber, S., Wang, Y. & Laan., M. J. v. d. 2010. Diagnosing and Responding to Violations in the Positivity Assumption. *U.C. Berkeley Division of Biostatistics Working Paper Series*.

Philips, Z., Bojke, L., Sculpher, M., Claxton, K. & Golder, S. 2006. Good practice guidelines for decision-analytic modelling in health technology assessment: a review and consolidation of quality assessment. *Pharmacoeconomics,* 24**,** 355-71.

Polsky, D. & Basu, A. 2006. Selection Bias in Observational Data. *The Elgar Companion to Health Economics*. Edward Elgar Publishing.

Robins, J., Rotnitzky, A. & Zhao, L. P. 1994. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association,* 89**,** 846-866.

Rosenbaum, P. R. 2002. *Observational studies,* New York ; London, Springer.

Rosenbaum , P. R. & Rubin, D. B. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika,* 70**,** 41-55.

Rubin, D. B. 2010. On the limitations of comparative effectiveness research. *Statistics in Medicine,* 29**,** 1991-1995.

Sekhon, J. S. & Grieve, R. D. 2011. A Matching Method for Improving Covariate Balance in Cost-Effectiveness Analyses. *Health Economics,* doi: 10.1002/hec.1748.

Sox, H., Helfand, M., Grimshaw, J., Dickersin, K., Tovey, D. & al., e. 2010. Comparative effectiveness research: challenges for medical journals. *Medical Decision Making,* 30**,** 301-303.

Stuart, E. A. 2010. Matching Methods for Causal Inference: A review and a look forward. *Statistical Science,* 25.

Thompson, S. G., Nixon, R. M. & Grieve, R. 2006. Addressing the issues that arise in analysing multicentre cost data, with application to a multinational study. *J Health Econ,* 25**,** 1015-28

Vansteelandt, S., Bekaert, M. & Claeskens, G. 2011. On model selection and model misspecification in causal inference. *Statistical Methods in Medical Research*.

Willan, A. R. & Briggs, A. H. 2006. *Statistical Analysis of Cost-effectiveness Data*, John Wiley & Sons Ltd.

Willan, A. R., Briggs, A. H. & Hoch, J. S. 2004. Regression methods for covariate adjustment and subgroup analysis for non-censored cost-effectiveness data. *Health Economics*, 13**,** 461-475.

Appendix 3.1 - Methodological guidance to the checklist

The aim of this guidance is to help a reviewer judge whether or not the published study assessed the main assumptions underlying the use of the statistical method for addressing selection bias. The guidance offers examples of how each criterion can be met, drawing on findings from the conceptual review. However, the guidance does not aim to be exhaustive. The checklist and guidance are intended for studies that meet the same inclusion criteria stipulated for the systematic review reported in the paper (Appendix 3.2).

Step 1: Identifying the primary statistical method

The checklist focuses on the statistical method that is used to address selection bias. A common example is where the study attempts to address selection bias with regression analysis, specifying a binary variable for the treatment and additional baseline covariates to adjust for observed differences in patient characteristics. By contrast, a regression model which aims to predict costs associated with the *outcome of interest* (e.g. the costs of atrial fibrillation event versus no event, as input to a decision analytical model) is not considered a relevant approach for this study.

Some questions of the checklist are applicable only for certain methods, therefore the primary statistical method needs to be identified and recorded. If several statistical methods are applied, the one which the authors rely on in the CEA is selected as primary method and any other method will count as part of the structural sensitivity analysis (Table 3.2, Question 5).

The classification of statistical methods used in this paper, with some accompanying examples, is as follows:

1. **Regression**: Examples are ordinary least squares (OLS), Generalised Linear Models (GLMs), panel data regression, difference-in-differences methods, Cox survival regression or net benefit regression.

2. **Matching methods:**

- Matching on the Pscore: for example, Pscore matching, inverse probability of treatment weighting (IPTW), kernel density matching, stratification/subclassification/blocking/interval matching, regression on the Pscore or full matching.
- Matching on individual covariates: for example, exact matching, Mahalanobis distance matching, or Genetic Matching.

3. **Instrumental variables (IV):** for example, two stage least squares, two stage residual inclusion or generalised method of moments.

Step 2: Applying the checklist

Question 1a: Did the study assess the "no unobserved confounding" assumption?

Applicability: A statistical method relies on the assumption of no unobserved confounding if regression or matching methods are used as a primary analysis of the study. If the primary method that is used allows for unobserved confounding (e.g. IV), the "Not applicable" option should be selected.

a) Yes, for example, with causal diagrams informed by external literature.

The assumption that needs to be assessed is that all potential confounders are accounted for in the statistical analysis, that is, baseline covariates that are associated with the treatment assignment and the cost or effectiveness endpoint. This assumption might also be referred to as "unconfoundedness", "strong ignorability", "exogeneity",

"selection on observables" or "conditional independence" (Imbens and Wooldridge, 2009a, Imbens and Wooldridge, 2009b, Jones and Rice, 2011). If a panel data regression or a difference-in-differences method is used, the assumption is somewhat weaker: instead the assumption is that there are no time-varying unobserved confounders correlated with treatment assignment and the endpoints (Imbens and Jeffrey, 2007).

A paper is defined to have fully assessed the assumption, if the causal relationships between covariates (observed and hypothesised unobserved) and endpoints (both cost and effectiveness) were assessed on the basis of *a priori* scientific knowledge (e.g. previous clinical studies on prognostic factors), using some of the following tools:

- Graphical representation with causal diagrams or directed acyclic graphs (Pearl, 2001).
- Mathematical description of the relationships, by structural equation models (Pearl, 2001).

These considerations might be complemented with placebo tests (Jones, 2007), which can detect violations of the assumption.

b) Partially, for example, with external literature/expert opinion.

The criterion would be partially met, if, for example:

- The authors justified the set of observed confounders used in their statistical methods with substantive *a priori* knowledge, for example of risk factors for the disease or mechanism of treatment assignment (Rubin, 2010), but did not use the tools mentioned in **a).**

- The paper justified the choice of specific covariates by commenting (for example, in the discussion) on the plausibility and sufficiency of the observed confounders.

c) No, if none of the above applies.

For example, if covariate selection is based solely on statistical tests (e.g. t-test of equality of covariate means between treatment groups) or automated model selection (e.g. stepwise) procedures (Brookhart et al., 2006). A general warning about unobserved confounding does not fulfil the criterion.

d) Not applicable - see applicability

Question 1b: Did the study assess the assumption that the IV was valid?

Applicability: This question is applicable if instrumental variable estimation was used as the primary statistical analysis. Otherwise, the "Not applicable" option should be selected.

The validity assumption consists of two untestable assumptions (Hernán and Robins, 2006): (i) the instrument only influences the outcome through treatment and (ii) the instrument is independent of the unobserved confounders. Although these assumptions cannot be tested directly, they can be assessed in similar ways to the no unobserved confounding assumption.

a) Yes, for example, with causal diagrams informed by external literature.

Similarly to Question 1a, previous empirical information and expert judgment need to be combined to assess the assumption that the instrument has a causal effect on the treatment, but does not have an independent casual effect on the outcome, nor is it associated with unobserved confounders. The causal diagrams can use this information

to formulate the relationships between the treatment, the instrument, observed and unobserved confounders and the endpoints. The use of causal diagrams can make the IV assumption explicit (Joffe and Mindell, 2006). If more than one instrument is available, tests of overidentifying restrictions can be used in addition (Jones and Rice, 2011).

b) Partially, for example, with external literature/expert opinion

- The authors justified the choice of the IV with substantive *a priori* knowledge of how the instrument is associated with treatment assignment, and is conditionally independent of endpoints without using the tools mentioned in **a).**
- The criterion can be partly met with commentary in the discussion justifying the validity of the instrument, if it uses prior scientific knowledge (e.g. other studies justifiying the use of the same instrument) or expert opinion.

c) No, if none of the above applies.

d) Not applicable - see applicability.

<u>Question 2: Did the study assess whether the baseline covariates (e.g. age and sex) had distributions that overlapped between the treatment groups?</u>

Applicability: This question is applicable for the same studies where **Q1a** is applicable. For IV, "Not applicable" should be selected.

a) Yes, for example with histograms and standardised differences.

This assumption is also referred to as the "common support" assumption (Rosenbaum and Rubin, 1983), the "experimental treatment assignment ", or the "positivity assumption" (Petersen et al., 2010). It is fully assessed if one of the following steps is taken, and the authors make explicit that the intention is assessing overlap:

- Histograms or smoothed density plots of the continuous covariates are plotted (Imbens and Wooldridge, 2009b), and areas of weak support in the densities are investigated. For binary variables, standardised differences are investigated.
- If there are many covariates, this criteria is met if the distribution of Pscore is inspected for both treatment groups, so as to reveal possible lack of overlap in the multivariate covariate distributions.
- Quantiles of the Pscore distributions are investigated.

b) Partially, for example, just with standardised differences.

Inspecting standardised differences of variables with the explicit objective of assessing overlap partially fulfills this requirement (Imbens and Wooldridge, 2009b).

c) No, if none of the above applies, for example if standardised differences are reported for the purposes of assessing balance.

d) Not applicable - see applicability.

Question 3: Did the study assess the specification of the regression model for (i) health outcomes and (ii) cost?

Applicability: This question is applicable for studies where the primary statistical analysis to address selection bias was regression adjustment. If regression adjustment was performed only for the cost or effectiveness endpoint, only the relevant part of the question needs to be answered. If regression on a univariate measure of net benefit was used (e.g. net benefit regression) the same answers should be given to both questions.

a) Yes, for example with statistical tests/plots.

Several options are available to statistically assess the specification of a regression model, for example,

- For GLMs, the correct specification of the link function and the outcome distribution can be tested, for example, with the the Hosmer–Lemeshow, Pregibon's link test, or the modified Park test (Basu et al., 2004, Jones, 2010).
- The fit of models estimated through a maximum likelihood method can be assessed using log-likelihood based fit statistics, for example, the Akaike Information Criterion or the Bayesian Information Criterion (Barber and Thompson, 2004, Jackson et al., 2011).
- Cross-validation is a general method to assess model fit based on predictive abilities (Buntin and Zaslavsky, 2004, Hill and Miller, 2010).
- If a linear model (e.g. OLS) is used, residual plots can be examined to detect misspecification of the functional form of the regression model or heteroscedastic errors (Jones, 2010). Multicollinearity in the models can also be assessed.
- Lag structure of time series models can be tested.
- If a semiparametric Cox proportional hazards model is used, the proportionality of the hazard can be tested (Hosmer et al., 2008).

b) Partially, with qualitative arguments.

If a study uses previous applied and methodological work to guide the choice of the regression model, the criterion can be considered partly met, for example,

- The distribution of an outcome can imply a modeling approach, e.g. logistic regression for binary data, or two-part models for costs with a large number of zeros (Buntin and Zaslavsky, 2004).
- There might be a consensus in the clinical literature that certain covariates have a nonlinear effect on the outcome, or interaction with the treatment.
- The linear net benefit regression is used and referenced (Hoch et al., 2002).

c) No, if none of the above applies.

d) Not applicable - see applicability.

Question 4: Was covariate balance assessed after applying a matching method?

Applicability: This question is applicable if the primary statistical method to address selection bias is a matching method.

a) Yes, with a measure that considered imbalance across different aspects of the distribution

This requirement is fulfilled if covariate balanced is assessed for aspects of the covariate distributions beyond the mean. Examples include:

- For continuous covariates quantile–quantile plots can be examined, which compare the empirical distributions of variables in the treatment and control groups. This can be also compared for second moments of the variables, and interactions (Stuart, 2010).
- Nonparametric tests of the equality of distributions can be performed, for example Kolmogorov-Smirnoff tests (Diamond and Sekhon, 2012).
- Five-number summaries (quantiles) of the distributions can be provided (Austin, 2009a).
- Side-by side boxplots (Austin, 2009a) are presented.
- Higher moments (variance, skewness, kurtosis) and cross-moments (covariance) of covariate distributions are compared (Jones and Rice, 2011). Variance can be compared using variance ratios (Austin, 2009a).

The assessment can be performed for any matching method. For methods that create matched treatment and control groups, the resultant groups can be compared. For Pscore subclassification, the assessment can be performed by the created strata; for IPTW, weighted boxplots can be used to assess balance (Stuart, 2010). If the Pscore is used for covariate adjustment, weighted conditional standardised absolute differences can be computed (Austin, 2009a).

b) Partially, for exmaple, by assessing mean differences.

Balance is partially assessed if covariate *means* of matched groups are compared, using the following tools, for example,

- *Standardised differences* (also referred to as normalised differences) are measures which express difference in means in units of the pooled standard deviation (Rosenbaum and Rubin, 1983) and are frequently used as balance diagnostics (Austin, 2009a). These measures are recommended for balance assessment since they are invariant to sample size (Stuart, 2010) and can be applied across a wide range of balancing methods (matching, IPTW, stratification). Using graphical displays (Austin, 2009a, Stuart, 2010) makes standardised differences on a large number of covariates easier to interpret.
- Comparing *mean differences* with for example t-tests carries some information on balance, therefore it is considered to partially fulfil the criterion of balance assessment. It is, however, not recommended because sample size can change during a matching process resulting in misleading tests statistics (Imbens and Wooldridge, 2009a).

c) No, if none of the above applies. For example, the c-statistic or area under the receiver operating characteristic curve of the Pscore model and balance assessed on the estimated Pscore is not regarded as informative (Austin, 2009a).

d) Not applicable – see applicability.

Question 5: Did the study consider structural uncertainty arising from the choice or specification of the statistical method for addressing selection bias?

a) Yes, the authors quantitatively assessed and interpreted the sensitivity of cost-effectiveness results to the choice of method.

This criterion is fully met if the authors conducted an additional statistical analysis beyond the primary method used to address selection bias, and interpreted how the results are altered by using the alternative method. Structural uncertainty stems from many sources (Jackson et al., 2011), and the particular form of structural uncertainty here is that pertaining to the method for handling selection bias. Even this specific form of structural uncertainty can take several forms. Some examples are as follows:

- Distinct methods based on different structural assumptions are applied, and results are reported and compared (e.g. instrumental variables versus Pscore, matching versus regression (Polsky and Basu, 2006)).
- Methods are combined to add robustness to the analysis. For example methods combining regression and Pscore (Robins et al., 1994), or matched data adjusted using regression models (Ho et al., 2007). Results of these analyses are reported, and compared to those obtained from just one method.
- Different specifications of the cost and effectiveness regressions are applied, results are reported and compared (e.g. with and without interactions; OLS versus gamma GLM).

- Structural uncertainty in the choice of parametric model for regression can be quantified, for example by Bayesian model averaging (Jackson et al., 2011).
- Assessing the sensitivity to the assumption of no unobserved confounders, by exploring the effect of potentially omitted confounders on the parameters of interest, using sensitivity analysis (Rosenbaum, 2002).

b) Partially, for example, additional statistical analysis with no interpetation, or commentary with no quantitave assessment.

The criterion is partially met if:

- Statistical analysis beyond the primary method was performed; however, the implications for cost-effectiveness results were not interpreted appropriately. Examples are:
  - Specification tests for regression model are conducted, but results obtained using different specifications are not contrasted.
  - Matched data is adjusted with a regression model.
  - Pscore is included as an additional covariate in the regression model.
  - As a sensitivity analysis, some covariates are omitted from the set of variables used in the statistical analysis.
- Commentary on the implications of the method choice is provided, without conducting a formal analysis, for example:
  - by discussing suspected bias due to unobserved confounders.
  - by outlining a possible instrumental variables analysis.

c) No, if none of the above applies. For example, conducting sensitivity analysis for other sources of structural uncertainty (e.g. Markov model structure) do not fulfil the criterion.

Appendix 3.2 – Systematic review search terms and inclusion criteria

Search terms

In order to minimise the risk of omitting potentially relevant studies, the search terms were broad, combining two requirements: the study is an economic evaluation and uses a statistical method. The search terms used for the MEDLINE and EMBASE databases are listed in Appendix 3.2 Table 1, and were adapted for the NHS EED (through Cochrane Library) and HEED (through Wiley Online library) databases.

Appendix 3.2 Table 1 - Search terms for NHS EED (adapted for HEED, MEDLINE and EMBASE databases)

#1 "economic evaluation" OR "cost effectiveness " OR "cost-effectiveness" OR "cost utility" OR "cost-benefit" in Economic Evaluations
#2 "regression" OR "covariate adjustment" OR "ordinary least square*" OR "OLS" or "generalised estimating equation*" OR "linear model*" OR "nonlinear model*" OR "logistic model*" in Economic Evaluations
#3 ("double robust" OR "doubly robust" OR "inverse probability weight*" OR "inverse probability of treatment") OR (weight* AND "propensity score*") in Economic Evaluations
#4 (stratification OR stratify OR stratified OR blocking OR block OR strata) AND "propensity score*" in Economic Evaluations
#5 "propensity score*" in Economic Evaluations
#6 "matching" or "matched" in Economic Evaluations
#7 "two stage least squares" OR "two-stage least squares" OR "2SLS" OR "instrumental variable*" in Economic Evaluations
#8 "panel data" OR "difference in differences" OR "repeated cross section" OR "repeated cross-section" OR "fixed effect*" in Economic Evaluations
#9 "regression discontinuity" in Economic Evaluations
#10 "control function" OR "Heckman selection" OR "selection model" in Economic Evaluations
#11 (#1 AND (#2 OR #3 OR #4 OR #5 OR #6 OR #7 OR #8 OR #9 OR #10)), from 2000 to 2010

Inclusion criteria

From the results of the broad search, relevant studies were narrowed down using the following inclusion criteria. Examples for excluded studies are provided.

1. *Individual patient level observational data are used* to calculate at least one of the following parameters: incremental cost, effectiveness or cost effectiveness parameters or relative surrogate outcomes, for example, relative risk of mortality.

Examples for excluded studies: when a decision analytical model uses aggregate inputs only, or when a study uses individual level RCT data only.

2. *The study is a full economic evaluation*: cost-effectiveness, cost-utility or cost-benefit analysis. Examples for excluded studies:
    - Cost-minimisation or cost-consequences analysis.
    - A study labelled as cost-benefit analysis which accounts cost-saving as benefits.
3. *A statistical method is used to address selection bias* when calculating at least one of the following parameters: incremental cost, incremental effectiveness or cost-effectiveness or relative surrogate outcomes, for example, relative risk of mortality. Statistical methods are defined as in Appendix 3.1.

    Examples for excluded studies are as follows:

    - No statistical adjustment when calculating incremental quantities (e.g. uncontrolled before and after analysis).
    - Statistical method is not used to address selection bias, but for other purposes (e.g. to create predictive equations stratified by risk factors).
4. Study is published in English.

Chapter 4 - Statistical methods for estimating subgroup effects in CEA that use patient-level observational data

4.1 Preamble to research paper 2

The conceptual review (chapter 2) found that an important challenge in CEA that use observational data is to estimate cost-effectiveness for patient subgroups. PS methods, such as PS matching and IPTW can estimate cost-effectiveness for subgroups, but can only provide unbiased estimates if they create balance between the distributions of confounders. The critical appraisal of applied studies (research paper 1) highlighted that CEA rarely assess balance appropriately. GM, a multivariate matching method, uses machine learning to directly balance the distributions of observed confounders, and provides a promising alternative. GM has not been used to estimate cost-effectiveness parameters for subgroups or compared to IPTW before. To help address these gaps in the methodological literature of CEA, research paper 2 compares the relative performance of GM, PS matching and IPTW for estimating cost-effectiveness in patient subgroups.

This paper first considers the methods in a motivating case study of the CEA. The subsequent simulation study is grounded in features of this case study and uses insights from the conceptual review to generate hypotheses (chapter 2). The paper provides guidance for choosing among the statistical methods considered, in order to obtain unbiased, precise estimates of cost-effectiveness by patient subgroup. In order to help the applied researcher, this paper provides sample code for implementing the methods (Appendix 4.2).

London School of Hygiene & Tropical Medicine
Keppel Street, London WC1E 7HT
www.lshtm.ac.uk

Registry
T: +44(0)20 7299 4646
F: +44(0)20 7299 4656
E: registry@lshtm.ac.uk

LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

COVER SHEET FOR EACH 'RESEARCH PAPER' INCLUDED IN A RESEARCH THESIS

Please be aware that one cover sheet must be completed for each 'Research Paper' included in a thesis.

1. **For a 'research paper' already published**

   1.1. Where was the work published? .....

   1.2. When was the work published? .....

   1.2.1. If the work was published prior to registration for your research degree, give a brief rationale for its inclusion

   .....

   .....

   .....

   1.3. Was the work subject to academic peer review? .....

   1.4. Have you retained the copyright for the work? Yes / No
   If yes, please attach evidence of retention.
   If no, or if the work is being included in its published format, please attach evidence of permission from copyright holder (publisher or other author) to include work

2. **For a 'research paper' prepared for publication but not yet published**

   2.1. Where is the work intended to be published? .....

   2.2. Please list the paper's authors in the intended authorship order

   .....

   2.3. Stage of publication – Not yet submitted / Submitted / Undergoing revision from peer reviewers' comments / In press

3. **For multi-authored work, give full details of your role in the research included in the paper and in the preparation of the paper. (Attach a further sheet if necessary)**

   .....

   .....

NAME IN FULL (Block Capitals) .....

STUDENT ID NO: .....

CANDIDATE'S SIGNATURE ..... Date .....

SUPERVISOR/SENIOR AUTHOR'S SIGNATURE (3 above) .....

Additional page for Question (3) on LSHTM cover sheet form:

The research question for this paper was linked to the ESRC project and identified by the principal investigator, RG. I designed the simulation study, with RG. I wrote the simulation code, with help from post-doctoral researchers employed by the ESRC project, R Ramsahai and R Radice. R Ramsahai helped me run simulations on the LSHTM high-performance computational cluster. I assisted ZS on the analysis of the motivating case study. I led on the reporting and interpretation of the results of the case study and the simulation studies, and wrote the first draft of the manuscript. I managed each round of comments and suggestions from the co-authors, in collaboration with RG. All authors read and approved the final draft prior to journal submission and inclusion in the dissertation. I built on insights from another study linked to the ESRC project (Radice et al., 2012). In this study, aimed at a biostatistics audience, I contributed to the design and the implementation of the simulations and to the interpretation of the results, as well as to writing sections of the manuscript.

4.2 Research paper 2 - Methods for estimating subgroup effects in cost-effectiveness analyses that use observational data

Noemi Kreif MSc[1], Richard Grieve PhD[1], Rosalba Radice PhD[1], Zia Sadique PhD[1], Roland Ramsahai PhD[1], Jasjeet S. Sekhon PhD[2]

[1]Department of Health Services Research and Policy, London School of Hygiene and Tropical Medicine, London, UK.

[2]Travers Department of Political Science, UC Berkeley, Berkeley, CA

Status: Published in Medical Decision Making, 2012, 32, 750-63

Contributions: The research question for this paper was linked to the ESRC project and identified by the principal investigator, RG. The candidate designed the simulation study, with RG. The candidate wrote the simulation code, with help from post-doctoral researchers employed by the ESRC project, R Ramsahai and R Radice. R Ramsahai helped the candidate run simulations on the LSHTM high-performance computational cluster. The candidate assisted ZS on the analysis of the motivating case study. The candidate led on the reporting and interpretation of the results of the case study and the simulation studies, and wrote the first draft of the manuscript. She managed each round of comments and suggestions from the co-authors, in collaboration with RG. All authors read and approved the final draft prior to journal submission and inclusion in the dissertation. The candidate built on insights from another study linked to the ESRC project (Radice et al., 2012). In this study, aimed at a biostatistics audience, the candidate contributed to the design and the implementation of the simulations and to the interpretation of the results, as well as to writing sections of the manuscript.

The candidate_____ The supervisor_____

Abstract

Decision makers require cost-effectiveness estimates for patient subgroups. In non-randomized studies, propensity score (PS) matching and inverse probability of treatment weighting (IPTW) can address overt selection bias, but only if they balance observed covariates between treatment groups. Genetic Matching (GM) matches on the PS and individual covariates using an automated search algorithm to directly balance baseline covariates. This paper compares these methods for estimating subgroup effects in cost-effectiveness analyses (CEA).The motivating case study is a CEA of a pharmaceutical intervention, Drotrecogin alfa (DrotAA) for patient subgroups with severe sepsis (n=2,726). Here GM reported better covariate balance than PS matching and IPTW. For the subgroup at a high level of baseline risk, the probability that DrotAA was cost effective ranged from 30% (IPTW) to 90% (PS matching and GM), at a threshold of £20,000 per QALY.

We then compared the methods in a simulation study, where initially the PS was correctly specified, and then misspecified, for example by ignoring the subgroup-specific treatment assignment. Relative performance was assessed as bias and root mean squared error (RMSE) in the estimated incremental net benefits. When the PS was correctly specified and inverse probability weights were stable, each method performed well; IPTW reporting the lowest RMSE. When the subgroup-specific treatment assignment was ignored, PS matching and IPTW reported covariate imbalance and bias; GM reported better balance, less bias, and more precise estimates. We conclude that if the PS is correctly specified and the weights for IPTW are stable, each method can provide unbiased cost-effectiveness estimates. However, unlike IPTW and PS matching, GM is relatively robust to PS misspecification.

Introduction[6]

Health care decision-makers often use cost-effectiveness information for overall populations, when setting priorities (Vanness and Mullahy, 2006). However, focusing on overall mean cost-effectiveness may hide important heterogeneity, and can lead to over (or under) treatment of particular subgroups (Coyle et al., 2003, Sculpher, 2008, Koerkamp et al., 2010). For cost-effectiveness analyses (CEA) to help maximize population health, they are required to provide results for patient subgroups (NICE, 2008). CEA ideally use evidence from pragmatic randomized controlled trials (RCTs) with broad entry criteria; these can provide unbiased estimates of relative cost-effectiveness for policy-relevant subgroups. However, for many decision problems appropriate RCT data are not available, for example because trials have excluded important subgroups, or have tightly-regulated protocols that hinder accurate cost estimation. In such circumstances, the best available data may come from non-randomized studies (NRS), such as prospective cohort studies (Deeks, 2003, Rubin, 2010).

In any NRS, the crucial concern is that treatment assignment is non-random leading to selection bias from confounding. If individual patient data are available for both treatment groups, statistical methods can tackle potential selection bias, but only when their key underlying assumptions are plausible (Rubin, 2010). Instrumental variable approaches can handle confounding and heterogeneity according to both observed and unobserved characteristics (Basu et al., 2007), but in some settings the assumptions required are implausible (Hernán and Robins, 2006). Instead, regression and matching

[6] US spelling conventions are used throughout this paper, due to the target journal criteria.

approaches warrant consideration, provided that the crucial assumption, that all potential confounders have been observed can be justified (Greenland et al., 1999).

Regression methods, recommended for subgroup analysis in CEA of RCTs (Nixon and Thompson, 2005, Willan et al., 2004), are common in CEA that use NRS (Kreif et al., 2012). Here, even if the assumption of no unobserved confounding is justified, cost-effectiveness estimates can be highly sensitive to the specification of the regression model (Grieve et al., 2008). If the model is misspecified, results may suffer from overt bias (Thompson and Nixon, 2005). Instead, propensity score (PS) approaches that aim to balance baseline covariates between treatment groups are advocated for estimating treatment effectiveness (Ho et al., 2007, Stuart, 2010), and cost-effectiveness (Mitra and Indurkhya, 2005, Pullenayegum and Willan, 2011). Austin (2009) has demonstrated that PS matching and inverse probability of treatment weighting (IPTW) can perform relatively well. IPTW has particular appeal for subgroup analysis; if the PS is correctly specified, IPTW can provide more precise estimates of treatment effects than matching (Hirano et al., 2003). A general concern is that these approaches assume the PS is correctly specified (Cole and Hernán, 2008). An alternative method, Genetic Matching (GM), harnesses an automated search algorithm to match on individual covariates as well as the PS. The explicit aim of GM is to balance distributions of observed covariates between the treatment groups. GM can provide some protection against PS misspecification (Sekhon and Grieve, 2011, Diamond and Sekhon, 2012) but has not previously been compared with IPTW.

A major gap in the CEA methods literature is that no previous study has compared alternative methods for subgroup analysis with data from NRS. A recent review of 80 published studies found most CEA that use NRS fail to balance baseline covariates for patient subgroups, potentially leading to biased cost-effectiveness estimates (Kreif et al.,

2012).[7] The aim of this paper is to compare the relative performance of alternative PS approaches for reporting subgroup effects in CEA.

We reanalyze a high profile case study, a CEA of Drotrecogin alfa (DrotAA) for patients with severe sepsis. The effectiveness of DrotAA may differ by subgroup (Bernard et al., 2001, Ely et al., 2003), but it is unclear whether the intervention is cost-effective for either subgroup; we consider this issue using data from a NRS (Rowan et al., 2008). This case study illustrates some general challenges that arise when aiming to report unbiased cost-effectiveness estimates by subgroup from an NRS. Here, statistical methods are required to balance baseline covariates between treatment and control groups within each subgroup. Balancing covariates at the subgroup level may prove particularly challenging if the treatment assignment mechanism differs systematically across subgroups. A PS approach has to then recognize the differential treatment assignment mechanism, for example by estimating separate PS models for each subgroup. We extend a previous study that matched on a single PS estimated across subgroups (Rowan et al., 2008), by estimating a separate PS for each subgroup. In the reanalysis, we employ PS matching, GM, and IPTW to report cost-effectiveness by subgroup.

We report a new Monte Carlo Simulation that builds on the case study by considering circumstances in which the treatment assignment mechanism differs by subgroup. The simulation study incorporates other features of the case study such as nonlinearities in the PS and unstable PS weights. The next section describes the statistical methods and the challenges they face when reporting CEA for different patient subgroups. We

[7] Note that other concerns such as missing data and non-compliance with treatment may also arise, and can lead to biased estimates if not handled appropriately. Methods for handling these issues are beyond the scope of this paper.

describe the case study methods and findings, then the design and results of the Monte Carlo simulations. The last section discusses the findings and outlines areas for further research.

Statistical methods

Each statistical method considered here assumes that there is no unobserved confounding (Greenland et al., 1999). This assumption implies that, conditional on the observed covariates, there are no differences in the distributions of unobserved confounders between treatment groups. As this assumption cannot be directly tested, it is important to draw on external evidence or expert opinion and consider *a priori* which baseline factors are potential confounders (Rubin, 2007). In the context of CEA, the study should carefully consider adjustment for baseline covariates that are potential confounders for either the cost or effectiveness endpoint (Hoch et al., 2002).

Each statistical method then aims to balance the distribution of those potential confounders that are observed. Balance can be achieved by matching (PS matching or GM) or by re-weighting the treatment and control samples (IPTW). Relative performance of weighting and matching methods can be assessed with weighted balance statistics (Stuart, 2010, Austin, 2009a). A recommended balance statistic (Austin, 2009a) is the weighted absolute standardized mean difference, often termed the *weighted standardized difference*.

When treatment effectiveness estimates for subgroups are required, the study should consider whether the treatment assignment mechanism differs by subgroup. For example, the relative influence of factors explaining treatment assignment may differ for high-risk versus low-risk patients. Hence, balancing baseline characteristics for overall samples of treated and control observations can leave potential confounders

imbalanced at the subgroup-level, possibly resulting in biased cost-effectiveness estimates for patient subgroups. An important aim of these methods is to balance baseline covariates in each subgroup of interest. The next sections describe the main distinguishing features of PS matching, GM, and IPTW.

PS matching

The true PS is the conditional probability of treatment assignment given observed baseline covariates (Rosenbaum and Rubin, 1983):

$$p_i = Pr(tx_i = 1|X_i) \qquad i = 1, \ldots, n$$

where tx_i is a binary treatment variable for the ith individual, X_i is a vector of measured baseline confounders and n is the sample size. The true PS is a balancing score: conditional on the PS, treatment and control groups are expected to have the same distribution of observed baseline characteristics. Matching on a correctly specified PS can therefore be expected to eliminate bias (Rubin and Thomas, 1992). However, in NRSs the specification of the true PS is generally unknown; i.e. just as the investigator does not know the specification of the relationship between covariates and endpoints they seldom know how covariates influence treatment receipt.

If the PS is misspecified, for example by disregarding differences between subgroups in the treatment assignment mechanism, then PS matching can lead to biased estimates of treatment effects (Cole and Hernán, 2008). It is unclear which forms of PS misspecification will lead to large biases when reporting subgroup results in CEA.

Methods guidance for estimating the PS suggests two ways of improving the resultant covariate balance. Firstly, the PS should be repeatedly reestimated, with balance reassessed until the analyst finds the best PS, the one that maximizes balance (Stuart, 2010). In this context, the PS is required to maximize balance at the subgroup level.

Second, to improve balance, matching on the PS should be combined with matching on individual covariates (Rosenbaum and Rubin, 1985). However, finding the correct PS specification and the best metric for matching on individual covariates is challenging (Austin, 2008), particularly when covariate balance at the subgroup level is required. Instead, a search algorithm can be used to help improve balance.

Genetic Matching

GM is a multivariate matching approach whose explicit aim is to optimize covariate balance (Diamond and Sekhon, 2012, Ramsahai et al., 2011, Sekhon, 2011, Sekhon and Grieve, 2011). GM extends standard PS matching in two ways. First, rather than the manual process of modifying the PS and balance-checking, GM harnesses an automated search algorithm that iteratively checks balance on observed confounders, and directs the search toward those matches that optimize balance (Diamond and Sekhon, 2012, Sekhon, 2011). Second, the GM algorithm can maximize covariate balance by matching on individual covariates as well as the PS. Hence, at the expense of computational time, the GM search algorithm optimizes covariate balance to the extent possible, given the data (Diamond and Sekhon, 2012, Sekhon, 2011). When the PS is misspecified, Sekhon and Grieve (2011) report that GM can improve covariate balance and reduce bias and variability in the cost-effectiveness estimates.

For subgroup analysis, the GM algorithm can be modified to maximize balance for each subgroup. A challenge GM shares with other multivariate matching estimators is that if required to balance covariates that are not true confounders, this increases the dimensionality of the matching problem. This can lead to loss of precision (Abadie and Imbens, 2006), which can be of particular concern when reporting cost-effectiveness results by subgroup as sample sizes can be relatively small.

The appendices offer further explanation (Appendix 4.1) and code for implementing the method (Appendix 4.2). Full details of the method are provided by Diamond and Sekhon (2012) and Sekhon (2011) .

Inverse probability of treatment weighting

In CEA, IPTW has been introduced for reducing selection bias in cost analysis (Basu et al., 2011) and for handling censored costs (Pullenayegum and Willan, 2011). IPTW has not previously been considered for addressing selection bias in CEA. In this context, IPTW can reweight the treatment and control samples, when estimating cost-effectiveness. The weight w_i is the inverse of the estimated probability of the treatment received, $w_i = \frac{tx_i}{\hat{p}_i} + \frac{1-tx_i}{1-\hat{p}_i}$ where $\hat{p}_i$ is the estimated PS. If the PS is correctly specified, IPTW will provide unbiased estimates of the ATEs and can reach semiparametric efficiency (Hirano et al., 2003). Covariate balance can be assessed following IPTW according to weighted standardized differences (Austin, 2009a), where the weights are the inverse probability weights.

A potential concern is that IPTW can be highly sensitive to PS misspecification; for example when treated observations have a true PS close to zero, even slight discrepancies in the estimated PS translate into large errors in the weights. This can lead to biased and inefficient estimates (Kang and Schafer, 2007). Unstable weights can arise with sparse data and lead to inefficient estimates even if the PS model is correctly specified. As Cole and Hernán (2008) demonstrate, even with good overlap between the treatment and control groups, estimated PS values close to zero can occur by chance. This tends to arise with small sample sizes or if the PS has many continuous covariates. Methods guidance suggest that extreme weights can be progressively truncated (Cole

and Hernán, 2008). The implications of unstable weights and weight truncation have not been reported before when using IPTW to reduce selection bias in CEA.

Issues arising when applying these methods to report cost-effectiveness by subgroup

Each approach can estimate average treatment effects (Polsky and Basu, 2006) - incremental costs (ΔC), incremental effects (ΔE), and incremental net monetary benefits (INBs) - for each prespecified subgroup. We identified 3 particular areas of potential concern when applying these methods to report cost-effectiveness by subgroup.

1. *Subgroup specific treatment assignment.*

When the treatment assignment mechanism differs by subgroup, the methods are required to balance baseline covariates in each subgroup of interest. In the case study we, illustrate how the methods can attempt to balance covariates in each subgroup. The simulation study then considers circumstances where the treatment assignment mechanisms differ by subgroup. The simulation study investigates how the statistical methods perform after incorrectly assuming that there is a single treatment assignment mechanism, rather than recognizing that treatment assignment differs by subgroup. The simulation study reports the relative bias and root mean squared error (RMSE) of the cost-effectiveness estimates across the methods.

2. *Sensitivity of estimates to misspecification of the PS.*

The methodological literature suggests that matching methods can be less sensitive to PS misspecification than IPTW (Lee et al., 2010). Specifically, if the estimated PS weights are unstable, IPTW can report biased and imprecise treatment effects. We consider issues raised by the unstable PS weights found in the motivating case study, when the true PS is not known. In the second simulation scenario, we compare the

relative performance of the methods when PS weights are extreme, for the first time in a CEA context.

3. *Different set of confounders in the cost and effectiveness endpoints.*

In CEA, different covariates can be potential confounders for the estimates of incremental costs versus effectiveness. One approach is to include in the PS all potential confounders for either endpoint. However, previous studies have shown that adjusting for a covariate which influences treatment assignment but not the endpoint can lead to estimates that are statistically inefficient (Brookhart et al., 2006). We designed a simulation scenario to consider the bias and RMSE across the methods of including a covariate in the PS, which is only a confounder for one of the endpoints.

Motivating case study

Overview

We present a CEA in which observational data are used to report cost-effectiveness by subgroup. The case study considers the implications for covariate balance of ignoring, then recognizing the subgroup–specific treatment assignment. We then report cost-effectiveness results for each subgroup. This case study also illustrates circumstances where the inverse PS weights are unstable.

This CEA evaluates DrotAA for patients with severe sepsis admitted to intensive care units (ICUs). The National Institute for Health and Clinical Excellence previously recommended DrotAA for severe sepsis patients with two or more organ systems failing (NICE, 2007), based on a CEA that used a US phase 3 RCT (Bernard et al., 2001). However, this trial, while not powered to detect treatment effects by subgroup, provided some evidence to suggest that the effectiveness of DrotAA may differ across patient

subgroups; the intervention was found to be relatively effective for patients at high levels of baseline risk (Bernard et al., 2001, Ely et al., 2003). Subsequent RCTs that only included low-risk patients were stopped early because of futility (Abraham et al., 2005, Barton et al., 2004) or lack of benefit (Silva et al., 2010). Given the controversy about the effectiveness, but also the infusion's costs (around £5,000 per patient) there is interest in the cost-effectiveness of DrotAA for patients with different risk profiles (NICE, 2007). A CEA for subgroups of patients defined according to baseline risk of death can help address this question. However, such analyses require the use of observational data, hence, the possibility of selection bias must be addressed.

We reanalyzed data from a large UK observational database (Rowan et al., 2008) which represents relevant real-world clinical practice for the subgroups of interest. Following previous analysis of this data (Rowan et al., 2008), we first used a single PS model estimated across all patient subgroups. Rowan et al. (2008) reported treatment effectiveness for subgroups defined according to high (3 to 5 organ failures at baseline) and low (2 organ failures) levels of baseline risk, and found that DrotAA reduced hospital mortality for high risk patients, but increased mortality for low risk patients. We extended this analysis by deploying the alternative statistical methods described to try and maximize covariate balance for each subgroup.

Data

We used the same data as the previous prospective cohort study (Rowan et al., 2008), from the UK Case-Mix Programme dataset co-ordinated by the Intensive Care National Audit and Research Centre (ICNARC). Of patients who met the study's inclusion criteria and were defined as having severe sepsis and multiple organ failures at

admission, 1,076 received DrotAA (treated) and 1,650 contemporaneous admissions did not receive DrotAA (controls).

The CEA estimated individual-level lifetime QALYs, based on individual patient's mortality data collected for a follow-up period of four years and, for those who survived, age- and gender-specific expected survival and quality of life. Costs of DrotAA and all hospitalizations were estimated at the patient level, over the same period of follow-up. QALYs and costs were discounted at the recommended rate of 3.5% (NICE, 2008). The subsequent statistical analysis used the individual level data on costs and lifetime QALYs. Further details on the CEA, including data sources are reported elsewhere (Sadique et al., 2011).

Statistical analysis of the case study

We extended the previous PS matching (Rowan et al., 2008) in creating subgroup-specific PS models, but also by considering GM and IPTW. The PS models were estimated by logistic regression and included the same potential confounders as the previous study. The baseline characteristics included were hospital type, number of critical care beds in the ICU, age, ICNARC model physiology score (IMscore), gender, number of organ systems failing, type of organ failures (cardiovascular, respiratory, renal, haematological, metabolic acidosis), source of admission to critical care (via the emergency department, theatre or recovery, ward, clinic or home), diagnostic category, and serious conditions in the past medical history. Age and IMscore were defined as nonlinear terms, fitted as smoothed functions using restricted cubic splines; other continuous measures were assumed to have a linear relationship with the logit of treatment assignment.

Each statistical method first aimed to maximize covariate balance in the overall sample by using a single PS model estimated across all subgroups (see example code for implementation in Appendix 4.1). The GM algorithm was required to minimize standardized differences for the overall sample. PS models were then estimated for each subgroup (2 or 3 to 5 organ failures) and GM was required to optimize balance for each subgroup. In a sensitivity analysis, GM optimized balance according to paired t-tests and Kolmogorov-Smirnoff tests (Sekhon, 2011).

We report lifetime incremental costs, QALYs and INBs of DrotAA versus control. Statistical uncertainty was considered by reporting 95% confidence intervals (CIs)(Davison and Hinkley, 1997), and cost-effectiveness acceptability curves (CEACs) using the nonparametric bootstrap to maintain the correlation between costs and QALYs (Fenwick et al., 2004). The resultant inferences should be regarded as conditional on the estimated PSs and the matched data (Hill and Reiter, 2005). For all statistical analyses, the R platform was used.

Case study results

Covariate balance

Before matching, the treatment groups were highly imbalanced; compared with controls, the DrotAA patients were on average younger, with a higher baseline probability of death (Table 4.1).

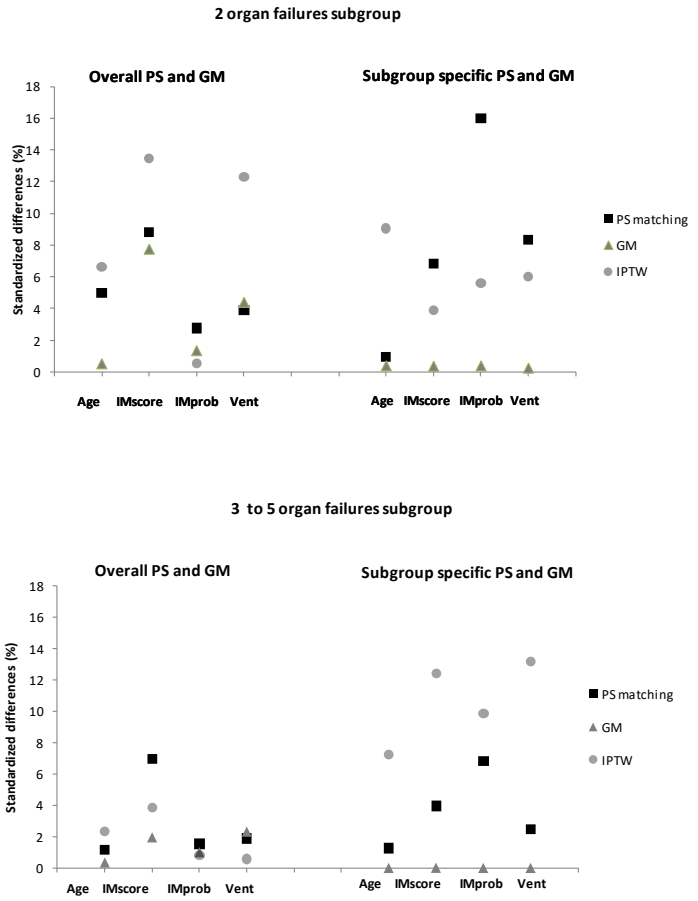
Table 4.1 - Case study results: baseline characteristics and covariate balance for DrotAA versus Control group, before matching or weighting.

Covariate	2 organ failures subgroup			3 to 5 organ failures subgroup		
	DrotAA (n=198)	Control (n=630)	Standardized difference (%)	DrotAA (n=878)	Control (n=1020)	Standardized difference (%)
Age	57.58	63.04	26.49	58.96	65.16	32.32
IMprob *	0.42	0.39	10.76	0.64	0.58	20.12
IMscore†	22.83	20.44	29.53	32.08	27.96	40.83
% vent. ‡	88.38	70.16	40.02	93.39	78.53	38.90

Abbreviations: *: ICNARC model predicted probability of acute hospital mortality † ICNARC model physiology score ‡ % of patients mechanically ventilated

Following PS matching and IPTW, standardized differences (%) remained large for both subgroups, both with the overall and the subgroup-specific PSs. GM reported better balance than the other methods, even when required to maximize balance across the overall sample. The lowest standardized differences were reported when GM was required to optimize balance for each subgroup (Figure 4.1). Subsequent results are reported just for the subgroup-specific PS models and GM algorithms.

Figure 4.1 - Case study: covariate balance reported as weighted standardized differences (%), after PS matching, GM and IPTW, for overall and subgroup specific PSs and GM algorithms



Abbreviations: IMprob - ICNARC model predicted probability of acute hospital mortality, IMscore - ICNARC model physiology score, vent - % of patients mechanically ventilated

Lifetime cost-effectiveness results

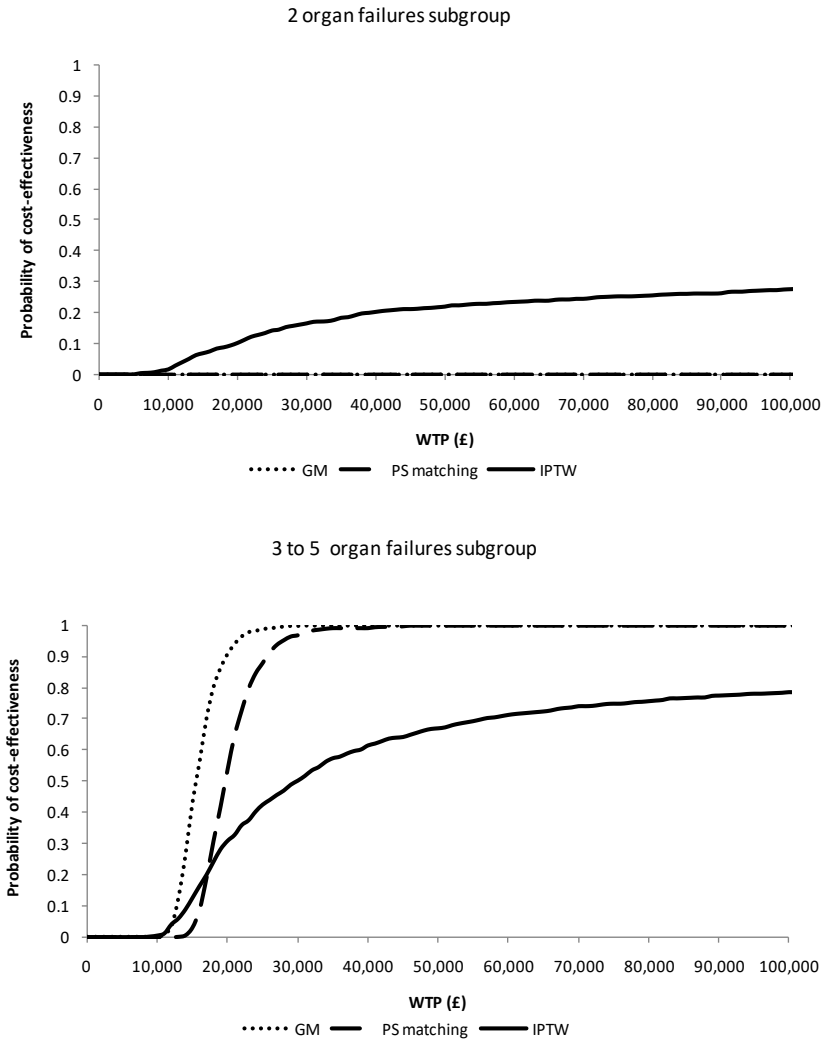
For patients with two organs failing, the INBs were all negative, but following IPTW, the 95% CI was relatively wide and included zero (Table 4.2). For the subgroup with 3 to 5 organs failing, the incremental QALYs were positive and relatively high for both matching methods. For IPTW, the QALY gain was smaller with 95% CI that included zero. The INB was lower following IPTW than for the matching methods, and again had a wide CI.

Table 4.2 - Case study: Lifetime incremental costs (£), QALYs and INBs (WTP=£20,000) for DrotAA versus Control group. Estimates are from subgroup specific PSs and GM algorithms.

	PS matching	GM	IPTW
2 organ failures subgroup			
Incremental costs (95%CI*)	12,710 (11,058 to 14,361)	14,703 (12,763 to 16,644)	13,750 (9,873 to 17,597)
Incremental QALYs (95% CI*)	-1.01 (-1.60 to -0.41)	-0.97 (-1.62 to -0.32)	-0.30 (-1.77 to 1.15)
INBs (95% CI*)	**-32,846 (-44,704 to -20,987)**	**-34,031 (-47,028 to -21,034)**	**-19,764 (-49,546 to 9,835)**
3-5 organ failures subgroup			
Incremental costs (95% CI*)	19,384 (17,696 to 21,071)	19,948 (17,610 to 22,286)	19,023 (15,636 to 22,102)
Incremental QALYs (95% CI*)	0.98 (0.65 to 1.33)	1.28 (0.86 to 1.70)	0.542 (-0.66 to 1.55)
INBs (95% CI*)	**391 (-6,350 to 7,133)**	**5,690 (-2,543 to 13,924)**	**-8,175 (-31,787 to 11,845)**

Notes: * Bootstrapped confidence intervals, conditional on the estimated PS and matched data.

Figure 4.2 - Case study: Cost-effectiveness acceptability curves for DrotAA versus Control group using subgroup specific PSs and GM algorithms



Notes: For the 2 organ failures subgroup, CEACs for GM and PS matching are indistinguishable

Figure 4.2 presents CEACs which suggest that DrotAA is not cost-effective for the 2 organ failures subgroup. The CEACs for the 3 to 5 organ failures subgroup differ somewhat by method; at realistic levels of WTP for a QALY gain in the UK (£20,000 to

£30,000) the probability that DrotAA is cost-effective is 30% following IPTW versus 90% for the other methods.

When the extreme inverse probability weights were truncated (Appendix 4.3, Figure 1), we found that covariate balance worsened (Appendix 4.3, Table 1), and the 95% CIs around the INBs were only slightly reduced (Appendix 4.3, Figure 2). When the GM algorithm was required to optimize alternative balance statistics (Kolmogorov-Smirnoff tests and paired t-tests), the results were similar to the base case.

Monte Carlo simulation study

Overview

Monte Carlo simulations were conducted to examine the relative performance of the methods, for estimating cost-effectiveness in prespecified subgroups. The study design extended previous simulations comparing PS matching with IPTW (Austin, 2009b) or GM (Sekhon and Grieve, 2011), to recognize the specific challenges that arise when reporting cost-effectiveness by subgroup. In particular, motivated by the case study, the treatment assignment mechanism was assumed to differ by subgroup. Cost and effectiveness data were simulated to recognize heterogeneity, and it was assumed that cost-effectiveness estimates were required by subgroup. The case study also illustrated that the weights for IPTW can be unstable. Here, we investigate the implications of such unstable weights by including a nonlinear term in the PS. We also consider an issue specific to CEA, which concerns the choice of covariates when attempting to address selection bias for both cost and effectiveness endpoints. The simulation study reported covariate balance, bias and RMSE of the estimated cost-effectiveness across the methods.

Description of scenarios

In the 3 scenarios, each estimation method was initially assumed to follow the true treatment assignment mechanism, and then a misspecified treatment assignment as described below. In the first scenario, each approach assumed that the same treatment assignment mechanism applied for both subgroups, but in fact it differed by subgroup. The second scenario recognized that treatment assignment was subgroup-specific, but each estimation approach was misspecified by excluding a nonlinear term. This scenario also considered the hypothesis that IPTW may provide inefficient estimates when, as in the case study, the estimated inverse probability weights are unstable (Cole and Hernán, 2008).

The third scenario considered the challenge of choosing the correct set of covariates when attempting to address selection bias for both cost and effectiveness endpoints. Here we build on a previous simulation (Brookhart et al., 2006) by introducing an additional covariate which is not a confounder for the cost endpoint, but does influence the effectiveness endpoint and the treatment assignment. We anticipate that conditioning on this variable will reduce bias in the estimated effectiveness, but will lead to more uncertainty in the estimation of incremental costs. The simulation scenarios are summarized in Table 4.3.

Data generating process

We simulated an observational dataset, extending previous data generating processes (DGPs) (Austin, 2009c, Austin, 2009b) to the context where cost-effectiveness estimates by subgroup are required. The main features of the DGP were grounded in the case study. In particular, there were two prespecified subgroups, heterogeneous treatment effects, a treatment assignment mechanism that differed by subgroup,

nonlinearities in the PS and unstable PS weights. For each subject, 3 confounders, 2 continuous (X_1 and X_2) and 1 binary (X_3), were generated from a bivariate normal and a Bernoulli distribution, respectively:

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim N\left\{ \begin{pmatrix} 2 \\ 4 \end{pmatrix}, \begin{pmatrix} 1 & 0.2 \\ 0.2 & 1 \end{pmatrix} \right\},$$

$$X_3 \sim Bern \begin{cases} 0.6 \text{ for } X_1 > 2 \\ 0.4 \text{ for } X_1 \leq 2 \end{cases},$$

where for example X_1 has a mean of 2, standard deviation of 1, and the covariance between X_1 and X_2 is 0.2. The variable X_3 defines the prespecified patient subgroup (subgroup 1 for X_3=0 and subgroup 2 for X_3=1). The treatment indicator, tx, was randomly generated from a Bernoulli distribution with parameter p, the PS, determined by a different logistic model for each pair of scenarios.

To reflect a typical CEA, costs and outcomes (QALYs) were drawn from a bivariate normal-gamma distribution, using a copula function (Trivedi and Zimmer, 2005, Mihaylova et al., 2010, Quinn, 2007), with the correlation coefficient equal to 0.4.[8] QALYs were drawn from a normal distribution

$$QALY \sim N(\mu_{QALY}, 0.2),$$

and costs from a gamma distribution with identity link, and shape and scale parameters defined as

$$Cost \sim \Gamma(10,\ \mu_{cost}/10).$$

The mean costs and QALYs, specific to each scenario are given below.

[8] The copula function can generate draws from a flexible multivariate distribution (in this case the bivariate) with different marginal distributions (here, the normal and the gamma).

Scenario 1

The true PS allowed the confounders to have a differential effect on treatment assignment according to subgroup, by including interaction terms in determining the logit of the PS as:

$$logit(p) = \ln(0.2) + 0.1X_1 + 0.2X_2 + 0.3X_3 + 0.2X_1X_3 - 0.2X_2X_3$$

The linear predictors for the cost and QALY endpoints also included interactions between X_3 and tx to ensure heterogeneity in the incremental costs and QALYs:

$$\mu_{cost} = 4000 + 4000tx + 5000X_1 + 4000X_2 + 3000X_3 - 1000X_3tx,$$

$$\mu_{QALY} = 9 + 0.25\text{tx} - 1X_1 - 0.6X_2 - 0.8X_3 + 0.5X_3tx.$$

In scenario 1a (correct specification), subgroup-specific PSs were used for matching and weighting. Similarly, for each subgroup, GM was required to match on and balance X_1, X_2 and the linear predictor of the estimated PS. In Scenario 1b, the PS and the GM algorithm were both misspecified; the PS was estimated for the overall sample (including X_1, X_2 and X_3 in the logistic regression). The GM was required to match on, and maximize balance for X_1, X_2 and X_3 across the overall sample.

Scenario 2

In Scenario 2 the term ${X_1}^2$ was added to the true PS model. The coefficient for the ${X_1}^2$ term was set to create unstable inverse probability weights for subgroup 2:

$$logit(p) = ln(0.2) + 0.1X_1 + 0.2X_2 + 0.3X_3 + 0.2{X_1}^2X_3.$$

In scenario 2a, we assumed correctly specified PS models: X_1, X_2 and ${X_1}^2$ were all included in separate logistic regression models for each subgroup. For each subgroup,

GM was required to maximize balance on each term contributing to the true PSs, including ${X_1}^2$. In scenario 2b, separate PS models and GM algorithms were specified as before. However, for subgroup 2, ${X_1}^2$ was excluded from the PSs and from the terms GM was required to match on and balance.

Scenario 3

This scenario extended scenario 1a by introducing a new continuous variable in the assignment model, $X_4 \sim N(3,1)$, which was a confounder for the QALY but not for the cost endpoint. The logit of the PS model and the linear predictor of the QALY were defined as:

$$logit(p) = \ln(0.2) + 0.1X_1 + 0.2X_2 + 0.3X_3 + 0.2X_1X_3 - 0.2X_2X_3 + 0.4X_4 - 0.2X_4X_3$$

$$\mu_{QALY} = 9 + 0.5tx - 1X_1 - 0.2X_2 - 0.8X_3 + 0.25X_3tx + 0.1X_4$$

In scenario 3a, the subgroup-specific PS models and the GM algorithms included X_4, while in scenario 3b, this covariate was excluded.

Table 4.3 - Monte Carlo simulations: summary of scenarios

	PS matching, IPTW	GM
Scenario 1 : Subgroup-specific treatment assignment		
1a Correct specification	Subgroup-specific PS	Subgroup-specific PS and GM algorithm
1b Misspecification	Overall PS	Overall PS and GM algorithm
Scenario 2: Nonlinear term in PS		
2a Correct specification	Nonlinear term included in PS	Nonlinear term included in PS and GM algorithm
2b Misspecification	Nonlinear term excluded from PS	Nonlinear term excluded from PS and GM algorithm
Scenario 3: Confounder for QALY		
3a Correct specification	Confounder included in PS	Confounder included in PS and GM algorithm
3b Misspecification	Confounder excluded from PS	Confounder excluded from PS and GM algorithm

Implementation

One thousand datasets of sample size 2000 were simulated. Both PS matching and GM matched one-to-one to the nearest neighbor, with replacement. GM was required to maximize balance according to weighted standardized differences. In scenario 2a, as a sensitivity analysis, IPTW results were also reported after progressively truncating the extreme PS weights. INBs were calculated at a societal WTP of £20,000 per QALY gained. For scenarios 1 and 2, the true INBs were set to £1,000 (subgroup 1) and £12,000 (subgroup 2), and for scenario 3 to £6,000 and £12,000. The methods were compared by calculating weighted standardized differences (reported as percentage), relative bias, and RMSE for the estimated incremental costs, QALYs and INBs. Appendix 4.2 provides sample R code for the data generating processes and for implementing each method.

Results of the Monte Carlo simulations

Covariate balance

Table 4.4 reports the weighted standardized differences averaged over 1000 replications. When each approach recognized the true treatment allocation mechanism (scenarios 1a, 2a, 3a), all the standardized differences were small. When the PS model was misspecified by fitting an overall PS, and the GM algorithm failed to match and balance at the subgroup level (scenario 1b), both PS matching and IPTW reported high standardized differences, whereas following GM, covariates were balanced. In scenario 2b, when the PS model for subgroup 2 excluded the nonlinear term $X_1{}^2$, the standardized differences for $X_1{}^2$ were 14% (IPTW), 4% (PS matching) and 3% (GM). In scenario 3b, when each method ignored X_4, the standardized differences for this variable were high.

Table 4.4 - Monte Carlo simulations: covariate balance reported as weighted standardized differences (%).

		Subgroup 1			Subgroup 2		
Scenario	**Method**	**X1**	**X2**	**X4**	**X1**	**X2**	**X4**
1a	PS matching	2.58	1.35		0.62	3.08	
	GM	0.10	0.12		0.19	0.15	
	IPTW	0.50	0.56		0.80	0.59	
1b	PS matching	7.91	8.54		8.36	8.55	
	GM	1.69	1.88		1.82	1.82	
	IPTW	8.09	8.12		8.51	8.12	
2a	PS matching	2.50	1.68		1.78	4.88	
	GM	0.20	0.24		1.43	1.87	
	IPTW	0.44	0.58		5.35	3.49	
2b	PS matching	2.58	1.35		2.21	4.97	
	GM	0.10	0.12		1.57	1.01	
	IPTW	0.50	0.56		5.09	2.06	
3a	PS matching	3.17	2.79	1.70	1.76	2.89	2.49
	GM	0.32	0.30	0.38	0.20	0.18	0.18
	IPTW	0.84	0.89	1.11	0.41	0.28	0.33
3b	PS matching	2.40	1.31	39.82	0.57	2.83	20.08
	GM	0.09	0.11	39.74	0.14	0.10	20.07
	IPTW	0.44	0.47	39.62	0.31	0.20	19.98

Bias and RMSE

Table 4.5 reports the relative bias and RMSE in the estimated INB, over 1000 replications for each scenario. The corresponding results for the incremental costs and QALYs are reported in Appendix 4.3. In scenario 1a, when the subgroup-specific treatment allocation was recognized, bias was low following each method, and IPTW reported the lowest RMSE for both subgroups. When the subgroup-specific assignment mechanism was ignored (scenario 1b), bias and RMSE were higher following PS matching and IPTW than for GM.

In scenario 2a, when the nonlinear term was correctly included in the PS for subgroup 2, biases were low but IPTW reported RMSE twice that of the matching methods. Insights as to why the precision for IPTW is worse can be gained from plotting the weights in a large sample (n=1,000,000), simulated by the same DGP. Visual inspection suggests these weights are highly variable for the controls in subgroup 2 (Appendix 4.3, Figure 3). When in the sensitivity analysis, weights are progressively truncated, the problem is not resolved; while the IPTW estimator for scenario 2a is less variable, bias increases (Appendix 4.3, Figure 4). In Scenario 2b after omitting the nonlinear term from the PS for subgroup 2, IPTW reported the highest bias and RMSE.

In scenario 3a, when each approach balanced all confounders including X_4, IPTW reported the lowest RMSE. In scenario 3b, the failure to balance X_4 resulted in biased estimates of the incremental QALY and the INB for all methods, with IPTW reporting the lowest RMSE.

Table 4.5 - Monte Carlo simulations: relative bias and RMSE for the INBs (WTP=£20,000)

Scenario	Method	Subgroup 1		Subgroup 2	
		Relative Bias (%)	RMSE	Relative Bias (%)	RMSE
1a					
	PS matching	4.4	961	0.6	1068
	GM	2.9	756	0.8	825
	IPW	1.5	675	0.3	782
1b					
	PS matching	58.0	1988	5.7	2031
	GM	11.6	1060	1.6	1088
	IPTW	71.3	1802	6.6	1875
2a					
	PS matching	5.0	989	3.2	1484
	GM	4.2	744	4.4	1267
	IPTW	2.2	676	2.2	2535
2b					
	PS matching	4.1	963	0.9	1306
	GM	2.4	756	3.7	1210
	IPTW	2.0	673	10.3	1839
3a					
	PS matching	0.6	1271	0.7	1043
	GM	0.7	699	0.3	784
	IPTW	0.2	686	0.0	689
3b					
	PS matching	12.5	1229	3.0	927
	GM	12.4	1036	3.0	885
	IPTW	12.7	999	3.3	802

Notes: For scenarios 1 and 2 the true INBs are £1,000 (subgroup 1) and £12,000 (subgroup 2), and the corresponding INBs for scenario 3 are £6,000 and £12,000.

Discussion

This paper compares alternative statistical methods for reducing selection bias when cost-effectiveness results are required for patient subgroups. The Monte Carlo simulation finds that if the treatment assignment mechanism ignores differential treatment allocation by subgroup, then cost-effectiveness estimates can be biased and inefficient. GM appears relatively robust to this misspecification, because it aims to balance confounders directly using an automated search algorithm. This is also highlighted in the case study, where GM achieves better balance than the other methods even if required to maximize balance across the overall treatment and control groups.

This paper extends the work of Sekhon and Grieve (2011), who showed that GM can create good balance and reduce bias in CEA, even if the PS model is misspecified. Our article considers the important context of subgroup analysis and includes IPTW as a comparator. We find that IPTW provides unbiased, precise cost-effectiveness results for subgroups, if the PS is correctly specified and the weights are stable. However, IPTW is sensitive to extreme probability weights (Kang and Schafer, 2007). In the case study, we find that IPTW has unstable weights and reports INBs with wider CIs than the matching approaches, and is anticipated to provide divergent estimates of the expected value of further information (Fenwick et al., 2004). In the simulation scenario with unstable weights (due to nonlinearity in the PS model), IPTW reports high RMSE compared to matching. Truncating the weights (Cole and Hernán, 2008) improves precision, but increases imbalance and bias.

Our work considers 2 distinct examples of PS misspecification: first, when the estimated PS disregards differences in the treatment assignment between subgroups, and second, when a nonlinear term is omitted. The simulations demonstrate that IPTW is more sensitive to either misspecification than the matching methods. In the case

study, IPTW reported poor balance and, for the high risk subgroup, divergent point estimates, compared to either matching approach. GM reported good balance for both subgroups and PS approaches, and hence a relatively sound basis for the ensuing cost-effectiveness estimates.

This article also contributes to the general methodological literature by considering the methods in a bivariate context. In CEA, potential confounders can differ between the cost and effectiveness endpoints; for example, baseline health-related quality of life might be associated with the QALY but not the cost endpoint. The simulations highlight that balance should be maximized on potential confounders for either endpoint. When a baseline covariate that influences just the QALY is left unbalanced, the estimates of the QALY gain are biased, and the only advantage is a slight improvement in the precision of the cost estimate. These findings extend previous univariate analyses showing that including covariates not associated with outcome reduces precision (Brookhart et al., 2006).

Our paper considers circumstances when subgroups are prespecified, informed by prior reasoning from the previous literature (Bernard et al., 2001, Ely et al., 2003). In other circumstances there may be insufficient information to predefine the subgroups of interest. Here, the optimal number and definition of subgroups could be established as part of the CEA, based on expected health benefits (Espinoza et al., 2011). Alternatively, to report subgroup-specific treatment effects, regression analysis with treatment by covariate interactions (Nixon and Thompson, 2005) could be applied to the matched or weighted data. The general requirement to choose an approach that minimizes selection bias is still of paramount importance, and can be informed by the results presented here.

This paper focuses on CEA that use patient-level data from a NRS. For many policy questions input parameters in decision analytical models are taken from studies that use either patient-level or aggregated estimates from NRS (Briggs et al., 2006, NICE, 2008, Kreif et al., 2012, Briggs et al., 2004). In this more general context to provide cost-effectiveness results by subgroup, heterogeneity may need to be recognized for a range of model input parameters (e.g. rates of adverse events, estimates of health-related quality of life and transition probabilities) (Sculpher, 2008, Koerkamp et al., 2010). The potential for selection bias must be recognized when estimating subgroup-specific input parameters from NRS whether using patient-level data or extracting aggregate input parameters from the literature. By highlighting the selection biases that can arise, this study provides important insights both for those doing CEA using patient-level data without a decision model and for those developing and interpreting decision models that report cost-effectiveness estimates by patient subgroup.

The major limitation of the methods described is that they all rely on the assumption of no unobserved confounding (Greenland et al., 1999). The case study followed recommendations by identifying potential confounders *a priori* (Rubin, 2007) and a rich set of measured confounders were selected for adjustment, based on previous literature (Rowan et al., 2008) and clinical expert opinion. As in any observational study, unmeasured confounding can still be present. In our case study, the potential for hidden bias is greatest in the 2 organ failures subgroup. Here, approximately one-third of DrotAA cases were treated with delay, possibly leading to unobserved differences in baseline severity between the comparison groups. However, as the simulations highlight, omitting a confounder can lead to similar levels of hidden bias for each method. Hence, unmeasured confounding is unlikely to drive any differences in cost-effectiveness results across the methods considered. Unobserved confounding can be

addressed with instrumental variables (Terza et al., 2008, Grootendorst, 2007) or control functions (Polsky and Basu, 2006). These approaches can potentially accommodate unobserved heterogeneity, and identify those patients who can make the largest health gains from treatment (Basu, 2011, Basu et al., 2007, Basu, 2009).

This article raises several areas for further research. The performance of the methods presented here can be improved by exploiting information on the data-generating process for the cost and effectiveness endpoints (Imbens and Wooldridge, 2009a). Regression models can then be applied to the matched data to adjust for any remaining imbalances in observed characteristics between the treatment groups (Abadie and Imbens, 2011). Regression post matching can be relatively insensitive to the choice of model specification (Ho et al., 2007). Doubly robust methods (Robins et al., 1995, Kang and Schafer, 2007, Robins et al., 2007, van der Laan and Gruber, 2010) can be deployed that use the estimated PS as weights (Hirano and Imbens, 2001) or adjustment terms (Glynn and Quinn, 2010) in the endpoint models.

We conclude that the key criterion for choosing amongst the proposed statistical methods is the level of covariate balance for each subgroup. IPTW can provide unbiased, precise cost-effectiveness estimates for patient subgroups, but only if the PS is correctly specified, and the PS weights are stable. If the inverse probability weights are unstable, IPTW estimates can be biased and imprecise. In most CEA that use observational data, the treatment assignment mechanism is unknown, and GM, which is an example of an automated approach, is relatively robust to PS misspecification. GM is publicly available in standard software packages (Sekhon, 2011, Hartman and Sekhon, 2011), and should be considered by future CEA that use NRS to report cost-effectiveness by patient subgroup.

Acknowledgements

We thank James Carpenter and Rhian Daniel (LSHTM) for valuable comments on the Monte Carlo simulations. We also thank David Harrison and Kathy Rowan (ICNARC) for access to the data used in the motivating case study. This work was funded by the Economic and Social Research Council (Grant no. RES-061-25-0343).

References

Abadie, A. & Imbens, G. W. 2006. Large Sample Properties of Matching Estimators for Average Treatment Effects. *Econometrica,* 74**,** 235-267.

Abadie, A. & Imbens, G. W. 2011. Bias-Corrected Matching Estimators for Average Treatment Effects. *Journal of Business & Economic Statistics,* 29**,** 1-11.

Abraham, E., Laterre, P., Garg, R. & al., e. 2005. Administration of Drotrecogin Alfa (Activated) in Early Stage Severe Sepsis (ADDRESS) Study Group: Drotrecogin alfa (activated) for adults with severe sepsis and a low risk of death. *New Engl J Med* 353**,** 1332–1341.

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

Austin, P. C. 2009a. Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. *Statistics in Medicine,* 28 3083-3107.

Austin, P. C. 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**,** 661-677.

Austin, P. C. 2009c. Type I Error Rates, Coverage of Confidence Intervals, and Variance Estimation in Propensity-Score Matched Analyses. *The International Journal of Biostatistics.,* 5.1.

Barton, P., Kalil, A., Nadel, S. & al, e. 2004. Safety, pharmacokinetics, and pharmacodynamics of drotrecogin alfa (activated) in children with severe sepsis. *Pediatrics,* 113**,** 7-17.

Basu, A. 2009. Individualization at the heart of comparative effectiveness research: The time for i-CER has come. *Medical Decision Making,* 29**,** N9-N11.

Basu, A. 2011. Economics of individualization in comparative effectiveness research and a basis for a patient-centered health care. *Journal of Health Economics,* 30**,** 549-559.

Basu, A., Heckman, J. J., Navarro-Lozano, S. & 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**,** 1133-1157.

Basu, A., Polsky, D. & Manning, W. 2011. Estimating treatment effects on healthcare costs under exogeneity: is there a 'magic bullet'? *Health Services and Outcomes Research Methodology,* 11**,** 1-26.

Bernard, G. R., Vincent, J.-L. L., P.-F. & Group, R. H. A. P. C. W. E. i. S. S. P. S. 2001. Efficacy and safety of recombinant human activated protein C for severe sepsis. *New Engl J Med***,** 699-709.

Briggs, A., Sculpher, M., Dawson, J., Fitzpatrick, R., Murray, D. & Malchau, H. 2004. The Use of Probabilistic Decision Models in Technology Assessment: The Case of Total Hip Replacement. *Applied Health Economics and Health Policy,* 3**,** 79-89.

Briggs, A., Sculpher, M. & Klaxton, K. (eds.) 2006. *Decision Modelling for Health Economic Evaluation*: Oxford University Press.

Brookhart, M. A., Schneeweiss, S., Rothman, K. J., Glynn, R. J., Avorn, J. & Sturmer, T. 2006. Variable selection for propensity score models. *Am J Epidemiol,* 163**,** 1149-56.

Cole, S. R. & Hernán, M. A. 2008. Constructing inverse probability weights for marginal structural models. *Am J Epidemiol,* 168**,** 656-64.

Coyle, D., Buxton, M. J. & O'Brien, B. J. 2003. Stratified cost-effectiveness analysis: a framework for establishing efficient limited use criteria. *Health Economics,* 12**,** 421-427.

Davison, A. & Hinkley, D. 1997. *Bootstrap Methods and Their Application,* New York, Cambridge University Press.

Deeks, J. J. 2003. *Evaluating non-randomised intervention studies* [Online]. Tunbridge Wells: published by Gray Pub. on behalf of NCCHTA. Available: http://www.hta.ac.uk/execsumm/summ727.htm [Accessed 23/05/2009.

Diamond, A. & Sekhon, J. S. 2012. Genetic Matching for Estimating Causal Effects: A General Multivariate Matching Method for Achieving Balance in Observational Studies. *Review of Economic and Statistics,* Forthcoming.
Ely, E., Laterre, P., Angus, D., Helterbrand, J., Levy, H., Dhainaut, J., JL, V., WL, M., GR, B. & Investigators., P. 2003. Drotrecogin alfa (activated) administration across clinically important subgroups of patients with severe sepsis. *Crit Care Med,* 31**,** 12-9.
Espinoza, M. A., Manca, A., Claxton, K. & Sculpher, M. J. 2011. The value of identifying heterogeneity: a framework for subgroup cost-effectiveness analysis. *HESG Conference presentation.* York.
Fenwick, E., O'Brien, B. & Briggs, A. 2004. Cost-effectiveness acceptability curves--facts, fallacies and frequently asked questions. *Health Economics,* 13**,** 405-415.
Glynn, A. N. & Quinn, K. M. 2010. An Introduction to the Augmented Inverse Propensity Weighted Estimator. *Political Analysis,* 18**,** 36-56.
Greenland, S., Pearl, J. & Robins, J. M. 1999. Confounding and Collapsibility in Causal Inference. *Statist. Sci. ,* 14**,** 29-46.
Grieve, R., Sekhon, J. S., Hu, T.-w. & 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**,** 1204-1222.
Grootendorst, P. 2007. A review of instrumental variables estimation in the applied health sciences. *Health Services and Outcomes Research Methodology* 7**,** 159-179.
Hartman, E. & Sekhon, J. S. 2011. *Matching: Stata version of Multivariate and Propensity Score Matching Software for Causal Inference.* [Online]. Available: http://ekhartman.berkeley.edu/stata/matching.html
Hernán, M. A. & Robins, J. M. 2006. Instruments for Causal Inference: An Epidemiologist's Dream? *Epidemiology,* 17**,** 360-372.
Hill, J. & Reiter, J. P. 2005. Interval estimation for treatment effects using propensity score matching. *Statistics in Medicine,* 25.
Hirano, K. & Imbens, G. W. 2001. Estimation of Causal Effects using Propensity Score Weighting: An Application to Data on Right Heart Catheterization. *Health Services and Outcomes Research Methodology,* 2**,** 259-278.
Hirano, K., Imbens, G. W. & Ridder, G. 2003. Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score. *Econometrica,* 71**,** 1161-1189.
Ho, D. E., Imai, K., King, G. & Stuart, E. A. 2007. Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference *Political Analysis,* 15**,** 199-236.
Hoch, J. S., Briggs, A. H. & Willan, A. R. 2002. Something old, something new, something borrowed, something blue: a framework for the marriage of health econometrics and cost-effectiveness analysis. *Health Economics,* 11**,** 415-430.
Imbens, G. M. & Wooldridge, J. M. 2009a. Recent Developments in the Econometrics of Program Evaluation. *Journal of Economic Literature,* 47**,** 5–86.
Kang, J. D. Y. & Schafer, J. L. 2007. Demystifying Double Robustness: A Comparison of Alternative Strategies for Estimating a Population Mean from Incomplete Data. *Statistical Science,* 22**,** 523-539.
Koerkamp, G. B., Weinstein, M. C., Stijnen, T., Heijenbrok-Kal, M. H. & Hunink, M. G. M. 2010. Uncertainty and Patient Heterogeneity in Medical Decision Models. *Medical Decision Making,* 30**,** 194-205.
Kreif, N., Grieve, R. & Sadique, Z. 2012. Statistical methods for cost-effectiveness analyses that use observational data: a critical appraisal tool and review of current practice. *Health Economics,* DOI: 10.1002/hec.2806.
Lee, B. K., Lessler, J. & Stuart, E. A. 2010. Improving propensity score weighting using machine learning. *Statistics in Medicine,* 29**,** 337-346.
Mebane, W. R. J. & Sekhon, J. S. 2011. Genetic optimization using derivatives: The rgenoud package for R. *Journal of Statistical Software,* 2011**,** 1-26.

Mihaylova, B., Briggs, A., O'Hagan, A. & Thompson, S. 2010. Review of statistical methods for analysing healthcare resources and costs. *Health Economics*, DOI: 10.1002/hec.1653.
Mitra, N. & Indurkhya, A. 2005. A propensity score approach to estimating the cost-effectiveness of medical therapies from observational data. *Health Economics,* 14**,** 805-15.
NICE. 2007. *Technology Appraisal Guidance 84. Drotrecogin alfa (activated) for severe sepsis* [Online]. Available: http://guidance.nice.org.uk/TA84 [Accessed 05/08/2011.
NICE. 2008. *Guide to the Methods of Technology Appraisal* [Online]. Available: http://www.nice.org.uk/media/B52/A7/TAMethodsGuideUpdatedJune2008.pdf [Accessed 24/10/2010.
Nixon, R. M. & Thompson, S. G. 2005. Methods for incorporating covariate adjustment, subgroup analysis and between-centre differences into cost-effectiveness evaluations. *Health Economics,* 14**,** 1217-1229.
Pearl, J. 2001. Causal Inference in the Health Sciences: A Conceptual Introduction. *Health Services and Outcomes Research Methodology,* 2**,** 189-220.
Polsky, D. & Basu, A. 2006. Selection Bias in Observational Data. *The Elgar Companion to Health Economics.* Edward Elgar Publishing.
Pullenayegum, E. M. & Willan, A. R. 2011. Marginal Models for Censored Longitudinal Cost Data: Appropriate Working Variance Matrices in Inverse-Probability-Weighted GEEs Can Improve Precision. *The International Journal of Biostatistics,* 7.
Quinn, C. 2007. *The health-economic applications of copulas: methods in applied econometric research* [Online]. HEDG, c/o Department of Economics, University of York. Available: http://ideas.repec.org/p/yor/hectdg/07-22.html [Accessed 10/08/2011.
Radice, R., Grieve, R., Ramsahai, R., Kreif, N., Sadique, Z. & 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).
Ramsahai, R., Grieve, R. & Sekhon, J. S. 2011. Extending Iterative Matching Methods: An Approach to Improving Covariate Balance that Allows Prioritisation. *Health Services and Outcomes Research Methodology*.
Robins, J., Sued, M., Lei-Gomez, Q. & Rotnitzky, A. 2007. Comment: Performance of Double-Robust Estimators When "Inverse Probability" Weights Are Highly Variable. *Statistical Science,* 22**,** 544-559.
Robins, J. M., Rotnitzky, A. & Zhao, L. P. 1995. Analysis of Semiparametric Regression Models for Repeated Outcomes in the Presence of Missing Data. *Journal of the American Statistical Association,* 90**,** 106-121.
Rosenbaum , P. R. & Rubin, D. B. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika,* 70**,** 41-55.
Rosenbaum , P. R. & Rubin, D. B. 1985. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician,* 39.
Rowan, K., Welch, C., North, E. & Harrison, D. 2008. Drotrecogin alfa (activated): real-life use and outcomes for the UK. *Critical Care,* 12.
Rubin, D. & Thomas, N. 1992. Characterizing the Effect of Matching Using Linear Propensity Score Methods with Normal Distributions. *Biometrika,* 79.
Rubin, D. B. 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. B. 2010. On the limitations of comparative effectiveness research. *Statistics in Medicine,* 29**,** 1991-1995.
Sadique, M. Z., Grieve, R., Harrison, D., Cuthbertson, B. & Rowan, K. 2011. Is Drotrecogin alfa (activated) for adults with severe sepsis, cost-effective in routine clinical practice? *Critical Care,* 15**,** R228.
Sculpher, M. 2008. Subgroups and Heterogeneity in Cost-Effectiveness Analysis. *Pharmacoeconomics,* 26**,** 799-806.

Sekhon, J. S. 2011. Matching: multivariate and propensity score matching with automated balance search. *Journal of Statistical Software*.
Sekhon, J. S. & Grieve, R. D. 2011. A Matching Method for Improving Covariate Balance in Cost-Effectiveness Analyses. *Health Economics,* doi: 10.1002/hec.1748.
Sekhon, J. S. & Mebane, W. R. J. 1998. Genetic optimization using derivatives: Theory and application to nonlinear models. *Political Analysis***,** 189-203.
Silva, E., de Figueiredo, L. F. P. & Colombari, F. 2010. Prowess-Shock Trial: A Protocol Overview and Perspectives. *Shock,* 34**,** 48-53 10.1097/SHK.0b013e3181e7e97b.
Stuart, E. A. 2010. Matching Methods for Causal Inference: A review and a look forward. *Statistical Science,* 25.
Terza, J. V., Basu, A. & Rathouz, P. J. 2008. Two-Stage Residual Inclusion Estimation: Addressing Endogeneity in Health Econometric Modeling. *J Health Econ,* 27**,** 531–543.
Thompson, S. G. & Nixon, R. 2005. How Sensitive Are Cost-Effectiveness Analyses to Choice of Parametric Distributions? *Medical Decision Making,* 25**,** 416-423.
Trivedi, P. K. & Zimmer, D. M. 2005. *Copula Modeling: An Introduction to Practitioners,* Delft, Now Publishing Inc.
van der Laan, M. J. & Gruber, S. 2010. Collaborative Double Robust Targeted Maximum Likelihood Estimation. *The International Journal of Biostatistics* 6.
Vanness, D. & Mullahy, J. 2006. Perspectives of mean -based evaluation of health care. *In:* JONES, A. M. (ed.) *The Elgar Companion to Health Economics.* Edward Elgar Publishing.
Willan, A. R., Briggs, A. H. & Hoch, J. S. 2004. Regression methods for covariate adjustment and subgroup analysis for non-censored cost-effectiveness data. *Health Economics,* 13**,** 461-475.

Appendix 4.1 - Genetic Matching

Overview

Genetic matching (GM) automates the process of maximizing balance on observed covariates in the matched sample by using an evolutionary search algorithm to determine the weight each individual covariate is given. As with any matching method, GM requires choices to be made *a priori* about which covariates to include in the matching and assessment of balance, and which balance statistic to use. In GM the key innovations are the generalised distance metric, and the use of an iterative search algorithm to maximize covariate balance. Full details of the method and its properties are covered in a general context elsewhere (Diamond and Sekhon, 2012, Sekhon, 2011), so here we summarize the key aspects.

Selection of covariates for matching algorithm

Before matching, it is necessary to choose which potential confounders to condition on.

The researcher should follow general guidance and only include those covariates anticipated to influence the endpoints (Rubin, 2007). This selection process should also consider interaction effects as well as main effects and nonlinear terms. The choice can be informed by previous empirical analyses, expert opinion, and causal diagrams (Rubin, 2007, Pearl, 2001). The GM algorithm will only use those matching variables that are pre-specified. The choice of variables for balance assessment should include those anticipated to be of high prognostic importance whether or not they are included in the matching. For example, a summary prognostic measure may be excluded from the matching because it is highly correlated with the underlying covariates, and better overall balance may be achieved by just matching on the covariates. However, balance

should be checked on both the underlying covariates and the summary measure. GM can also be tailored to prioritize achieving covariate balance on particular covariates designated as "high priority", for further details see Ramsahai et al. (2011) (Ramsahai et al., 2011).

Covariate balance statistics

A recommended statistic for checking covariate balance is the weighted standardized mean difference:

$$d = \frac{\bar{x}_{treatment} - \bar{x}_{control}}{\sqrt{\frac{s^2 treatment + s^2 control}{2}}},$$

where for continuous covariates, $\bar{x}$ and s^2 denote the covariate's weighted means and variances. This balance statistic allows matching methods to be compared to IPTW, by using the appropriate weights. For matching these are the frequency weights from the matched datasets, and for IPTW the weights calculated from the PS. This balance statistic can be adapted for binary variables (Austin, 2009a).

In some circumstances, the weighted standardized mean differences are an insufficient measure of balance as they are insensitive to imbalances in aspects of the covariate distribution beyond the mean (e.g., variance, maximum, skew, kurtosis). To address imbalances beyond differences in means for linear terms, matching methods can consider standardized differences for higher order terms, but also alternative balance statistics such as Kolmogorov-Smirnov (KS) tests and empirical quantile-quantile plots (Austin, 2009a). The drawback with these non-parametric measures is that weighted versions that would enable comparisons between matching and IPTW are not currently available.

Distance metric

The Mahalanobis distance (MD) between any two observations (one from treatment and the other from control) is:

$$MD(\mathbf{X}_i, \mathbf{X}_j) = \left\{(\mathbf{X}_i - \mathbf{X}_j)^{\mathrm{T}} (\mathbf{S}^{-1/2})^{\mathrm{T}} \mathbf{S}^{-1/2} (\mathbf{X}_i - \mathbf{X}_j)\right\}^{\frac{1}{2}}$$

where **S** is the sample covariance matrix of **X** and $\mathbf{X}^{\mathrm{T}}$ is the transpose of the matrix **X**. Using this metric, the distance between individual covariates is collapsed into a single scalar. The PS can be combined with MD by, for example, including the PS as a variable in the **X** matrix. GM generalizes the MD by including an additional weight matrix W:

$$GMD(\mathbf{X}_i, \mathbf{X}_j) = \left\{(\mathbf{X}_i - \mathbf{X}_j)^{\mathrm{T}} (\mathbf{S}^{-1/2})^{\mathrm{T}} \mathbf{W} \mathbf{S}^{-1/2} (\mathbf{X}_i - \mathbf{X}_j)\right\}^{\frac{1}{2}}$$

where **W** is a $k \times k$ positive definite weight matrix with k being the number of matching covariates , and $\boldsymbol{S}^{-\frac{1}{2}}$ is the Cholesky decomposition of **S**. GM essentially matches by minimizing the generalized version of MD. **W** is chosen to be the weight matrix that minimizes covariate imbalance according to the balance statistics the user chooses (e.g., standardized differences, KS statistics). The GM algorithm uses the distance measure, GMD, in which (by default) all elements of **W** are zero except down the main diagonal. The main diagonal is the vector of weights chosen by the algorithm. If each of the weights for the covariates are set equal to one and the weight for the PS is zero, GMD is the same as MD. That is, GM will converge to the MD if this is the optimal distance metric. If the PS contains all the information required to maximize covariate balance, the algorithm will converge to the corresponding distance metric, that is, the PS will be given full weight, and the other elements in **W** will be given zero weight. Hence, both PS and MD matching can be considered as limiting cases of GM.

The inclusion of individual covariates in the **X** matrix, rather than relying solely on the specification of the PS, helps ensure covariate balance when the PS is misspecified. In this sense, GM is robust to misspecifications in the PS.

The iterative search algorithm

Here we provide an overview of the optimization algorithm. Further details are available elsewhere (Sekhon and Mebane, 1998, Mebane and Sekhon, 2011). The aim of the GM algorithm is to find the optimal weights, **W**, that is the weight matrix which produces the matched sample with the best balance. GM uses a genetic search algorithm to search the weight matrices **W**, where each possible **W** corresponds to a different distance metric. The algorithm proposes batches of weights, **W**s and moves towards the batch which contains the optimal weights. Each batch is a $generation$ and is used iteratively to produce a subsequent generation with better candidate Ws. The size of each generation is the $population\ size$ (e.g., 1000) and is constant for all generations. For each generation the sample is matched according to each metric, corresponding to each **W**, to produce as many matched samples as the population size. Balance is evaluated for each matched sample and the algorithm identifies the weights corresponding to the best balance. The generation of candidate **W**s evolves towards those containing, on average, better **W** and asymptotically converges to contain the optimal **W**: the one which maximizes balance.

The X matrix includes all variables which are matched on and is used to define the GMD between units. The balance matrix consists of columns of data for each variable used to measure balance. By default, the balance matrix is identical to the X matrix. Optimization can be stopped either if there is no significant improvement in the minimum loss over a specified number of generations or after a fixed number of

generations (e.g., 200). The algorithm will optimize whichever balance statistics are chosen, recommended statistics include t-statistics from paired t-tests, D-statistics from Kolmogorov-Smirnov tests (Sekhon, 2011, Diamond and Sekhon, 2012) and weighted standardized differences (Austin, 2009a, Stuart, 2010).

Previous simulation evidence

Diamond and Sekhon (2008) (Sadique et al., 2011) conducted an extensive simulation study to compare the performance of GM to other matching methods (PS matching, MD matching, PS and MD matching combined). The results showed that GM produced better covariate balance. Where the PS was correctly specified and the covariates were multivariate normal, GM dominated the other multivariate matching methods in terms of bias and RMSE, and reported lower MSE than PS matching. When the PS was misspecified, GM reported lower bias and RMSE than the other estimators. Sekhon and Grieve (2011) (Sekhon and Grieve, 2011) compared GM to PS matching in a challenging setting where some covariates were discrete, and others continuous but with highly skewed distributions. The simulation reported that GM achieved better covariate balance, lower bias and MSE, compared with PS matching.

Diamond and Sekhon (2010) (Diamond and Sekhon, 2012) also compared the performance of GM to PS matching, where the PS was estimated by a linear logistic regression model, random forests and boosted Classification and Regression Trees (CART). The simulations considered scenarios that differed in the degree of linearity and additivity in the true PS model, that is the extent to which the PS model included quadratic and interaction terms. GM reported the smallest MSE and bias, apart from one scenario where matching on the correctly specified PS model gave least bias.

Implementation

Various matching options can be implemented in the software for GM (Sekhon, 2011). For example, matching can be performed with or without replacement, with calipers, 1:1 or 1:n, with or without ties. Software and further details can be found at http://sekhon.berkeley.edu/matching.

Appendix 4.2 - Illustrative R code for the simulation study of research paper 2

For scenario 1a the following code was used for the data generating process, and to report balance and average treatment effects (ATEs). For the remaining scenarios, the code was modified accordingly.

Data generating process

The following libraries are required for the code:

```
library(Rlab)
library(Matching)
library(stats)
library(boot)
library(copula)
```

A simulated dataset, including the covariates `X1,X2,X3` the treatment variable `tx` and the endpoints `cost` and `Y` (denoting QALY) was created using the following commands:

```
Sigma<-matrix(c(1,0.2,0.2,1),2,2)
                  Data generating
X12<-mvrnorm(n,c(2, 4), Sigma)

X1<-X12[,1]

X2<-X12[,2]

X3<-rbern(n,0.5+ifelse(X1>2,0.1,-0.1))

psc_logit<-log(0.2)+(0.1*X1)+(0.2*X2)+(0.3*X3)+(0.2*X1*X3)-
(0.2*X2*X3)

psc<-inv.logit(psc_logit)
tx<-rbern(n,psc)

E.cost<- 4000+ 4000*tx+5000*X1+4000*X2+3000*X3-1000*X3*tx
E.cost=ifelse(E.cost<=0,0.1,E.cost)
E.Y <- 9+0.25*tx-(1*X1)-(0.6*X2)-(0.8*X3)+(0.5*X3*tx)

ngmvdc <- mvdc(normalCopula(0.4), c("norm", "gamma"),
          list(list(mean = E.Y, sd =0.2),
list(shape=10,rate=10/E.cost)))
rng <- rmvdc(ngmvdc, n)

Y <- rng[,1]
cost <- rng[,2]
```

```
dataset<-cbind(X1,X2,X3,Y,cost,tx)
dataset<-as.data.frame(dataset)

dataset.X3 <- dataset[dataset$X3==1,]
dataset.noX3 <- dataset[dataset$X3==0,]
```

The object `dataset` denotes the whole sample, `dataset.noX3` and `dataset.X3` are the subsamples for subgroup 1 and subgroup 2, respectively.

PS model fitting

Two PS models were fitted for subgroup 1 and subgroup 2. Subgroup specific propensity scores (`pscore.noX3` and `pscore.X3`), linear predictors (`pscore.lin.noX3` and `pscore.lin.X3`) and inverse probability weights (pscorwght.noX3 and pscorwght.X3) were calculated, and attached to the datasets:

```
pmodel.noX3<-glm(tx~X1+X2+X3,family=binomial,data=dataset.noX3)
pscore.lin.noX3<-pmodel.noX3$linear.predictor
pscore.noX3<-pmodel.noX3$fitted.values
pscorwght.noX3<-(dataset.noX3$tx/pscore.noX3)+
((1-dataset.noX3$tx)/(1-pscore.noX3))
      dataset.noX3<-
cbind(dataset.noX3,pscore.noX3,pscore.lin.noX3,
pscorwght.noX3)
rm(pscore.lin.noX3,pscore.noX3,pscorwght.noX3)
pmodel.X3<-glm(tx~X1+X2+X3,family=binomial,data=dataset.X3)
pscore.lin.X3<-pmodel.X3$linear.predictor
      pscore.X3<-pmodel.X3$fitted.values
pscorwght.X3<-(dataset.X3$tx/pscore.X3)+((1-dataset.X3$tx)/
(1-pscore.X3))
dataset.X3<-
cbind(dataset.X3,pscore.X3,pscore.lin.X3,pscorwght.X3)
      rm(pscore.lin.X3,pscore.X3,pscorwght.X3)
```

In the following sections, the implementation of PS matching, GM and IPTW is described, for subgroup 1. For subgroup 2, the code was modified accordingly.

PS matching

First, matched datasets were created, using the `Match()` function from the Matching library.

```
attach(dataset.noX3)

mtchout.Y.noX3<-Match(Y=Y,Tr=tx,
```

```
X=cbind(pscore.lin.noX3),exact=c(FALSE),
estimand="ATE")
mtchout.cost.noX3<-
Match(Y=cost,Tr=tx,X=cbind(pscore.lin.noX3),exact=c(FALSE),estim
and="ATE")

detach(dataset.noX3)

mtch.data.noX3<-
rbind(dataset.noX3[mtchout.Y.noX3$index.treated,]
,dataset.noX3[mtchout.Y.noX3$index.control,])
mtch.data.noX3<-
cbind(mtch.data.noX3,weights=c(mtchout.Y.noX3$weights,mtchout.Y.
noX3$weights))
```

After matching, ATEs for the QALY and cost endpoint can be extracted as follows:

```
Y_ps.noX3<-mtchout.Y.noX3$est
cost_ps.noX3<-mtchout.cost.noX3$est
```

The covariate balance measured as weighted standardized differences can be reported, for example for the covariate X1 as follows:

```
attach(mtch.data.noX3)

X1.ps.sdiff.X3<-100*abs(weighted.mean(x=as.matrix(X1[tx==1]),
w=weights[tx==1])-
weighted.mean(x=as.matrix(X1[tx==0]),w=weights[tx==0]))
/sqrt((cov.wt(x=as.matrix(X1[tx==1]),
wt=weights[tx==1])$cov+
cov.wt(x=as.matrix(X1[tx==0]),wt=weights[tx==0])$cov)/2)

detach(mtch.data.noX3)
```

Genetic Matching

The automated GM algorithm was run using the GenMatch() function from the Matching library:

```
genmtchout.noX3<-GenMatch(Tr=tx,X=cbind(pscore.lin.noX3,X1,X2),
estimand="ATE",
fit.func = my.fitfunc_sdiff,
starting.values=c(10000,0,0),exact=c(FALSE,FALSE,
FALSE),pop.size=gpop,unif.seed=seedin,
int.seed=seedin)
```

Matched datasets were then created:

```
attach(dataset.noX3)
```

```
gmtchout.Y.noX3<Match(Y=Y,Tr=tx,X=cbind(pscore.lin.noX3,X1,X2),
exact=c(FALSE,FALSE,FALSE),Weight.matrix=diag(w8s),
estimand="ATE")

gmtchout.cost.noX3<-
Match(Y=cost,Tr=tx,X=cbind(pscore.lin.noX3,X1,X2),
exact=c(FALSE,FALSE,FALSE),Weight.matrix=diag(w8s),
estimand="ATE")
detach(dataset.noX3)
mtch.data.noX3<rbind(dataset.noX3[gmtchout.Y.noX3$index.treated,
],
dataset.noX3[gmtchout.Y.noX3$index.control,])

mtch.data.noX3<cbind(mtch.data.noX3,
weights=c(gmtchout.Y.noX3$weights,
gmtchout.Y.noX3$weights))
```

ATEs can then be extracted as previously:

```
Y_gn.noX3<-gmtchout.Y.noX3$est
cost_gn.noX3<-gmtchout.cost.noX3$est
```

The covariate balance for the covariate X1 can be calculated as follows:

```
attach(mtch.data.noX3)
X1.gn.sdiff.noX3<-
100*abs(weighted.mean(x=as.matrix(X1[tx==1]),
w=weights[tx==1])-
weighted.mean(x=as.matrix(X1[tx==0]),w=weights[tx==0]))/
sqrt((cov.wt(x=as.matrix(X1[tx==1]),
wt=weights[tx==1])$cov
+cov.wt(x=as.matrix(X1[tx==0]),
wt=weights[tx==0])$cov)/2)

detach(mtch.data.noX3)
```

Inverse probability of treatment weighting

The inverse probability weights (`variable pscorwght.noX3`) calculated previously were used in a weighted mean difference of the respective endpoints, to calculate the average treatment effects.

```
attach(dataset.noX3)

Y_ipw.noX3<-weighted.mean(x=Y[tx==1],
w=pscorwght.noX3[tx==1])-
weighted.mean(x=Y[tx==0],w=pscorwght.noX3[tx==0])

cost_ipw.noX3<-weighted.mean(x=cost[tx==1],
w=pscorwght.noX3[tx==1])-
weighted.mean(x=cost[tx==0],w=pscorwght.noX3[tx==0])
```

Balance can be calculated as standardized mean difference, where the treated and control samples are weighted with the inverse probability weights:

```
X1.ipw.sdiff.noX3<-100*abs(weighted.mean(x=as.matrix(X1[tx==1]),
            w=pscorwght.noX3[tx==1])-
            weighted.mean(x=as.matrix(X1[tx==0]),w=pscorwght.noX
            3[tx==0]))/sqrt((cov.wt(x=as.matrix(X1[tx==1]),wt=ps
            corwght.noX3[tx==1])$cov+cov.wt(x=as.matrix(X1[tx==0
            ]),wt=pscorwght.noX3[tx==0])$cov)/2)

detach(dataset.noX3)
```

Appendix 4.3 - Supplementary tables and figures for research paper 2

Appendix 4.3 Table 1 - Sensitivity analysis for the case study. Covariate balance (% weighted standardized differences) after IPTW with weights truncated to different percentiles.

Covariate	Percentile	2 organ failures	3 to 5 organ failures
Age			
	0,100	9.06	7.24
	1,99	9.13	2.72
	5,95	14.83	8.33
	10,90	20.33	13.29
	25,75	26.19	24.52
IMprob			
	0,100	5.58	9.86
	1,99	7.49	0.27
	5,95	8.17	4.88
	10,90	9.04	7.75
	25,75	10.47	14.01
IMscore			
	0,100	3.88	12.41
	1,99	9.97	4.93
	5,95	16.31	12.89
	10,90	21.66	18.45
	25,75	27.92	30.61
% Ventilated			
	0,100	5.99	13.19
	1,99	13.39	5.54
	5,95	20.23	10.55
	10,90	24.89	14.60
	25,75	30.74	24.44

Appendix 4.3 Table 2 - Monte Carlo simulations: relative bias for estimated incremental costs (ΔC), QALYs (ΔE) and INBs (WTP=£20,000)

Scenario	Method	Relative bias (%) ΔC		Relative bias (%) ΔE		Relative bias (%) INB	
		Subgr. 1	Subgr. 2	Subgr. 1	Subgr. 2	Subgr. 1	Subgr. 2
1a							
	PS matching	0.3	0.0	0.6	0.5	4.4	0.6
	GM	0.1	1.2	0.5	0.4	2.9	0.8
	IPTW	0.2	0.0	0.1	0.2	1.5	0.3
1b							
	PS matching	1.3	2.5	10.6	4.0	58.0	5.7
	GM	0.7	1.2	1.8	1.0	11.6	1.6
	IPTW	2.1	3.0	12.6	4.7	71.3	6.6
2a							
	PS matching	0.4	3.0	0.7	1.9	5.0	3.2
	GM	0.3	3.9	0.6	2.8	4.2	4.4
	IPTW	0.2	2.5	0.3	1.3	2.2	2.2
2b							
	PS matching	0.3	0.4	0.6	0.8	4.1	0.9
	GM	0.1	2.0	0.4	2.5	2.4	3.7
	IPTW	0.3	8.7	0.2	6.5	2.0	10.3
3a							
	PS matching	0.8	0.0	0.0	0.5	0.6	0.7
	GM	0.4	0.6	0.2	0.3	0.7	0.3
	IPTW	0.3	0.3	0.0	0.1	0.2	0.0
3b							
	PS matching	0.5	0.5	7.7	2.3	12.5	3.0
	GM	0.3	0.0	7.6	2.4	12.4	3.0
	IPTW	0.3	0.4	7.8	2.6	12.7	3.3

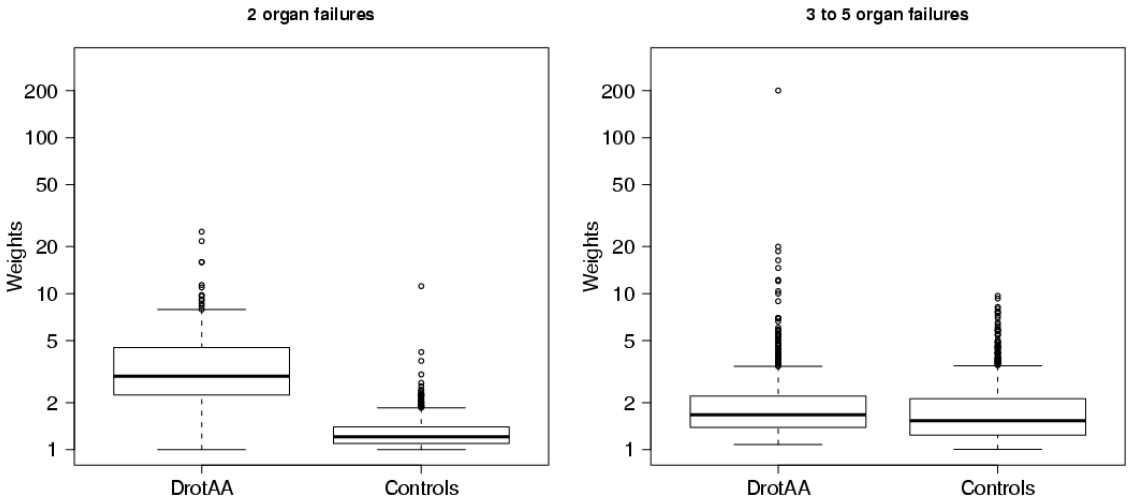
Notes: True value of incremental costs is £3,000 (Subgroup 1) and £4,000 (Subgroup 2), for all Scenarios. True value of incremental QALYs for Scenarios 1 and 2 are 0.25 (Subgroup 1) and 0.75 (Subgroup 2), and for Scenario 3, 0.5 and 0.75. The true INBs are therefore £1,000 and £12,000 for Scenarios 1 and 2, £6,000 and £12,000 for Scenario 3.

Appendix 4.3 Table 3 - Monte Carlo simulations: RMSE for estimated incremental costs (ΔC), QALYs (ΔE) and INBs (WTP=£20,000)

Scenario	Method	RMSE: ΔC		RMSE: ΔE		RMSE: INB	
		Subgr. 1	Subgr. 2	Subgr. 1	Subgr. 2	Subgr. 1	Subgr. 2
1a							
	PS matching	788	915	0.03	0.03	961	1,068
	GM	819	889	0.02	0.02	756	825
	IPTW	679	773	0.02	0.02	675	782
1b							
	PS matching	853	953	0.08	0.08	1,988	2,031
	GM	818	888	0.03	0.03	1,060	1,088
	IPTW	765	841	0.07	0.07	1,802	1,875
2a							
	PS matching	787	1,231	0.03	0.04	989	1,484
	GM	790	1,173	0.02	0.03	744	1,267
	IPTW	681	1,155	0.02	0.09	676	2,535
2b							
	PS matching	788	1,224	0.03	0.03	963	1,306
	GM	820	1,192	0.02	0.03	756	1,210
	IPTW	680	920	0.02	0.07	673	1,839
3a							
	PS matching	812	866	0.04	0.03	1,271	1,043
	GM	755	845	0.02	0.01	699	784
	IPTW	682	730	0.02	0.01	686	689
3b							
	PS matching	769	835	0.05	0.03	1,229	927
	GM	771	856	0.04	0.02	1,036	885
	IPTW	668	724	0.04	0.02	999	802

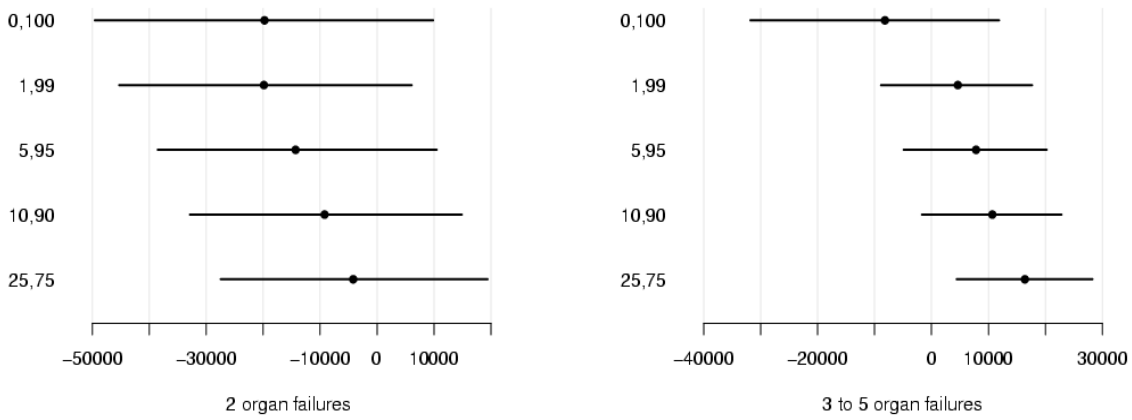
Notes: True value of incremental costs is £3,000 (subgroup 1) and £4,000 (subgroup 2), for all scenarios. True value of incremental QALYs for scenarios 1 and 2 are 0.25 (subgroup 1) and 0.75 (subgroup 2), and for scenario 3, 0.5 and 0.75. The true INBs are therefore £1,000 and £12,000 for Scenarios 1 and 2, £6,000 and £12,000 for scenario 3.

Appendix 4.3 Figure 1 - Sensitivity analysis for the case study. Distribution of inverse probability weights for DrotAA and control groups



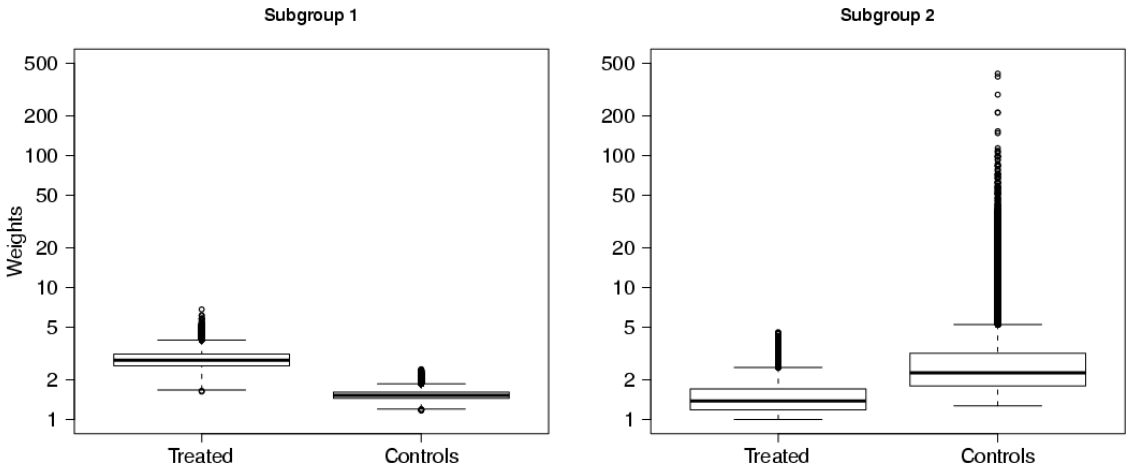
Notes: The boxplots show the median, interquartile distance and extreme values of the inverse probability weights.

Appendix 4.3 Figure 2 - Sensitivity analysis for case study. 95% bootstrapped CIs following IPTW, after truncation of the weights according to different percentiles



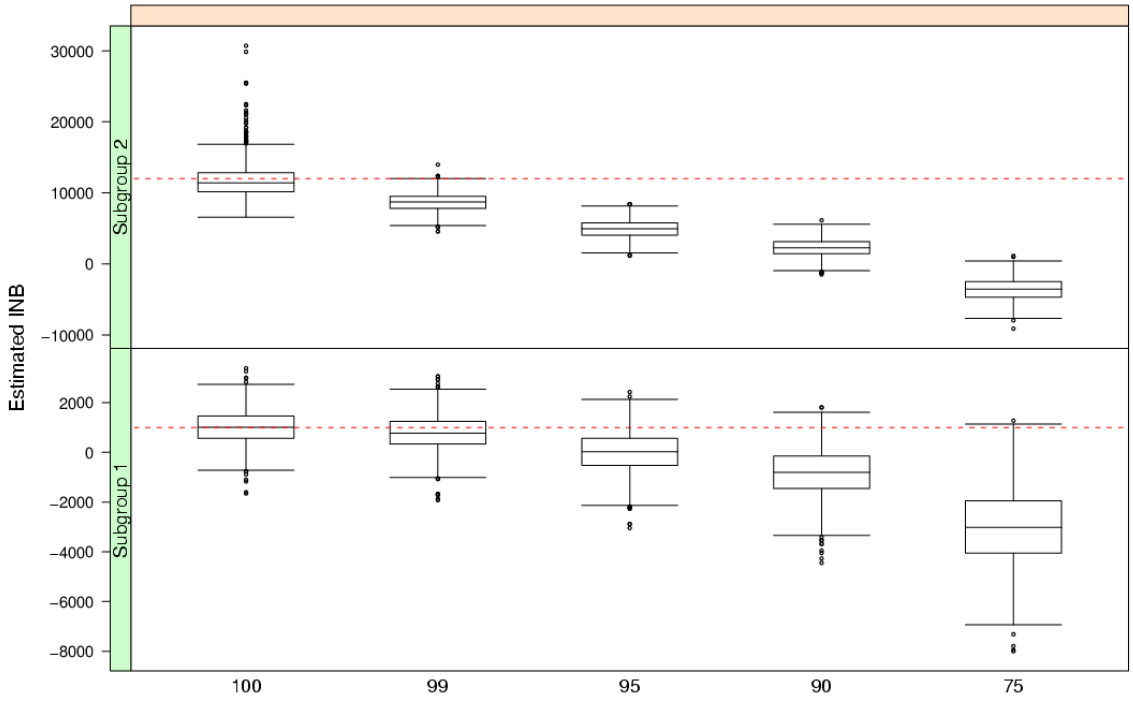
Notes: 0,100 corresponds to no truncation, 1,99 corresponds to the case where weights are truncated at the first and 99th percentiles

Appendix 4.3 Figure 3 - Sensitivity analysis for the Monte Carlo simulations, Scenario 2a. Distribution of inverse probability weights for treated and control observations, generated for a typical sample (n=1,000,000)



Notes: The boxplots show the median, interquartile distance and extreme values of the inverse probability weights.

Appendix 4.3 Figure 4 - Sensitivity analysis for the Monte Carlo simulations, scenario 2a. Boxplots of the INBs (WTP= £20,000) after IPTW with truncated weights.



Notes: Dotted lines indicate the true values of the INBs (1,000 for subgroup 1 and 12,000 for subgroup 2). 100 corresponds to no truncation, 99 corresponds to the case where weights are truncated at the first and 99th percentiles. Results are across 1000 replications.

Chapter 5 - Statistical methods that combine the PS with endpoint regression models, for estimating cost-effectiveness

5.1 Preamble to research paper 3

The conceptual review (chapter 2) highlighted the challenges in CEA of correctly specifying the PS and the regression models for the cost and effectiveness endpoints. The critical appraisal of the applied literature (research paper 1) found that most applied CEA used regression or PS matching for addressing selection bias; however they did not carefully assess whether it was plausible to assume that the PS and the endpoint regression models were correctly specified.

Research paper 2 proposed the use of GM to protect against misspecification of the PS. The conceptual review (chapter 2) also suggested methods that combine the PS with endpoint regression models: DR methods and regression-adjusted matching. These methods can provide unbiased estimates even when either the PS or the endpoint models is misspecified. These methods have not been considered for addressing selection bias in CEA before. Research paper 3 aims to address this gap in the literature.

This paper compares DR with regression-adjusted PS matching, traditional PS and regression approaches, for estimating incremental cost-effectiveness. The simulation study is grounded in a motivating CEA. The paper provides insights on the relative performance of the methods across typical CEA settings. To assist applied researchers who wish to implement the proposed methods, the paper provides sample software code (Appendix 5.1)

London School of Hygiene & Tropical Medicine
Keppel Street, London WC1E 7HT
www.lshtm.ac.uk

Registry
T: +44(0)20 7299 4646
F: +44(0)20 7299 4656
E: registry@lshtm.ac.uk

LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

COVER SHEET FOR EACH 'RESEARCH PAPER' INCLUDED IN A RESEARCH THESIS

Please be aware that one cover sheet must be completed for each 'Research Paper' included in a thesis.

1. **For a 'research paper' already published**

   1.1. Where was the work published? .....

   1.2. When was the work published? .....

   1.2.1. If the work was published prior to registration for your research degree, give a brief rationale for its inclusion

   .....

   .....

   .....

   1.3. Was the work subject to academic peer review? .....

   1.4. Have you retained the copyright for the work? **Yes / No**
   If yes, please attach evidence of retention.
   If no, or if the work is being included in its published format, please attach evidence of permission from copyright holder (publisher or other author) to include work

2. **For a 'research paper' prepared for publication but not yet published**

   2.1. Where is the work intended to be published? .....

   2.2. Please list the paper's authors in the intended authorship order

   .....

   2.3. Stage of publication – Not yet submitted / Submitted / Undergoing revision from peer reviewers' comments / In press

3. **For multi-authored work, give full details of your role in the research included in the paper and in the preparation of the paper. (Attach a further sheet if necessary)**

   .....

   .....

NAME IN FULL (Block Capitals) .....

STUDENT ID NO: .....

CANDIDATE'S SIGNATURE ..... Date .....

SUPERVISOR/SENIOR AUTHOR'S SIGNATURE (3 above) .....

Additional page for Question (3) on LSHTM cover sheet form:

The candidate led on the conception of the research question for research paper 3, in collaboration with RG and an external collaborator, JS. The candidate led on the choice of statistical methods under comparison, and the design of the simulation study, in collaboration with RG, RR and JS. RR contributed to the design and implementation of the simulation scenarios. The candidate wrote the code for the simulation study, with help from RR. She conducted the statistical analysis for the motivating case study, and interpreted the results of the paper, with RR and RG. The candidate wrote the first draft of the manuscript. She managed each round of comments and suggestions from the co-authors, in collaboration with RG. All authors read and approved the final draft prior to inclusion in the thesis.

5.2 Research paper 3 - Regression-adjusted matching and double-robust methods for estimating average treatment effects in health economic evaluation

Noémi Kreif[1], Richard Grieve[1], Rosalba Radice[2] and Jasjeet S. Sekhon[3]

[1] Department of Health Services Research & Policy, LSHTM, University of London;

[2] Department of Economics, Mathematics and Statistics, Birkbeck, London, UK;

[3] Department of Political Science and Department of Statistics, UC Berkeley, Berkeley, CA, US

Status: Rejected from Health Economics, 2012. To be submitted to Health Services and Outcomes Research Methodology, December 2012.

Contributions: The candidate led on the conception of the research question for research paper 3, in collaboration with RG and an external collaborator, JS. The candidate led on the choice of statistical methods under comparison, and the design of the simulation study, in collaboration with RG, RR and JS. RR contributed to the design and implementation of the simulation scenarios. The candidate wrote the code for the simulation study, with help from RR. She conducted the statistical analysis for the motivating case study, and interpreted the results of the paper, with RR and RG. The candidate wrote the first draft of the manuscript. She managed each round of comments and suggestions from the co-authors, in collaboration with RG. All authors read and approved the final draft prior to inclusion in the thesis.

The candidate_____ The supervisor_____

Abstract

Regression and propensity score (PS) methods can reduce selection bias when estimating average treatment effects (ATEs), if their underlying models are correctly specified. In cost-effectiveness analysis (CEA), the correct specification of these models can be challenging, due to potential nonlinear functional form relationships. Double-robust (DR) methods and regression-adjusted matching can protect against bias from model misspecification, but their relative performance has not been previously assessed. This paper compares selected DR methods (weighted regression and augmented inverse probability of treatment weighting), regression-adjusted matching, regression and PS methods for addressing selection bias in CEA.

We contrast the methods in a CEA of a pharmaceutical intervention, Drotrecogin alfa, for severe sepsis. We find that cost-effectiveness estimates differ across methods, and methods that combine the PS with endpoint regression report narrower confidence intervals than methods that use the PS alone. Motivated by the case study, our simulation study compares the methods in scenarios with estimated PSs close to 0 or 1, that have unstable inverse probability of treatment (IPT) weights. The simulations include settings with functional form misspecification in the PS and endpoint regression models (e.g. cost model with log instead of identity link). Measures of relative performance include bias and root mean squared error (RMSE) of the incremental net benefit.

We found that combining PS methods with endpoint regression reduced bias and RMSE compared to using PS only. With unstable IPT weights and misspecifications to the PS and regression models, regression-adjusted matching reported less bias than DR methods, and the lowest RMSE of all the approaches considered.

Introduction

Recent investments in large observational datasets offer new opportunities for comparative effectiveness research, including cost-effectiveness analysis (CEA). CEA ideally use evidence from pragmatic randomised controlled trials (RCTs) which include patients, centres and comparators appropriate to the decision context (Willan and Briggs, 2006, Glick et al., 2007, Gray et al., 2010). For many decision problems RCTs may be unavailable or insufficient, and so the CEA may be reliant partly or entirely on non-randomised studies (NRS) (Kreif et al., 2012b). While instrumental variable methods (Basu et al., 2007) can remove selection bias due to observed and unobserved confounding, in many circumstances plausible instruments are not available. The majority of CEA that use individual patient data from NRS rely on regression and propensity score (PS) methods (Kreif et al., 2012b). These approaches assume "unconfoundedness", but also that the functional form of the regression model or the PS is correctly specified. Alternatively, regression and PS approaches can be combined, for example in double-robust (DR) estimation (Bang and Robins, 2005), or regression-adjusted matching (Ho et al., 2007). Both "combined approaches" can protect against bias from misspecification of the PS or regression models, but they have not been compared before.

DR methods combine inverse probability of treatment (IPT) weights with endpoint regression models. Under the "double-robust" property, unbiased estimates of average treatment effects (ATEs) can be obtained if *either* one of the regression or PS models is correctly specified (Robins et al, 1994). However, when estimated propensity scores are close to 0 or 1, the IPT weights can be unstable. In circumstances where there is dual misspecification and unstable IPT weights, certain DR approaches have been reported to

be more biased and less efficient than ordinary least squares (OLS) regression (Kang and Schafer, 2007, Freedman and Berk, 2008).

An alternative "combined approach" is regression-adjusted matching (Hill and Reiter, 2005, Ho et al., 2007), which aims to create balanced comparison groups before regression adjustment. This approach purports to reduce the sensitivity of the regression estimates to the choice of model specification (Dehejia and Wahba, 2002, Ho et al., 2007).

Regression-adjusted matching and DR methods warrant particular consideration for CEA that use data from NRS. A recent systematic review found that most studies use regression or PS matching, but do not carefully assess model specification (Kreif et al., 2012b). DR methods have been proposed for addressing censoring (Pan and Zeng, 2011, Bang and Tsiatis, 2000), or selection bias in cost analyses (Basu et al., 2011), but have not been considered for CEA. Previous findings on the relative merits of alternative DR methods (Porter et al., 2011, van der Laan and Gruber, 2010, Robins et al., 2007) may not translate to the CEA setting. Here typical circumstances include: baseline covariates that are widely imbalanced between the treatment groups, unstable IPT weights, and misspecified parametric models for both costs (Jones, 2010) and health outcomes (Basu and Manca, 2011).

The aim of this paper is to compare selected DR estimators with a regression-adjusted PS matching estimator, and common PS or regression approaches for estimating ATEs in CEA. We illustrate the approaches with a case study, a CEA of a pharmaceutical intervention, Drotrecogin alfa (DrotAA) for patients with severe sepsis. We consider the relative performance of the methods in a simulation study that extends an influential methodological paper in medical statistics (Kang and Schafer, 2007, Robins et al., 2007, Porter et al., 2011) to a bivariate CEA context. The simulation study design is grounded

in the characteristics of the case study, for example by including scenarios with unstable IPT weights.

In the next section, we outline the statistical methods under comparison. The following section presents the motivating example. We then report the design and results of the simulation study. The last section discusses the findings and suggests areas for further research.

Statistical methods

The methods considered assume no unobserved confounding, and require choices about potential measured confounders, x, to be made in advance, for example drawing on theory (Rubin, 2007), published literature, expert opinion or causal diagrams (Pearl, 1995). We denote by Y_{ik} the observed outcome (cost if $k = C$ and effectiveness if $k = E$), for individual $i = 1, \ldots, n$, where n is the sample size. The parameter of interest is the average treatment effect (ATE) of a binary treatment t, which in CEA corresponds to the incremental cost and effectiveness parameters.

Regression adjustment

Incremental cost, effectiveness and cost-effectiveness can be modelled with simple generalised linear models (GLM) (Barber and Thompson, 2004), two-part models (Buntin and Zaslavsky, 2004, Basu, 2011), semi-parametric methods such as extended estimating equations (Basu and Rathouz, 2005), or flexible parametric methods such as beta-type size distributions (Jones et al., 2011). These approaches have the potential to address skewness, heavy tails and nonlinear relationships between covariates and endpoints. A general concern is that, even a flexible parametric approach is not a substitute for finding the correct model specification (Manning et al., 2005).

We consider common GLMs for estimating incremental cost and effectiveness. Following Barber and Thompson (2004), GLMs for Y_{ik} can be written as

$$g_k(\mu_{i,k}) = \gamma_k t_i + x_i \beta_k; \quad Y_{ik} \sim F_k. \tag{1}$$

Here $\mu_{i,k} = E(Y_{i,k})$ is the expectation of $Y_{i,k}$, g_k is the link function which describes the scale on which x_i are related to $Y_{i,k}$, γ_k and β_k are the regression coefficients, and F_k is an exponential family distribution. Parameters can be estimated via maximum likelihood (ML), quasi ML or Bayesian methods (Basu and Manca, 2011). The joint uncertainty in the estimates of the incremental cost-effectiveness can be recognised by bivariate models (Nixon and Thompson, 2005) or with the nonparametric bootstrap (Davison and Hinkley, 1997). A common GLM for estimating incremental costs uses the gamma distribution with a log link (Buntin and Zaslavsky, 2004), and assumes that the covariates have multiplicative effects on the endpoint. A general way to obtain ATEs, which can handle such nonlinearities is with the method of recycled predictions (Basu and Rathouz, 2005) :

$$\frac{1}{n}\sum_{i=1}^{n}\{\hat{\mu}_{i,k}(x_{i,}, t_i = 1) - \hat{\mu}_{i,k}(x_{i,}, t_i = 0)\}, \tag{2}$$

where $\hat{\mu}_{i,k}(.)$ is the predicted mean of Y_{ik} from the GLM (1) given x_i, and t_i is set to 1 and 0 for the whole sample.

Propensity score methods

Propensity score matching

The PS is the conditional probability of treatment assignment given x_i (Rosenbaum and Rubin, 1983):

$$p_i = Pr(t_i = 1|x_i).$$

Consistent estimates of the ATE can be obtained by creating matched treated and control comparison groups, using the estimated PS, $\hat{p}_i$ as a distance metric (Rosenbaum and Rubin, 1983). However, in finite samples even a correctly specified PS can leave some baseline covariates imbalanced, which can lead to bias if these variables are highly prognostic (Stuart, 2010). Implementations of PS matching include pair matching, nearest neighbour or kernel matching (Imbens and Wooldridge, 2009a, Stuart, 2010, Basu et al., 2011). Here we consider nearest neighbour 1:1 matching with replacement, without callipers[9] (Austin, 2008, Caliendo and Kopeinig, 2008, Stuart, 2010).

The matching estimator is the weighted mean difference between matched treated and control groups, which can be written as:

$$\frac{1}{n}\sum_{i=1}^{n}\{(2t_i - 1)(1 + K_i)Y_{i,k}\},$$

where K_i is the sum of the frequency weights unit i has as a match for other units (Abadie et al., 2004a).

[9] Here a calliper is defined as the pre-specified amount by which propensity scores of matched pairs are allowed to differ.

Inverse probability of treatment weighting

IPTW can estimate ATEs, by reweighting the observed cost and effectiveness endpoints for treatment and control samples[10]. The IPT weight w_i is the inverse of the estimated probability of the observed treatment, $w_i = \frac{t_i}{\hat{p}_i} + \frac{1-t_i}{1-\hat{p}_i}$. If the PS is correctly specified, IPTW can provide consistent estimates and reach semi-parametric efficiency (Hirano et al., 2003). However, even when the PS model is correctly specified, practical violations of the positivity assumption (Westreich and Cole, 2010) resulting in unstable weights, can lead to estimates of ATEs that are biased and inefficient (Kang and Schafer, 2007, Lee et al., 2010, Busso et al., 2011). Here we implement the normalised IPTW estimator (Hirano and Imbens, 2001, Kang and Schafer, 2007), defined as:

$$\frac{\sum_{i=1}^{n} t_i w_i Y_{i,k}}{\sum_{i=1}^{n} t_i w_i} - \frac{\sum_{i=1}^{n} (1-t_i) w_i Y_{i,k}}{\sum_{i=1}^{n} (1-t_i) w_i}$$

Combining regression and PS adjustment

Double-robust methods

Double-robust (DR) methods combine models for the PS and for the endpoint. The distinctive property of DR estimators is that they are consistent if either (but not necessarily both) the PS or the regression model is correctly specified (Robins et al., 1994, Robins et al., 1995, Bang and Robins, 2005). If both components are correct, the DR estimator is a semiparametric efficient estimator (Robins et al., 2007). Compared to IPTW, DR methods can increase efficiency, by stabilising the IPT weights (Glynn and Quinn, 2010). However when both the PS and the endpoint model are misspecified, DR

[10] Further possible ways of balancing with the PS include stratification (blocking) by the quintiles of the PS and adding the PS as a covariate (Rosenbaum and Rubin, 1983). They have been demonstrated to be dominated by IPTW and matching (Lunceford and Davidian, 2004, Austin, 2009b).

estimators generally provide biased and inefficient estimates of ATEs (Kang and Schafer, 2007, Porter et al., 2011, Freedman and Berk, 2008, Basu et al., 2011).

Here, we consider DR methods that are commonly used in the causal inference literature (Lunceford and Davidian, 2004, Funk et al., 2011, Freedman and Berk, 2008). The augmented IPTW (AIPTW) (Robins et al., 1994, Basu et al., 2011) estimator weights residuals from a regression model. The AIPTW estimator is:

$$\frac{\sum_{i=1}^{n} t_i w_i (Y_{i,k} - \hat{\mu}_{i,k}(x_i))}{\sum_{i=1}^{n} t_i w_i} - \frac{\sum_{i=1}^{n} (1 - t_i) w_i (Y_{i,k} - \hat{\mu}_{i,k}(x_i))}{\sum_{i=1}^{n} (1 - t_i) w_i} +$$

$$\frac{1}{n} \sum_{i=1}^{n} \{\hat{\mu}_{i,k}(x_i, t_i = 1) - \hat{\mu}_{i,k}(x_i, t_i = 0)\},$$

where $\hat{\mu}_{i,k}(.)$ is the predicted endpoint from the GLMs defined in equation (1), and w_i is the IPT weight.

One alternative is the weighted regression estimator (Freedman and Berk, 2008, Kang and Schafer, 2007), which can be constructed by combining w_i with the GLMs for Y_{ik}. ATEs can be obtained using the method of recycled predictions:

$$\frac{1}{n} \sum_{i=1}^{n} \{\hat{\mu}_{i,k,wreg}(x_{i,}, t_i = 1) - \hat{\mu}_{i,k,wreg}(x_{i,}, t_i = 0)\},$$

where $\hat{\mu}_{k,wreg}(.)$ is the predicted endpoint from a weighted GLM.

Regression-adjusted matching

It is generally recommended that matching is followed by regression adjustment (Rubin, 1973, Rubin and Thomas, 2000, Abadie and Imbens, 2006). The idea is similar to double- robustness, and also to regression-adjustment in randomised trials: regression is used to "clean up" imbalances between treatment groups after matching (Stuart, 2010). We consider regression-adjusted matching undertaken as a two stage process: matching,

that forms the design stage of the analysis, is followed with regression modelling using the matched data (Ho et al., 2007). This approach can reduce the sensitivity of the estimated ATEs to the specification of the endpoint model (Hill and Reiter, 2005, Ho et al., 2007), and can reduce finite sample bias and increase efficiency compared to matching alone. We implement this approach by undertaking PS matching, and using the frequency weights from the matching to weight the GLMs (1). The regression-adjusted matching estimator of the ATEs for each endpoint can be obtained as:

$$\frac{1}{n}\sum_{i=1}^{n}\left\{\hat{\mu}_{i,k,reg_{match}}(x_{i,}, t_i = 1) - \hat{\mu}_{i,k,reg_{match}}(x_{i,}, t_i = 0)\right\},$$

where $\hat{\mu}_{i,k,reg_{match}}(.)$ is the predicted endpoint obtained from applying GLMs to the matched data.

Motivating case study

Case study overview

We compared the methods in a CEA that evaluated Drotrecogin alpha (activated) (DrotAA), a pharmaceutical for patients with severe sepsis admitted to intensive care units (ICUs), using observational data from the UK Case-Mix Programme dataset coordinated by the Intensive Care National Audit and Research Centre (ICNARC) (Rowan et al., 2008). We revisit a previous CEA (Kreif et al., 2012a, Sadique et al., 2011) and consider high-risk patients defined as having 3, 4 or 5 organ systems failing at ICU admission (n=878 DrotAA and n=1020 controls). The CEA estimated individual-level lifetime quality-adjusted life years (QALYs), based on individual patient's mortality data collected for a follow-up period of four years, and for those who survived, age- and gender-specific expected survival and quality of life. Costs of

DrotAA and all hospitalisations were estimated at the patient level, over the same period of follow-up. QALYs and costs were discounted at 3.5% (NICE, 2008).

Statistical analysis

We used a previously published PS (Rowan et al., 2008), which included the following baseline covariates: hospital type, number of critical care beds in the ICU, age, ICNARC model physiology score (IMscore), gender, number of organ systems failing, type of organ failures (cardiovascular, respiratory, renal, haematological, metabolic acidosis), source of admission to critical care (via the emergency department, theatre or recovery, ward, clinic or home), diagnostic category, and serious conditions in the past medical history. The PS was estimated by logistic regression, the potential nonlinear effects of age and IMscore on the logit of treatment assignment were recognised with restricted cubic splines.

Regression models were developed for the cost and QALY endpoints drawing on the literature (Rowan et al., 2008). Linear predictors included a treatment indicator, treatment by covariate interaction terms, and cubic splines of age and IMscore to take into account possible nonlinearities. Model fit was evaluated using the Akaike information criterion and split sample cross validation (Buntin and Zaslavsky, 2004). A gamma GLM with log link, and a normal model with identity link were selected for the cost and QALY endpoints, respectively.

Both treated and control individuals were matched to their nearest neighbour in the comparison group, one-to-one, with replacement, based on the linear predictor of the PS, using the "Matching" package (Sekhon, 2011). Balance on those potential confounders anticipated to be most important (Sadique et al., 2011), was reported with

weighted standardised differences (Austin, 2009a)[11]. The AIPTW estimator used predictions from the GLMs described above, with both IPTW and AIPTW using normalised weights (Kang and Schafer, 2007). Weighted regression applied the GLMs on data with IPT weights, while regression-adjusted matching applied the same regression models to the matched data, using the frequency weights from the matching.

We reported lifetime incremental costs, QALYs and INBs (£20,000 per QALY) of DrotAA versus control. Statistical uncertainty was considered by reporting 95% confidence intervals (CIs), and cost-effectiveness acceptability curves (CEACs) (Fenwick et al., 2004). For variance estimation, the non-parametric bootstrap (Davison and Hinkley, 1997) was used to maintain correlation between incremental costs and QALYs. After all methods except for regression alone, inferences should be regarded as conditional on the estimated PS and, for the matching estimators, inferences were also conditional on the matched data (Hill and Reiter, 2005). For all statistical analyses the R platform was used (R Development Core Team, 2011).

Case study results

Before adjustment, potential confounders were highly imbalanced between the treatment groups; compared with controls, DrotAA patients were on average younger, with a higher baseline probability of death (Table 5.1). Following PS matching and IPTW, standardised differences somewhat decreased, but PS matching and particularly IPTW still reported high imbalances for important potential confounders such as the proportion of patients with five organ systems failing, and the baseline probability of death (IMprob).

[11] Standardised differences are weighted using matching frequency weights and IPT weights.

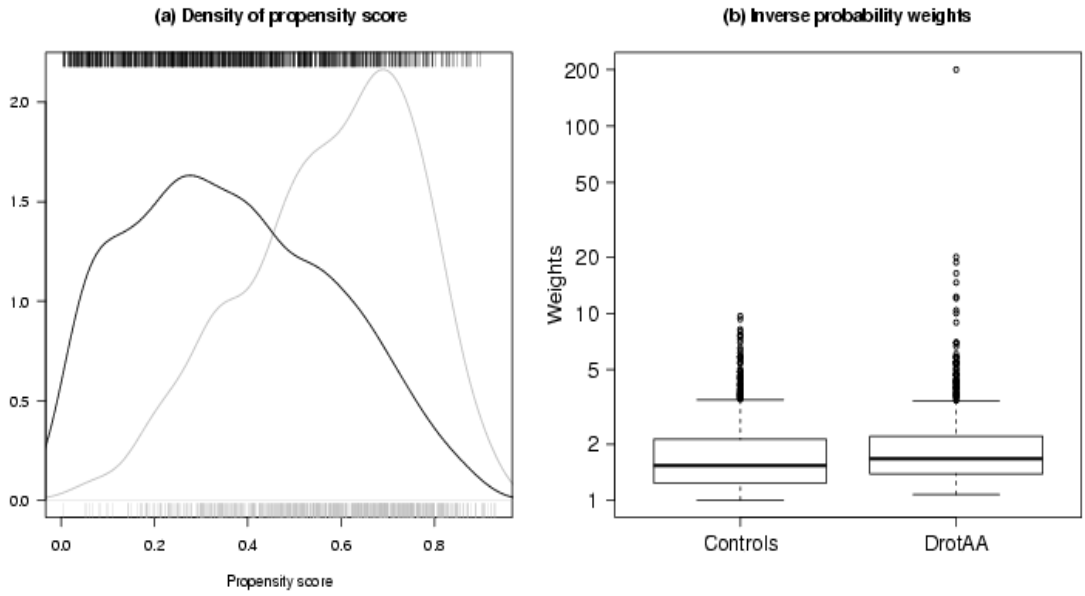
Table 5.1 - Case study results: baseline characteristics and covariate balance for DrotAA versus control group, before and then after matching or weighting

Covariate	Unadjusted			PS matching	IPTW
	DrotAA (n=878)	Control (n=1020)	Standardised difference (%)	Standardised difference (%)	Standardised difference (%)
Age	58.96	65.16	32.32	1.25	7.24
IMprob	0.64	0.58	20.12	6.82	9.86
IMscore	32.08	27.96	40.83	3.95	12.41
% vent.	93.39	78.53	38.90	2.48	13.19
Medical history:					
Cardiovascular	0.34	1.86	13.58	7.05	5.08
Respiratory	1.60	2.94	7.78	1.53	1.30
Renal	0.91	1.47	4.39	0.97	2.45
Liver	0.68	1.77	8.70	0.17	29.30
	7.29	12.75	15.55		
Immunosuppressed				3.93	4.33
Number of organ systems failing:					
3	49.09	62.94	22.88	2.98	4.60
4	41.12	29.31	20.07	9.3	1.98
5	9.80	7.75	5.82	10.62	11.55

Abbreviations: IMprob: ICNARC model predicted probability of acute hospital mortality, IMscore: ICNARC model physiology score, %vent: % of patients mechanically ventilated

A comparison of the distribution of the estimated PSs shows good overlap between the treatment groups, however there are some values close to 0 and 1 (e.g. minimum 0.014 for DrotAA), resulting in unstable IPT weights (Figure 5.1).

Figure 5.1 - Panel (a): Distribution of the estimated PSs for DrotAA (grey line) and control (black line) observations. Panel (b): IPT weights for DrotAA and control observations in the case study.



Notes: The rug plots, at the top and bottom of panel (a), shows the values of the PS.

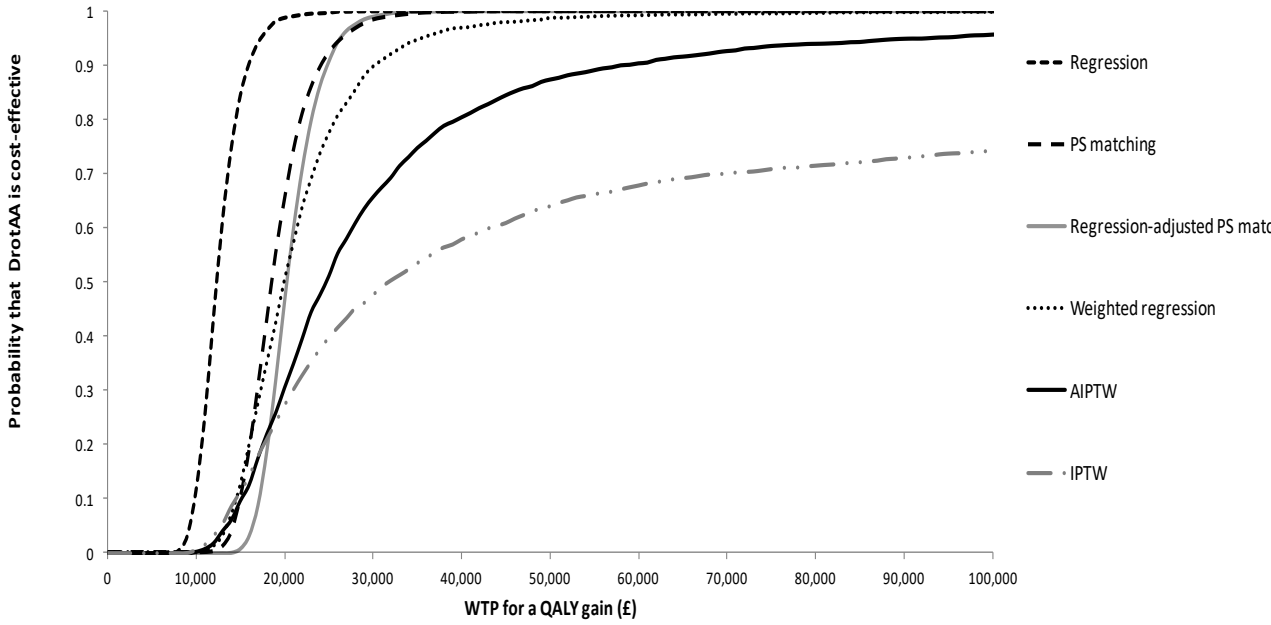
After regression adjustment, the INB was large, with 95% CI that excluded zero. After PS matching and IPTW, the point estimates were much lower, with CIs that included zero, IPTW reporting the widest CIs (Table 5.2). Combined methods reported somewhat differing point estimates, and narrower CIs than their PS method counterparts. AIPTW reported narrower CIs than IPTW, but less certain estimates than the other combined approaches. The probability that DrotAA is cost-effective at the WTP threshold of £20,000 per QALY is 30% following IPTW versus 90% for regression (Figure 5.2). The corresponding probability that DrotAA is cost-effective is around 50% following weighted regression and regression-adjusted matching.

Table 5.2 - Case study: Lifetime incremental costs (£), QALYs and INBs (WTP=£20,000) for DrotAA versus control group.

	Incremental costs (£) (95% CI*)	Incremental QALYs (95% CI*)	INBs (£) (95% CI*)
Regression	19,390 (16,086 to 22,084	1.47 (0.98 to 1.93)	9997 (762 to 19,353)
PS matching	19,384 (17843 to 21,061)	0.99 (0.65 to 1.33)	391 (-6,413 to 6,911)
IPTW	19,023 (15946 to 22,506)	0.54 (-0.97 to 1.44)	-8175 (-36,763 to 9,547**)**
AIPTW	20,262 (16,723 to 25,026)	0.77 (-0.01 to 1.36)	-4861 (-22,251 to 7887)
Weighted regression	19,728 (16587 to 22,951)	0.96 (0.37 to 1.48)	-430 (-13,304 to 9,920)
Reg. -adjusted matching	19,705 (18012 to 21,286)	0.93 (0.68 to 1.18)	-1147 (-6,072 to 3,820)

Notes: * The non-parametric bootstrap is used for all CIs. After all methods except for regression alone, inferences should be regarded as conditional on the estimated PS and, for matching methods, on the matched data.

Figure 5.2 - Cost-effectiveness acceptability curves for DrotAA versus control groups in the case study



In summary, while the estimated incremental costs were similar across approaches, there were differences in the incremental QALYs leading the estimated INBs to differ somewhat by method. For each PS method, adding regression led to narrower CIs. To assess the relative bias and efficiency of the methods in a CEA context, we now use salient features of this case study, such as nonlinear relationships between covariates and endpoints, and unstable IPT weights in the subsequent Monte Carlo simulation study.

Monte Carlo simulation study

Simulation overview

Monte Carlo simulations were conducted to examine the relative performance of single and combined methods, for estimating cost-effectiveness. We extended previous simulation studies (Kang and Schafer, 2007, Porter et al., 2011) to generate cost-

effectiveness data that reflected the settings typified by the case study. In particular, key features of the case study were that there were estimated PSs close to 0 and 1 i.e. unstable IPT weights, and some continuous confounders were modelled nonlinearly in the PS model (Rowan et al., 2008). The study also built on other simulation studies which compared alternative single PS approaches in estimating incremental effectiveness (Austin, 2009b, Radice et al., 2012) and cost-effectiveness (Kreif et al., 2012a).

The simulation study aimed to investigate the following main hypotheses:

1. Combined methods can increase precision and reduce bias compared to single PS methods only, even if the PS is correctly specified. Regression adjustment using a correctly specified model is expected to correct for finite sample imbalances and therefore reduce finite sample bias, and increase efficiency compared to PS matching alone (Rubin and Thomas, 2000). Adding a misspecified regression model after applying PS matching can also reduce finite sample bias (Abadie and Imbens, 2011). With stable weights, IPTW can be consistent and asymptotically efficient (Hirano et al., 2003), but with unstable weights IPTW is expected to be inefficient (Kang and Schafer, 2007). Under such circumstances AIPTW and weighted regression are expected to increase efficiency, by stabilising IPT weights (Glynn and Quinn, 2010).

2. Compared to using regression alone, combined methods can reduce bias due to the misspecification of the regression models. The DR property ensures that AIPTW and weighted regression with correctly specified IPT weights can protect from bias due to a misspecified regression model. Using PS matching for balancing the data before regression adjustment can reduce bias compared to regression alone (Dehejia and Wahba, 2002, Ho et al., 2007), even when the PS is misspecified (Busso et al., 2011).

3. When IPT weights are unstable, regression-adjusted matching can be less sensitive to PS misspecification than weighted regression or AIPTW. Weighted regression and AIPTW are expected to be sensitive to unstable IPT weights and misspecification of the PS (Basu et al., 2011, Kang and Schafer, 2007, Freedman and Berk, 2008). Here regression-adjusted matching might perform better (Busso et al., 2011) than the DR methods considered. AIPTW can report higher bias than weighted regression when both models are misspecified (Kang and Schafer, 2007).

We considered four scenarios (Table 5.3). We assumed a PS mechanism that generates stable IPT weights (Scenario 1 and 2) and one that generates unstable weights (Scenario 3 and 4). We considered a "mild" (Scenario 1) and "major" (Scenario 2 and 3) misspecification of the PS and regression models. In Scenario 4 we also considered a further regression misspecification where the wrong link function was chosen. In each scenario, four different settings were considered: when the PS and regression models were correct (a), when one of the two was misspecified (b and c), and when neither model was correctly specified (d).

Bias and RMSE were obtained to provide information about the accuracy and the precision of the estimated incremental costs, effectiveness and INB across the methods. Relative bias was calculated as the difference between the true parameter value and the mean of the estimated parameter, expressed as a percentage of the true value. The RMSE was taken as the square root of the mean squared differences between the true and estimated parameter values.

Table 5.3 - Monte Carlo simulations, summary of scenarios

Scenario 1: Stable IPT weights + "mild" misspecification		
	Regression correctly specified	Regression misspecified
PS correctly specified	1a	1b
PS misspecified	1c	1d
Scenario 2: Stable IPT weights + "major" misspecification		
	Regression correctly specified	Regression misspecified
PS correctly specified	2a (= 1a)	2b
PS misspecified	2c	2d
Scenario 3: Unstable IPT weights + "major" misspecification		
	Regression correctly specified	Regression misspecified
PS correctly specified	3a	3b
PS misspecified	3c	3d
Scenario 4: Unstable IPT weights + "major" misspecification (as in Scenario 3, but regression misspecification also includes log instead of identity link for costs)		
	Regression correctly specified	Regression misspecified
PS correctly specified	4a (= 3a)	4 b
PS misspecified	4c (= 3c)	4d

Data generating process

1000 CEA datasets were simulated, each with 2000 subjects. For each subject, continuous confounders (Z_1, Z_2, Z_3, Z_4) were generated from bivariate normal distributions with the following means, standard deviations and correlation:

$$\begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix} \sim N\left\{\begin{pmatrix} 2 \\ 4 \end{pmatrix}, \begin{pmatrix} 1 & 0.2 \\ 0.2 & 1 \end{pmatrix}\right\}$$

$$\begin{pmatrix} Z_3 \\ Z_4 \end{pmatrix} \sim N\left\{\begin{pmatrix} 2 \\ 4 \end{pmatrix}, \begin{pmatrix} 1 & 0.2 \\ 0.2 & 1 \end{pmatrix}\right\}.$$

The treatment indicator t was randomly generated from a Bernoulli distribution with parameter p (the PS), determined by a logistic model with a nonlinear term in the logit, as seen in the case study:

$$logit(p_{Scen\ 1,2}) = 0.4 - 1Z_1 + 0.5Z_2 + 0.025{Z_2}^2 - 0.25Z_3 - 0.1Z_4 \qquad (3)$$

Costs and effectiveness endpoints were drawn from a bivariate gamma-normal distribution, with a copula function[12] (Trivedi and Zimmer, 2005, Mihaylova et al., 2010, Quinn, 2007). The dependence between the two endpoints was recognised by setting the correlation between the error terms equal to 0.4[13]. The effectiveness endpoint was drawn from a normal distribution:

$$Y_E \sim N(\mu_E, 0.2)\ ,$$

and costs from a gamma distribution with identity link, shape and scale parameters[14]:

$$Y_C \sim \Gamma(10,\ \mu_C/10).$$

The relationship between the covariates with μ_C and μ_E was assumed linear, with a constant treatment effect, as:

$$\mu_C\ = 10000 + 6000t - 2000Z_1 + 2000Z_2 - 2000Z_3 + 2000Z_4\ ,$$

$$\mu_E = 9 + 0.4t + 0.1Z_1 - 0.05Z_2 + 0.05Z_3 - 0.05Z_4\ . \qquad (4)$$

Simulation scenarios

Scenarios 1 and 2 considered stable IPT weights (Figure 5.3). Scenarios 3 and 4 used unstable weights similar to the case study, in that a large portion of the true PSs were

[12] The copula function can generate draws from a flexible multivariate distribution (in this case the bivariate) with different marginal distributions (here, the gamma and the normal).
[13] This resulted in a correlation of 0.34 between the cost and QALY variable, which reflects the correlation (0.22) found in the case study.
[14] The choice of normal distribution for Y_E and the identity link function for Y_C was made for transparency reasons and to facilitate replication.

close to 0 and 1 (Figure 5.3). This was achieved by modifying the coefficients in the PS model (equation (3)) as:

$$logit\left(p_{Scen\,3,4}\right) = 1.5 - 2Z_1 + 1Z_2 + 0.05{Z_2}^2 - 0.5Z_3 - 0.2Z_4.$$

In each scenario a common functional form misspecification for the PS and the endpoint model was defined as in Kang and Schafer (2007): we assumed that instead of the true confounders Z_1 to Z_4, their nonlinear functions, X_1 to X_4 were observed, both when modelling the PS and the endpoints. Scenario 1 had a "mild" functional form misspecification, with the nonlinear functions defined as:

$$\begin{aligned} X_1 &= \exp\left(\frac{Z_1}{10}\right), \\ X_2 &= Z_2 * (1 + Z_1) + 10, \\ X_3 &= (Z_3/25 + 0.6)^2, \\ X_4 &= (Z_4 + 20)^2. \end{aligned}$$

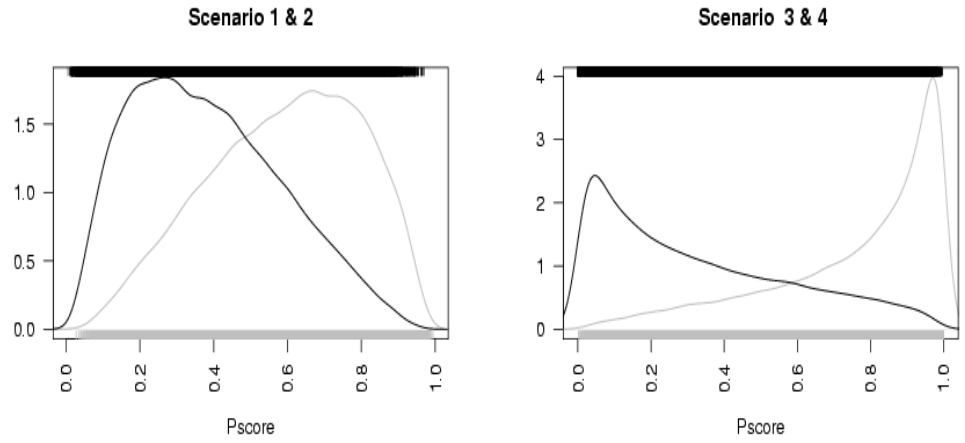
In addition, when the PS was misspecified, the squared term was omitted from the logistic regression. Scenarios 2-4 had "major" misspecifications of the PS and endpoint models. Here, the misspecification of X_1 and X_4 was increased:

$$\begin{aligned} X_1 &= \exp\left(\frac{Z_1}{3}\right), \\ X_4 &= (Z_2/10 + Z4 + 20)^2. \end{aligned}$$

Scenario 4 extended scenario 3, here the cost regression model had a further misspecification in that a log rather than the correct identity link function (equation 4) was chosen. This setting was again motivated by the case study[15].

[15] The proportion of individuals in the treatment group were typically around 50% (scenarios 1 and 2) and 60% (scenarios 3 and 4), compared with 46% in the case study.

Figure 5.3 - Densities of the true PS in the Monte Carlo simulation study, using data from a typical sample (n = 1,00,000) for treated (grey line) and control (black line). The rug plots, at the top and bottom of each graph show the values of the PS.



Simulation results

Table 5.4 reports the relative bias (%) and RMSE of the estimated INB, over 1000 replications for the different scenarios. Tables containing simulation results for the incremental costs and QALYs are presented in Appendix 5.1.

When the IPT weights were stable and both the PS and the endpoint models were correctly specified (scenario 1a), all methods provided relatively low bias, with regression reporting the lowest RMSE. In scenario 2d with stable IPT weights and a "major" misspecification of the regression and the PS models, PS matching and regression-adjusted matching performed best, with lower bias and RMSE then DR methods (relative bias of 11% after regression-adjusted matching versus 53% after AIPTW and 29% after weighted regression).

Under unstable IPT weights (scenario 3), combined methods outperformed correctly specified PS methods. In general, DR methods reported lower bias and RMSE than IPTW, and regression-adjusted matching performed better than PS matching alone.

These combined methods performed better even when a misspecified regression model was applied, after matching or weighting with a correctly specified PS (scenario 3b). Under dual misspecification (scenario 3d), all methods reported bias. Regression-adjusted matching again provided estimates with the lowest relative bias (20% versus 50% with regression and 89% with IPTW) and RMSE.

When the regression for the cost endpoint used a log rather than the correct identify link function (scenario 4d), PS matching reported slightly lower bias than regression-adjusted matching, however regression-adjusted matching again reported the lowest RMSE.

Across the scenarios, regression-adjusted matching reported lower RMSE than PS matching alone, and in most scenarios lower bias. DR methods always reduced RMSE compared to IPTW only, and under unstable IPT weights they also reduced bias. However in these scenarios (3d and 4d), both DR methods were outperformed by regression-adjusted matching.

Table 5.4 - Monte Carlo simulations results: relative bias and RMSE for the INBs (WTP=£20,000)

Scenarios	1: Stable IPT weights, "mild" misspec.		2: Stable IPT weights, "major" misspec.		3: Unstable IPT weights, "major" misspec.		4: Unstable IPT weights, "major" misspec. (link function misspecified)	
	Relative bias	**RMSE**	**Relative bias**	**RMSE**	**Relative bias**	**RMSE**	**Relative bias**	**RMSE**
a: PS and regression correctly specified								
Regression	0.2%	330	0.2%	330	0.9%	376	0.9%	376
PS matching	1.9%	439	1.9%	439	9.4%	860	9.4%	860
IPTW	0.0%	402	0.0%	402	3.1%	1372	3.1%	1372
AIPTW	0.2%	348	0.2%	348	1.0%	795	1.0%	795
Weighted regression	0.2%	348	0.2%	348	0.7%	660	0.7%	660
Reg.-adjusted PS matching	0.3%	406	0.3%	406	1.0%	735	1.0%	735
b: PS correct and regression misspecified								
Regression	11.1%	396	34.8%	770	51.3%	1087	52.0%	1102
PS matching	1.9%	439	1.9%	439	9.4%	860	9.4%	860
IPTW	0.0%	402	0.0%	402	3.1%	1372	3.1%	1372
AIPTW	0.2%	361	0.1%	381	2.3%	872	2.1%	875
Weighted regression	0.2%	359	0.3%	377	2.1%	735	10.8%	725
Reg.-adjusted PS matching	0.3%	408	0.1%	414	3.3%	748	12.2%	760
c: PS misspecified and regression correct								
Regression	0.2%	330	0.2%	330	0.9%	376	0.9%	376
PS matching	6.1%	440	12.4%	485	26.0%	897	26.0%	897
IPTW	4.7%	621	44.4%	2161	88.6%	4069	88.6%	4069
AIPTW	0.4%	370	1.1%	784	2.6%	2105	2.6%	2105
Weighted regression	0.2%	359	0.2%	445	0.7%	863	0.7%	863
Reg.-adjusted PS matching	0.2%	393	0.0%	400	0.8%	684	0.8%	684
d: PS misspecified and regression misspecified								
Regression	11.1%	396	34.8%	770	51.3%	1087	52.0%	1102
PS matching	6.1%	440	12.4%	485	26.0%	897	26.0%	897
IPTW	4.7%	621	44.4%	2161	88.6%	4069	88.6%	4069
AIPTW	10.3%	446	53.6%	1966	69.0%	3012	79.5%	3140
Weighted regression	9.4%	409	29.0%	713	34.2%	986	44.2%	1165
Reg.-adjusted PS matching	4.6%	408	11.2%	459	19.5%	769	30.2%	881

Notes: The RMSE was taken as the square root of the mean squared differences between the true and estimated parameter values.

Discussion

This paper finds that regression-adjusted matching can reduce overt selection bias and increase precision in the estimates of ATEs versus alternative approaches such as regression, PS matching and IPTW, across a range of scenarios typical of CEA. We show that regression-adjusted PS matching is relatively insensitive to functional form misspecifications of both the PS and the regression models, even when there are positivity violations, i.e. estimated PS values close to 0 and 1, resulting in unstable IPT weights. Here, both DR approaches reported higher bias and RMSE than regression-adjusted matching. In the case study, the cost-effectiveness estimates and the reported CIs differed across methods. The differences in the cost-effectiveness estimates were driven by the estimated incremental QALY which may reflect imbalances in important potential confounders that remained after PS matching, and especially after IPTW.

This paper builds on previous simulation studies that compared alternative methods for addressing overt selection bias in cost and cost-effectiveness analysis (Basu et al., 2011, Sekhon and Grieve, 2011, Kreif et al., 2012a). Our paper considers combined methods for the first time in a CEA setting. This setting provoked the new simulation scenarios considered. The case study had typical characteristics of CEA that use observational data (Kreif et al., 2012a) and helped ground the simulation study. While the example used previously published PS models, and regression models were selected based on model fit, there were concerns about model misspecification and unstable IPT weights. Hence the simulation scenarios most relevant to applied studies are when both the PS and the regression models are misspecified. We find that even with dual misspecification, weighted regression and regression-adjusted matching can reduce bias and RMSE compared to using a single method alone. Amongst the combined methods

considered, regression-adjusted matching appears the least sensitive to misspecification, and this advantage was most apparent with unstable IPT weights.

Our simulation scenarios considered settings characteristic of CEA. Here, while GLM approaches have been recommended (Barber and Thompson, 2004, Glick et al., 2007), the analyst has to choose the correct functional form of the linear predictors and the appropriate link function; in practice these are seldom known. We found that using PS matching in conjunction with GLMs can mitigate bias due to either misspecification. To help analysts consider these approaches further in different settings, we append R and Stata code for each combined method (see Appendix 5.2), and the code for simulating the data (Appendix 5.3).

Our work also builds on previous simulation studies in the more general methodological literature (Kang and Schafer, 2007, Porter et al., 2011, Basu et al., 2011) by considering a range of scenarios with stable and unstable IPT weights. Similarly to previous studies (Kang and Schafer, 2007, Basu et al., 2011, Radice et al., 2012) we find that IPTW can be inefficient under unstable IPT weights, even if the PS is correctly specified, and can be highly biased due to the misspecification of the PS. In contrast to Kang and Schafer (2007), we find that under dual misspecification and unstable IPT weights, weighted regression and regression-adjusted matching can outperform regression. The current paper finds that regression-adjusted matching, an estimator not considered in previous comparisons (Basu et al., 2011), can be relatively robust to misspecification of the PS and the regression models. Related simulation work reports that another implementation of matching combined with regression, "bias-corrected matching" (Abadie and Imbens, 2011) also performed well compared to DR methods (Busso et al., 2011).

This work has some limitations. The methods and the simulation settings all assume no unobserved confounding. This paper could not consider all circumstances that may arise in practice; for example in CEA, further endpoint model misspecifications can arise with quality of life data (Basu and Manca, 2011). To aid transparency we considered relatively simple matching and regression estimators. Rather than nearest neighbour PS matching one could use other multivariate matching approaches such as Genetic Matching (GM), (Diamond and Sekhon, 2012, Sekhon, 2011), previously demonstrated to reduce selection bias in a range of applications (Grieve et al., 2008, Sekhon and Grieve, 2011, Radice et al., 2012, Kreif et al., 2012a). More flexible regression models could also be compared, including extended estimating equations (Basu et al., 2011) or beta-type size distributions (Jones et al., 2011), which can outperform GLMs. These methods can also be combined with PS methods such as matching.

This work also opens up areas for further research. Further studies could consider alternative DR methods such as targeted-maximum likelihood estimation (van der Laan, 2010, van der Laan and Gruber, 2010, Gruber and van der Laan, 2010). When IPT weights are unstable, these approaches can perform better than the DR approaches considered, but they have not been compared to the regression-adjusted matching estimators described (Porter et al., 2011). New developments in machine learning methods for the estimation of the PS and the endpoint regression (Lee et al., 2010, Austin, 2012, van der Laan, 2007, Westreich et al., 2010) can further reduce bias due to functional form misspecification. These methods warrant careful consideration in further simulation studies relevant to health economic evaluations.

Acknowledgements

We thank Zia Sadique (LSHTM) for help in the motivating case study, Roland Ramsahai (University of Cambridge) for valuable comments on the Monte Carlo simulations, Manuel Gomes, Karla Diaz-Ordaz, Adam Steventon, Rhian Daniel (all LSHTM) and Susan Gruber (Harvard School of Public Health) for comments on the manuscript. We also thank David Harrison and Kathy Rowan (ICNARC) for access to the data used in the case study. Funding from the Economic and Social Research Council (Grant no. RES-061-25-0343) is greatly appreciated.

References

Abadie, A., Drukker, D., Herr, J. L. & Imbens, G. 2004a. Implementing matching estimators for average treatment effects in Stata. The Stata Journal 4, 4, 290-311.

Abadie, A., Herr, J. L., Imbens, G. W. & Drukker, D. M. 2004b. NNMATCH: Stata module to compute nearest-neighbor bias-corrected estimators [Online]. Boston College Department of Economics. Available: http://fmwww.bc.edu/repec/bocode/n/nnmatch.hlp.

Abadie, A. & Imbens, G. W. 2006. Large Sample Properties of Matching Estimators for Average Treatment Effects. Econometrica, 74, 235-267.

Abadie, A. & Imbens, G. W. 2011. Bias-Corrected Matching Estimators for Average Treatment Effects. Journal of Business & Economic Statistics, 29, 1-11.

Austin, P. C. 2008. A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003. Statistics in Medicine, 27, 2037-2049.

Austin, P. C. 2009a. Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. Statistics in Medicine, 28 3083-3107.

Austin, P. C. 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, 661-677.

Austin, P. C. 2012. Using Ensemble-Based Methods for Directly Estimating Causal Effects: An Investigation of Tree-Based G-Computation. Multivariate Behavioral Research, 115-135.

Bang, H. & Robins, J. M. 2005. Doubly Robust Estimation in Missing Data and Causal Inference Models. Biometrics 962–972.

Bang, H. & Tsiatis, A. A. 2000. Estimating medical costs with censored data. Biometrica 87, 329-343.

Barber, J. & Thompson, S. G. 2004. Multiple regression of cost data: use of generalised linear models. Journal of Health Services Research Policy, 9, 197-204.

Basu, A. 2011. Economics of individualization in comparative effectiveness research and a basis for a patient-centered health care. Journal of Health Economics, 30, 549-559.

Basu, A., Heckman, J. J., Navarro-Lozano, S. & 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, 1133-1157.

Basu, A. & Manca, A. 2011. Regression Estimators for Generic Health-Related Quality of Life and Quality-Adjusted Life Years. Med Decis Making, 2011 Oct 18. [Epub ahead of print].

Basu, A., Polsky, D. & Manning, W. 2011. Estimating treatment effects on healthcare costs under exogeneity: is there a 'magic bullet'? Health Services and Outcomes Research Methodology, 11, 1-26.

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

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

Busso, M., DiNardo, J. & McCrary, J. 2011. New Evidence on the Finite Sample Properties of Propensity Score Reweighting and Matching Estimators. Working paper.

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

Davison, A. & Hinkley, D. 1997. Bootstrap Methods and Their Application, New York, Cambridge University Press.

Dehejia, R. H. & Wahba, S. 2002. Propensity Score-Matching Methods For Nonexperimental Causal Studies. The Review of Economics and Statistics, 84, 151-161.

Diamond, A. & Sekhon, J. S. 2012. Genetic Matching for Estimating Causal Effects: A General Multivariate Matching Method for Achieving Balance in Observational Studies. Review of Economic and Statistics, Forthcoming.
Fenwick, E., O'Brien, B. & Briggs, A. 2004. Cost-effectiveness acceptability curves--facts, fallacies and frequently asked questions. Health Economics, 13, 405-415.
Freedman, D. & Berk, R. A. 2008. Weighting regression by propensity score. Evaluation Review, 32, 392-409.
Funk, M. J., Westreich, D., Wiesen, C., Stürmer, T., Brookhart, M. A. & Davidian, M. 2011. Doubly Robust Estimation of Causal Effects. American Journal of Epidemiology, 173, 761-767.
Glick, H., Doshi, J., Sonnad, S. & Polsky, D. 2007. Economic evaluation in clinical trials, Oxford, Oxford University Press.
Glynn, A. N. & Quinn, K. M. 2010. An Introduction to the Augmented Inverse Propensity Weighted Estimator. Political Analysis, 18, 36-56.
Gray, A. M., Clarke, P. M., Wolstenholme, J. L. & Wordsworth, S. 2010. Applied Methods of Cost-Effectiveness Analysis in Healthcare, Oxford University Press.
Grieve, R., Sekhon, J. S., Hu, T.-w. & 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, 1204-1222.
Gruber, S. & van der Laan, M. J. 2010. An application of collaborative targeted maximum likelihood estimation in causal inference and genomics. The International Journal of Biostatistics, 6.
Hill, J. & Reiter, J. P. 2005. Interval estimation for treatment effects using propensity score matching. Statistics in Medicine, 25.
Hirano, K. & Imbens, G. W. 2001. Estimation of Causal Effects using Propensity Score Weighting: An Application to Data on Right Heart Catheterization. Health Services and Outcomes Research Methodology, 2, 259-278.
Hirano, K., Imbens, G. W. & Ridder, G. 2003. Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score. Econometrica, 71, 1161-1189.
Ho, D. E., Imai, K., King, G. & Stuart, E. A. 2007. Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference Political Analysis, 15, 199-236.
Imbens, G. M. & Wooldridge, J. M. 2009a. Recent Developments in the Econometrics of Program Evaluation. Journal of Economic Literature, 47, 5–86.
Jones, A., Lomas, J. & Rice, N. 2011. Applying Beta-type Size Distributions to Healthcare Cost Regressions. HEDG Working Papers. HEDG, c/o Department of Economics, University of York.
Jones, A. M. 2010. Models For Health Care. HEDG Working Papers. HEDG, c/o Department of Economics, University of York.
Kang, J. D. Y. & Schafer, J. L. 2007. Demystifying Double Robustness: A Comparison of Alternative Strategies for Estimating a Population Mean from Incomplete Data. Statistical Science, 22, 523-539.
Kreif, N., Grieve, R., Radice, R., Sadique, Z., Ramsahai, R. & Sekhon, J. S. 2012a. Methods for Estimating Subgroup Effects in Cost-Effectiveness Analyses That Use Observational Data. Medical Decision Making, 32, 750-63.
Kreif, N., Grieve, R. & Sadique, Z. 2012b. Statistical methods for cost-effectiveness analyses that use observational data: a critical appraisal tool and review of current practice. Health Economics, DOI: 10.1002/hec.2806.
Lee, B. K., Lessler, J. & Stuart, E. A. 2010. Improving propensity score weighting using machine learning. Statistics in Medicine, 29, 337-346.
Lunceford, J. K. & Davidian, M. 2004. Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. Statistics in Medicine, 23, 2937-2960.
Manning, W. G., Basu, A. & Mullahy, J. 2005. Generalized modeling approaches to risk adjustment of skewed outcomes data. Journal of Health Economics, 24, 465-488.

Mihaylova, B., Briggs, A., O'Hagan, A. & Thompson, S. 2010. Review of statistical methods for analysing healthcare resources and costs. Health Economics, DOI: 10.1002/hec.1653.
NICE. 2008. Guide to the Methods of Technology Appraisal [Online]. Available: http://www.nice.org.uk/media/B52/A7/TAMethodsGuideUpdatedJune2008.pdf [Accessed 24/10/2010.
Nixon, R. M. & Thompson, S. G. 2005. Methods for incorporating covariate adjustment, subgroup analysis and between-centre differences into cost-effectiveness evaluations. Health Economics, 14, 1217-1229.
Pan, W. & Zeng, D. 2011. Estimating Mean Cost Using Auxiliary Covariates. Biometrics, 67, 996-1006.
Pearl, J. 1995. Causal Diagrams for Empirical Research. Biometrika, 82 669-688.
Porter, K. E., Gruber, S., Laan, M. J. v. d. & Sekhon, J. S. 2011. The Relative Performance of Targeted Maximum Likelihood Estimators. The International Journal of Biostatistics, 7.
Quinn, C. 2007. The health-economic applications of copulas: methods in applied econometric research [Online]. HEDG, c/o Department of Economics, University of York. Available: http://ideas.repec.org/p/yor/hectdg/07-22.html [Accessed 10/08/2011.
R Development Core Team 2011. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing.
Radice, R., Grieve, R., Ramsahai, R., Kreif, N., Sadique, Z. & 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).
Robins, J., Rotnitzky, A. & Zhao, L. P. 1994. Estimation of regression coefficients when some regressors are not always observed. Journal of the American Statistical Association, 89, 846-866.
Robins, J., Sued, M., Lei-Gomez, Q. & Rotnitzky, A. 2007. Comment: Performance of Double-Robust Estimators When "Inverse Probability" Weights Are Highly Variable. Statistical Science, 22, 544-559.
Robins, J. M., Rotnitzky, A. & Zhao, L. P. 1995. Analysis of Semiparametric Regression Models for Repeated Outcomes in the Presence of Missing Data. Journal of the American Statistical Association, 90, 106-121.
Rosenbaum , P. R. & Rubin, D. B. 1983. The central role of the propensity score in observational studies for causal effects. Biometrika, 70, 41-55.
Rowan, K., Welch, C., North, E. & Harrison, D. 2008. Drotrecogin alfa (activated): real-life use and outcomes for the UK. Critical Care, 12.
Rubin, D. B. 1973. The Use of Matched Sampling and Regression Adjustment to Remove Bias in Observational Studies. Biometrics, 29.
Rubin, D. B. 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. B. & Thomas, N. 2000. Combining propensity score matching with additional adjustments for prognostic covariates. Journal of the American Statistical Association, 95, 573–585.
Sadique, M. Z., Grieve, R., Harrison, D., Cuthbertson, B. & Rowan, K. 2011. Is Drotrecogin alfa (activated) for adults with severe sepsis, cost-effective in routine clinical practice? Critical Care, 15, R228.
Sekhon, J. S. 2011. Matching: multivariate and propensity score matching with automated balance search. Journal of Statistical Software.
Sekhon, J. S. & Grieve, R. D. 2011. A Matching Method for Improving Covariate Balance in Cost-Effectiveness Analyses. Health Economics, doi: 10.1002/hec.1748.
StataCorp 2011. Stata Statistical Software: Release 12. College Station, TX: StataCorp LP.
Stuart, E. A. 2010. Matching Methods for Causal Inference: A review and a look forward. Statistical Science, 25.
Trivedi, P. K. & Zimmer, D. M. 2005. Copula Modeling: An Introduction to Practitioners, Delft, Now Publishing Inc.

van der Laan, M. J. 2010. Targeted Maximum Likelihood Based Causal Inference: Part I. The International Journal of Biostatistics.
van der Laan, M. J. & Gruber, S. 2010. Collaborative Double Robust Targeted Maximum Likelihood Estimation. The International Journal of Biostatistics 6.
van der Laan, M. J. P., Eric C.; and Hubbard, Alan E. 2007. Super Learner Statistical Applications in Genetics and Molecular Biology: , Vol. 6
Westreich, D., Lessler, J. & Funk, M. 2010. Propensity score estimation: neural networks, support vector machines, decision trees (CART), and meta-classifiers as alternatives to logistic regression. Journal of Clinical Epidemiology, 63, 826-33.
Willan, A. R. & Briggs, A. H. 2006. Statistical Analysis of Cost-effectiveness Data, John Wiley & Sons Ltd.

Appendix 5.1 - Additional tables for research paper 3

Appendix 5.1 Table 1 - Monte Carlo simulation results: relative bias and RMSE of the estimated incremental cost

Scenarios	1: Stable IPT weights, "mild" misspec.		2: Stable IPT weights, "major" misspec.		3: Unstable IPT weights, "major" misspec.		4: Unstable IPT weights, "major" misspec (link function misspecified)	
	Relative bias	**RMSE**	**Relative bias**	**RMSE**	**Relative bias**	**RMSE**	**Relative bias**	**RMSE**
a: PS and regression correctly specified								
Regression	0.1%	341	0.1%	341	0.3%	390	0.3%	390
PS matching	0.5%	433	0.5%	433	1.8%	826	1.8%	826
IPTW	0.1%	380	0.1%	380	0.8%	1004	0.8%	1004
AIPTW	0.1%	358	0.1%	358	0.4%	821	0.4%	821
Weighted regression	0.1%	357	0.1%	357	0.2%	685	0.2%	685
Reg.-adjusted PS matching	0.2%	416	0.2%	416	0.2%	768	0.2%	768
b: PS correct and regression misspecified								
Regression	2.5%	372	6.9%	534	10.0%	702	10.2%	718
PS matching	0.5%	433	0.5%	433	1.8%	826	1.8%	826
IPTW	0.1%	380	0.1%	380	0.8%	1004	0.8%	1004
AIPTW	0.1%	365	0.1%	372	0.7%	822	0.6%	827
Weighted regression	0.1%	363	0.1%	370	0.5%	719	3.4%	699
Reg.-adjusted PS matching	0.2%	416	0.2%	419	0.7%	765	3.7%	762
c: PS misspecified and regression correct								
Regression	0.1%	341	0.1%	341	0.3%	390	0.3%	390
PS matching	1.6%	437	2.3%	442	4.7%	723	4.7%	723
IPTW	0.5%	485	7.1%	1347	14.6%	2812	14.6%	2812
AIPTW	0.0%	379	0.3%	776	1.3%	2071	1.3%	2071
Weighted regression	0.1%	369	0.1%	457	0.3%	891	0.3%	891
Reg.-adjusted PS matching	0.2%	413	0.1%	415	0.3%	723	0.3%	723
d: PS misspecified and regression misspecified								
Regression	2.5%	372	6.9%	534	10.0%	702	10.2%	718
PS matching	1.6%	437	2.3%	442	4.7%	779	4.7%	779
IPTW	0.5%	485	7.1%	1347	14.6%	2812	14.6%	2812
AIPTW	2.4%	410	10.2%	1273	12.3%	2356	15.9%	2454
Weighted regression	2.2%	389	5.8%	528	6.5%	817	9.8%	974
Reg.-adjusted PS matching	2.5%	372	2.2%	431	3.8%	732	7.4%	799

Appendix 5.1 Table 2 - Monte Carlo simulation results: relative bias and RMSE of the estimated incremental QALYs

Scenarios	1: Stable IPT weights , "mild" misspec.		2: Stable IPT weights, "major" misspec.		3 and 4: Unstable IPT weights, "major" misspec.	
	Relative bias	RMSE	Relative bias	RMSE	Relative bias	RMSE
a: PS and regression correctly specified						
Regression	0.1%	0.01	0.1%	0.01	0.0%	0.01
PS matching	0.1%	0.01	0.1%	0.01	1.0%	0.02
IPTW	0.1%	0.01	0.1%	0.01	0.2%	0.04
AIPTW	0.1%	0.01	0.1%	0.01	0.1%	0.02
Weighted regression	0.1%	0.01	0.1%	0.01	0.0%	0.02
Reg.-adjusted PS matching	0.1%	0.01	0.1%	0.01	0.1%	0.02
b: PS correct and regression misspecified						
Regression	0.9%	0.01	3.5%	0.02	5.3%	0.02
PS matching	0.1%	0.01	0.1%	0.01	1.0%	0.02
IPTW	0.1%	0.01	0.1%	0.01	0.2%	0.04
AIPTW	0.1%	0.01	0.1%	0.01	0.1%	0.03
Weighted regression	0.1%	0.01	0.1%	0.01	0.1%	0.02
Reg.-adjusted PS matching	0.1%	0.01	0.1%	0.01	0.3%	0.02
c: PS misspecified and regression correct						
Regression	0.1%	0.01	0.1%	0.01	0.0%	0.01
PS matching	0.3%	0.01	1.3%	0.01	3.0%	0.02
IPTW	0.8%	0.02	5.8%	0.05	11.2%	0.10
AIPTW	0.1%	0.01	0.0%	0.03	0.3%	0.06
Weighted regression	0.1%	0.01	0.1%	0.01	0.1%	0.03
Reg.-adjusted PS matching	0.2%	0.01	0.0%	0.01	0.0%	0.02
d: PS misspecified and regression misspecified						
Regression	0.9%	0.01	3.5%	0.02	5.3%	0.02
PS matching	0.3%	0.01	1.3%	0.01	3.0%	0.02
IPTW	0.8%	0.02	5.8%	0.05	11.2%	0.10
AIPTW	0.8%	0.01	5.7%	0.05	7.9%	0.08
Weighted regression	0.7%	0.01	2.9%	0.02	3.7%	0.03
Reg.-adjusted PS matching	0.9%	0.01	1.2%	0.01	2.0%	0.02

Appendix 5.2 – Code for implementing the methods in research paper 3

This section provides code for the implementation of the combined statistical approaches proposed in the paper, using the R (R Development Core Team, 2011) and Stata statistical softwares (StataCorp, 2011). The user-written functions implemented here call some pre-written R routines, for example "glm" for generalised linear models, or the "Matching" library (Sekhon, 2011). When implementing the methods in Stata, we use the NNMATCH routine (Abadie et al., 2004b) for matching.

R code for AIPTW

First, the regression models for the cost and effectiveness endpoints need to be defined:

```
rmodel_Y<-glm(Y~tx+Z1+Z2+Z3+Z4,family=gaussian,data=dataset)

rmodel_cost<-glm(cost~tx+Z1+Z2+Z3+Z4,family=gaussian,data=dataset)
```

The IPT weights (`pscw`) need to be created:

```
ps.formula<- as.formula(tx~Z1+Z2+I(Z2^2)+Z3+Z4)

pmodel_func=function(data, formula) {

              pmodel<-glm(formula,family=binomial,data=data)
              pscore.lin<-pmodel$linear.predictor
                     pscore<-pmodel$fitted.values

              pscw=data$tx/pscore+(1-data$tx)/(1-pscore)
              return(cbind(pscore,pscore.lin,pscw)
              rm(pscore,pscore.lin,pscw)
                                          }
        dataset=cbind(dataset,pmodel_func(dataset,ps.formula))
```

The function for AIPTW will take the above objects as in:

```
ate_aipw=function(data, model, weight, endpoint) {
data_new0=data
data_new0$tx=0
data_new1=data
data_new1$tx=1
m0=predict(model, newdata=data_new0, type="response")
m1=predict(model, newdata=data_new1, type="response")
m=predict(model, type="response")
```

```
mu1 <- sum(data$tx * weight * (endpoint-m) )/sum(data$tx
* weight) + mean(m1)
mu0 <- sum((1-data$tx)* weight* (endpoint-m))/sum((1-data$tx) * weight)
+ mean(m0)
              my.ate.aiptw=mu1-mu0
              return(my.ate.aiptw)

                                    }
```

To obtain the estimate for example for the incremental QALYs, we call the above function:

```
aipw_Y=ate_aipw(dataset,rmodel_Y, dataset$pscw,dataset$Y)
```

R code for weighted regression

Weighted GLMs are constructed, using the IPT weights (`pscw`) defined above:

```
wrmodel_Y<glm(Y~tx+Z1+Z2+Z3+Z4,family=gaussian,data=dataset,weight=pscw)
wrmodel_cost<glm(cost~tx+Z1+Z2+Z3+Z4,family=gaussian,data=dataset,weight=pscw)
```

The function ate_reg is called to obtain the estimates of ATE, for example for the incremental QALYs:

```
ate_reg=function(data, model) {
                    data_new0=data
                    data_new0$tx=0
                    data_new1=data
                    data_new1$tx=1
                    m0=predict(model, newdata=data_new0, type="response")
                    m1=predict(model, newdata=data_new1, type="response")
                    mu1 <- mean(m1)
                    mu0 <- mean(m0)
                    my.ate.reg=mu1-mu0
                    return(my.ate.reg)
                                }
wreg_Y=ate_reg(dataset,wrmodel_Y)
```

R code for regression-adjusted matching

We create the matched datasets, using the linear predictor of the estimated PS (pscore.lin) as a proximity measure.

```
PSmatch_data=function(data){
attach(data)
mtchout.Y=Match(Y=Y,Tr=tx,X=cbind(pscore.lin),estimand="ATE")
detach(data)
mtch.data<rbind(data[mtchout.Y$index.treated,],data[mtchout.Y$index.control,])
mtch.data<cbind(mtch.data,weights=c(mtchout.Y$weights,mtchout.Y$weights))
```

```
                return(mtch.data)
                                                }
```

The matched data is stored in the object `mdataset`, where the matching frequency weights are stored as `mtch.data$weights`.

```
            mdataset=PSmatch_data(dataset)
```

The PS matching estimator of ATE is the following:

```
      rmatch_ate=function(data,endpoint,formula){
                  model=glm(formula,family=gaussian,data=data,weights=weights)
                  data_new0=data
                  data_new0$tx=0
                  data_new1=data
                  data_new1$tx=1
                  m0=predict(model, newdata=data_new0, type="response")
                  m1=predict(model, newdata=data_new1, type="response")
                  mu1 <- mean(m1)
                  mu0 <- mean(m0)
                  my.ate.match.reg=mu1-mu0
                  return(my.ate.match.reg)
                                                }
```

This function is called to obtain the estimates of the ATE, for example:

```
rPSmatch_Y=rmatch_ate(mdataset,mdataset$Y,reg.formula_y)
```

Stata code for AIPTW

```
use "L:\mylibrary\statasim.dta", clear
```

First, run an unweighted regression:

```
            reg Y tx Z1 Z2 Z3 Z4

            gen truetx = tx
```

Generate the predicted potential outcomes:

```
            replace tx=1
            predict mu1
            replace tx=truetx

            replace tx=0
            predict mu0
            replace tx=truetx
```

Generate the predicted observed outcome:

```
            predict mu
```

Calculate the mean the predicted potential outcomes:

```
summ mu1
scalar Ytreated=r(mean)

summ mu0
scalar Ycontrol=r(mean)
```

Estimate the PS:

```
logit tx Z1 Z2 Z3 Z4
predict pscorelin, xb
predict pscore, pr
```

Construct IPT weights:

```
gen pscw=1
replace pscw=1/invlogit( pscorelin) if tx==1
replace pscw=1/invlogit(-pscorelin) if tx==0
```

Generate the sum of these weights, in order to obtain a normalised estimator:

```
egen sumpscw1= total(pscw) if tx==1
egen sumpscw0= total(pscw) if tx==0
```

Generate augmented mean potential outcomes:

```
egen mu1Y = total(pscw*(Y-mu)/sumpscw1) if tx==1
egen mu0Y = total(pscw*(Y-mu)/sumpscw0) if tx==0

summ mu1Y
scalar meanmu1Y=r(mean)+ Ytreated
summ mu0Y
scalar meanmu0Y=r(mean) + Ycontrol
scalar aiptw_Y=meanmu1Y-meanmu0Y
di aiptw_
```

Stata code for weighted regression

Again, the method of recycled predictions is used, using regression models weighted with the IPT weights (`pscw`):

```
reg Y tx Z1 Z2 Z3 Z4
reg Y tx Z1 Z2 Z3 Z4 [pweight= pscw]
gen truetx = tx
replace tx=1
predict mu1
replace tx=truetx
replace tx=0
predict mu0
replace tx=truetx
summ mu1
scalar Ytreated=r(mean)
summ mu0
scalar Ycontrol=r(mean)
scalar wreg_Y  = Ytreated-Ycontrol
dis wreg_Y

drop mu1 mu0
scalar drop Ytreated Ycontrol wreg_Y
```

Stata code for regression-adjusted matching

First we create an id which can be used to merge the main dataset the matched data:

```
egen id= seq()
sort id
```

We save the data before the matching:

```
save "L:\mylibrary\statasim1.dta", replace
```

We perform nearest neighbor, 1:1 PS matching, using the NNmatch package by Abadie et al. (2004) (Abadie et al., 2004b):

```
nnmatch Y tx pscorelin,  keep(matchdata_Y) replace
use matchdata_Y.dta, clear
```

In order to be able to use the frequency weights for later analysis, the weight variable needs to be generated as follows:

```
gen km_mod=km+1
```

We store these frequency weights and merge it with the unmatched dataset:

```
keep id km_mod
rename km_mod weight
save  "L:\mylibrary\matchweight.dta", replace
merge id using "L:\mylibrary\statasim1.dta"
```

We save the combined dataset:

```
save  "L:\mylibrary \statasim2.dta", replace
```

Now we use the same approach as in the case of weighted regression, but instead of using inverse probability weights, we use frequency weights from the matching (`weight`).

```
reg Y tx Z1 Z2 Z3 Z4 [fweight= weight]
replace tx=1
predict mu1
replace tx=truetx
replace tx=0
predict mu0
replace tx=truetx
summ mu1
scalar Ytreated=r(mean)
summ mu0
scalar Ycontrol=r(mean)
scalar rPSmatch_Y  = Ytreated-Ycontrol
dis rPSmatch_Y
drop mu1 mu0
scalar drop Ytreated Ycontrol rPSmatch_Y
```

Appendix 5.3 - R code for generating data in the simulations of research paper 3

Here we provide the code that was used to generate data for Scenario 1 of the simulations. First, the necessary libraries need to be loaded:

```
library(Rlab)
library(Matching)
library(stats)
library(boot)
library(copula)
```

A simulated dataset, including the covariates `Z1,Z2,Z3,Z4,` the treatment variable `tx` and the endpoints `cost` and `Y` (denoting QALY) was created. First we generated the confounders from correlated normal distributions:

```
Sigma<-matrix(c(1,0.2,0.2,1),2,2)
Z12<-mvrnorm(n,c(2, 4), Sigma)
Z1<-Z12[,1]
Z2<-Z12[,2]
Z34<-mvrnorm(n,c(2, 4), Sigma)
Z3<-Z34[,1]
Z4<-Z34[,2]
```

Then we drew the treatment variable `tx` from a Bernoulli distribution with parameter `psc` (the true PS):

```
psc_logit<- 0.4 + (-1*Z1) + (0.5*Z2) + (0.025*Z2^2 )+ (-0.25*Z3) -
(0.1*Z4)
psc<-inv.logit(psc_logit)
tx<-rbern(n,psc)
```

We generated the cost and QALY endpoints using the `copula` package in R:

```
E.cost<-10000+ 6000*tx-2000*Z1+2000*Z2-2000*Z3+2000*Z4
E.cost=ifelse(E.cost<=0,0.1,E.cost)
E.Y <- 9+0.4*tx+(0.1*Z1)-(0.05*Z2)+(0.05*Z3) -(0.05*Z4)
ngmvdc <- mvdc(normalCopula(0.4), c("norm", "gamma"),
list(list(mean = E.Y, sd =0.2), list(shape=10,rate=10/E.cost)))
rng <- rmvdc(ngmvdc, n)
Y <- rng[,1]
cost <- rng[,2]
```

The misspecified variables were defined as follows:

```
X1 = exp(Z1/10)
X2 = Z2*(1+Z1)+10
X3 = (Z3/25+0.6)^2
X4 = (Z4+20)^2
```

The final dataset was created as follows:

```
dataset<-as.data.frame(cbind(X1,X2,X3,X4,Z1,Z2,Z3,Z4,Y,cost,tx))
```

Chapter 6 - Estimating treatment effectiveness under model misspecification: a comparison of targeted maximum likelihood estimation with bias-corrected matching

6.1 Preamble to research paper 4

The critical appraisal of the applied literature (research paper 1) showed that the specification of the PS and cost and effectiveness regression models is rarely appropriately assessed, raising concerns about biased parameter estimates in CEA. One area which received relatively little attention in the methodological literature in CEA is estimating incremental effectiveness parameters from HRQoL data (Basu and Manca, 2011). The conceptual review identified machine learning techniques for estimating the PS and the endpoint regression, which can reduce bias due to functional form misspecification. Research paper 4 contrasts two approaches, TMLE and BCM, which combine the PS with endpoint regression, and can be coupled with machine learning estimation techniques. These methods are flexible extensions of the DR and regression-adjusted matching approaches presented in research paper 3. These methods have not been used to estimate the effectiveness of treatment on HRQoL, and TMLE has not been compared to BCM in the general literature. Research paper 4 aims to address these gaps in the literature.

The motivating case study of this paper extends research papers 2 and 3 in using a contrasting case study, where relative effectiveness parameters need to be estimated in order to populate a decision-analytical model. A related simulation study compares the relative performance of TMLE and BCM, alongside traditional PS, regression and DR methods, for estimating incremental effectiveness. I contrast these methods when using

fixed parametric models for the PS and the endpoint regression, and when using machine-learning estimation. This paper focuses on the realistic scenario when the true parametric models are unknown.

This paper provides recommendations to help future studies choose more appropriate methods for estimating treatment effects in realistic circumstances, and includes software code to help implement the proposed methods (Appendix 6.1).

London School of Hygiene & Tropical Medicine
Keppel Street, London WC1E 7HT
www.lshtm.ac.uk

Registry
T: +44(0)20 7299 4646
F: +44(0)20 7299 4656
E: registry@lshtm.ac.uk

LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

COVER SHEET FOR EACH 'RESEARCH PAPER' INCLUDED IN A RESEARCH THESIS

Please be aware that one cover sheet must be completed for each 'Research Paper' included in a thesis.

1. **For a 'research paper' already published**

   1.1. Where was the work published? .....

   1.2. When was the work published? .....

   1.2.1. If the work was published prior to registration for your research degree, give a brief rationale for its inclusion

   .....

   .....

   .....

   1.3. Was the work subject to academic peer review? .....

   1.4. Have you retained the copyright for the work? **Yes / No**
   If yes, please attach evidence of retention.
   If no, or if the work is being included in its published format, please attach evidence of permission from copyright holder (publisher or other author) to include work

2. **For a 'research paper' prepared for publication but not yet published**

   2.1. Where is the work intended to be published? .....

   2.2. Please list the paper's authors in the intended authorship order

   .....

   2.3. Stage of publication – Not yet submitted / Submitted / Undergoing revision from peer reviewers' comments / In press

3. **For multi-authored work, give full details of your role in the research included in the paper and in the preparation of the paper. (Attach a further sheet if necessary)**

   .....

   .....

NAME IN FULL (Block Capitals) .....

STUDENT ID NO: .....

CANDIDATE'S SIGNATURE ..... Date .....

SUPERVISOR/SENIOR AUTHOR'S SIGNATURE (3 above) .....

Additional page for Question (3) on LSHTM cover sheet form:

I led the design of the research question, in collaboration with RG and an external collaborator, SG. I developed the simulation scenarios, with help from RR and RG. I wrote the simulation code and implemented the statistical methods in the motivating case study, with help from RR and SG. I led on the interpretation of the case study and simulation results, with contributions from SG, RR, RG and JS. I wrote the first draft of the manuscript. I managed each round of comments and suggestions from the co-authors, in collaboration with RG. All authors read and approved the final draft prior to inclusion in the thesis.

6.2 Research paper 4 - Evaluating treatment effectiveness under model misspecification: a comparison of targeted maximum likelihood estimation with bias-corrected matching

Noemi Kreif Msc[1], Susan Gruber PhD[2], Rosalba Radice PhD[3], Richard Grieve PhD[1], Jasjeet S. Sekhon PhD[4]

[1]Department of Health Services Research and Policy, London School of Hygiene and Tropical Medicine, London, UK;

[2]Department of Epidemiology, Harvard School of Public Health, Boston, MA, USA;

[3]Department of Economics, Mathematics and Statistics, Birkbeck, London, UK;

[4]Travers Department of Political Science, UC Berkeley, Berkeley, CA, USA

Status: To be submitted to Statistics in Medicine, December 2012.

Contributions: The candidate led the design of the research question, in collaboration with RG and an external collaborator, SG. The candidate developed the simulation scenarios, with help from RR and RG. The candidate wrote the simulation code and implemented the statistical methods in the motivating case study, with help from RR and SG. She led on the interpretation of the case study and simulation results, with contributions from SG, RR, RG and JS. The candidate wrote the first draft of the manuscript. She managed each round of comments and suggestions from the co-authors, in collaboration with RG. All authors read and approved the final draft prior to inclusion in the thesis.

The candidate_____ The supervisor_____

Abstract

This paper considers two approaches which combine the propensity score (PS) with an endpoint regression, when the functional forms of both models are misspecified (dual misspecification). Targeted maximum likelihood estimation (TMLE) is a double-robust (DR) approach designed to reduce bias in the estimate of the parameter of interest. Bias-corrected matching (BCM) adjusts covariate imbalance between matched pairs using regression predictions. For both methods, we consider machine learning techniques such as boosted classification and regression trees and "super learning", as well as using fixed parametric approaches to estimate the PS and endpoint regression.

We contrast TMLE and BCM, alongside PS, regression and DR approaches, in a motivating example evaluating the effect of different types of hip prosthesis on health-related quality of life (HRQoL) of patients with osteoarthritis. We find that the estimated effect of prosthesis type on HRQoL was similar across methods.

In a related simulation study we generated HRQoL data, with nonlinear functional form relationships, and good and poor overlap of the PS. Performance metrics included bias, root mean squared error (RMSE) and the 95% confidence interval (CI) coverage.

In the scenarios when either the PS or regression model was correct, both TMLE and BCM remained unbiased and reported CI coverage close to nominal levels. With misspecified, fixed models for the PS and the endpoint, all methods reported relatively high bias. With machine learning estimation, this bias was reduced. When overlap of the PS was good, TMLE provided estimates with the lowest bias and RMSE, and with poor overlap BCM performed best.

TMLE and BCM, when coupled with machine learning, are both appropriate methods for estimating of treatment effectiveness across the circumstances considered.

Introduction

Health policy-makers require unbiased, precise estimates of the effectiveness and cost-effectiveness of health interventions to inform the allocation of scarce health care resources (Rubin, 2010, Tunis et al., 2010, Fung et al., 2011). Observational studies are widely used to estimate average treatment effects (ATEs), but the major concern is selection bias due to confounding. Researchers often model the conditional expectation of the endpoint given covariates, with parametric regression models. An alternative is using the propensity score (PS), defined as the probability of treatment assignment, given the covariates. These methods however assume that the endpoint regression or PS model is correctly specified (Rubin, 1997). When estimating ATEs on health endpoints, such as health-related quality of life (HRQoL), correctly specifying a regression model can be challenging, particularly if the distribution of the endpoint is skewed, and the relationship between the covariates and the endpoint is nonlinear (Basu and Manca, 2011). In an observational setting, the PS is unknown and needs to be estimated, taking into account potential nonlinearities and interactions between covariates that can predict the treatment assignment (Dehejia and Wahba, 2002, Westreich et al., 2010).

Double-robust (DR) methods (Robins et al., 1994, Robins et al., 2007) combine endpoint regression models with the PS, and can be consistent if either the endpoint regression or the PS is correctly specified. However, for practical applications, there is an interest in the performance of these methods when both the endpoint regression and PS models are misspecified (dual misspecification) (Kang and Schafer, 2007a, Waernbaum, 2011, Gruber and van der Laan, 2012a). A further challenge is posed by poor overlap, also referred to as violation of the positivity assumption, which occurs when a certain set of baseline covariates is almost completely predictive of the

treatment within the sample (Westreich and Cole, 2010). Poor overlap can lead to unstable inverse probability of treatment (IPT) weights (Petersen et al., 2010). In circumstances where there is dual misspecification and unstable IPT weights, common DR approaches such as weighted regression can be more biased and less efficient than ordinary least squares (OLS) regression (Kang and Schafer, 2007a, Freedman and Berk, 2008).

An innovative DR method, targeted maximum likelihood estimation (TMLE) (van der Laan, 2010, van der Laan and Rubin, 2006) can outperform conventional DR methods when there is poor overlap (Porter et al., 2011, Stitelman and van der Laan, 2010, Gruber and van der Laan, 2010a). TMLE is a two-stage estimator, which fluctuates an initial regression prediction of the endpoint, using a function of the estimated PS. Another approach which can exploit information from the PS and the endpoint regression is bias-corrected matching (BCM) (Rubin, 1973, Abadie and Imbens, 2011). The idea behind this method is to subtract the estimate of the asymptotic bias from the nearest neighbour matching estimator, using regression predictions of the endpoint.

Both TMLE and BCM have the potential to reduce bias under dual misspecification. For both methods, estimates of the endpoint regression function and the PS can be obtained with fixed parametric models, i.e. when the functional form and distribution of the endpoint is chosen by the analyst. However these approaches can also accommodate machine learning techniques, where the best fitting model is selected by an algorithm. Simulation studies have demonstrated that TMLE, coupled with machine learning estimation for the endpoint and the PS, can report low bias when the correct parametric models are unknown (Porter et al., 2011, Gruber and van der Laan, 2012c). BCM can be relatively robust under dual misspecification. When the response surface, defined as the functional form relationship between the covariates and the endpoint (Rubin, 1979), is

moderately nonlinear, bias due to a misspecified PS can be eliminated even if adjustment is performed with an OLS regression model (Abadie and Imbens, 2011, Busso et al., 2011, Rubin and Thomas, 2000). Adjustment with OLS might be insufficient with highly nonlinear response surfaces, and so recommendations for BCM suggest that flexible approaches such as series regression are used for the bias-adjustment (Abadie and Imbens, 2011). However no previous studies have formally considered flexible regression methods for BCM.

There is limited work on the relative performance of matching and reweighting estimators such as inverse probability of treatment weighing (IPTW) and DR methods (Busso et al., 2011, Busso et al., 2009, Waernbaum, 2011, Radice et al., 2012). These studies found that with correctly specified PS and good overlap, reweighting estimators can be less biased and more efficient than matching. However, nearest neighbour matching estimators are less sensitive to misspecification of the PS: while for matching it is sufficient for the estimated PS to be a balancing score, for weighting, the PS needs to be the correct conditional probability of treatment assignment (Busso et al., 2009, Waernbaum, 2011). The only study that, to our knowledge, considered both DR and BCM estimators found that, with poor overlap and correctly specified PS, BCM provided less biased and more efficient estimates of the average treatment effect on the treated (ATT) than reweighing estimators (Busso et al., 2011). Machine learning estimation approaches for estimating the PS (Lee et al., 2010, Westreich et al., 2010, Setoguchi et al., 2008), and the endpoint regression function (Austin 2012) have been shown to reduce bias due to model misspecification, but few studies have investigated machine learning in the context of DR approaches (Porter et al., 2011, Ridgeway and McCaffrey, 2007) and no study has considered machine learning for BCM.

The objective of this paper is to compare the relative performance of TMLE and BCM in estimating the ATE of a binary treatment on a continuous endpoint, focusing on dual functional form misspecification of the PS and the endpoint regression. We also compare TMLE and BCM to other commonly applied DR (Kang and Schafer, 2007a), PS matching (Austin, 2008, Caliendo and Kopeinig, 2008) and regression (Basu and Manca, 2011) approaches. We consider each method with fixed parametric regression models but also with machine learning techniques, such as super learning for estimating the endpoint regression function (van der Laan et al., 2007) and boosted regression trees for the PS (McCaffrey et al., 2004, Lee et al., 2010, Xu et al., 2010).

We consider the methods in a motivating case study and in a simulation study. The case study considers the relative effectiveness of alternative types of total hip replacement on post-operative HRQoL of patients with osteoarthritis. This study exploited a large UK survey, which collects patient reported outcome measures (PROMs). The resulting observational database includes HRQoL data on all patients who have had elective surgical procedures provided by the NHS in England (PROMs, 2010, Ousey and Cook, 2011), with measurements of a rich set of pre-operative characteristics. While there was a good overlap in the distributions of potential confounders, important prognostic variables such as age and pre-operative health status were imbalanced. Correctly specifying the endpoint regression function was challenging in this dataset: the distribution of the HRQoL endpoint was bounded between a small negative value (indicating a health state worse than death) (Dolan et al., 1995) and 1 (perfect health), and in addition, there was a large spike at 1 in the distribution. We reported ATEs with each approach.

The simulation study was grounded in the case study, and compared the relative performance of the methods for a range of data-generating processes (DGPs) typical of

HRQoL data: in that the data were assumed to follow normal, gamma and two-part distributions, and the response surface was nonlinear. We considered scenarios of good overlap, as in the case study, but also scenarios with poor overlap, to explore the performance of the methods in the most challenging settings. We compare the relative performance of the methods according to bias, root mean squared error (RMSE), and coverage rates of nominal 95% confidence intervals (CIs).

In the next section, we outline the statistical methods under comparison. The following section presents the motivating example. Then we report the design and results of the simulation study. The last section discusses the findings and suggests areas for further research.

Statistical methods

The parameter of interest is the ATE of a binary treatment A, defined as

$$\psi = E[Y(1) - Y(0)] ,$$

where $Y(1)$ is the potential outcome under treatment, i.e. the endpoint that would be observed under treatment state, and $Y(0)$ is the potential outcome under control state. The vector of confounding factors, that is all factors that influence the potential outcomes and treatment assignment, is defined as *W*. Under unconfoundedness (Greenland et al., 1999), also known as conditional exchangeability, all elements of W are observed, and the mean of the conditional distribution of the potential outcomes corresponds with the mean of the conditional distribution of the observed endpoint *Y*:

$$E[Y(1)|W] = E[Y|A = 1, W] \text{ and } \quad E[Y(0)|W] = E[Y|A = 0, W].$$

Under the additional assumptions of consistency and positivity, the ATE can be identified as

$$\psi = E\big[[Y|A=1,W] - E[Y|A=0,W]|W\big].$$

The consistency assumption states that an individual's potential outcome under the observed treatment is exactly the observed endpoint (Cole and Frangakis, 2009). The positivity assumption requires that there are both treated and control individuals at each combination of the values of observed confounders in the population (Westreich and Cole, 2010), formally, $0 < g(A,W) < 1$, for any stratum defined by W, where $g(A,W) = P(A|W)$ is the model for the treatment assignment. In finite samples, even in the absence of structural violations, practical positivity violations often arise; in particular covariate strata there might be few or no individuals from either treatment group (Westreich and Cole, 2010), and so the estimated $\hat{g}(A,W)$ can be close to 0 or 1. The econometric literature on matching methods refers to positivity violations as "poor overlap" (Imbens, 2004). Here we define both structural and practical violations of the positivity assumption as poor overlap, and use this terminology throughout.

Regression estimators

We consider a general regression estimator, also known as the G-computation estimator (Robins, 1986) , which uses estimates of the expected potential outcomes, conditional on observed characteristics, defined as $Q(A,W) = E[Y|A,W]$.

The estimator for the ATE is given by:

$$\hat{\psi}^{reg} = \frac{1}{N}\sum_{i=1}^{N}\big\{\hat{Q}(1,W_i) - \hat{Q}(0,W_i)\big\}, \qquad (1)$$

where $\hat{Q}(1,W)$ and $\hat{Q}(0,W)$ are the estimated potential outcomes for each individual under treatment and control states, respectively, and N is the number of subjects in the sample.

$\hat{Q}(0, W)$ and $\hat{Q}(1, W)$ can be obtained by fitting a regression model that includes the observed covariates and a treatment variable, for example OLS or a generalised linear model (GLM). A more flexible method is to fit separate models for the treated and control samples (Imbens and Wooldridge, 2009b). Unbiased estimates of the ATE can only be achieved if $Q(1, W)$ and $Q(0, W)$ are estimated consistently. When there is poor overlap, regression estimators extrapolate, which can lead to large biases if the regression model is misspecified (Ho et al., 2007, Rubin, 1997).

Flexible estimation techniques of the endpoint regression function include series estimation (Imbens and Wooldridge, 2009b, Abadie and Imbens, 2011) and machine learning (Austin, 2012). Both approaches can reduce bias which results from model misspecification. Here we consider a machine learning approach, the "super learning" algorithm (van der Laan et al., 2007). Super learning uses a collection of prediction algorithms pre-selected by the user, potentially including parametric and non-parametric regression models. The algorithm uses cross-validation to select a weighted combination of estimates given by the prediction procedures (Polley and van der Laan, 2010b). Asymptotically, the super learner algorithm performs as well as the best possible combination of the candidate estimators (van der Laan and Dudoit, 2003).

Propensity score methods

The propensity score (PS) is defined as the conditional probability of treatment assignment given W, $g(1|W) = Pr(A = 1|W)$. Using the estimated PS, $\hat{g}(.)$, as a distance metric, matched treated and control comparison groups can be created (Rosenbaum and Rubin, 1983). The matching estimator imputes the missing potential outcomes, $Y(0)$ and $Y(1)$, using the observed endpoints of the closest *M* individuals:

$$\hat{Y}(0, W_i) = \begin{cases} Y_i & \text{if } A_i = 0 \\ \frac{1}{M} \sum_{j \epsilon \zeta_M(i)} Y_j & \text{if } A_i = 1 \end{cases},$$

$$\hat{Y}(1, W_i) = \begin{cases} \frac{1}{M} \sum_{j \epsilon \zeta_M(i)} Y_j & \text{if } A_i = 0 \\ Y_i & \text{if } A_i = 1 \end{cases},$$

where $\zeta_M(i)$ is the set of *M* individuals matched to unit i. The estimator for the ATE is the mean of the estimated individual-level treatment effects:

$$\hat{\psi}^{match} = \frac{1}{N} \sum_{i=1}^{N} \{\hat{Y}(1, W_i) - \hat{Y}(0, W_i)\}$$

Inverse probability of treatment weighting (IPTW) reweights the treated and control samples using inverse weights $\frac{A_i}{\hat{g}(1|W_i)}$ for the treated and $\frac{1-A_i}{1-\hat{g}(1|W_i)}$ for the control observations. The normalised IPTW estimator (Hirano and Imbens, 2001, Kang and Schafer, 2007a) is defined as:

$$\hat{\psi}^{IPTW} = \frac{\sum_{i=1}^{N} A_i \frac{Y_i}{\hat{g}(1|W_i)}}{\sum_{i=1}^{N} \frac{A_i}{\hat{g}(1|W_i)}} - \frac{\sum_{i=1}^{N} (1 - A_i) \frac{Y_i}{1 - \hat{g}(1|W_i)}}{\sum_{i=1}^{N} \frac{1 - A_i}{1 - \hat{g}(1|W_i)}}$$

Matching estimators are consistent if $\hat{g}(.)$ is correctly specified (Waernbaum, 2011), but have larger finite sample bias and are less precise than a correctly specified regression estimator (Abadie and Imbens, 2006, Rubin, 1973). With a correctly specified $\hat{g}(.)$, IPTW can also provide consistent and efficient estimates (Hirano et al., 2003). However, poor overlap can result in unstable IPT weights, and biased or inefficient estimates of the ATEs (Kang and Schafer, 2007a, Lee et al., 2010, Busso et al., 2011, Radice et al., 2012). In these settings, recommended approaches include truncating IPT weights at fixed levels (Elliott, 2008) or percentiles of $\hat{g}(.)$ (Cole and Hernán, 2008), and estimating ATEs for a subsample with good overlap (Crump et al., 2009).

Instead of estimating the PS with a fixed parametric model, such as a logistic regression including main terms, flexible approaches have been proposed to help correctly specify $g(.)$. These include the series regression estimator (Hirano et al., 2003), and methods from the machine learning literature, including decision trees, neural networks, linear classifiers and boosting (McCaffrey et al., 2004, Setoguchi et al., 2008, Westreich et al., 2010). This paper considers the machine learning approach of boosted classification and regression trees (CART). This approach has been shown to reduce bias in the estimated ATE compared to a misspecified logistic regression, under circumstances of unstable IPT weights (Lee et al., 2010), and outperformed alternative machine learning methods such as pruned CARTs. Boosted CART fits regression trees on random subsets of the data, and in each iteration, the data points that were incorrectly classified with the previous trees receive greater priority. According to general recommendations (Stuart, 2010), the algorithm can be set to select the final PS model that maximises covariate balance (McCaffrey et al., 2004, Lee et al., 2010).

Double-robust methods

Double-robust (DR) methods (Bang and Robins, 2005, Robins et al., 1994) combine models for $Q(.)$ and $g(.)$, with most estimators using $\hat{g}(.)$ as IPT weights (Kang and Schafer, 2007b). The distinctive property of DR estimators is that they are consistent if either (but not necessarily both) $g(.)$ or $Q(.)$ is correctly specified (Robins et al., 1994). If both components are correct, the DR estimator can be a semiparametric efficient estimator (Robins et al., 2007, van der Laan and Rubin, 2006). A commonly used DR method is the weighted regression (Freedman and Berk, 2008, Kang and Schafer, 2007a), which weights the covariates in an endpoint regression, using IPT weights.

In realistic settings such as when there is poor overlap and dual misspecification, weighted regression can report biased and inefficient estimates of ATEs (Kang and Schafer, 2007a, Porter et al., 2011, Freedman and Berk, 2008, Basu et al., 2011). An ongoing debate discusses the relative merits of different DR estimators in these circumstances (Porter et al., 2011, van der Laan and Gruber, 2010, Robins et al., 2007). One recommendation when faced with an unknown PS and unstable IPT weights is to use machine learning methods to estimate $g(.)$ (Ridgeway and McCaffrey, 2007). Here we consider the weighted least squares estimator (WLS) and implement it with weights obtained from a fixed logistic regression but also using boosted CART.

It has been suggested that DR estimators should have a "boundedness property": they should respect the known bounds of the endpoint, for example that an HRQoL endpoint is between small negative values and 1, so that the estimated parameter will always fall into the parameter space (Robins et al., 2007, Rotnitzky et al., 2012). This property can reduce bias and increase precision when the PS is used as weights, where large weights can lead to estimated values of the endpoint falling outside of a plausible range (Gruber and van der Laan, 2010a). Below we discuss a DR estimator, TMLE, that can have this boundedness property (Rotnitzky et al., 2012), and is therefore appealing for settings of poor overlap (Gruber and van der Laan, 2010a, Gruber and van der Laan, 2012b).

Targeted maximum likelihood estimation

While standard maximum likelihood estimation aims to find parameter values that maximise the likelihood function for the whole distribution of the data, TMLE is concerned with a particular feature of the distribution such as the ATE (van der Laan and Rubin, 2006, Moore and van der Laan, 2009). Maximising a global likelihood function may not yield the least biased estimate of the target parameter, so TMLE is

designed to target the initial estimate to reduce bias in the estimate of the parameter of interest. The TMLE estimator solves the efficient influence curve estimating equation, where an influence curve describes the behaviour of an estimator under slight changes of the data distribution (Hampel, 1974). Performing TMLE involves two stages (Gruber and van der Laan, 2012c), which, for estimating the ATE, are:

1: To obtain an initial estimate of the conditional mean of Y given A and W by using regression to predict the potential outcomes $\hat{Q}(1,W)$ and $\hat{Q}(0,W)$.

2: To fluctuate this initial estimate, $\hat{Q}^0(A,W)$, exploiting the information in the treatment assignment mechanism, $g(.)$.

Here, the fluctuation corresponds to extending the parametric model for $Q(A,W)$ with "clever covariates" (h):

$$h_0(A,W) = \frac{1-A}{1-g(A=1|W)}$$

$$h_1(A,W) = \frac{A}{g(A=1|W)}$$

For continuous endpoints, it is recommended (Gruber and van der Laan, 2012b, Gruber and van der Laan, 2010a) that known bounds of the endpoint are exploited by rescaling Y to between 0 and 1, to ensure the boundedness of the TMLE estimator. The rescaled endpoint is defined as $Y^* = \frac{Y-a}{b-a}$, where a and b are known limits of Y. Using Y^*, $Q^*(A,W) = \frac{Q(A,W)-a}{b-a}$ can be defined. The fluctuation can then be performed on the logistic scale:

$$logit\left(\widehat{Q^*}^1(0,W)\right) = logit\left(\widehat{Q^*}^0(0,W)\right) + \hat{\varepsilon}_0\hat{h}_0(0,W) \text{ and}$$

$$logit(\widehat{Q^*}^1(1,W)) = logit(\widehat{Q^*}^0(1,W)) + \hat{\varepsilon}_1\hat{h}_1(1,W).$$

Here, $\hat{\varepsilon}_0$ and $\hat{\varepsilon}_1$ can be estimated by logistic regression with quasi-binomial distribution of Y^* on $\hat{h}_0$ and $\hat{h}_1$, and offset $logit(\widehat{Q^*}^0(A,W))$. h_0 and h_1 are constructed to solve

the efficient influence curve estimating equation for the ATE. This regression can be interpreted as explaining the residual variability of the predicted endpoint, using information in the treatment assignment mechanism. $\hat{Q}^1(A,W)$ can be obtained by back-transforming $\widehat{Q^*}^1(A,W)$ to the original scale.

The resulting targeted estimates of the potential outcomes, $\hat{Q}^1(0,W)$ and $\hat{Q}^1(1,W)$ are applied in the G-computation formula in order to obtain the TMLE estimator:

$$\hat{\psi}^{TMLE} = \frac{1}{N}\sum_{i=1}^{N} \hat{\mathrm{Q}}^1(1,W_i) - \hat{\mathrm{Q}}^1(0,W_i)$$

TMLE has the property of double-robustness: if either the initial estimate of $Q(.)$ or $g(.)$ are correctly specified, the estimator is consistent. TMLE is also an asymptotically efficient estimator, and if both $Q(.)$ or $g(.)$ are correct, it reaches the semiparametric efficiency bound (van der Laan, 2010). The estimator can use predictions from any fixed parametric model for the initial $Q(.)$ (for example OLS or GLM) and $g(.)$ (for example logistic regression). However, with machine learning methods, TMLE has been shown to reduce bias when the models for the assignment mechanism and the endpoint are unknown (Porter et al., 2011). As in the previous sections, we consider super learning for the initial $Q(.)$ and boosted CARTs for $g(.)$.

Bias-corrected matching

It is generally recommended that matching methods are followed by regression adjustment (Rubin, 1973, Rubin and Thomas, 2000, Abadie and Imbens, 2006). The idea is similar to regression-adjustment in randomised trials: regression is used to "clean up" residual imbalances between treatment groups after matching (Stuart, 2010). BCM (Abadie et al., 2004, Abadie and Imbens, 2011) adjusts the imputed potential outcome with the difference in the predicted endpoint that can be attributed to covariate

imbalance between the matched pairs. These predictions are obtained by a regression of the endpoint on covariates, stratified by treatment assignment. The bias-corrected predictions of the potential outcomes are obtained as follows:

$$\hat{Y}(0,W_i) = \begin{cases} Y_i & \text{if } A_i = 0 \\ \frac{1}{M}\sum_{j\epsilon\zeta_M(i)} Y_j + \hat{Q}(0,W_i) - \hat{Q}(0,W_j) & \text{if } A_i = 1 \end{cases},$$

$$\hat{Y}(1,W_i) = \begin{cases} \frac{1}{M}\sum_{j\epsilon\zeta_M(i)} Y_j + \hat{Q}(1,W_i) - \hat{Q}(1,W_j) & \text{if } A_i = 0 \\ Y_i & \text{if } A_i = 1 \end{cases},$$

For example, for an individual i who receives control, the imputed potential outcome under the treatment state is the average outcome of the *M* closest matches from the treatment group (indexed by *j*), adjusted with the difference between the predicted outcomes under treatment, when covariate values are set to those of its own values, $\hat{Q}(1,W_i)$ and the covariate values of the match, $\hat{Q}(1,W_j)$. The corresponding estimator is the mean difference of these predictions:

$$\hat{\psi}^{BCM} = \frac{1}{N}\sum_{i=1}^{N} \hat{Y}(1,W_i) - \hat{Y}(0,W_i)$$

BCM is consistent if $Q(0,W)$ and $Q(1,W)$ are consistently estimated (Abadie and Imbens, 2011). Matching can decrease the sensitivity of estimates to the misspecification of the endpoint regression model (Ho et al., 2007) and, for moderately nonlinear response surfaces, adjustment even with a misspecified OLS model can reduce bias (Rubin, 1973, Rubin and Thomas, 2000, Abadie and Imbens, 2011, Busso et al., 2011). Because an OLS regression, even including nonlinear terms, might not capture highly nonlinear response surfaces, we consider super learning for predicting the potential outcomes, as well as fixed parametric models. Following recommendations, we select the number of matches (*M*) to 1 (Stuart, 2010, Caliendo and Kopeinig, 2008), and match on the liner predictor of PS with replacement, allowing

for ties. As for the DR methods, we estimate the PS using logistic regression and also using boosted CARTs.

Motivating case study

Overview

We consider the methods in a case study that evaluates the effect of alternate hip prosthesis types on the HRQoL of patients with osteoarthritis, using an observational database on patients with total hip replacement (THR). THR is one of the most common surgical procedures in the UK, with over 50,000 hip procedures performed in the NHS in England and Wales in 2011 (NICE, 2012). With a large number of brands of prosthesis in use, with differing costs, UK health care decision makers have a considerable interest in evaluating the clinical and cost-effectiveness of different prosthesis types in routine care (NICE, 2012). A large scale UK survey that collects patient-reported outcome measures (PROMs) on all patients who undergo elective surgery in the NHS provides a key data source for this evaluation. The resulting observational dataset, used in this case study, includes pre- and post-operative HRQoL of patients with THR procedures (PROMs, 2010, Ousey and Cook, 2011).

The dataset measures the HRQoL endpoint as EQ-5D-3L scores (EuroQol Group, 1990). The EQ-5D-3L is a generic instrument with five dimensions of health (mobility, self care, usual activities, pain and discomfort, anxiety and depression) and three levels (no problems, some problems, severe problems). The EQ-5D-3L profiles were combined with health state preference values from the UK general population, to give utility index scores on a scale ranging from 1 (perfect health), through 0 (death) to the worst possible health state, -0.59 (Dolan et al., 1995). This results in a bounded

distribution of the endpoint that exhibited a spike at 1, posing a challenge for the specification of the regression model (Basu and Manca, 2011).

A previous analysis of this dataset (Pennington et al., 2012) reported the relative effectiveness on EQ-5D-3L scores of common prosthesis types, such as cemented, cementless, and "hybrid" prostheses. The analysis used multivariate matching and linear regression to adjust for confounding, and found a small but statistically significant advantage of hybrid compared to cementless prostheses.

The objective of this case study is to estimate the ATE on EQ-5D-3L, 6 months after operation, in patients with hybrid hip prosthesis, compared to cementless hip prosthesis. For this motivating example, we constrained the population of interest to be UK males patients with osteoarthritis, aged 65-74 (n = 3583). We contrast TMLE and BCM, and also compare them to standard statistical approaches such as regression, matching, IPTW and WLS. We implement each method with fixed parametric models and machine learning estimation techniques.

Statistical methods in the case study

Measured potential confounders included patient characteristics such as age, sex, body mass index, pre-operative heath status ("Oxford Hip score" and HRQoL), comorbidities, disability, index of multiple deprivation, and characteristics related to the intervention, such as surgeon experience (senior surgeon or not) and hospital type (NHS, private sector hospital, or treatment centre). Of the 3,583 patients included in the analysis, 32% had missing data on post-operative HRQoL and 39% on BMI. Other covariates were complete for over 90% of the sample. Multiple imputation using chained equations was applied to impute missing covariate and endpoint values (Pennington et al., 2012). Following recommendations (Rubin, 1996), five multiply

imputed datasets were created, and the analysis described below was performed on each dataset. Point estimates and variances were combined using Rubin's rule (Rubin, 1996). Fixed parametric approaches for estimating $Q(.)$ included OLS regression and a two-part model which can account for the spike in the observed distribution of the endpoint (Buntin and Zaslavsky, 2004, Basu and Manca, 2011). Here the continuous part ($Y' = 1 - Y$ when $Y < 1$) was modelled with a gamma regression, and logistic regression was used for modelling $P(Y < 1)$. The PS was estimated using logistic regression. In order to allow for nonlinearities, for each model, continuous variables were fitted with smoothing splines (using default degrees of freedom of 4).

For machine learning estimation of $Q(.)$, we used the R package "Super Learner" (Polley and van der Laan, 2010a), where the user-defined library included the following prediction algorithms: "glm" (main terms linear regression), "glm.interaction" (glm with covariate interactions), and a package that implements multivariate adaptive polynomial spline regression methods, "polymars" (Kooperberg, 2010). Machine learning estimation of $g(.)$ relied on boosted (logistic) CARTs, using the R package "twang" (Ridgeway et al., 2006), with tuning parameters recommended by the developers (McCaffrey et al., 2004, Lee et al., 2010). This implementation aimed to minimise mean covariate imbalance measured using Kolmogorov-Smirnov tests, reweighed by the estimated IPT weights.

We applied WLS using smoothing splines, with IPT weights obtained from the logistic model and also from the boosted CARTs. TMLE used the known minimum and maximum values of the endpoint as bounds, -0.59 and 1 (Dolan et al., 1995). Standard errors and 95% CIs were calculated using the sandwich estimator for IPTW and WLS, and using the influence curve (van der Laan, 2010), for TMLE. For matching and BCM, estimated standard errors took into account the matching process, conditional on the

estimated PS (Abadie and Imbens, 2011, Abadie and Imbens, 2006). For the two-part model and the super learning regression estimator, we used the non-parametric bootstrap (Davison and Hinkley, 1997) to obtain standard errors.

Case study results

Table 6.1 shows balance on the main pre-operative characteristics of patients with hybrid and cementless hip replacement, reported as absolute standardised mean differences. Patients with hybrid hip replacement were slightly older, had more co-morbidities (measured as a co-morbidity score), worse social status, and were less likely to have been treated by a consultant or in a treatment centre.

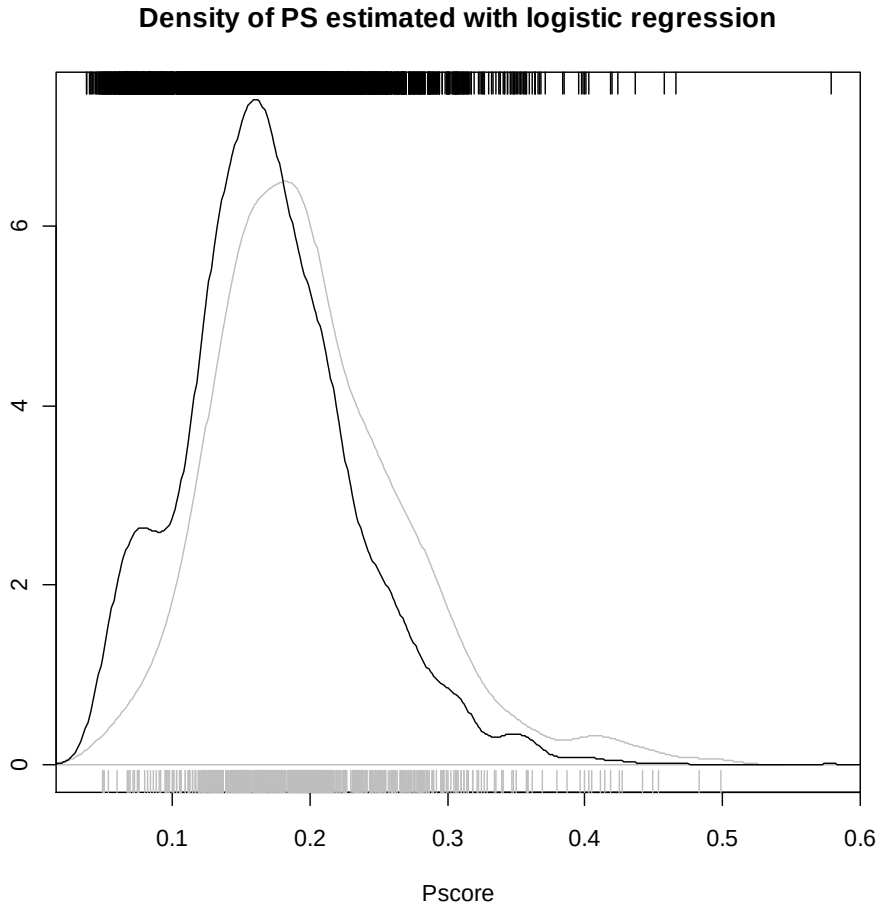
Table 6.1 - Balance on pre-operative characteristics, means and % standardised mean differences

Covariate	Mean hybrid (n=631)	Mean cementless (n=2952)	SMD (%)
Age	69.71	69.25	15.98
Oxford hip score[1]	20.17	19.93	2.83
Pre-operative EQ-5D[1]	0.40	0.40	0.63
Index of deprivation[1]	3.26	3.03	15.92
ASA grade 1 (%)[1]	1.00	0.96	4.14
ASA grade 2 (%)[1]	0.09	0.12	9.55
Disability (%)	0.74	0.74	0.52
Obese (%)[1]	0.27	0.27	0.69
Morbidly obese (%)[1]	0.10	0.11	4.30
Nr of co-morbidities	1.00	0.96	4.14
Co-morbidities			
Heart disease	0.18	0.15	7.86
High bp	0.40	0.42	4.55
Stroke	0.03	0.02	7.78
Circulation	0.08	0.07	4.08
Lung disease	0.06	0.06	3.61
Diabetes	0.13	0.12	2.20
Kidney disease	0.01	0.02	6.24
Nervous system	0.01	0.01	5.20
Liver disease	0.01	0.00	7.65
Cancer	0.06	0.05	3.80
Depression	0.05	0.04	5.84
Consultant (%)	0.80	0.87	17.64
Treatment centre (%)	0.05	0.12	26.16

Notes: SMD - standardised mean difference. SMD was calculated as $d = 100 * \frac{|\bar{x}_h - \bar{x}_c|}{\sqrt{\frac{s^2_h + s^2_c}{2}}}$, where $\bar{x}_h$ and $\bar{x}_c$ are the means for the hybrid and cementless group, while in the denominator includes the pooled standard deviation of the two groups, for a given covariate. Variables are dichotomous, with the exception of age, Oxford hip score, pre-operative EQ-5D score, index of deprivation and number of co-morbidities.
[1] Variables with missing values. Here, SMDs were combined using Rubin's rule.

There was good overlap between the densities of the estimated PSs for hybrid and cementless groups, when $g(.)$ was obtained using logistic regression (Figure 6.1). The plots obtained using boosted CART for estimating the PS were similar.

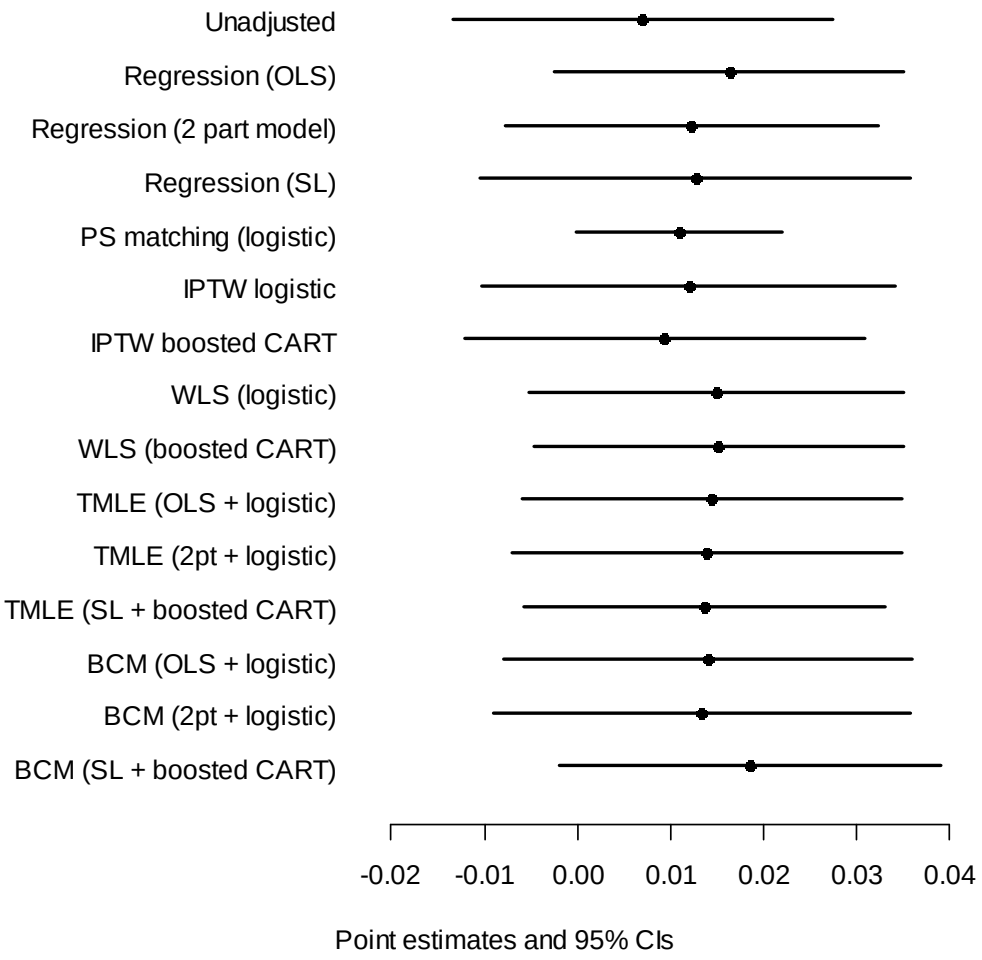
Figure 6.1 - Estimated PS using logistic regression, hybrid vs. cementless hip prosthesis



Notes: Hybrid (grey line) vs. cementless (black line). The rug plots, at the top and bottom, show the corresponding values of the PS. Estimates are based on the first imputed dataset.

Figure 6.2 shows the point estimates and 95% CIs after combining the estimates obtained for the imputed datasets. All methods reported a small positive advantage in mean EQ-5D-3L scores for hybrid versus cementless prostheses, but with CIs including zero.

Figure 6.2 - Point estimates and 95% CIs of ATE in terms of EQ-5D-3L score, hybrid vs. cementless hip prosthesis, across statistical methods.



Notes: SL- super learner

Simulation study

Overview

The simulation study aimed to compare the performance of BCM and TMLE, in estimating the ATE of a binary treatment on a continuous endpoint with a nonlinear response surface. As in the case study, we compared these methods to regression, PS

matching, IPTW and WLS, and for each method, we considered fixed parametric models and machine learning estimation for $Q(.)$ and $g(.)$. Motivated by the case study and previous methodological work, we designed a range of scenarios with typical characteristics of HRQoL data. Irregular distributions (Manning et al., 2005, Basu, 2011) of HRQoL data were recognised by generating endpoints from distributions such as the gamma and two part distributions. Following previous simulation studies (Basu and Manca, 2011, Porter et al., 2011, Austin, 2012), we considered data-generating processes (DGPs) with nonlinear response surfaces, good and poor overlap, and with moderate and strong association between confounders and the endpoints. These DGPs were selected to highlight the differences between the performance of the methods under realistic circumstances, by investigating the following hypotheses:

1. *Reweighting methods are anticipated to outperform BCM* when overlap is good (Busso et al., 2011). In such scenarios, TMLE is expected to outperform BCM in terms of bias and efficiency.

2. *When overlap is poor, BCM is expected to outperform TMLE*, because matching can be less sensitive than weighting to extreme PS values and to the misspecification of $g(.)$ (Lee et al., 2010, Busso et al., 2011, Waernbaum, 2011).

3. *Using appropriate machine learning methods is anticipated to reduce bias* compared to using misspecified parametric models for $Q(.)$ and $g(.)$ (Porter et al., 2011, Austin, 2012), across all methods considered.

Table 6.2 summarises the selected DGPs. We assumed a PS mechanism that generated good overlap of the densities of the true PS (DGP 1 and 2) and one that generated poor overlap (DGP 3 to 5). We considered moderate (DGP 1) and strong (DGP 2 to 5) association between the confounders and the endpoints. DGPs 1-3 considered a normal endpoint with an identity link function between the linear predictor and the endpoint, DGP 4 considered an endpoint which followed a gamma distribution with a log link,

while DGP 5 considered a two part data-generating distribution, with a mixture of a beta-distributed random variable and values of 1.

For each DGP, five scenarios were considered: (a) when fixed parametric models were used for both the PS and the endpoint regression, and these were correctly specified, (b and c) when one of the two was misspecified and (d) when neither model was correctly specified. In scenario (e) we considered machine learning estimates of $Q(.)$ and $g(.)$ for each method, under the assumption that the investigator, similarly to (d), does not know the correct parametric models, hence the correct model is not included among the candidate prediction algorithms. For DGP 1, we report results from each of the five scenarios, while for DGP 2 to 5, we only report the results for scenarios (d) and (e), as these were a priori judged the most realistic. The results for the remaining scenarios are available upon request.

Bias, variance, root mean squared error (RMSE) and the coverage rate for nominal 95% CIs of the estimated ATEs were obtained. Relative bias was calculated as the percentage of the absolute bias of the true parameter value, where absolute bias is the difference between the true parameter value and the mean of the estimated parameter. The RMSE was taken as the square root of the mean squared differences between the true and estimated parameter values.

Table 6.2 - Summary of DGPs used in the simulation study

	Overlap	**Confounder-endpoint association**	**Endpoint distribution**
DGP 1	Good	Moderate	Normal
DGP 2	Good	Strong	Normal
DGP 3	Poor	Strong	Normal
DGP 4	Poor	Strong	Gamma
DGP 5	Poor	Strong	Two part

Data generating process

For each DGP, we generated 1000 datasets of n=1000, with binary (W_1 to W_5) and standard normally distributed covariates (W_6 to W_8). This mix of binary and continuous covariates reflects the case study. The correlation coefficients between the covariates were set between 0.075 and 0.6. All covariates were true confounders, i.e. they influenced both the treatment assignment and the endpoint. Treatment was assigned according to a true PS that, like previous simulation studies, included main terms, higher order terms and interactions (Austin, 2012, Waernbaum, 2011).

For DGP 1, the PS model resulted in a good overlap of the true PS (see Figure 6.3):

$$logit(PS) = -1 + k_1(0.3W_1 - 0.1W_2 - 0.2W_3 + 0.4W_4 + 0.7W_5 + 0.2W_6 + 0.2W_7 - 0.25W_8 + 0.8{W_6}^2 - 0.3{W_7}^2 - 0.3{W_8}^2 - 0.05W_1W_2 - 0.05W_1W_3)\ ,$$

where $k_1 = 0.3$.

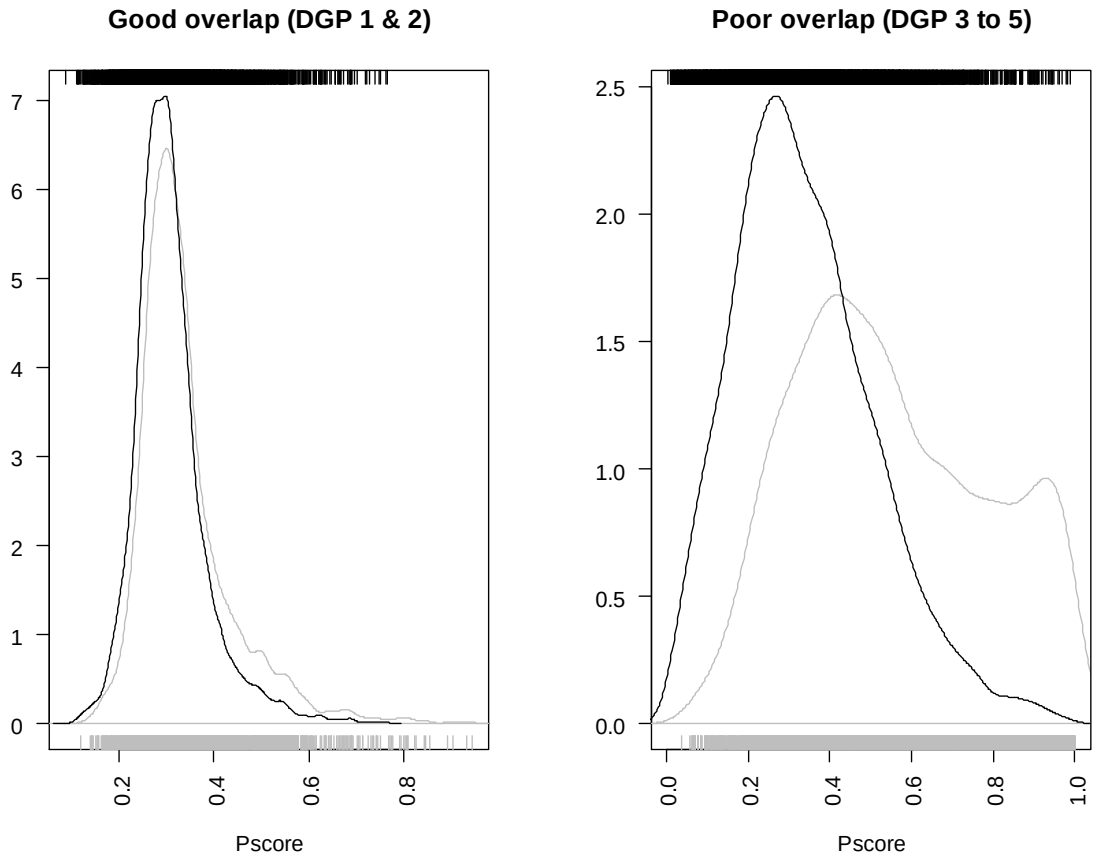
The treatment variable A was drawn from a Bernoulli distribution, using the PS as the parameter of success probability. The endpoint was drawn from a normal distribution with mean

$$\mu_{norm} = 15 + 0.4A + k_2(1W_1 - 0.1W_2 + 0.1W_3 - 0.1W_4 + 0.1W_5 - 0.1W_6 + 0.1W_7 + 0.1W_8 - 0.2{W_6}^2 - 0.1{W_7}^2 - 0.1{W_8}^2 + 0.2{W_6}^3 + 0.1{W_7}^3 + 0.1{W_8}^3 - 0.1W_1W_2 + 0.5W_1W_7),$$

standard deviation of 1 and $k_2 = 1$.

In DGP 2, setting k_2 to 4 increased the strength of the confounder-endpoint association. In DGP 3, changing k_1 to 1 created a poor overlap of the true PS distributions (see Figure 6.3).

Figure 6.3 - Densities of the true PS in the simulations for a typical sample (n =10,000)



Notes: Treated (grey line) vs. control (black line). The rug plots at the top and bottom show the corresponding values of the PS.

In DGP 4, the endpoint was drawn from a gamma distribution, with a log link, shape parameter of 100 and a scale parameter of $\frac{\mu_{\text{gam}}}{100}$, where the linear predictor was

$$log(\mu_{gam}) = 3 + 0.2A - 0.2W_1 + 0.2W_2 - 0.2W_3 + 0.5W_4 - 1W_5 + 0.5W_6 - 0.5W_7 + 0.2W_8 - 0.2{W_6}^2 - 0.01{W_7}^2 - 0.01{W_8}^2 - 0.01{W_6}^3 - 0.01{W_7}^3 - 0.01{W_8}^3 - 0.01W_1W_2 - 0.4W_6W_7.$$

In DGP 5, motivated by the case study and previous simulation studies (see Basu and Manca, 2011), the endpoint was generated as a mixture of a beta distributed continuous variable Y′ and 1, using a Bernoulli distribution with parameter *p* to select between values from the two distributions :

$$Y \sim (1-p) * 1 + p(1-Y'),$$

where

$$logit(p) = 4 - 1A - 0.2W_1 + 0.5W_2 - 0.5W_3 - 1W_4 - 0.3W_5 + 0.2W_6 + 0.5W_7 - 0.5W_8,$$

$$Y' \sim Beta(\mu_{beta} * phi,\ \mu_{beta} * (1-phi)),$$

$$logit(\mu_{beta}) = -1 - 0.2A - 0.5W_1 - 0.5W_2 - 0.5W_3 + 0.5W_4 - 0.5W_5 - 0.5W_6 - 0.5W_7 - 0.5W_8 - 0.2{W_6}^2 - 0.2{W_7}^2 - 0.2{W_8}^2 - 0.2{W_6}^3 - 0.2{W_7}^3 - 0.2{W_8}^3 - 0.2W_1W_2 - 0.2W_6W_7).$$

The resulting distribution with a spike at 1 reflects the observed endpoint in the case study. The true ATE was 0.4 in DGP 1 to 3, it was 9.98 for DGP 4 and 0.062 for DGP 5. While in DGP 1 to 3 the treatment effect was constant across individuals, for DGP 4 and 5, the true ATE was obtained by simulating both potential outcomes for each individual, and taking the mean of the individual-level additive treatment effects.

Implementation of the methods

Correct specification was defined as applying a fixed parametric model with the knowledge of features of the true DGP, such as the link function, the functional form between the covariates and the linear predictor, and the error distribution. For each DGP, the misspecified parametric $g(.)$ and $Q(.)$ models were logistic and OLS regressions with main terms only. Machine learning estimation of $g(.)$ and $Q(.)$ was as described in the case study section. The WLS estimator was implemented with main terms only, hence in this estimator the $Q(.)$ component is misspecified. For the DGPs with poor overlap, in a sensitivity analysis we modified the IPTW, WLS and TMLE estimators, and used weights based on $g(.)$ truncated at fixed levels of 0.025 and 0.975. For calculating coverage rates of nominal 95% CIs, standard errors were obtained as described in the case study section.

Simulation study results

Tables 6.3-6.5 report the relative bias (%), variance, RMSE and 95% CI coverage for the estimators considered, and Figure 6.4 presents quintiles of the estimated ATE.

Table 6.3 reports results for DGP 1, when there was good overlap, with a moderate association between the confounders and a normally distributed endpoint. When both $Q(.)$ and $g(.)$ were correctly specified, all methods reported minimal bias, with parametric regression (OLS with nonlinear terms) and TMLE reporting the lowest RMSE. Regression, TMLE and BCM all provided coverage at the nominal 95%, while IPTW and PS matching reported coverage rates higher (98% and 99%) than the nominal level. When only one of the PS or endpoint model was misspecified, BCM and both DR methods (WLS and TML) remained unbiased. With dual misspecification, each method reported moderate levels of bias, but when machine learning estimation was used for $Q(.)$ and $g(.)$, bias was reduced to close to zero for all the methods that combined these components, with WLS and TMLE providing estimates with the lowest RMSE.

For DGPs 2-5, results showed a similar pattern to DGP 1 when either $g(.)$ or $Q(.)$ was correctly specified, hence we only report result with dual misspecification. In DGP 2, with misspecified fixed parametric methods, stronger association between the confounders and the endpoint led to higher biases, but with machine learning estimation the bias for the methods that combined $g(.)$ and $Q(.)$ again decreased to below 10%. WLS and TMLE reported lower bias and RMSE than BCM. In DGPs 3-5, where there was poor overlap, with misspecified fixed parameteric models, each method reported high bias. For each of these DGPs, machine learning estimation improved performance of the methods that combined $g(.)$ and $Q(.)$. In DGP 3, TMLE provided the lowest bias and RMSE, albeit with CI coverage that was lower than the nominal level. In DGP 4

where we considered an endpoint with a gamma distribution, with machine learning approaches BCM showed less relative bias (2.5%) than TMLE (20.7%). In DGP 5, where we considered an endpoint with a two-part distribution, TMLE and BCM with machine learning estimation performed best; BCM gave the lowest relative bias (1.1% vs. 7.2%) and best CI coverage whereas TMLE reported the lowest RMSE.

IPTW using machine learning weights often reported high bias: for example for DGP 5, it reported higher bias than using a misspecified, fixed logistic regression to obtain the PS. This indicated that using machine learning estimation for the PS alone was insufficient to eliminate bias. For DGPs 3 to 5, where overlap was poor, truncating the IPT weights for IPTW and TMLE for either the logistic or the boosted PS models did not change the results.

Table 6.3 - Simulation results for DGP 1, over 1000 replications: normal endpoint, moderate association confounder-endpoint association, good overlap

Scenario	Relative bias	Variance	RMSE	95 % CI coverage
(a) Q correct - g correct				
OLS	-0.1%	0.005	0.070	95%
IPTW	0.5%	0.008	0.091	99%
PS matching	1.2%	0.011	0.106	98%
TMLE	-0.1%	0.005	0.071	95%
BCM	-0.1%	0.007	0.082	95%
(b) Q correct - g misspecified				
OLS	-0.1%	0.005	0.070	95%
IPTW	-15.0%	0.008	0.110	97%
PS matching	-8.1%	0.013	0.117	96%
TMLE	-0.2%	0.005	0.070	94%
BCM	0.7%	0.007	0.085	93%
(c) Q misspecified - g correct				
OLS	-11.7%	0.008	0.098	90%
IPTW	0.5%	0.008	0.091	99%
PS matching	1.2%	0.011	0.106	98%
WLS	0.6%	0.008	0.087	95%
TMLE	0.6%	0.008	0.087	95%
BCM	0.7%	0.009	0.097	95%
(d) Q misspecified - g misspecified				
OLS	-11.7%	0.008	0.098	90%
IPTW	-15.0%	0.008	0.110	97%
PS matching	-8.1%	0.013	0.117	96%
WLS	-12.7%	0.008	0.103	90%
TMLE	-12.9%	0.008	0.104	90%
BCM	-7.4%	0.011	0.108	93%
(e) Machine learning				
Regression (Q super learner)	-3.1%	0.006	0.079	95%
IPTW (g boosted CART)	10.2%	0.007	0.091	98%
WLS (Q OLS, g boosted CART)	0.5%	0.006	0.076	97%
TMLE (Q SL, g boosted CART)	1.1%	0.006	0.074	94%
BCM (Q SL, g boosted CART)	2.1%	0.008	0.092	95%

Notes: In DGP 1 the true ATE was 0.4 and the bias using a naive estimator based on the mean difference was 20%. WLS is implemented as main terms only regression; hence it is reported as a misspecified estimator.

Table 6.4 - Simulation results for DGP 2 and 3, over 1000 replications: normal endpoint, strong confounder-endpoint association, good and poor overlap

	Relative bias	Variance	RMSE	95 % CI coverage
DGP 2: Normally distributed endpoint, strong confounder-endpoint association, good overlap				
(d) Q misspecified - g misspecified				
OLS regression	-45.9%	0.052	0.292	86%
IPTW	-59.1%	0.067	0.350	98%
PS matching	-34.0%	0.099	0.342	96%
WLS	-50.2%	0.059	0.315	87%
TMLE	-45.7%	0.041	0.272	86%
BCM	-31.4%	0.074	0.299	90%
(e) Machine learning				
Regression (Q super learner)	-8.6%	0.025	0.162	96%
IPTW (g boosted CART)	41.0%	0.036	0.251	99%
WLS (Q OLS, g boosted CART)	2.6%	0.022	0.149	100%
TMLE (Q SL, g boosted CART)	3.1%	0.011	0.106	95%
BCM (Q SL, g boosted CART)	9.8%	0.029	0.174	98%
DGP 3: Normally distributed endpoint, strong confounder-endpoint association, poor overlap				
(d) Q misspecified - g misspecified				
OLS regression	-119.2%	0.050	0.527	40%
IPTW	-160.6%	0.082	0.703	71%
PS matching	-81.1%	0.100	0.453	84%
WLS	-137.9%	0.063	0.606	39%
TMLE	-129.7%	0.046	0.561	35%
BCM	-73.8%	0.072	0.399	74%
(e) Machine learning				
Regression (Q super learner)	-22.0%	0.046	0.233	94%
IPTW (g boosted CART)	100.6%	0.034	0.442	82%
WLS (Q OLS, g boosted CART)	-12.8%	0.025	0.165	99%
TMLE (Q SL, g boosted CART)	5.6%	0.019	0.139	87%
BCM (Q SL, g boosted CART)	12.3%	0.034	0.191	98%

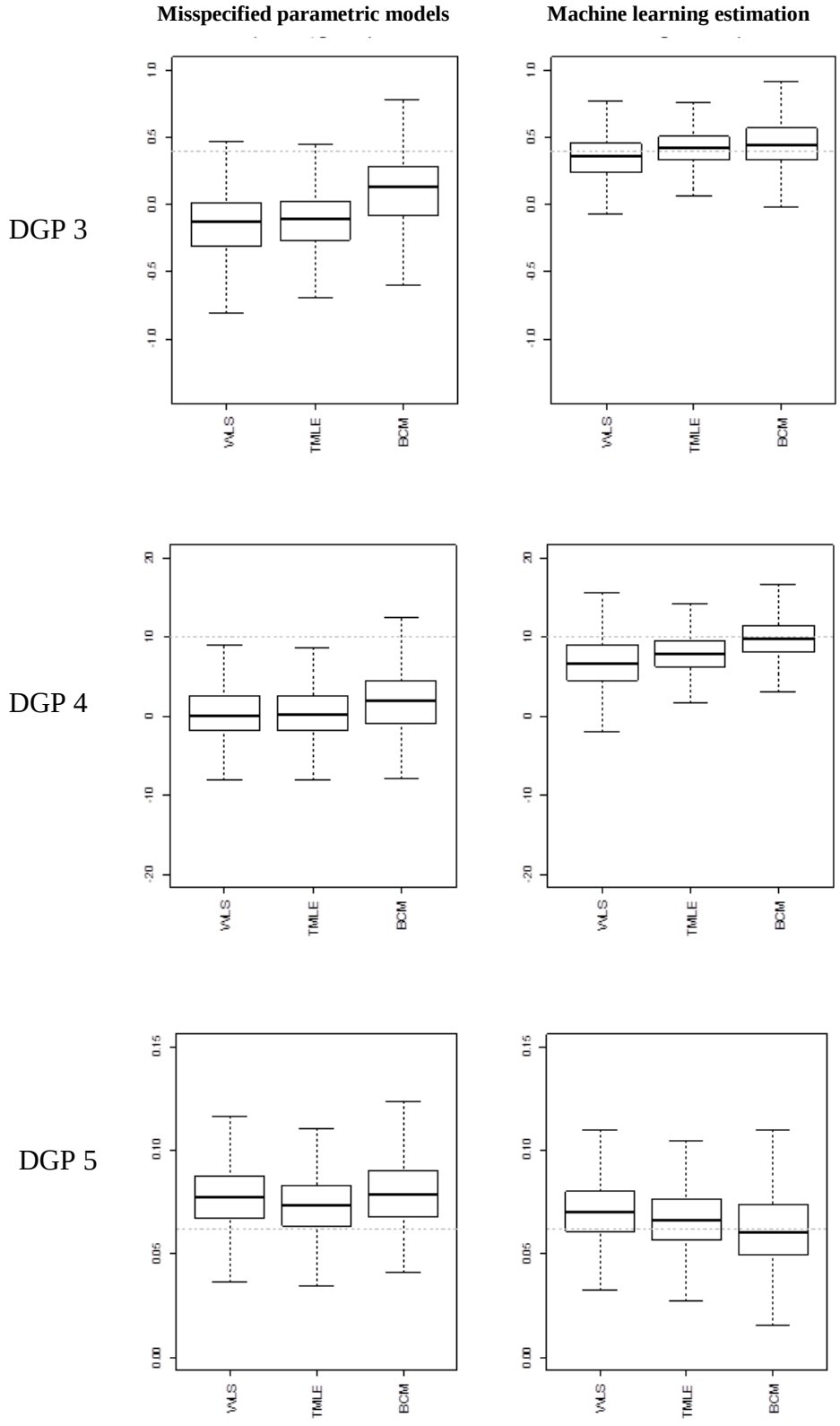
Notes: In DGPs 2 and 3, the true ATE was 0.4 and the biases, using a naive estimator based on the mean difference, were 80% and 190%, respectively.

Table 6.5 - Simulation results for DGP 4 and 5, over 1000 replications: Normal and gamma endpoints, strong confounder-endpoint relationship, poor overlap

	Relative bias	Variance	RMSE	95 % CI coverage
DGP 4: Gamma endpoint, strong confounder-endpoint association, poor overlap				
(d) Q misspecified - g misspecified				
OLS	-93.3%	10.175	9.843	16%
IPTW	-102.7%	11.850	10.817	34%
PS matching	-85.6%	19.120	9.595	59%
WLS	-96.9%	11.475	10.252	19%
TMLE	-96.4%	10.303	10.140	17%
BCM	-80.7%	17.642	9.085	37%
(e) Machine learning				
Regression (Q super learner)	-11.8%	7.600	2.998	90%
IPTW (g boosted CART)	-80.1%	16.585	8.974	62%
WLS (Q OLS, g boosted CART)	-32.1%	11.024	4.612	81%
TMLE (Q SL, g boosted CART)	-20.7%	6.115	3.224	70%
BCM (Q SL, g boosted CART)	-2.5%	6.755	2.610	98%
DGP 5: Two part endpoint, strong confounder-endpoint association, poor overlap				
(d) Q misspecified - g misspecified				
OLS	26.0%	0.0002	0.022	78%
IPTW	15.0%	0.0003	0.019	99%
PS matching	26.9%	0.0004	0.026	93%
WLS	23.9%	0.0003	0.022	83%
TMLE	17.9%	0.0002	0.019	90%
BCM	27.1%	0.0003	0.024	82%
(e) Machine learning				
Regression (Q super learner)	13.5%	0.0002	0.017	91%
IPTW (g boosted CART)	59.4%	0.0003	0.041	72%
WLS (Q OLS, g boosted CART)	12.9%	0.0002	0.017	90%
TMLE (Q SL, g boosted CART)	7.2%	0.0002	0.016	87%
BCM (Q SL, g boosted CART)	-1.1%	0.0004	0.019	95%

Notes: In DGPs 4 and 5, the true ATE was 9.98 and 0.062, respectively. The bias using a naive estimator based on the mean difference was 170% and 150%, respectively.

Figure 6.4 - Estimated ATEs in the simulations



Notes: The boxplots show bias and variation, as median, quartiles and 1.5 times interquartile range for the estimated ATEs across 1,000 replications. The dashed lines are the true values. The left panel provides results for when the PS model and endpoint were estimated with misspecified fixed parametric methods, the right panel for when machine learning estimation was used.

Discussion

This paper finds that combining information from the conditional distribution of the endpoint and the treatment assignment mechanism can reduce bias due to observed confounding. Both methods under comparison, TMLE and BCM, can exploit machine learning estimation of the endpoint regression function and the PS, and can be robust even when the true parametric models are unknown.

We considered these methods, alongside more traditional PS, regression and DR methods in a case study of evaluating the effect of alternative types of hip prosthesis on HRQoL for patients with osteoarthritis. Here, a major challenge was to specify a regression model for an endpoint with a spike at 1, and bounded at a small negative value. This case study motivated the simulation studies, where we generated HRQoL data with skewed distributions and nonlinear response surfaces, in order to create settings where the correct specification of an endpoint regression and PS model is challenging. In the simulations, when machine learning techniques were used to estimate the endpoint regression function and the PS, both TMLE and BCM could almost fully eliminate bias, in contrast to the high bias where misspecified fixed parametric models were used. We found that the relative advantage of TMLE vs. BCM was dependent on the features of the DGPs considered. Confirming the first hypothesis of the simulation study, in favourable settings such as good overlap and moderate association between the confounder and the endpoint, TMLE outperformed BCM in terms of bias and precision. This result corresponds to previous work that found that reweighting estimators outperformed BCM under good overlap (Busso et al., 2011). In a more challenging setting, when overlap was poor, and there was a strong association between the confounders and the endpoint, we found a bias-variance trade off between the methods under comparison: BCM showed less bias, but was more variable than

TMLE. We also found that another DR method, WLS, performed similarly well to TMLE in the less challenging settings such as normally distributed endpoint and good overlap. However, similarly to findings from previous studies (Porter et al., 2011), WLS was outperformed by TMLE in the more challenging DGPs. We followed recent recommendations when reporting CIs after matching estimators (Abadie and Imbens, 2006), and like previous studies, we found that they reported somewhat higher than nominal coverage (Abadie and Imbens, 2011).

Our work extends the previous literature in several aspects. First, this is the first paper that compares the relative performance of BCM and TMLE, and also compares these methods to traditional approaches. Second, while BCM has been proposed with flexible approaches for estimating the endpoint regression function, previous studies used OLS for adjustment (Abadie and Imbens, 2011, Busso et al., 2011). This study considers super learning, a machine learning method for bias correction, and finds that when matching is based on a PS that was also estimated using machine learning, the bias due to model misspecification was almost fully eliminated. We find this result across a range of DGPs including highly nonlinear response surfaces. Third, unlike previous studies that used machine learning only for selected combined methods such as TMLE (Porter et al., 2011), this paper took a systematic approach, and evaluate the impact of using machine learning estimation for single methods, such as regression and IPTW, and for combined methods, such as TMLE and BCM. Our main finding is that combining the PS and endpoint regression from misspecified fixed parametric models does not in itself provide an advantage compared to using these models in single methods such as IPTW. This corresponds to the findings on Kang and Schafer (2007). Similarly, using a machine learning approach alone, for example boosted CART for IPTW is not sufficient to reduce bias. Possible remaining misspecification of the PS

using the boosted CART is indicated by the low coverage rates reported by TMLE, where the nominal standard errors, obtained using the influence curve, are only expected to be valid when $g(.)$ is correct. In the scenarios considered in this study, it was the combined use of machine learning approaches for estimating the endpoint regression and the PS, that helped eliminate most of the bias due to observed confounding.

This work has some caveats. The methods considered and the simulation settings all assume no unobserved confounding. Machine learning methods cannot replace subject matter knowledge when selecting the set of confounders that need to be controlled for (Rubin, 2007). In the case study, while we used a rich set of measured cofounders suggested by previous literature and clinical expert opinion (Pennington et al., 2012), some unobserved confounding such as unobserved patient preferences may prevail.

This paper did not have the scope to compare alternative machine learning approaches. We found that boosted CARTs for estimating the PS, a method that has been found to outperform logistic regression and alternative machine learning approaches (Lee et al., 2010), did not always reduce bias compared to misspecified logistic regression. Hence further machine learning approaches may be considered for the PS, such as random forests (Lee et al., 2010) or neural networks (Westreich et al., 2010). These approaches also have promising application for estimating the endpoint regression function (Austin, 2012).

Any machine learning method relies on subjective choices of the user. For boosted CARTs, tuning parameters such as the shrinkage parameter had to be selected (McCaffrey et al., 2004). When applying the super learner, subject-matter knowledge can be used to select a wide range of prediction algorithms. A richer set of prediction

algorithms, while subject to constraints in computational resources, can facilitate the consistent estimation of the regression function (Polley and van der Laan, 2010b).

This paper considered an innovative DR method, TMLE, alongside a more commonly used DR approach, WLS (Kang and Schafer, 2007a, Freedman and Berk, 2008). We did not consider another standard DR method, augmented inverse probability of treatment weighting (Glynn and Quinn, 2010), because previous studies demonstrated that it can be particularly biased and inefficient under circumstances of poor overlap (Porter et al., 2011, van der Laan and Gruber, 2010). A recently developed improved DR estimator (Rotnitzky et al., 2012), similarly to TMLE, is proposed to have the boundedness property and may be of interest in further methodological comparisons.

This work also opens up areas for further research. In the common settings of poor overlap, an extension of TMLE, collaborative maximum likelihood estimation (C-TMLE) (van der Laan and Gruber, 2010, Gruber and van der Laan, 2010b) can outperform TMLE. C-TMLE uses machine learning to select a sufficient set of covariates for inclusion in $g(.)$ that reduces bias while minimising overall mean squared error. Furthermore, rather than PS matching, multivariate matching approaches such as Genetic Matching warrant consideration (Diamond and Sekhon, 2012, Sekhon, 2011). Genetic Matching uses machine learning to directly maximise covariate balance in the matched data (Grieve et al., 2008, Sekhon and Grieve, 2011, Ramsahai et al., 2011, Kreif et al., 2012, Radice et al., 2012), and can be combined with regression-adjustment.

We conclude that both TMLE and BCM have the potential to reduce bias due to observed confounding, in common settings of dual misspecification, if coupled with machine learning methods for estimating the PS and the endpoint regression function. With the increasing interest in using observational data for deriving measures of

effectiveness of health interventions it is crucial that statistical methods make plausible underlying assumptions (Rubin, 2010), and are relatively robust in challenging settings such as dual misspecification and poor overlap. The methods considered in this paper have the potential to provide robust estimates to inform clinical and policy decisions. TMLE is implemented as a readily available software package (Gruber and van der Laan, 2012c). For BCM, the available packages currently allow for regression adjustment using OLS only (Abadie et al., 2004, Sekhon, 2011). In order to facilitate the uptake of the methods, Appendix 6.1 provides code for the implementation of TMLE and BCM with machine learning.

Acknowledgements

We thank Jan vanderMeulen and Nick Black (LSHTM) for access to the PROMs data, and the Department of Health for funding the primary analysis of the PROMs data. We are grateful to Mark Pennington (LSHTM) for advice on the motivating case study. We thank Rhian Daniel and Karla Diaz-Ordaz (LSHTM) for valuable comments on the manuscript. Funding from the Economic and Social Research Council (Grant no. RES-061-25-0343) is greatly appreciated.

References

Abadie, A., Herr, J. L., Imbens, G. W. & Drukker, D. M. 2004. *NNMATCH: Stata module to compute nearest-neighbor bias-corrected estimators* [Online]. Boston College Department of Economics. Available: http://fmwww.bc.edu/repec/bocode/n/nnmatch.hlp.

Abadie, A. & Imbens, G. W. 2006. Large Sample Properties of Matching Estimators for Average Treatment Effects. *Econometrica,* 74**,** 235-267.

Abadie, A. & Imbens, G. W. 2011. Bias-Corrected Matching Estimators for Average Treatment Effects. *Journal of Business & Economic Statistics,* 29**,** 1-11.

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

Austin, P. C. 2012. Using Ensemble-Based Methods for Directly Estimating Causal Effects: An Investigation of Tree-Based G-Computation. *Multivariate Behavioral Research***,** 115-135.

Bang, H. & Robins, J. M. 2005. Doubly Robust Estimation in Missing Data and Causal Inference Models. *Biometrics* 962–972.

Basu, A. 2011. Economics of individualization in comparative effectiveness research and a basis for a patient-centered health care. *Journal of Health Economics,* 30**,** 549-559.

Basu, A. & Manca, A. 2011. Regression Estimators for Generic Health-Related Quality of Life and Quality-Adjusted Life Years. *Med Decis Making,* 2011 Oct 18. [Epub ahead of print].

Basu, A., Polsky, D. & Manning, W. 2011. Estimating treatment effects on healthcare costs under exogeneity: is there a 'magic bullet'? *Health Services and Outcomes Research Methodology,* 11**,** 1-26.

Buntin, M. B. & Zaslavsky, A. M. 2004. Too much ado about two-part models and transformation?: Comparing methods of modeling Medicare expenditures. *Journal of Health Economics,* 23**,** 525-542.

Busso, M., DiNardo, J. & McCrary, J. 2009. Finite Sample Properties of Semiparametric Estimators of Average Treatment Effects.

Busso, M., DiNardo, J. & McCrary, J. 2011. New Evidence on the Finite Sample Properties of Propensity Score Reweighting and Matching Estimators. *Working paper.*

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

Cole, S. R. & Frangakis, C. E. 2009. The Consistency Statement in Causal Inference: A Definition or an Assumption? *Epidemiology,* 20**,** 3-5 10.1097/EDE.0b013e31818ef366.

Cole, S. R. & Hernán, M. A. 2008. Constructing inverse probability weights for marginal structural models. *Am J Epidemiol,* 168**,** 656-64.

Crump, R. K., Hotz, V. J., Imbens, G. W. & Mitnik, O. A. 2009. Dealing with limited overlap in estimation of average treatment effects. *Biometrika,* 96**,** 187-199.

Davison, A. & Hinkley, D. 1997. *Bootstrap Methods and Their Application,* New York, Cambridge University Press.

Dehejia, R. H. & Wahba, S. 2002. Propensity Score-Matching Methods For Nonexperimental Causal Studies. *The Review of Economics and Statistics,* 84**,** 151-161.

Diamond, A. & Sekhon, J. S. 2012. Genetic Matching for Estimating Causal Effects: A General Multivariate Matching Method for Achieving Balance in Observational Studies. *Review of Economic and Statistics,* Forthcoming.

Dolan, P., Gudex, C., Kind, P. & Williams, A. 1995. A social tariff for EuroQol: results from a UK general population survey. *In:* CENTRE FOR HEALTH ECONOMICS, U. O. Y. (ed.) *Working Papers.*

Elliott, M. R. 2008. Model Averaging Methods for Weight Trimming. *J Off Stat,* 24**,** 517–540.

EuroQol Group 1990. EuroQol--a new facility for the measurement of health-related quality of life. *Health Policy* 16**,** 199–208.

Freedman, D. & Berk, R. A. 2008. Weighting regression by propensity score. *Evaluation Review,* 32**,** 392-409.
Fung, V., Brand, R. J., Newhouse, J. P. & Hsu, J. 2011. Using Medicare Data for Comparative Effectiveness Research: Opportunities and Challenges. *Am J Manag Care,* 17**,** 489-496.
Glynn, A. N. & Quinn, K. M. 2010. An Introduction to the Augmented Inverse Propensity Weighted Estimator. *Political Analysis,* 18**,** 36-56.
Greenland, S., Pearl, J. & Robins, J. M. 1999. Confounding and Collapsibility in Causal Inference. *Statist. Sci. ,* 14**,** 29-46.
Grieve, R., Sekhon, J. S., Hu, T.-w. & 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**,** 1204-1222.
Gruber, S. & van der Laan, M. 2010a. A targeted maximum likelihood estimator of a causal effect on a bounded continuous outcome. *Int J Biostat. ,* 6.
Gruber, S. & van der Laan, M. 2012a. Consistent causal effect estimation under dual misspecification and implications for confounder selection procedures. *Stat Methods Med Res*.
Gruber, S. & van der Laan, M. J. 2010b. An application of collaborative targeted maximum likelihood estimation in causal inference and genomics. *The International Journal of Biostatistics,* 6.
Gruber, S. & van der Laan, M. J. 2012b. Targeted minimum loss based estimation of causal effects on an outcome with known conditional bounds. *International Journal of Biostatistics,* in press.
Gruber, S. & van der Laan, M. J. 2012c. tmle: An R Package for Targeted Maximum Likelihood Estimation. *Journal of Statistical Software,* 51**,** 1-35.
Hampel, F. R. 1974. The influence curve and its role in robust estimation. *Journal of the Americal Statistical Association,* 69**,** 383-393.
Hirano, K. & Imbens, G. W. 2001. Estimation of Causal Effects using Propensity Score Weighting: An Application to Data on Right Heart Catheterization. *Health Services and Outcomes Research Methodology,* 2**,** 259-278.
Hirano, K., Imbens, G. W. & Ridder, G. 2003. Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score. *Econometrica,* 71**,** 1161-1189.
Ho, D. E., Imai, K., King, G. & Stuart, E. A. 2007. Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference *Political Analysis,* 15**,** 199-236.
Imbens, G. & Wooldridge, J. M. 2009b. New Developments in Econometrics. *Lecture Notes, CEMMAP, UCL.*
Imbens, G. W. 2004. Nonparametric Estimation of Average Treatment Effects Under Exogeneity: A Review. *The Review of Economics and Statistics,* 86**,** 4-29.
Kang, J. D. Y. & Schafer, J. L. 2007a. Demystifying Double Robustness: A Comparison of Alternative Strategies for Estimating a Population Mean from Incomplete Data. *Statistical Science,* 22**,** 523-539.
Kang, J. D. Y. & Schafer, J. L. 2007b. Rejoinder: Demystifying Double Robustness: A Comparison of Alternative Strategies for Estimating a Population Mean from Incomplete Data. *Statistical Science,* 22**,** 574-580.
Kooperberg, C. 2010. polspline: Polynomial spline routines.
Kreif, N., Grieve, R., Radice, R., Sadique, Z., Ramsahai, R. & Sekhon, J. S. 2012. Methods for Estimating Subgroup Effects in Cost-Effectiveness Analyses That Use Observational Data. *Medical Decision Making,* 32**,** 750-63.
Lee, B. K., Lessler, J. & Stuart, E. A. 2010. Improving propensity score weighting using machine learning. *Statistics in Medicine,* 29**,** 337-346.
Lumley, T. 2012. mitools: Tools for multiple imputation of missing data.
Manning, W. G., Basu, A. & Mullahy, J. 2005. Generalized modeling approaches to risk adjustment of skewed outcomes data. *Journal of Health Economics,* 24**,** 465-488.

McCaffrey, D., Ridgeway, G. & Morral, A. 2004. Propensity Score Estimation with Boosted Regression for Evaluating Causal Effects in Observational Studies. *Psychol Methods,* 9**,** 403-425.

Moore, K. L. & van der Laan, M. J. 2009. Covariate adjustment in randomized trials with binary outcomes: Targeted maximum likelihood estimation. *Statistics in Medicine,* 28**,** 39-64.

NICE. 2012. *Health Technology Appraisal. Total hip replacement and resurfacing arthroplasty for the treatment of pain or disability resulting from end stage arthritis of the hip (Review of technology appraisal guidance 2 and 44). Final scope.* [Online]. Available: http://www.nice.org.uk/nicemedia/live/13690/61348/61348.pdf.

Ousey, K. & Cook, L. 2011. Understanding patient reported outcome measures (PROMs). . *British Journal of Community Nursing 2011,* 16**,** 80-82

Pennington, M., Grieve, R., Sekhon, J. S., Gregg, P., Black, N. & van der Meulen, J. 2012. Cemented, cementless and hybrid prostheses for total hip replacement: a cost-effectiveness analysis. *British Medical Journal,* submitted.

Petersen, M. L., Porter, K., Gruber, S., Wang, Y. & Laan., M. J. v. d. 2010. Diagnosing and Responding to Violations in the Positivity Assumption. *U.C. Berkeley Division of Biostatistics Working Paper Series*.

Polley, E. C. & van der Laan, M. J. 2010a. *R package "SuperLearner"* [Online]. Available: http://cran.r-project.org/web/packages/SuperLearner/index.html.

Polley, E. C. & van der Laan, M. J. 2010b. Super Learner In Prediction. *U.C. Berkeley Division of Biostatistics Working Paper Series.*

Porter, K. E., Gruber, S., Laan, M. J. v. d. & Sekhon, J. S. 2011. The Relative Performance of Targeted Maximum Likelihood Estimators. *The International Journal of Biostatistics,* 7.

PROMs 2010. Provisional Monthly Patient Reported Outcome Measures (PROMs) in England. April 2009 - April 2010: Pre- and post-operative data: Experimental Statistics. The Health and Social Care Information Centre.

R Development Core Team 2011. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing.

Radice, R., Grieve, R., Ramsahai, R., Kreif, N., Sadique, Z. & 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).

Ramsahai, R., Grieve, R. & Sekhon, J. S. 2011. Extending Iterative Matching Methods: An Approach to Improving Covariate Balance that Allows Prioritisation. *Health Services and Outcomes Research Methodology,* 11**,** 95-114.

Ridgeway, G., McCaffrey, D., Morral, A., Griffin, B. A. & Burgette, L. 2006. twang: Toolkit for Weighting and Analysis of Nonequivalent Groups. 1.2-5 ed.

Ridgeway, G. & McCaffrey, D. F. 2007. Comment: Demystifying Double Robustness: A Comparison of Alternative Strategies for Estimating a Population Mean from Incomplete Data. *Statist. Sci.,* 22**,** 540-543.

Robins, J. 1986. A new approach to causal inference in mortality studies with a sustained exposure period – application to control of the healthy worker survivor effect. . *Mathematical Modelling.* , 7**,** 1393–1512.

Robins, J., Rotnitzky, A. & Zhao, L. P. 1994. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association,* 89**,** 846-866.

Robins, J., Sued, M., Lei-Gomez, Q. & Rotnitzky, A. 2007. Comment: Performance of Double-Robust Estimators When "Inverse Probability" Weights Are Highly Variable. *Statistical Science,* 22**,** 544-559.

Rosenbaum , P. R. & Rubin, D. B. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika,* 70**,** 41-55.

Rotnitzky, A., Lei, Q., Sued, M. & Robins, J. M. 2012. Improved double-robust estimation in missing data and causal inference models. *Biometrika,* 99**,** 439-456.

Rubin, D. 1996. Multiple Imputation after 18+ years. *Journal of the Americal Statistical Association,* 91**,** 473-489.

Rubin, D. B. 1973. The Use of Matched Sampling and Regression Adjustment to Remove Bias in Observational Studies. *Biometrics,* 29.
Rubin, D. B. 1979. Using multivariate matched sampling and regression adjustment to control bias in observational studies. *Journal of the American Statistical Association***,** 318–328.
Rubin, D. B. 1997. Estimating Causal Effects from Large Data Sets Using Propensity Scores. *Annals of Intenal Mediciene***,** 757-763.
Rubin, D. B. 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. B. 2010. On the limitations of comparative effectiveness research. *Statistics in Medicine,* 29**,** 1991-1995.
Rubin, D. B. & Thomas, N. 2000. Combining propensity score matching with additional adjustments for prognostic covariates. *Journal of the American Statistical Association,* 95**,** 573–585.
Sekhon, J. S. 2011. Matching: multivariate and propensity score matching with automated balance search. *Journal of Statistical Software*.
Sekhon, J. S. & Grieve, R. D. 2011. A Matching Method for Improving Covariate Balance in Cost-Effectiveness Analyses. *Health Economics,* doi: 10.1002/hec.1748.
Setoguchi, S., Schneeweiss, S., Brookhart, M., Glynn, R. & EF, C. 2008. Evaluating uses of data mining techniques in propensity score estimation: a simulation study. *Pharmacoepidemiol Drug Saf,* 17**,** 546-55.
Stitelman, O. M. & van der Laan, M. J. 2010. Collaborative targeted maximum likelihood for time to event data. *Int J Biostat,* 6.
Stuart, E. A. 2010. Matching Methods for Causal Inference: A review and a look forward. *Statistical Science,* 25.
Tunis, S. R., Benner, J. & McClellan, M. 2010. Comparative effectiveness research: Policy context, methods development and research infrastructure. *Statistics in Medicine,* 29**,** 1963-1976.
van der Laan, M. J. 2010. Targeted Maximum Likelihood Based Causal Inference: Part I. *The International Journal of Biostatistics*.
van der Laan, M. J. & Dudoit, S. 2003. Unified cross-validation methodology for se- lection among estimators and a general cross-validated adaptive epsilon-net estimator: Finite sample oracle inequalities and examples. *Technical report, Division of Biostatistics, University of California, Berkeley*.
van der Laan, M. J. & Gruber, S. 2010. Collaborative Double Robust Targeted Maximum Likelihood Estimation. *The International Journal of Biostatistics* 6.
van der Laan, M. J., Polley, E. C. & Hubbard, A. E. 2007. Super Learner. *Statistical Applications in Genetics and Molecular Biology,* 6.
van der Laan, M. J. & Rubin, D. 2006. Targeted maximum likelihood learning. *The International Journal of Biostatistics,* 2.
Waernbaum, I. 2011. Model misspecification and robustness in causal inference: comparing matching with doubly robust estimation. *Stat Med,* DOI: 10.1002/sim.4496.
Westreich, D. & Cole, S. R. 2010. Invited Commentary: Positivity in Practice. *American Journal of Epidemiology*.
Westreich, D., Lessler, J. & Funk, M. 2010. Propensity score estimation: neural networks, support vector machines, decision trees (CART), and meta-classifiers as alternatives to logistic regression. *Journal of Clinical Epidemiology,* 63**,** 826-33.
Xu, S., Ross, C., Raebel, M. A., Shetterly, S., Blanchette, C. & Smith, D. 2010. Use of Stabilized Inverse Propensity Scores asWeights to Directly Estimate Relative Risk and Its Confidence Intervals. *Value in Health,* 13**,** 273-277.

Appendix 6.1- R code for the implementation of TMLE and BCM

This section provides code for the implementation of TMLE and BCM, coupled with machine learning estimation approaches proposed in the paper, using the R statistical software (R Development Core Team, 2011). We also present code for the machine learning methods: "super learning" for predicting the endpoint and boosted CARTs for estimating the PS. The user-written functions implemented here call some pre-written R routines, for example the `tmle` (Gruber and van der Laan, 2012), `Matching` (Sekhon, 2011), `Super Learner` (Polley and van der Laan, 2010) and `twang` (Ridgeway et al., 2006) packages. These packages need to be installed and loaded in the R workspace order to use the functions presented here.

First, the necessary libraries need to be loaded:

```
library(splines)
library(twang)
library(SuperLearner)
library(Matching)
```

Then, we define the data frame, `data`, used in the analysis. This dataset includes the endpoint `q2_eq5d_index`, the treatment indicator `Hyb` and the covariates.

```
data=as.data.frame(dataset_men_M1)
```

We create the design matrix that will be used by some of the corresponding user defined R functions:

```
designreg=glm( Hyb ~ age
         +  q1_score
         +  q1_eq5d_index
         +  factor(IMD)
         +  ASAgrade1
         +  ASAgrade2
         +  q1_disability
         +  obese
```

```
          +  morbobese
          +  heart_disease
          +  high_bp
          +  stroke
          +  circulation
          +  lung_disease
          +  diabetes
          +  kidney_disease
          +  nervous_system
          +  liver_disease
          +  cancer
          +  depression
          +  Consultant
          +  TC, family=binomial(link="logit"), data=data)

design=model.matrix(designreg)
```

We also define the formula for the PS:

```
boost.CART.form <- as.formula(Hyb ~ age
          +  q1_score
          +  q1_eq5d_index
          +  IMD
          +  ASAgrade1
          +  ASAgrade2
          +  q1_disability
          +  obese
          +  morbobese
          +  heart_disease
          +  high_bp
          +  stroke
          +  circulation
          +  lung_disease
          + diabetes
          +  kidney_disease
          +  nervous_system
          +  liver_disease
          +  cancer
          +  depression
          +   Consultant
          +   TC)
```

Estimating the PS

The function named `boost.CART.func` estimates a PS using boosted logistic CARTs. The function takes the arguments `formula`, which specifies the variables the user wants to include in the PS model, and `data`. The function sets the tuning parameters to the values recommended by the developers (McCaffrey et al., 2004), and

maximises balance based on the mean of the KS statistic. The function returns the estimated PS, the linear predictor of the estimated PS, and the IPT weight.

```
boost.CART.func = function(formula, data) {

            boost.CART.ps=ps(formula=formula,
            data=data,
            shrinkage = 0.0005,
            n.trees=10000,
            interaction.depth=2,
            iterlim=20000,
            stop.method="ks.mean")

            w.boost.CART = boost.CART.ps$w
            ps.boost.CART = boost.CART.ps$ps
                  linpred.boost.CART=
predict.gbm(boost.CART.ps$gbm.obj,            data,
boost.CART.ps$n.trees)
            return(list(w.boost.CART=w.boost.CART,
ps.boost.CART=ps.boost.CART,
linpred.boost.CART=linpred.boost.CART))

}
```

By calling the function, we can obtain the estimated PS, and attach it to `data`:

```
res.boost<-boost.CART.func(boost.CART.form, data=data)

ps.boost = unlist(res.boost$ps.boost.CART)

data=cbind(data, ps.boost)
rm(ps.boost)
```

Creating PS matched data

Before BCM is performed, a matched dataset needs to be created. The function named `PSmatch.function` calls the `Match()` function (Sekhon, 2011), taking the following arguments: `data,` and `pscore`, the estimated propensity score. The function returns the original return object of the `Match()` function (Sekhon, 2011), which includes the matched dataset, here named `mtchout.Y`. The function also returns the matching frequency weights `K`, as well as transformed version of this vector, `Kprime`, later used for calculating the standard errors around for matching estimator (Abadie and Imbens, 2011).

```
PSmatch.function=function(data,pscore)  {

      mtchout.Y=Match(Y=data$q2_eq5d_index,Tr=data$Hyb,X=pscore,

            estimand="ATE", ties=TRUE)
            n <- length(data$q2_eq5d_index)
            K <- rep(0, n)
            names(K) <- 1:n
            Kprime <- K
            extra <- by(mtchout.Y$MatchLoopC[,3],
                   mtchout.Y$MatchLoopC[,2], sum)
            K[rownames(extra)] <-  extra
            Kprime.extra <- by(mtchout.Y$MatchLoopC[,3],
                        mtchout.Y$MatchLoopC[,2],
                        function(x){sum(x^2)})
            Kprime[rownames(Kprime.extra)] <-  Kprime.extra
            return(list(mtchout.Y=mtchout.Y,K=K, Kprime=Kprime))
                                                                  }
```

Now the function can be called to obtain the matched dataset, `ps.match.data.w.boost`, and the vectors of frequency weights (`K.boost` and `K.prime.boost`).

```
PS.match.object.boost=PSmatch.function(data,data$ps.boost)

ps.match.data.boost=PS.match.object.boost$mtchout.Y
ps.match.data.w.boost <-
rbind(data[ps.match.data.boost$index.treated,],
data[ps.match.data.boost$index.control,])

ps.match.data.w.boost <- cbind(ps.match.data.w.boost,
weights=c(ps.match.data.boost$weights,ps.match.data.boost$weight
s))

K.boost=PS.match.object.boost$K
K.prime.boost=PS.match.object.boost$Kprime
```

Predicting the expected potential outcome with the super learner

The user defined function named `my.SL.ate` predicts the expected potential outcomes under treated and control states, using the Super Learner (Polley and van der Laan, 2010). The function takes the arguments `W` (the covariates), `A` (the observed treatment), and `Y`, the observed endpoint. We use super learner for estimating two regressions functions, stratified by treatment, as suggested by Abadie and Imbens (2011). This provides the algorithm flexibility to select different models for estimating

the potential outcomes under treatment and control states. The function returns the original "Super Learner" object that includes information on the final models selected, `m0` for the regression function selected to estimate the potential outcome under control and `m1` under treatment. The predictions for the potential outcomes are stored in the matrix `Q.SL.object`, which includes two vectors, the predicted potential outcomes under control and treatment, each with the length of the number of individuals in the sample.

```
my.SL.ate=function(W,A,Y) {

      matrix <- data.frame(W)

       m0 <- SuperLearner(Y[A==0], matrix[A==0,], newX = matrix,
            SL.library =  my.SL.library.short,
            family = gaussian())

      m1 <- SuperLearner(Y[A==1], matrix[A==1,], newX = matrix,
             SL.library = my.SL.library.short,
            family = gaussian())

      Yhat.0 <- m0$SL.predict
      Yhat.1 <- m1$SL.predict

      Q.SL.object=cbind(Yhat.0,Yhat.1)
      ate_SL=mean(Yhat.1-Yhat.0)
      return(list(Q.SL.object=Q.SL.object,m0=m0,m1=m1))
                                                    }
```

This function can be extended to incorporate weights, the modified function is named `my.SL.ate.matchw`. This is necessary, because for the bias-corrected matching estimator it is recommended that regression predictions are obtained using data weighted with the matching frequency weights, `K` (Abadie and Imbens, 2011).

```
my.SL.ate.matchw=function(W,A,Y,K) {

 matrix <- data.frame(W)

 m0 <- SuperLearner(Y[A==0], matrix[A==0,], newX = matrix,
       SL.library =  my.SL.library.short,
       family = gaussian(),obsWeights=K[A==0])

 m1 <- SuperLearner(Y[A==1], matrix[A==1,], newX = matrix,
       SL.library = my.SL.library.short,
      family = gaussian(),obsWeights=K[A==1])
```

```
        Yhat.0 <- m0$SL.predict
        Yhat.1 <- m1$SL.predict

        Q.SL.object=cbind(Yhat.0,Yhat.1)
        ate_SL=mean(Yhat.1-Yhat.0)

        return(list(Q.SL.object=Q.SL.object,ate_SL=ate_SL,m0=m0,m1=m1))

}
```

Before calling the function, the super learner library, including all the prediction algorithms selected by the user, needs to be defined.

```
my.SL.library.short <-c("SL.glm","SL.glm.interaction", "SL.polymars")
```

Here we include the algorithms `"SL.glm", "SL.glm.interaction"` and `"SL.polymars".`

```
        SL.object=my.SL.ate(design,data$Hyb,data$q2_eq5d_index),
```

then with matching frequency weights.

```
        SL.object.BCM.boost    <-  my.SL.ate.matchw(design,data$Hyb,
                                data$q2_eq5d_index,K=K.boost)
```

Implementing BCM

The function `BCM.AI` implements the BCM estimator proposed by Abadie and Imbens (2011). The function takes the following objects: `Y` (the observed endpoint), `A` (the observed treatment), `d.match`, the matched data obtained from the PS matching, `Yhat.0`, the vector of predicted potential outcomes under control, and `Yhat.1`, the vector of predicted potential outcome under treatment, as well as the `K` and `Kprime` vectors, describing the matching frequency weights. The function returns the point estimate of the ATE, `tau`, and the estimated variance of the ATE, `AIvar`.

```
BCM.AI <- function(Y,A, d.match,Yhat.0,Yhat.1, K, Kprime) {
      Y_j.0 <- Y_j.1 <- Y
      Ycounterfactual <- by(Y[d.match$MatchLoopC[,2]],
                        d.match$MatchLoopC[,1], mean)
      Y_j.1[A==0] <- Ycounterfactual[A==0]
      Y_j.0[A==1] <- Ycounterfactual[A==1]

      mu_0.Xi <- Yhat.0
      mu_0.Xj <- by(Yhat.0[d.match$MatchLoopC[,2]],
                  d.match$MatchLoopC[,1], mean)
      mu_1.Xi <- Yhat.1
      mu_1.Xj <- by(Yhat.1[d.match$MatchLoopC[,2]],
                  d.match$MatchLoopC[,1], mean)

      Ytilde.0 <- Y_j.0
      Ytilde.1 <- Y_j.1
      Ytilde.0[A==1] <- Y_j.0[A==1] + mu_0.Xi[A==1] - mu_0.Xj[A==1]
      Ytilde.1[A==0] <- Y_j.1[A==0]  + mu_1.Xi[A==0] - mu_1.Xj[A==0]

      tau.bcm <- mean(Ytilde.1 - Ytilde.0)
      n <- length(Y)
      sigmasq.X <- 1/(2*n) * sum((Ytilde.1      - Ytilde.0 -
      tau.bcm)^2)

      var.SATE <- 1/n^2 *  sum(( 1 + K)^2 * sigmasq.X)
      var.PATE <- 1/n^2 * sum ( (Ytilde.1 - Ytilde.0 - tau.bcm)^2 +
                  ( K^2 + 2*K - Kprime)*sigmasq.X)

      return(list(tau = tau.bcm, AIvar=max(var.SATE, var.PATE),
                                          var.PATE = var.PATE))

}
```

The point estimate and CI around the ATE is estimated by calling the function:

```
BCM.SL.boost <- BCM.AI(dataset_men_M1$q2_eq5d_index,
                     dataset_men_M1$Hyb,
                     ps.match.data.boost,
                     SL.object.BCM.boost$Q.SL.object[,1],
                     SL.object.BCM.boost$Q.SL.object[,2],
                     PS.match.object.boost$K,
                     PS.match.object.boost$Kprime)
```

The estimated ATE, with its standard error can be obtained as follows:

```
coef.BCM.SL.boost <- BCM.SL.boost$tau
se.BCM.SL.boost <- sqrt(BCM.SL.boost$AIvar)
```

Implementing TMLE

The R package `tmle()` offers an accessible implementation of TMLE (Gruber and van der Laan, 2012). The `tmle()` function takes the arguments `Y` (the observed

endpoint), `A` (the observed treatment),`W` (the design matrix), and `Q` (the two vectors of potential outcomes). As a default, the `tmle()` function applies logistic fluctuation, and bounds the endpoint between the observed minimum and maximum values (here, between -0.59 and 1).

```
tmle.SL.boost=tmle(Y=data$q2_eq5d_index,A=data$Hyb,W=design, Q=Q.SL,
g1W=data$pscore.boost)
```

The estimated ATE and its confidence intervals can be then obtained:

```
coef.tmle.SL.boost <- tmle.SL.boost$estimates$ATE$psi
ciU.tmle.SL.boost <- summary(tmle.SL.boost)$estimates$ATE$CI[2]
ciL.tmle.SL.boost <- summary(tmle.SL.boost)$estimates$ATE$CI[1]
```

Chapter 7 – Discussion

7.1 Introduction

Cost-effectiveness analyses (CEA) often make use of non-randomised studies (NRS), when randomised controlled trials (RCTs) are inappropriate or insufficient to provide the evidence required to inform decisions (NICE, 2008). Here the main methodological challenge is to address potential selection bias, due to confounding. Where individual patient data (IPD) from NRS is available for estimating parameters for CEA, selection bias can be addressed with appropriate statistical methods (Polsky and Basu, 2006). At the outset of this thesis, there was no comprehensive guidance on using statistical methods for CEA that use NRS, which was raised as a priority for methodological research in a recent review of NICE methods for health technology assessment (HTA) (Longworth et al., 2009). This thesis helped address this gap in the literature.

The overall objective of this thesis was to consider alternative statistical methods for addressing selection bias in CEA that use patient-level observational data. The specific objectives were:

1. To develop and apply a new checklist for assessing the underlying assumptions made by statistical methods for addressing selection bias in CEA, that use patient-level observational data;
2. To consider which further statistical methods from the general causal inference literature may be appropriate for addressing selection bias in CEA;
3. To compare the relative performance of propensity score (PS) approaches and Genetic Matching (GM), a multivariate matching method, for estimating subgroup-effects in CEA;

4. To compare methods that combine regression with PS approaches for addressing selection bias when estimating incremental effectiveness and cost-effectiveness parameters.

The next section discusses the overall findings from the thesis. Sections 7.3 and 7.4 highlight the contributions of the thesis to the methodological literature. Sections 7.5 and 7.6 summarise the limitations and identify areas for future research. Section 7.7 and 7.8 discuss the implications for applied researchers and policy making. The last section concludes.

7.2 Overall findings of the thesis

The methods currently recommended in CEA for addressing selection bias make some key underlying assumptions, which the conceptual review (chapter 2) examined. The unconfoundedness assumption implies that all variables which are prognostic for the cost or effectiveness endpoints, and also influence treatment assignment, are observed. The assumption of good overlap across covariate distributions between the treatment groups requires that there are no combinations of observed covariates which fully predict assignment to the treated or control group. Regression and PS approaches also assume that the relationship between the covariates and the endpoints, or the covariates and the treatment assignment is correctly specified. The conceptual review found that for CEA that use NRS, the correct specification of endpoint regression models and the PS can be challenging, especially when there is an interest in cost-effectiveness estimates at the subgroup-level. Due to these challenges, structural uncertainty from the choice or specification of the statistical method needs to be acknowledged when presenting and interpreting results from a CEA that use patient-level observational data.

I developed a new checklist (research paper 1, chapter 3) for critical appraisal of applied CEA, informed by the findings from the conceptual review. I then applied the checklist in a systematic review of published CEA. A key finding was that the majority of the 81 studies reviewed relied on the unconfoundedness assumption, and used regression or matching to try and address selection bias, without appropriately assessing their underlying assumptions. Half of the studies did not consider structural uncertainty from the choice of statistical method.

The conceptual review (chapter 2) identified alternative statistical methods that have the potential to make less restrictive assumptions about the specification of the PS or the endpoint regression model. These methods are GM, a multivariate matching method that uses a machine learning algorithm to directly maximise covariate balance; double-robust (DR) methods and regression-adjusted matching. I also found that machine learning techniques hold promise for estimating the PS and the endpoint regression. I contrasted the relative performance of these methods in simulation studies, informed by case studies that represented typical circumstances faced by CEA.

In research paper 2 (chapter 4) I compared GM, PS matching and inverse probability of treatment weighting (IPTW) for estimating cost-effectiveness for patient subgroups. In the motivating CEA of Drotrecogin alfa activated (DrotAA) in patients with severe sepsis, I found that covariate balance for the subgroups of interest improved, when each method aimed to optimise balance by subgroup. GM achieved the best balance, and the cost-effectiveness results for subgroups differed by method, with IPTW producing poor covariate balance and reporting the widest confidence intervals (CIs) of the estimated incremental net benefit (INB). The simulations demonstrated that the key criterion for choosing among statistical methods is the covariate balance created for each subgroup.

I found that GM, unlike PS matching or IPTW, was relatively robust to functional form misspecification of the PS, such as the omission of nonlinear terms.

In research paper 3 (chapter 5) I considered methods that combine the PS with endpoint regression (combined methods), such as DR methods and regression-adjusted matching, and compared them to regression, PS matching and IPTW. When contrasting these methods in the CEA of DrotAA, I found that combined methods reported differing point estimates and narrower CIs of the INB than methods that relied on the estimated PS only. In the simulations, I found that using combined methods could reduce bias and root mean squared error (RMSE) in the estimated INB when compared to using PS matching or IPTW, across a range of scenarios characteristic of CEA. In the realistic scenario of functional form misspecification of both the PS and the endpoint regression (dual misspecification), and unstable IPT weights, regression-adjusted matching reported lower bias and RMSE than the DR methods considered.

Research paper 4 (chapter 6) considered extensions of these combined methods, for estimating incremental effectiveness parameters. The motivating case study was an evaluation of the effect of alternative hip prostheses on patients' health related quality of life (HRQoL), where the HRQoL endpoint had a skewed distribution with a spike at 1. I considered an innovative DR method, targeted maximum likelihood estimation (TMLE), and compared it to bias-corrected matching (BCM), where initially both methods were implemented with fixed parametric models. I then coupled both methods with using machine-learning techniques to estimate the PS and the endpoint regression. In the simulation study I found that both TMLE and BCM reported relatively robust estimates of treatment effects when coupled with machine-learning techniques, as opposed to when using fixed parametric models that were misspecified. When overlap

between the covariate distributions was good, TMLE reported the lowest bias and RMSE, and BCM performed best when overlap was poor.

7.3 Main contributions of the thesis

This thesis contributed to the literature on analytical methods for CEA (Hoch et al., 2002, Willan et al., 2004, Nixon and Thompson, 2005, Polsky and Basu, 2006, Sekhon and Grieve, 2011), drawing on insights from the causal inference (Rosenbaum and Rubin, 1983, Robins et al., 2000, Imbens and Wooldridge, 2009a) and health econometrics literature (Jones, 2007, Jones, 2010, Jones and Rice, 2011). An important contribution of this thesis is that it directly compares methods across these strands of literature which tend to progress independently. For example research paper 4 contrasts a two-part model, recommended in the health econometrics and CEA literature for handling HRQoL data (Buntin and Zaslavsky, 2004, Basu and Manca, 2011), with TMLE, a recently recommended DR method from the causal inference literature. In the research papers contrasting statistical methods, I generated hypotheses based on insights from the general causal inference literature, but grounded in typical features of CEA data (chapter 2). The simulation scenarios were informed by the systematic review of applied CEA (research paper 1) and by the motivating examples (research paper 2, 3 and 4). The following sections describe the specific contributions of this thesis.

7.3.1 Developing a new checklist for critical appraisal of statistical methods for addressing selection bias in CEA

I developed a critical appraisal tool for assessing the underlying assumptions made by statistical methods for addressing selection bias in CEA that use patient-level observational data (research paper 1). This checklist complements previous quality-assessment tools and methodological guidance (Drummond et al., 2005, Philips et al.,

2006, Glick et al., 2007), which did not include specific criteria for the statistical analysis of patient-level data from observational studies. The systematic review presented in research paper 1 was the first study that assessed an important aspect of the quality of CEA which use patient-level observational data: the statistical methods used to address selection bias. I found that the underlying assumptions of statistical methods were not appropriately assessed, and statistical approaches that have the potential to make less restrictive assumptions were not used in practice. The main contribution of the new checklist is that it can raise awareness of the assumptions underlying alternative statistical methods. The checklist should prove helpful for the applied researcher conducting statistical analysis, for reviewers and journal editors considering future CEA articles, and for decision makers appraising and using published CEA.

7.3.2 Methodological insights on statistical approaches for estimating subgroup effects in CEA that use observational data

This thesis provided the first simulation study which compared alternative statistical methods for reducing selection bias when cost-effectiveness results are required for patient subgroups. Previous methodological guidance recommended regression methods for estimating cost-effectiveness parameters for subgroups (Nixon and Thompson, 2005, Willan et al., 2004). These methods can, however, be sensitive to the choice of model specification in a NRS setting (Ho et al., 2007). In research paper 2, I considered alternative methods: PS matching, IPTW, and GM for estimating subgroup effects in CEA. This paper extended previous work by Sekhon and Grieve (2011) who compared GM to PS matching in reporting overall cost-effectiveness parameters. Research paper 2 considered the context of subgroup analysis, and included IPTW as a methodological comparator.

7.3.3 Considering methods that combine the PS and endpoint regression for estimating parameters for CEA

This thesis considered approaches that can combine information from the treatment assignment mechanism with that from the cost and effectiveness endpoint models for the first time in CEA. Research paper 3 considered DR methods such as weighted regression and augmented inverse probability of treatment weighting (AIPTW), and regression-adjusted matching. Previous simulation studies considered DR methods for a generic continuous endpoint (Porter et al., 2011, Kang and Schafer, 2007), and for cost analysis (Basu et al., 2011). Research paper 3 extended these studies to the bivariate CEA context, where the correct specification of regression models for both the cost and effectiveness endpoints is a concern.

Research paper 4 extended this work by investigating combined approaches on the forefront of causal inference research, TMLE and BCM, for estimating treatment effectiveness. This study also extended a previous simulation study which estimated parameters for HRQoL data using regression methods (Basu and Manca, 2011). This is the first study to use machine learning estimation techniques to estimate incremental effectiveness.

7.4 Other general methodological contributions emerging from the thesis

Findings from this thesis also contributed to the general causal inference literature on estimating treatment effects, by providing new methodological comparisons, as well as by considering statistical methods in the bivariate context of CEA for the first time.

7.4.1 New methodological comparisons

This thesis contributed to the limited comparative work on the relative performance of DR and matching approaches (Waernbaum, 2011, Busso et al., 2011, Busso et al., 2009, Basu et al., 2011), by presenting two simulation studies that compared regression-adjusted matching to DR methods. In research paper 3, I considered the implementation of regression-adjusted matching as "non-parametric pre-processing", proposed by Ho et al. (2007). This approach has not been considered in simulation studies before. I extended previous simulation studies (Kang and Schafer, 2007, Porter et al., 2011, Basu et al., 2011, Freedman and Berk, 2008), which found that, with unstable IPT weights and dual misspecification, DR methods can report more biased and less efficient results than ordinary least squares (OLS) regression. I found that with a more severe misspecification, weighted regression could outperform a misspecified regression estimator. I also found that regression-adjusted matching can be more robust to misspecification of the PS and the endpoint regression, than DR methods.

Research paper 4 compared TMLE with BCM for the first time. I extended the findings of Busso et al. (2011), who showed that with correctly specified PS and poor overlap, BCM provides less biased estimates than reweighting estimators. I find similar results, that for the more realistic scenario of misspecified PS, BCM outperformed TMLE. The main finding of research paper 4 was that when coupled with machine learning estimation methods, both TMLE and BCM could reduce bias due to dual misspecification, as opposed to using OLS for adjustment, considered in previous studies (Abadie and Imbens, 2011, Busso et al., 2011).

7.4.2 Insights from using machine learning methods for estimating treatment effects

This thesis followed recent recommendations that suggest machine learning approaches for estimating the PS (Westreich et al., 2010, Lee et al., 2010) and the endpoint regression function (Austin, 2012, Porter et al., 2011). An important contribution of research paper 4 was that it took a systematic approach in comparing advanced, combined approaches to traditional methods, therefore extended previous studies which used machine learning only for selected methods such as TMLE (Porter et al., 2011). Research paper 4 demonstrated the impact of moving from single (such as IPTW) to combined methods (DR methods or BCM), using both fixed parametric methods and machine learning techniques. In addition, both within the single and combined approaches, I looked at the impact of moving from fixed to machine learning approaches. I found that in challenging circumstances such as dual misspecification and poor overlap, combined methods using fixed parametric models reported high bias, and when machine learning techniques were used, this bias was much reduced.

7.4.3 Considering statistical methods in a bivariate context

This thesis considers IPTW, PS matching, GM (research paper 2), DR methods and regression-adjusted matching (research paper 3) in the bivariate context of CEA. Here a general challenge is that statistical methods need to recognise that the endpoints of interest can be correlated (O'Hagan and Stevens, 2001, Nixon et al., 2010). The CEA of DrotAA (research papers 2 and 3) demonstrated that for each of the statistical approaches considered, the non-parametric bootstrap can be used to calculate uncertainty around the estimates of incremental cost-effectiveness, while recognising the correlation between the endpoints. The simulation study in research paper 2

highlighted that in CEA, potential confounders can differ between the cost and effectiveness endpoints, for example, baseline HRQoL might influence the QALY but not the cost endpoint. The simulations showed that in order to reduce selection bias, balance needs to be maximised for potential confounders for both endpoints.

7.5 Limitations

While this thesis presented a comprehensive assessment and comparison of alternative statistical methods for addressing selection bias in CEA that use patient-level observational data, it has some limitations. In this section, I acknowledge general weaknesses regarding the scope of thesis, the range of statistical methods considered and the circumstances considered for the methodological comparisons.

7. 5.1 Scope of the thesis

Alternative use of observational data for CEA

Observational data can be used to estimate a wide range of parameters in CEA, including incremental cost and effectiveness endpoints, but also other parameters such as relative risks (Drummond, 1998, Philips et al., 2006). The type of observational data available to obtain these parameters also varies, including IPD, the focus of this thesis, but studies commonly use aggregate data from published observational studies (Cooper et al., 2007). Examples of settings that this thesis did not cover include using published, aggregate data to derive parameters such as effectiveness or baseline probabilities (Philips et al., 2006), or where patient-level data is used to develop risk equations to populate decision models (Caro et al., 2012).

While the focus of the simulation studies (research paper 2 and 3) was using IPD from a single observational study to calculate incremental effectiveness and cost-effectiveness

parameters, the case studies made more general use of observational data. The DrotAA case study (research paper 2 and 3) combined patient-level mortality data with aggregate estimates of long term survival and quality of life. Research paper 4 reanalysed a large observational dataset on health outcomes following total hip replacement, where estimates of relative treatment effectiveness on HRQoL provided input parameters for a decision analytical model (Pennington et al., 2012). Here, when applying the estimated parameters in the cost-effectiveness model, hybrid hip prosthesis remained the dominant alternative compared to uncemented prosthesis, with lower mean costs, and positive incremental QALYs (0.16 for OLS, 0.11 for PS matching, and 0.19 for BCM with machine learning). While research paper 4 did not consider methods in a bivariate setting, recommendations apply to the general context when patient-level observational data is used to estimate input parameters for CEA. The checklist (research paper 1) also pertains to this more general use of observational data, for example when only one incremental parameter is estimated using patient-level observational data.

Further statistical challenges in CEA that use patient-level observational data

The focus of this thesis is to investigate statistical methods that can address selection bias in CEA. The use of patient-level data in CEA often poses further statistical challenges. Statistical analysis may also need to recognise the data hierarchy in multicenter trials (Grieve et al., 2007, Manca et al., 2007) and cluster-randomised trials (Gomes et al., 2012, Grieve et al., 2010), as well as missing data (Noble et al., 2012), non-compliance to randomised treatment (Hughes et al., 2001), censoring (Willan et al., 2002, Willan et al., 2005, Raikou and McGuire, 2004) or measurement error (Marschner, 2006). These issues are beyond the scope of the thesis.

Some of the methods considered in this thesis, for example IPTW (Willan et al., 2002) and DR methods (Bang and Robins, 2005, Bang and Tsiatis, 2000) can be applied to account for censoring and missing data in CEA. However, it is unknown whether the findings from this thesis in the context of addressing selection bias would translate directly to the context of censored or missing CEA data, and hence further research is warranted.

7.5.2 Range of statistical methods considered for this thesis

Methods assume no unobserved confounding

The statistical methods that were contrasted in the case studies and simulation studies all assumed no unobserved confounding. The conceptual review (chapter 2) highlighted the importance of this assumption, and proposed instrumental variable (IV) estimation as an alternative. IV methods can potentially reduce selection bias due to both observed and unobserved confounding, however, they make alternative untestable assumptions that may be unrealistic in a CEA setting (Polsky and Basu, 2006). The critical appraisal tool provides some guidance in assessing these assumptions (research paper 1). After careful assessment I found that neither of the case studies considered in the thesis had an appropriate IV. The systematic review of applied CEA (research paper 1) identified only two studies which used IV methods.

The critical appraisal tool presented in research paper 1 provides detailed guidance on how to appropriately assess the plausibility of the unconfoundedness assumption in the CEA context. Following these suggestions, the case studies carefully considered the previous clinical literature on prognostic factors, before selecting the potential confounders for adjustment. A further recommendation, not covered in this thesis, is to use placebo tests (Imbens, 2004, Jones, 2007, Abadie et al., 2010). Placebo tests offer

an indirect way to use the data to assess the validity of the no unobserved confounding assumption through using the set of measured confounders to estimate a treatment effect on a variable, where it is known to be zero, for example on the pre-treatment health status.

Alternative implementations of statistical methods not considered in the thesis

Regression methods

This thesis considered regression methods recommended for estimating parameters of cost and effectiveness endpoints, such as GLMs (Barber and Thompson, 2004) and two-part models (Buntin and Zaslavsky, 2004, Basu and Manca, 2011). More flexible regression approaches have recently been proposed for skewed cost and effectiveness data, such as extended estimating equations (Basu and Rathouz, 2005), the use of beta-distributions with quasi-likelihood estimation (Basu and Manca, 2011), or beta-type size distributions (Jones et al., 2011). This thesis did not consider these methods, but took the approach of machine learning estimation for handling skewed HRQoL data. The super learner approach considered in research paper 4 is a flexible prediction method which can also incorporate the above regression approaches (Polley and van der Laan, 2010).

DR methods

This thesis did not compare all currently available implementations of DR methods. I implemented methods that are commonly used in the general causal inference literature, such as AIPTW (research paper 3) and weighted regression (research papers 3 and 4); as well as a recently proposed DR method, TMLE. Further approaches such as an improved DR substitution estimator (Rotnitzky et al., 2012), or an extension of TMLE,

collaborative maximum likelihood estimation (C-TMLE) (van der Laan and Gruber, 2010) are promising alternatives, and warrant further consideration.

Machine learning estimation techniques for estimating the PS and the endpoint

This thesis highlights the potential for machine learning estimation techniques to reduce bias due to functional form misspecification, compared to using fixed parametric models. Following recommendations from previous simulation studies, this thesis considered boosted CARTs for estimating the PS, and super learning to estimate the endpoint regression function. This thesis did not compare alternative machine learning approaches from the computer science and data mining literature, such as bagged regression trees, random forests (Austin, 2012), decision trees, neural networks or linear classifiers (Westreich et al., 2010).

Considering GM with bias-adjustment

Like previous studies (Diamond and Sekhon, 2012, Sekhon and Grieve, 2011), this thesis found that GM can provide excellent balance and unbiased estimates of treatment effects, even if the PS is misspecified (research paper 2). Research papers 3 and 4 take a further approach for reducing bias in matching estimators, by using regression-adjustment after matching. In order to allow for a systematic comparison across methods, in these papers I used the estimated PS to create matched data. The methodological literature suggests bias correction for a general family of nearest neighbour matching estimators, including PS matching and multivariate matching (Abadie and Imbens, 2011). Hence the bias-reduction reported when using regression-adjustment after matching (research papers 3 and 4) is expected to hold in the case of the multivariate matching approach of GM as well.

7.5.4 Range of circumstances considered for the methodological comparisons

Types of misspecification considered in the simulations

The simulation scenarios considered in research papers 2, 3 and 4 focused on functional form misspecification of the PS and endpoint regression models, following previous simulation studies identified in the conceptual review (chapter 2), and motivated by the case studies. Types of misspecifications included ignoring differential treatment assignment by subgroup, misspecifying the linear predictor in the PS model, ignoring nonlinear functional form relationships between the covariates and the endpoints, misspecifying the link function of the cost endpoint, as well as misspecifying a two-part data generating distribution.

A further type of misspecification often considered in the general causal inference literature (e.g. Glynn and Quinn, 2010, Gruber and van der Laan, 2010) is omission of confounders, i.e. in simulation studies ignoring variables that are known to influence the treatment assignment and the endpoint. This misspecification was not the focus of this thesis. All the methods considered rely on the assumption of no unobserved confounding, and so it can be anticipated that all methods are biased when influential confounders are omitted. This was confirmed in the simulation study of research paper 2, which demonstrated that unless confounders influential for both the cost and effectiveness endpoints are adjusted for, each method reported biased estimates of the INB.

Types of heterogeneity in cost-effectiveness considered

Research paper 2 considered heterogeneous treatment effects and heterogeneous assignment to treatment, across subgroups defined by an observed confounder, baseline

disease severity. Here, patient subgroups of interest for CEA were pre-specified using reasoning from the clinical literature. The optimal number and definition of subgroups could be also established as part of the CEA, using for example expected health benefits (Espinoza et al., 2011). Heterogeneity in cost-effectiveness can also stem from other sources (Sculpher, 2008), including unobserved patient characteristics such as preferences for treatment. For decision makers implementing personalised medicine, accounting for such heterogeneity can be relevant (Basu, 2011, Ioannidis and Garber, 2011). The conceptual review identified statistical methods that have potential to handle unobserved heterogeneity, such as instrumental variables (Basu et al., 2007, Evans and Basu, 2011) and control functions (Basu, 2011), this thesis however did not cover these approaches.

A further form of heterogeneous treatment effects comes from non-linear response surfaces for cost and effectiveness endpoints (Basu et al., 2011, Basu et al., 2008). While applying traditional regression approaches such as OLS regression might mask this heterogeneity, the method of recycled predictions, considered in this thesis, can account for it.

7.6 Areas of further research

This thesis identified the following areas for further investigation: applying formal methods of sensitivity analysis to assess the impact of potential violations of statistical assumptions, further examination of methods for estimating the variance of treatment effects in a bivariate context of CEA, and extending the methods to estimate parameters other than the additive treatment effects, such as odds ratios or hazard ratios.

7.6.1 Using formal tools of sensitivity analysis to address structural uncertainty

Structural uncertainty is a relatively under-researched area of uncertainty in CEA (Gray et al., 2010), hence contributing to developing methodological guidance in this area is warranted. Specifically, this thesis highlighted that uncertainty due to the possible violations of underlying assumptions of statistical methods can be characterised as a source of structural uncertainty in CEA (Jackson et al., 2011). The case studies presented in this thesis (research papers 2 to 4) acknowledged structural uncertainty due to the choice or specification of statistical method by presenting a range of estimates obtained with different statistical approaches, and interpreting the differences in the estimated cost-effectiveness.

The conceptual review (chapter 2) also identified quantitative approaches that can help acknowledge the uncertainty, due to possible violations of statistical assumptions. A recommended approach for assessing the sensitivity of estimated treatment effects to the potential for unobserved confounding is to use Rosenbaum's method of sensitivity analysis (Rosenbaum, 2002). This method provides a statement on the strength of unobserved confounding, which is necessary to change the conclusions regarding the estimated treatment effect. In the CEA context, this approach could be combined with reporting cost-effectiveness acceptability curves, to provide information on the necessary strength of unobserved confounding to alter the estimated probability that the intervention is cost-effective. Software implementation of this approach for matching estimators is available (Keele, 2011), and its use has been demonstrated for clinical evaluations (Noah et al., 2011).

A further source of structural uncertainty stems from the unknown nature of the correct endpoint regression model. Here, uncertainty in the choice of regression model

specification can be quantified by using Bayesian model averaging (Hoeting et al., 1999). This approach combines estimates from competing regression models, using weights derived from some measure of model appropriateness, for example the Akaike information criterion. Bayesian model averaging has been proposed for the more general context of decision models in CEA (Jackson et al., 2009).

7.6.2 Estimating uncertainty for cost-effectiveness parameters

In the case studies presented in this thesis, I used the non-parametric bootstrap (Davison and Hinkley, 1997) for estimating the standard errors for the estimated INB (Nixon et al., 2010). Previous studies indicated that the nonparametric bootstrap can report valid confidence intervals around the INB (Nixon et al., 2010). Bivariate regression models such as "seemingly unrelated regressions" or Bayesian bivariate models (Willan et al., 2004, Nixon and Thompson, 2005, Manca and Austin, 2008) provide alternative ways of estimating standard errors. This thesis did not consider bivariate approaches, because implementing DR methods or regression on matched data using bivariate modelling may prove complex (Manca and Austin, 2008). Future simulation studies can provide additional information on the performance of the bootstrapped variance estimator, by also reporting the coverage properties of the 95% CIs, and where feasible, comparing it to bivariate modelling approaches.

The estimation of variance for matching approaches has been widely debated in the methodological literature (Hill, 2008, Hill and Reiter, 2005, Abadie and Imbens, 2006b, Austin, 2008a, Stuart, 2010). Inference after PS matching needs to account for several sources of uncertainty: from using the estimated PS instead of the true PS, as well as uncertainty from the matching process. There is a consensus that under relatively general circumstances, using the estimated PS instead of the true PS provides

conservative variance estimates (Stuart, 2010). Analytical variance formulas which can account for the matching process for certain matching estimators are available, however they are subject to ongoing research (Abadie and Imbens, 2009, Abadie and Imbens, 2006a).

In research papers 2 and 3 I followed the suggestion of estimating bootstrapped standard errors conditional on the matched data (Ho et al., 2007). In research paper 4, when considering a univariate endpoint, I applied recommended analytical formulas for variance estimators for PS matching and BCM (Abadie and Imbens, 2011, Abadie and Imbens, 2006a). The extension of these analytical formulas for a bivariate context of CEA is a potential subject of further methodological investigation.

7.6.3 Extending statistical methods for different types of data

This thesis focused on statistical methods that can address selection bias in CEA that use IPD to estimate incremental parameters of continuous endpoints, such as incremental costs, QALYs or HRQoL. Each method proposed by this thesis can be extended for endpoints such as binary, count or event time data, and corresponding estimands such as odds ratios (Radice et al., 2012, Moore and van der Laan, 2009), relative risks (Austin, 2008b, Austin, 2010a, Austin, 2010b) or hazard ratios (Stitelman and van der Laan, 2010, Thompson et al., 2010).

This thesis compared methods that can estimate the effect of a time constant treatment. IPTW and DR methods can be extended to handle treatment and confounders that vary over time (Robins et al., 2000). Such methods can be useful when estimating parameters for decision models which needs to allow for cross-over between treatments, or treatment starting at different time points (Caro et al., 2012). Exploring these alternative

methods in settings characteristic of CEA is a subject of further methodological research.

7.7 Recommendations for applied researchers

Findings from this thesis can help the applied researcher conducting CEA, when applying statistical methods to address selection bias. This thesis recommends that the applied researcher follows the general steps below.

1. To assess the plausibility of the fundamental assumptions of unconfoundendess and overlap.

The checklist and accompanying guidance presented in research paper 1 suggest appropriate methods to assess the plausibility of these assumptions (checklist questions 1a and 2). For example it is recommended that researchers carefully use subject matter knowledge to assess whether all potential confounders have been observed, both for the cost and the effectiveness endpoints.

2. To use statistical methods which are potentially robust to the misspecification of the endpoint regression models and the PS.

This thesis identified a number of methods that can be appropriate for estimating parameters in realistic CEA circumstances. In general, this thesis found that matching methods can help provide robust estimates of cost-effectiveness, as they are relatively insensitive to PS misspecification, as opposed to other methods such as IPTW.

GM does not rely on the correctly specified PS, and can directly maximise balance in the matched data using machine learning. When cost-effectiveness for patient subgroups are of interest, this thesis suggests that covariate balance is assessed for each subgroup of policy relevance. This thesis recommends that GM is applied to directly

maximise balance for patient subgroups. Research papers 1 and 2 provide guidance on appropriate assessment of balance, for overall populations and for subgroups of interest. To assist the applied researcher implement GM, sample code is provided in Appendix 4.2.

This thesis recommends that matching is followed by regression-adjustment, in order to reduce bias due to finite sample imbalance, and to increase efficiency. Research paper 3 proposes a straightforward two-step approach for performing regression-adjustment on the matched data (for software code, see Appendix 5.2). This approach reduces the sensitivity of the estimates to the regression specification, by first improving covariate balance.

As an alternative remedy for the challenge of model misspecification, research paper 4 proposes machine learning techniques for estimating the PS and the endpoint regression. In particular, this paper proposes another implementation of regression-adjusted matching, BCM, which can be used with machine-learning. Amongst the DR methods examined in this thesis, TMLE is recommended when coupled with machine learning estimation techniques. I provide the applied researcher with guidance for implementing these methods in Appendix 6.1.

One consideration for the choice of methods is the computing time and resources involved. TMLE and BCM run instantly when using fixed parametric approaches, however when coupled with machine learning, each approach can take more than 3 hours with a standard PC. Therefore the use of high performance computing (HPC) is recommended, for example the LSHTM HPC cluster (http://wiki.lshtm.ac.uk/hpc/index.php5/Main_Page). GM exploits the parallel computing abilities of an HPC cluster, however depending on the dimensions of the dataset and the number of variables to balance on, can be computationally intensive. So

for example for the analysis of the case study in research paper 2, running GM on the HPC cluster took between 3 and 10 hours.

3. To report structural uncertainty according to the choice or specification of the statistical method.

This thesis also recommends that researchers acknowledge structural uncertainty from the choice or specification of the statistical approach used for addressing selection bias. The quality assessment tool provides suggestions on ways to account for this structural uncertainty (research paper 1, checklist question 5). For example, following the example of the case studies presented in this thesis, the applied researcher is advised to implement several statistical approaches which rely on different assumptions, and then present cost-effectiveness results after each approach, interpreting potential differences in the resulting point estimates, confidence intervals and CEACs.

7.8 Implications for policy making

Observational data can provide a valuable source of evidence for health care decision makers who aim to allocate scarce resources. Current methodological guidelines propagate the use of patient-level data for deriving parameters for CEA (NICE, 2008, Briggs et al., 2012). While developing methods for incorporating observational data was raised as a priority for methodological research in CEA (Longworth et al., 2009, Kearns et al., 2012), there is currently no detailed guidance for critical appraisal of CEA that use observational data. The checklist developed in this thesis (research paper 1) provides decision makers with a critical appraisal tool for evaluating an important aspect of the quality of CEA that use observational data: the statistical methods that are used to address selection bias.

While this thesis focused on statistical methods, it also provides some insights for the design of CEA that use observational data. With large investments in observational databases worldwide, it is desirable that observational studies are designed so as to help subsequent statistical analysis make more plausible assumptions (Rubin, 2010, Rubin, 2008). For the purposes of CEA, observational data collected on health care interventions should ideally include the potential confounders that are judged to be prognostic of either the effectiveness or the cost endpoint. If it is unlikely that all the confounders can be observed, the researcher is recommended to consider whether there are plausible IVs that could be measured (Grootendorst, 2007).

7.9 Conclusion

This thesis aimed to address the relative lack of methodological guidance on statistical methods for CEA that use patient-level observational data. The critical appraisal of applied CEA highlighted that studies using observational data did not appropriately assess the underlying assumptions their statistical methods make. The conceptual review drew on insights from the causal inference literature, and identified promising further methods for CEA.

This thesis found that methods that can avoid assuming that the endpoint regression and the PS are correctly specified, can give less biased and more precise estimates of cost-effectiveness than methods previously recommended for CEA. In particular, combining matching methods with regression is a robust, appropriate and accessible method that should be adopted in future studies. This thesis presents methods that can improve the quality of CEA that use patient-level observational data, to help future studies provide a sounder basis for policy making.

References

Abadie, A., Diamond, A. & Hainmueller, J. 2010. Synthetic Control Methods for Comparative Case Studies:Estimating the Effect of California's Tobacco Control Program. *Journal of the American Statistical Association***,** 493-505.

Abadie, A. & Imbens, G. W. 2006a. Large Sample Properties of Matching Estimators for Average Treatment Effects. *Econometrica,* 74**,** 235-267.

Abadie, A. & Imbens, G. W. 2006b. On the Failure of the Bootstrap for Matching Estimators. *National Bureau of Economic Research Technical Working Paper Series,* No. 325.

Abadie, A. & Imbens, G. W. 2009. Matching on the Estimated Propensity Score. National Bureau of Economic Research, Inc.

Abadie, A. & Imbens, G. W. 2011. Bias-Corrected Matching Estimators for Average Treatment Effects. *Journal of Business & Economic Statistics,* 29**,** 1-11.

Austin, P. C. 2008a. A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003. *Statistics in Medicine,* 27**,** 2037-2049.

Austin, P. C. 2008b. Inverse probability weighted estimation of the marginal odds ratio. *Statistics in Medicine,* 27**,** 5560–5563.

Austin, P. C. 2010a. Absolute risk reductions, relative risks, relative risk reductions, and numbers needed to treat can be obtained from a logistic regression model. *Journal of Clinical Epidemiology***,** 2-6.

Austin, P. C. 2010b. The performance of different propensity-score methods for estimating differences in proportions (risk differences or absolute risk reductions) in observational studies. *Stat Med,* 29**,** 2137-48.

Austin, P. C. 2012. Using Ensemble-Based Methods for Directly Estimating Causal Effects: An Investigation of Tree-Based G-Computation. *Multivariate Behavioral Research***,** 115-135.

Bang, H. & Robins, J. M. 2005. Doubly Robust Estimation in Missing Data and Causal Inference Models. *Biometrics* 962–972.

Bang, H. & Tsiatis, A. A. 2000. Estimating medical costs with censored data. *Biometrica* 87**,** 329-343.

Barber, J. & Thompson, S. 2004. Multiple regression of cost data: use of generalised linear models. *J Health Serv Res Policy,* 9**,** 197-204.

Basu, A. 2011. Economics of individualization in comparative effectiveness research and a basis for a patient-centered health care. *Journal of Health Economics,* 30**,** 549-559.

Basu, A., Heckman, J. J., Navarro-Lozano, S. & 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**,** 1133-1157.

Basu, A. & Manca, A. 2011. Regression Estimators for Generic Health-Related Quality of Life and Quality-Adjusted Life Years. *Med Decis Making,* 2011 Oct 18. [Epub ahead of print].

Basu, A., Polsky, D. & Manning, W. 2011. Estimating treatment effects on healthcare costs under exogeneity: is there a 'magic bullet'? *Health Services and Outcomes Research Methodology,* 11**,** 1-26.

Basu, A., Polsky, D. & Manning, W. G. 2008. Use of propensity scores in non-linear response models: The case for health care expenditures. HEDG, c/o Department of Economics, University of York.

Basu, A. & Rathouz, P. J. 2005. Estimating marginal and incremental effects on health outcomes using flexible link and variance function models. *Biostatistics,* 6**,** 93-109.

Briggs, A. H., Weinstein, M. C., Fenwick, E. A. L., Karnon, J., Sculpher, M. J. & Paltiel, A. D. 2012. Model Parameter Estimation and Uncertainty Analysis. *Medical Decision Making,* 32**,** 722-732.

Buntin, M. B. & Zaslavsky, A. M. 2004. Too much ado about two-part models and transformation?: Comparing methods of modeling Medicare expenditures. *Journal of Health Economics,* 23**,** 525-542.

Busso, M., DiNardo, J. & McCrary, J. 2009. Finite Sample Properties of Semiparametric Estimators of Average Treatment Effects.
Busso, M., DiNardo, J. & McCrary, J. 2011. New Evidence on the Finite Sample Properties of Propensity Score Reweighting and Matching Estimators. *Working paper.*
Caro, J. J., Briggs, A. H., Siebert, U. & Kuntz, K. M. 2012. Modeling Good Research Practices—Overview. *Medical Decision Making,* 32**,** 667-677.
Cooper, N. J., Sutton, A. J., Ades, A. E., Paisley, S. & Jones, D. R. 2007. Use of evidence in economic decision models: practical issues and methodological challenges. *Health Economics,* 16**,** 1277-1286.
Davison, A. & Hinkley, D. 1997. *Bootstrap Methods and Their Application,* New York, Cambridge University Press.
Diamond, A. & Sekhon, J. S. 2012. Genetic Matching for Estimating Causal Effects: A General Multivariate Matching Method for Achieving Balance in Observational Studies. *Review of Economic and Statistics,* Forthcoming.
Drummond, M., Sculpher, M., Torrance, G., O'Brien, B. & Stoddart, G. 2005. *Methods for the Economic Evaluation of Health Care Programmes,* Oxford, Oxford University Press.
Drummond, M. F. 1998. Experimental versus observational data in the economic evaluation of pharmaceuticals. *Medical Decision Making,* 18.
Espinoza, M. A., Manca, A., Claxton, K. & Sculpher, M. J. 2011. The value of identifying heterogeneity: a framework for subgroup cost-effectiveness analysis. *HESG Conference presentation.* York.
Evans, H. & Basu, A. 2011. Exploring comparative effect heterogeneity with instrumental variables: prehospital intubation and mortality. HEDG, c/o Department of Economics, University of York.
Freedman, D. & Berk, R. A. 2008. Weighting regression by propensity score. *Evaluation Review,* 32**,** 392-409.
Glick, H., Doshi, J., Sonnad, S. & Polsky, D. 2007. *Economic evaluation in clinical trials,* Oxford, Oxford University Press.
Glynn, A. N. & Quinn, K. M. 2010. An Introduction to the Augmented Inverse Propensity Weighted Estimator. *Political Analysis,* 18**,** 36-56.
Gomes, M., Grieve, R., Nixon, R., Ng, E. S. W., Carpenter, J. & Thompson, S. G. 2012. Methods for covariate adjustment in cost-effectiveness analysis that use cluster randomised trials *Health Economics***,** n/a-n/a.
Gray, A. M., Clarke, P. M., Wolstenholme, J. L. & Wordsworth, S. 2010. *Applied Methods of Cost-Effectiveness Analysis in Healthcare*, Oxford University Press.
Grieve, R., Nixon, R. & Thompson, S. G. 2010. Bayesian Hierarchical Models for Cost-Effectiveness Analyses that Use Data from Cluster Randomized Trials. *Medical Decision Making,* 30**,** 163-175.
Grieve, R., SG, T., Nixon, R. M. & Cairns, J. 2007. Multilevel models for estimating incremental net benefits in multinational studies. *Health Economics,* 16**,** 815–26.
Grootendorst, P. 2007. A review of instrumental variables estimation in the applied health sciences. *Health Services and Outcomes Research Methodology* 7**,** 159-179.
Gruber, S. & van der Laan, M. 2010. A targeted maximum likelihood estimator of a causal effect on a bounded continuous outcome. *Int J Biostat. ,* 6.
Hill, J. 2008. Discussion of research using propensity-score matching: Comments on 'A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003' by Peter Austin. *Statistics in Medicine,* 27 2055–2061.
Hill, J. & Reiter, J. P. 2005. Interval estimation for treatment effects using propensity score matching. *Statistics in Medicine,* 25.
Ho, D. E., Imai, K., King, G. & Stuart, E. A. 2007. Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference *Political Analysis,* 15**,** 199-236.
Hoch, J. S., Briggs, A. H. & Willan, A. R. 2002. Something old, something new, something borrowed, something blue: a framework for the marriage of health econometrics and cost-effectiveness analysis. *Health Economics,* 11**,** 415-430.

Hoeting, J. A., Madigan, D., Raftery, A. E. & Volinsky, C. T. 1999. Bayesian model averaging: a tutorial (with discussion). . *Statist. Sci.*, 14**,** 382–401.
Hughes, D. A., Bagust, A., Haycox, A. & Walley, T. 2001. The impact of non-compliance on the cost-effectiveness of pharmaceuticals: a review of the literature. *Health Economics,* 10**,** 601-615.
Imbens, G. M. & Wooldridge, J. M. 2009a. Recent Developments in the Econometrics of Program Evaluation. *Journal of Economic Literature,* 47**,** 5–86.
Imbens, G. W. 2004. Nonparametric Estimation of Average Treatment Effects Under Exogeneity: A Review. *The Review of Economics and Statistics,* 86**,** 4-29.
Ioannidis, J. P. A. & Garber, A. M. 2011. Individualized Cost-Effectiveness Analysis. *PLoS Med,* 8**,** e1001058.
Jackson, C., Bojke, L., Thompson, S., Claxton, K. & Sharples, L. 2011. A Framework for Addressing Structural Uncertainty in Decision Models. *Medical Decision Making,* 31**,** 662–674.
Jackson, C., Thompson, S. & LD, S. 2009. Accounting for uncertainty in health economic decision models by using model averaging. *Journal of the Royal Statistical Society, Series A,* 383--404.
Jones, A., Lomas, J. & Rice, N. 2011. Applying Beta-type Size Distributions to Healthcare Cost Regressions. *HEDG Working Papers.* HEDG, c/o Department of Economics, University of York.
Jones, A. M. 2007. Identification of treatment effects in Health Economics. *Health Economics,* 16**,** 1127-1131.
Jones, A. M. 2010. Models For Health Care. *HEDG Working Papers.* HEDG, c/o Department of Economics, University of York.
Jones, A. M. & Rice, N. 2011. Econometric Evaluation of Health Policies. *In:* GLIED, S. & SMITH, P. (eds.) *The Oxford handbook of health economics.* Oxford: Oxford University Press.
Kang, J. D. Y. & Schafer, J. L. 2007. Demystifying Double Robustness: A Comparison of Alternative Strategies for Estimating a Population Mean from Incomplete Data. *Statistical Science,* 22**,** 523-539.
Kearns, B., Ara, R. & Wailoo, A. 2012. *A review of the use of statistical regression models to inform cost-effectiveness analyses withing the NICE Technology Appraisals Programme* [Online]. ScHARR, University of Sheffield. Available: http://www.nicedsu.org.uk/FINAL%20DSU%20Regressions%20report_09.10.12.pdf.
Keele, L. J. 2011. *Rbounds: An R Package For Sensitivity Analysis with Matched Data.* [Online]. Available: http://cran.r-project.org/web/packages/rbounds/index.html.
Lee, B. K., Lessler, J. & Stuart, E. A. 2010. Improving propensity score weighting using machine learning. *Statistics in Medicine,* 29**,** 337-346.
Longworth, L., Bojke, L., Tosh, J. & Sculpher, M. 2009. *MRC-NICE Scoping Project: Identifying the National Institute For Health And Clinical Excellence's Methodological Research Priorities and an Initial Set of Priorities* [Online]. University of York. [Accessed CHE Research Paper 51.
Manca, A. & Austin, P. C. 2008. *Using propensity score methods to analyse individual patient-level cost-effectiveness data from observational studies* [Online]. Available: http://www.york.ac.uk/res/herc/documents/wp/08_20.pdf.
Manca, A., Lambert, P., Sculpher, N. & Rice, N. 2007. Cost-effectiveness analysis using data from multinational trials: the use of Bayesian hierarchical modelling. *Medical Decision Making***,** 471–90.
Marschner, I. C. 2006. Measurement error bias in pharmaceutical cost-effectiveness analysis. *Applied Stochastic Models in Business and Industry,* 22**,** 621-630.
Moore, K. L. & van der Laan, M. J. 2009. Covariate adjustment in randomized trials with binary outcomes: Targeted maximum likelihood estimation. *Statistics in Medicine,* 28**,** 39-64.
NICE. 2008. *Guide to the Methods of Technology Appraisal* [Online]. Available: http://www.nice.org.uk/media/B52/A7/TAMethodsGuideUpdatedJune2008.pdf [Accessed 24/10/2010.

Nixon, R., Wonderling, D. & Grieve, R. 2010. Non-parametric methods for cost-effectiveness analysis: the central limit theorem and the bootstrap compared. *Health Economics,* 19**,** 316-33.

Nixon, R. M. & Thompson, S. G. 2005. Methods for incorporating covariate adjustment, subgroup analysis and between-centre differences into cost-effectiveness evaluations. *Health Economics,* 14**,** 1217-1229.

Noah, M. A., Peek, G. J., Finney, S. J., Griffiths, M. J., Harrison, D. A., Grieve, R., Sadique, M. Z., Sekhon, J. S., McAuley, D. F., Firmin, R. K., Harvey, C., Cordingley, J. J., Price, S., Vuylsteke, A., Jenkins, D. P., Noble, D. W., Bloomfield, R., Walsh, T. S., Perkins, G. D., Menon, D., Taylor, B. L. & Rowan, K. M. 2011. Referral to an Extracorporeal Membrane Oxygenation Center and Mortality Among Patients With Severe 2009 Influenza A(H1N1). *JAMA: The Journal of the American Medical Association,* 306**,** 1659-1668.

Noble, S., Hollingworth, W. & Tilling, K. 2012. Missing data in trial-based cost-effectiveness analysis: the current state of play. *Health Economics,* 21**,** 187-200.

O'Hagan, A. & Stevens, J. 2001. A framework for cost-effectiveness analysis from clinical trial data. *Health Economics,* 10**,** 303-15.

Pennington, M., Grieve, R., Sekhon, J. S., Gregg, P., Black, N. & van der Meulen, J. 2012. Cemented, cementless and hybrid prostheses for total hip replacement: a cost-effectiveness analysis. *British Medical Journal,* submitted.

Philips, Z., Bojke, L., Sculpher, M., Claxton, K. & Golder, S. 2006. Good practice guidelines for decision-analytic modelling in health technology assessment: a review and consolidation of quality assessment. *Pharmacoeconomics,* 24**,** 355-71.

Polley, E. C. & van der Laan, M. J. 2010. Super Learner In Prediction. *U.C. Berkeley Division of Biostatistics Working Paper Series.*

Polsky, D. & Basu, A. 2006. Selection Bias in Observational Data. *The Elgar Companion to Health Economics.* Edward Elgar Publishing.

Porter, K. E., Gruber, S., Laan, M. J. v. d. & Sekhon, J. S. 2011. The Relative Performance of Targeted Maximum Likelihood Estimators. *The International Journal of Biostatistics,* 7.

Radice, R., Grieve, R., Ramsahai, R., Kreif, N., Sadique, Z. & 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).

Raikou, M. & McGuire, A. 2004. Estimating medical care costs under conditions of censoring. *J Health Econ. ,* 23**,** 443-70.

Robins, J., Hernán, M. A. & Brumback, B. 2000. Marginal structural models and causal inference in epidemiology. *Epidemiology***,** 550-60.

Rosenbaum , P. R. & Rubin, D. B. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika,* 70**,** 41-55.

Rotnitzky, A., Lei, Q., Sued, M. & Robins, J. M. 2012. Improved double-robust estimation in missing data and causal inference models. *Biometrika,* 99**,** 439-456.

Rubin, D. 2008. For Objective Causal Inference, Design Trumps Analysis. *The Annals of Applied Statistics,* 2**,** 808-840.

Rubin, D. B. 2010. On the limitations of comparative effectiveness research. *Statistics in Medicine,* 29**,** 1991-1995.

Sculpher, M. 2008. Subgroups and Heterogeneity in Cost-Effectiveness Analysis. *Pharmacoeconomics,* 26**,** 799-806.

Sekhon, J. S. & Grieve, R. D. 2011. A Matching Method for Improving Covariate Balance in Cost-Effectiveness Analyses. *Health Economics,* doi: 10.1002/hec.1748.

Stitelman, O. M. & van der Laan, M. J. 2010. Collaborative targeted maximum likelihood for time to event data. *Int J Biostat,* 6.

Stuart, E. A. 2010. Matching Methods for Causal Inference: A review and a look forward. *Statistical Science,* 25.

Thompson, S., Kaptoge, S., White, I., Wood, A., Perry, P., Danesh, J. & Collaboration, T. E. R. F. 2010. Statistical methods for the time-to-event analysis of individual participant data from multiple epidemiological studies. *International Journal of Epidemiology*.

van der Laan, M. J. & Gruber, S. 2010. Collaborative Double Robust Targeted Maximum Likelihood Estimation. *The International Journal of Biostatistics* 6.

Waernbaum, I. 2011. Model misspecification and robustness in causal inference: comparing matching with doubly robust estimation. *Stat Med,* DOI: 10.1002/sim.4496.

Westreich, D., Lessler, J. & Funk, M. 2010. Propensity score estimation: neural networks, support vector machines, decision trees (CART), and meta-classifiers as alternatives to logistic regression. *Journal of Clinical Epidemiology,* 63**,** 826-33.

Willan, A. R., Briggs, A. H. & Hoch, J. S. 2004. Regression methods for covariate adjustment and subgroup analysis for non-censored cost-effectiveness data. *Health Economics,* 13**,** 461-475.

Willan, A. R., Lin, D. Y., Cook, R. J. & Chen, E. B. 2002. Using inverse-weighting in cost-effectiveness analysis with censored data. *Statistical Methods in Medical Research,* 11**,** 539-551.

Willan, A. R., Lin, D. Y. & Manca, A. 2005. Regression methods for cost-effectiveness analysis with censored data. *Statistics in Medicine,* 24**,** 131-145.

Willan, A. R., Briggs, A. H. & Hoch, J. S. 2004. Regression methods for covariate adjustment and subgroup analysis for non-censored cost-effectiveness data. *Health Economics,* 13**,** 461-475.

Willan, A. R., Lin, D. Y., Cook, R. J. & Chen, E. B. 2002. Using inverse-weighting in cost-effectiveness analysis with censored data. *Statistical Methods in Medical Research,* 11**,** 539-551.

Willan, A. R., Lin, D. Y. & Manca, A. 2005. Regression methods for cost-effectiveness analysis with censored data. *Statistics in Medicine,* 24**,** 131-145.