
# Statistical methods for cost-effectiveness analysis that use cluster randomised trials

Manuel Gomes

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

London School of Hygiene and Tropical Medicine

February 2012

I, Manuel António de Oliveira Gomes, 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: 13/02/2012

# Abstract

This thesis considers alternative statistical methods for cost-effectiveness analysis (CEA) that use cluster randomised trials (CRTs). The thesis has four objectives: firstly to develop criteria for identifying appropriate methods for CEA that use CRTs; secondly to critically appraise the methods used in applied CEAs that use CRTs; thirdly to assess the performance of alternative methods for CEA that use CRTs in settings where baseline covariates are balanced; fourthly to compare statistical methods that adjust for systematic covariate imbalance in CEA that use CRTs.

The thesis developed a checklist to assess the methodological quality of published CEAs that use CRTs. This checklist was informed by a conceptual review of statistical methods, and applied in a systematic literature review of published CEAs that use CRTs. The review found that most studies adopted statistical methods that ignored clustering or correlation between costs and health outcomes.

A simulation study was conducted to assess the performance of alternative methods for CEA that use CRTs across different circumstances where baseline covariates are balanced. This study considered: seemingly unrelated regression (SUR) and generalised estimating equations (GEEs), both with a robust standard error; multilevel models (MLMs) and a non-parametric 'two-stage' bootstrap (TSB). Performance was reported as, for example, bias and confidence interval (CI) coverage of the incremental net benefit. The MLMs and the TSB performed well across all settings; SUR and GEEs reported poor CI coverage in CRTs with few clusters.

The thesis compared methods for CEA that use CRTs when there are systematic differences in baseline covariates between the treatment groups. In a case study and further simulations, the thesis considered SUR, MLMs, and TSB combined with SUR to adjust for covariate imbalance. The case-study showed that cost-effectiveness results can differ according to adjustment method. The simulations reported that MLMs performed well across all settings, and unlike the other methods, provided CI coverage close to nominal levels, even with few clusters and unequal cluster sizes.

The thesis concludes that MLMs are the most appropriate method across the circumstances considered. This thesis presents methods for improving the quality of CEA that use CRTs, to help future studies provide a sound basis for policy making.

# Contents

# Acknowledgements

First and foremost I want to thank my supervisor Richard Grieve. It has been an honour to be his first doctoral student. Richard has provided me with immense intellectual and practical support. His vast experience and excellent guidance were vital throughout this journey. I am also thankful for the excellent example he has provided with his successful achievements, both professional and personal.

I am very grateful to my advisors, Richard Nixon, James Carpenter and John Edmunds, who offered me their continuous support and constructive comments at each stage of the thesis. In particular, I would like to thank Richard Nixon for giving me the opportunity to work in the Modelling and Simulation Group at Novartis. I am also very grateful to Edmond Ng for his statistical advice and practical help with the simulations. Sincere thanks go to Simon Thompson, Zaid Chalabi and John Cairns for their enlightening thoughts and advices. I also thank Max Bachmann and Simon Dixon for providing full access to the datasets of the case studies.

Warm thanks go numerous friends and colleagues at LSHTM who contributed to an inspiring environment for research. In particular, I would like to thank Noemi Kreif and Tazio Vanni for their stimulating comments, discussions and brainstorming. Financial support from the Fundação para a Ciência e a Tecnologia is greatly appreciated.

Lastly, special thanks go to my family for their invaluable support. In particular, I am deeply indebted to my parents for their eternal love and for passing on to me exceptional values. I dedicate this work to Carina for her infinite love, admirable encouragement, and unwavering fortitude. Thank you!

# Abbreviations

**AIC** – Akaike information criterion
**ATE** – Average treatment effect
**BVN** – Bivariate Normal
**CADTH** – Canadian Agency for Drugs and Technology in Health
**CEA** – Cost-effectiveness analysis
**CEAC** – Cost-effectiveness acceptability curve
**CI** – Confidence Interval
**CLT** – Central limit theorem
**CRT** – Cluster randomised trial
**CV** – Coefficient of variation
**DGP** – Data generating process
**EVPI** – Expected value of perfect information
**GEE** – Generalised estimating equations
**GLM** – Generalised linear models
**GLS** – Generalised linear squares
**ICC** – Intra-cluster correlation coefficient
**ICER** – Incremental cost-effectiveness ratio
**INB** – Incremental net-benefit
**IPD** – Individual patient data
**IQR** – Interquartile range
**IQWIG** – Institute for Quality and Efficiency in Health Care
**MCMC** – Markov chain Monte Carlo
**ML** – Maximum likelihood
**MLM** – Multilevel model
**MRC** – Medical Research Council
**NB** – Net-benefit
**NICE** – National Institute for Health and Clinical Excellence
**OLS** – Ordinary least squares
**PBCA** – Pharmaceutical Benefits Advisory Committee
**PoNDER** – Psychological interventions for postnatal depression
**QALY** – Quality-adjusted life year
**RCT** – Randomised controlled trial
**RMSE** – Root mean square error
**SE** – Standard error
**SUR** – Seemingly unrelated regression
**TSB** – Two-stage bootstrap
**VOI** – Value of information

# List of tables

# List of Figures

# List of Appendices

# Chapter 1

## Introduction

## 1.1 Economic evaluation of health care

Health economic evaluation aims to assist policy making by assessing the relative value of alternative health care technologies, public health interventions, and ways of organising health services (Gold, 1996; Drummond et al., 2005; Morris et al., 2007). It provides a structured framework which can help address the goal of maximising population health subject to available resources (Weinstein and Stason, 1977). Policy makers worldwide (e.g. CADTH, 2006; NICE, 2008; PBCA, 2008; IQWIG, 2009) now use economic evaluation studies to inform decisions about, for example, which health interventions to fund. Methodological guidelines for economic evaluation of health care programmes are relatively well established and encourage the use of individual patient data (IPD) from randomised controlled trials (RCTs) (Gold, 1996; Willan and Briggs, 2006; Glick et al., 2007). It has become increasingly common for RCTs to collect IPD on resource use alongside health outcomes to help evaluate which interventions offer the best value for money.

The availability of patient-level cost-effectiveness data has led to a greater focus on using appropriate statistical methods in cost-effectiveness analysis (CEA)[1] (Briggs et al., 2002). For example, methodological guidance emphasise the need for studies to use methods that can address the correlation between individual costs and health outcomes (O'Hagan and Stevens, 2001; Hoch et al., 2002; Willan et al., 2004). Another important issue is that costs and outcomes may be collected from different health care settings (e.g. different countries). In these circumstances, methods must recognise that cost and outcome data may be more similar within than across settings due to, for example, patient case-mix and cost variation across countries

[1] The term 'cost-effectiveness analysis' has a narrower application than 'economic evaluation' (Gold, 1996; Drummond et al., 2005), but these terms are used interchangeably throughout this thesis.

(Willke et al., 1998; Grieve et al., 2005). An additional concern in CEA is that cost data are typically skewed, with a small proportion of individuals incurring high costs (Barber and Thompson, 1998; Briggs and Gray, 1998; Thompson and Barber, 2000). The skewed nature of cost data may raise issues for methods which assume that data are Normally distributed. These methodological challenges have encouraged the development of statistical methods for CEA conducted alongside RCTs where individuals are randomly allocated to alternative treatment groups.

To address some of the methodological challenges raised in CEA, statistical methods have been successfully transferred from other areas such as biostatistics and econometrics. For example, methods that respect the correlation between costs and outcomes have adapted bivariate models and seemingly unrelated regressions (SUR) (O'Hagan and Stevens, 2001; Willan et al., 2004; Nixon and Thompson, 2005) originally proposed in medical statistics (Anderson, 1984; Timm, 2002) and econometrics (Wooldridge, 2002; Greene, 2003). More generally, these methods offer additional appeal for CEA in that they allow for covariate adjustment to help increase the precision of the estimates or perform subgroup analyses (Willan et al., 2004; Nixon and Thompson, 2005). Other studies have adapted multilevel models, originally developed for education and health research (Leyland and Goldstein, 2001; Goldstein, 2003), to acknowledge the hierarchical nature of multicentre and multinational cost and cost-effectiveness data (Manca et al., 2005; Nixon and Thompson, 2005; Grieve et al., 2007). To handle skewed costs, methods that avoid distributional assumptions such as non-parametric bootstrapping (Efron and Tibshirani, 1993), or methods that can allow for a range of realistic parametric distributions such as generalised linear models (McCullagh and Nelder, 1989), have been adopted in CEA (Barber and Thompson, 2000; Manning, 2006; Mihaylova et al., 2011).

Research funders such as the UK Medical Research Council and National Institute for Health Research have encouraged methodological developments in CEA to help studies provide sound evidence for policy making. It is anticipated that the methods developed can then 'feed into' methodological guidelines for CEA used by health decision makers such as NICE[2] (NICE, 2008), and lead to improvements in practice. A number of other areas have received relatively little attention in CEA and further methodological development has been advocated. For example, commentators have highlighted the need for additional work in methods for CEA that use cost-effectiveness data that are subject to censoring or missingness (Young, 2005; Noble et al., 2010). Also, the characterisation of structural uncertainty in CEA is receiving increasing attention, and studies are advised to carefully address structural/model uncertainty in addition to parameter uncertainty (Claxton, 2008; Jackson et al., 2009). Another important area where the need for further methodological improvement has been recognised is in CEA that use cost-effectiveness data from cluster randomised trials (Flynn and Peters, 2005a; Willan and Briggs, 2006). This thesis will focus its attention on improving statistical methods in this area.

## 1.2 Cluster randomised trials

Economic evaluations of public health interventions often use cost-effectiveness data from cluster randomised trials (CRTs). In CRTs, the unit of randomisation is the cluster, for example the hospital, not the individual patient. The cluster design is preferred in many situations (Donner and Klar, 2000; Hayes and Moulton, 2009). For instance, the intervention may be delivered at the group-level such as professional training for general practitioners in order to change their

[2] National Institute for Health and Clinical Excellence, UK.

behaviour towards their patients, or a prevention programme may be delivered at the hospital-level. Cluster designs are also common when it is important to avoid contamination between individual patients of different treatment groups. For example, in a CRT to evaluate a smoking cessation intervention, it is important to prevent individuals in the treatment group telling the control group about the preventive strategy, otherwise dilution bias may arise.

While CRTs can provide an appropriate design for many interventions, their analysis can pose specific challenges. A fundamental issue in the analysis of CRTs is that patients are more homogeneous in their outcomes within than between clusters. Hence, the use of standard methods such as OLS regression, that assume individual observations are independent, will be incorrect (Donner, 1998; Murray et al., 1998). Cornfield who first brought to light the analytical implications of cluster randomisation stated (Cornfield, 1978):

"*Randomisation by cluster accompanied with an analysis appropriate to randomisation by individual is an exercise in self deception, however, and should be discouraged*"

In other words, methods that ignore the clustering assume they are using more information than they actually have, and will understate the uncertainty (Donner, 1998; Murray et al., 1998). In circumstances where there is a relationship between the size of the cluster and the endpoints, ignoring clustering may lead to bias (Panageas et al. 2007).

Another key concern with CRTs is that the cluster design may be prone to systematic differences in baseline covariates between treatment groups (Donner and Klar, 2000; Puffer et al., 2005; Carter, 2010).

In CRTs, confounding[3] can arise because many CRTs are unblinded, and hence, the recruiting centre is aware of the treatment assignment and patients' characteristics prior to their inclusion. Therefore, the CRT design can yield systematic imbalances in baseline characteristics; for example, there may be circumstances where older patients are less likely to participate in the treatment than the control group (Hahn et al 2005). Studies that fail to account for potential confounding on these observed factors will provide biased results (Puffer et al., 2003; Hahn et al., 2005; Eldridge et al., 2008).

Despite general improvements in methods for CEA alongside RCTs (Gold, 1996; Willan and Briggs, 2006; Glick et al., 2007), methods for CEA that use data from CRTs have received relatively little attention (Flynn and Peters, 2005a). Commentators have recognised that this area can raise additional challenges for analysts and highlighted that additional methodological development is required (Klar and Donner, 2001; Flynn and Peters, 2005a; Willan, 2006). In CEA that use CRTs, it is crucial that methods address the issues that can arise with the cluster design such as the clustering and covariate imbalance, while acknowledging other important concerns in CEA such as the correlation between costs and outcomes. For example, Grieve and others (2010) suggest that ignoring the clustering can lead to inaccurate estimates of incremental cost-effectiveness and the accompanying uncertainty. While some methods, such as the non-parametric bootstrap (Flynn and Peters, 2005b) and multilevel models (Grieve et al., 2010), have been proposed for CEA that use CRTs, there is a lack of work comparing these alternative approaches. Only one study (Bachmann et al., 2007) considers alternative methods for CEA that

---

[3] Here 'confounding' is defined as when there is an observed or unobserved baseline characteristic, which is correlated with both the treatment and the endpoint. This can lead to a biased estimate of the true treatment effect. In the health economics and econometrics literatures this problem is also referred as 'selection bias' or 'endegoneity'. This thesis considers approaches for handling this aspect of confounding due to observed characteristics (Jones and Rice 2011).

use cluster trials. This paper compares the methods in a single case-study with relatively ideal characteristics such as large number of clusters and equal cluster sizes. The study finds small differences across methods and fails to provide general insights on the relative merits of alternative methods. There have been no previous simulation studies that have compared the performance of alternative statistical methods across typical circumstances faced by CEA that use CRTs. Furthermore, it is unknown whether applied CEAs that use CRTs use appropriate methods and thus whether they can provide sound evidence for policy making.

## 1.3 Aims and objectives

The overall aim of this thesis is to identify appropriate statistical methods for CEA that use CRTs and to assess their relative performance across a wide range of realistic scenarios typically faced by CEA that use CRTs. The specific objectives are:

1. To develop criteria for identifying appropriate statistical methods for CEA that use CRTs.
2. To critically appraise the methods used in applied CEAs that use CRTs.
3. To assess the relative performance of alternative statistical methods for CEA that use CRTs in settings where baseline covariates are balanced.
4. To compare alternative methods to adjust for systematic covariate imbalance in CEA that use CRTs.

## 1.4 Conceptual framework

To address these objectives, the thesis requires a clear conceptual framework, drawing on insights from general biostatistics and medical statistics. An important conceptual point, well established in the literature, is that the form of clustering inherent in CRTs is distinct (Donner, 1998). Unlike individually-randomised multicentre trials, where individuals within each centre receive different treatments, in CRTs all patients within the cluster are assigned to the same treatment group. Another important concern is that cluster randomisation is more vulnerable to systematic imbalances in baseline covariates between treatment groups than individual randomisation (Puffer et al., 2003; Hahn et al., 2005; Eldridge et al., 2008). This means that covariate adjustment may be required if potential prognostic factors are anticipated to confound the treatment effect. In addition to these, CRTs typically have few clusters and unequal numbers of individuals per cluster, which raises important considerations for the statistical analyses (Ukoumunne et al., 1999; Eldridge et al., 2004; Campbell et al., 2007). While these fundamental issues are well recognised in the statistics literature, they have received little attention in the context of CEA that use CRTs.

To develop appropriate methods for CEA that use CRTs it is important to combine these methodological insights from biostatistics and medical statistics with conceptual ideas from the health economics literature. To transfer methods directly from one context to another is insufficient (Briggs et al., 2002). Undertaking CEA raises issues beyond those that arise in the analysis of clinical outcomes, which need to be carefully addressed. Firstly, as cost function theory suggests, costs may be associated with high levels of within-cluster correlation due to large heterogeneity across clusters in resource use, unit costs, efficiency and patient case-mix (Raikou et al., 2000; Morris et al., 2007). The anticipation of relatively large heterogeneity

further encourages the use of methods that appropriately account for the clustering. Secondly, methodological guidelines for CEA emphasise that individual costs are often correlated with individual outcomes (Willan and Briggs, 2006). Hence, methods that allow for the joint estimation of endpoints while addressing the clustering are required. Thirdly, unlike clinical outcomes, cost data are typically highly skewed, with heavy right tails and bounded by zero. In these circumstances, assuming Normality may not be plausible, and using methods that allow for an appropriate joint distribution of costs and outcomes may be preferred (O'Hagan and Stevens, 2001; Nixon and Thompson, 2005).

This thesis will undertake a conceptual review to examine methodological guidance in the medical statistics and health economics literature. The review will identify key methodological issues that analytical methods need to address in CEA that use CRTs. These criteria will then be used for two main purposes. Firstly they will help identify appropriate statistical methods for the empirical investigations, and secondly to inform the development of a checklist for critical appraisal of CEA that use data from cluster trials.

The empirical investigation will include simulations and case-studies that draw on the concepts from the methods review. The simulations will allow a clear assessment of the statistical performance of each method against the true parameter values, for example the true incremental net benefit (INB). Methods will be compared across realistic scenarios, which will be informed by the conceptual review, a systematic review of the applied literature and available case-studies. Analysis of the case-studies will provide insights on whether the empirical differences between methods identified in the simulations can lead to differences in the cost-effectiveness results used to inform policy making.

## 1.5 Overall contribution of the thesis

This thesis develops specific criteria for critical appraisal of CEA that use CRTs. General checklists and methodological guidelines for CEA do not include specific criteria for critically appraising CEA that use cluster trials (Ofman et al., 2003; Drummond et al., 2005; Evers et al., 2005; Philips et al., 2006). In addition, it is unknown whether applied CEAs that use CRTs use appropriate statistical methods. Research paper 1 addresses these gaps in the literature by identifying fundamental statistical issues that need to be addressed in CEA that use CRTs. These criteria are used to help develop a checklist for critical appraisal of published CEAs that use CRTs. This checklist is applied in a systematic review of the applied literature and finds that applied CEAs that use CRTs often fail to use appropriate statistical methods. In particular, the review shows that most studies fail to account for the clustering or correlation between costs and health outcomes, possibly resulting in misleading inferences about the cost-effectiveness of health care interventions. The new checklist aims to complement more generic checklists and methodological guidance for CEA (Drummond et al., 2005; Evers et al., 2005).

This thesis provides the first comparison of the relative performance of alternative statistical methods for CEA that use CRTs. Research paper 2 considers seemingly unrelated regression (SUR) and generalised estimating equations (GEE), both with robust standard errors, multilevel models (MLMs), and a non-parametric two-stage bootstrap (TSB). This paper firstly shows that methods which fail to account for key statistical issues such as clustering perform poorly, for example confidence interval (CI) coverage is below 0.9, for a nominal level of 0.95. The paper considers SUR and GEEs with robust variance estimators for the first time in this context, and

finds that these may be inappropriate when there are few clusters. Unlike relatively complex MLMs proposed previously for CEA that use CRTs (Grieve et al., 2010), the simulation study demonstrates that a simpler bivariate Normal MLM performed well across the scenarios considered. In addition, the paper extends the non-parametric TSB considered in previous studies (Flynn and Peters, 2004; 2005b) to recognise circumstances where there are unequal cluster sizes, and shows that this method performs relatively well throughout.

Research paper 3 extends seminal work on covariate adjustment in CEA (Willan et al., 2004; Nixon and Thompson, 2005) by investigating the relative merits of SUR, MLMs and non-parametric TSB to address systematic covariate imbalance in CEA that use CRTs. This paper demonstrates that failing to adjust for confounding, even if small, leads to biased results. Unlike research paper 2, this study finds that SUR with robust variance may not perform well even with a moderate number of clusters. To handle the covariate adjustment, this paper extends the original TSB routine (Davison and Hinkley, 1997) and combines it with SUR to adjust for the covariates. This new TSB approach provides unbiased estimates, but it gives poor CI coverage across the scenarios considered. The paper shows that MLMs provide good CI coverage (close to nominal level), even in scenarios with few clusters, unequal cluster sizes and highly skewed data.

The thesis concludes that methods which fail to account for important statistical issues in CEA that use CRTs can provide misleading cost-effectiveness results. It raises awareness of the poor methods used in practice, and provides methodological insights on the relative merits of alternative methods for CEA that use CRTs across a large number of realistic scenarios. The three research papers presented here provide methods for improving CEA that use CRTs to help future studies provide a stronger basis for decision making.

## 1.6 Structure of the thesis

The remaining chapters of the thesis are as follows. Chapter 2 describes the conceptual review, which identifies the key statistical issues in CEA that use CRTs and assesses the appropriateness of prospective statistical methods against those criteria. This chapter then considers each of the statistical methods judged appropriate for CEA that use CRTs. Finally, the chapter examines the plausibility of the assumptions underlying each of these statistical methods and discusses their anticipated performance across different circumstances, in order to help inform the empirical investigation.

Chapters 3 to 5 comprise the three research papers, each prefaced with a brief preamble. Research paper 1 develops a checklist for critical appraisal of the methodological quality of CEAs that use CRTs, and applies this checklist in a systematic review of published studies. Research paper 2 uses simulations and a case study to assess the relative performance of alternative statistical methods for CEA that use cluster trials in setting where baseline covariates are balanced. Research paper 3 evaluates the performance of alternative methods for covariate adjustment in CEA that use CRTs. The paper considers a motivating example with covariate imbalance and conducts further simulations to compare the methods in circumstances where systematic imbalances in baseline covariates can arise.

Chapter 6 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 for applied researchers and policy making.

## 1.7 Contribution of the candidate to the thesis

The work conducted on this thesis was linked to a research grant funded by the Medical Research Council (MRC) methodology programme on improving analytical methods for CEA that use CRTs. Research paper 1 was designed by the candidate in collaboration with his supervisor Richard Grieve, and conducted independently from the project. In this study, the candidate carried out a conceptual review to develop a checklist for critical appraisal of CEA that use CRTs, applied this checklist in a systematic review of applied studies, and interpreted the findings.

The research question for research paper 2 was linked to the MRC project and identified by the principal investigator, Richard Grieve. The candidate led the design of the simulations conducted for this paper while visiting the Modelling and Simulation group at Novartis Pharma (Switzerland), and was guided by Richard Nixon, researcher at Novartis and collaborator on the MRC project. Edmond Ng, the lecturer in statistics working on the project, helped the candidate write code for implementing the statistical methods. The candidate led on the interpretation of the results.

The candidate led on the conception of the research question for research paper 3 in collaboration with his supervisor, Richard Grieve. The candidate was responsible for designing the simulations, writing additional code to implement the statistical methods, and conducting and interpreting the analyses.

Other grantholders in the MRC project (James Carpenter and Simon Thompson) also contributed to the analyses and interpretation of the empirical findings. Further details on more specific

contributions of the candidate and co-authors are described in the cover page of each research paper. The remaining chapters of the thesis are the sole work of the candidate.

## References


Anderson, T. W. 1984. *An introduction to multivariate statistical analysis,* New York ; Chichester, Wiley.

Bachmann, M. O., Fairall, L., Clark, A. & Mugford, M. 2007. Methods for analyzing cost effectiveness data from cluster randomized trials. *Cost Eff Resour Alloc,* 5, 12.

Barber, J. A. & Thompson, S. G. 1998. Analysis and interpretation of cost data in randomised controlled trials: review of published studies. *British Medical Journal,* 317, 1195-1200.

Barber, J. A. & Thompson, S. G. 2000. Analysis of cost data in randomized trials: an application of the non-parametric bootstrap. *Statistics in Medicine,* 19, 3219-3236.

Briggs, A. & Gray, A. 1998. The distribution of health care costs and their statistical analysis for economic evaluation. *J Health Serv Res Policy,* 3, 233-45.

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

Cadth 2006. Guidelines for the Economic Evaluation of Health Tecnologies: Canada. 3rd Ed. *Canadian Agency for Drugs and Technologies in Health.* , Ottawa, Canada.

Campbell, M. J., Donner, A. & Klar, N. 2007. Developments in cluster randomized trials and Statistics in Medicine. *Stat Med,* 26, 2-19.

Carter, B. 2010. Cluster size variability and imbalance in cluster randomized controlled trials. *Stat Med,* 29, 2984-93.

Claxton, K. 2008. Exploring uncertainty in cost-effectiveness analysis. *Pharmacoeconomics,* 26, 781-798.

Cornfield, J. 1978. Randomization by Group - Formal Analysis. *American Journal of Epidemiology,* 108, 100-102.

Davison, A. C. & Hinkley, D. V. 1997. *Bootstrap methods and their application,* Cambridge, UK, Cambridge University Press.

Donner, A. 1998. Some aspects of the design and analysis of cluster randomization trials. *Applied Statistics,* 47, 95-113.

Donner, A. & Klar, N. 2000. *Design and analysis of cluster randomization trials in health research,* London, UK, Hodder Arnold Publishers.

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

Efron, B. & Tibshirani, R. 1993. *An introduction to Bootstrap,* New York, US, Chapman and Hall.

Eldridge, S., Ashby, D., Bennett, C., Wakelin, M. & Feder, G. 2008. Internal and external validity of cluster randomised trials: systematic review of recent trials. *British Medical Journal,* 336, 876-880.

Eldridge, S. M., Ashby, D., Feder, G. S., Rudnicka, A. R. & Ukoumunne, O. C. 2004. Lessons for cluster randomized trials in the twenty-first century: a systematic review of trials in primary care. *Clin Trials,* 1, 80-90.

Evers, S., Goossens, M., De Vet, H., Van Tulder, M. & Ament, A. 2005. Criteria list for assessment of methodological quality of economic evaluations: Consensus on Health Economic Criteria. *Int J Technol Assess Health Care,* 21, 240-5.

Flynn, T. & Peters, T. 2005a. Conceptual issues in the analysis of cost data within cluster randomized trials. *J Health Serv Res Policy,* 10, 97-102.

Flynn, T. N. & Peters, T. J. 2004. Use of the bootstrap in analysing cost data from cluster randomised trials: some simulation results. *Bmc Health Services Research,* 4, 33-43.

Flynn, T. N. & Peters, T. J. 2005b. Cluster randomized trials: Another problem for cost-effectiveness ratios. *International Journal of Technology Assessment in Health Care,* 21, 403-409.

Glick, H. A., Doshi, J. A., Sonnad, S. S. & Polsky, D. 2007. *Economic Evaluation in Clinical Trials,* Oxford, UK, Oxford University Press.

Gold, M. R. 1996. *Cost-effectiveness in health and medicine,* New York, Oxford University Press.

Goldstein, H. 2003. *Multilevel Statistical Models,* Oxford, UK, Oxford University Press.

Greene, W. H. 2003. *Econometric analysis,* Upper Saddle River, N.J., Great Britain, Prentice Hall.

Grieve, R., Nixon, R. & Thompson, S. G. 2010. Bayesian hierarchical models for cost-effectiveness analyses that use data from cluster randomized trials. *Med Decis Making,* 30, 163-75.

Grieve, R., Nixon, R., Thompson, S. G. & Cairns, J. 2007. Multilevel models for estimating incremental net benefits in multinational studies. *Health Econ,* 16, 815-26.

Grieve, R., Nixon, R., Thompson, S. G. & Normand, C. 2005. Using multilevel models for assessing the variability of multinational resource use and cost data. *Health Econ,* 14, 185-96.

Hahn, S., Puffer, S., Torgerson, D. J. & Watson, J. 2005. Methodological bias in cluster randomised trials. *BMC Med Res Methodol,* 5, 10.

Hayes, R. & Moulton, L. 2009. *Cluster Randomised Trials,* Boca Raton - Florida, US, CRC Press, Taylor & Francis Group.

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 Econ,* 11, 415-30.

Iqwig 2009. Methods for assessment of the relation of Benefits to Costs in the German Statutory Health Care System. *Institute for Quality and Efficiency in Health Care.,* Cologne, Germany.

Jackson, C. H., Thompson, S. G. & Sharples, L. D. 2009. Accounting for uncertainty in health economic decision models by using model averaging. *Journal of the Royal Statistical Society Series a-Statistics in Society,* 172, 383-404.

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

Klar, N. & Donner, A. 2001. Current and future challenges in the design and analysis of cluster randomization trials. *Stat Med,* 20, 3729-40.

Leyland, A. & Goldstein, H. 2001. *Multilevel Modelling of Health Statistics,* Chichester, UK, John Wiley & Sons, Ltd.

Manca, A., Rice, N., Sculpher, M. J. & Briggs, A. H. 2005. Assessing generalisability by location in trial-based cost-effectiveness analysis: the use of multilevel models. *Health Econ,* 14, 471-85.

Manning, W. 2006. Dealing with skewed data on costs and expenditures. *In:* JONES, A. (ed.) *The Elgar Companion to Health Economics.* Cheltenham, UK: Edward Elgar.

Mccullagh, P. & Nelder, J. A. 1989. *Generalized linear models,* London, Chapman and Hall.

Mihaylova, B., Briggs, A., O'H<agan, A. & Thompson, S. G. 2011. Review of statistical methods for analysing healthcare resources and costs. *Health Econ,* 20, 897-916.

Morris, S., Devlin, N. & Parkin, D. 2007. *Economic analysis in health care,* Chichester, Wiley.

Murray, D. M., Hannan, P. J., Wolfinger, R. D., Baker, W. L. & Dwyer, J. H. 1998. Analysis of data from group-randomized trials with repeat observations on the same groups. *Stat Med,* 17, 1581-600.

Nice 2008. Methods for Technology Appraisal. *National Institute for Health and Clinical Excellence,* London, UK.

Nixon, R. M. & Thompson, S. G. 2005. Methods for incorporating covariate adjustment, subgroup analysis and between-centre differences into cost-effectiveness evaluations. *Health Econ,* 14, 1217-29.

Noble, S. M., Hollingworth, W. & Tilling, K. 2010. Missing data in trial-based cost-effectiveness analysis: the current state of playdagger. *Health Econ.*

O'Hagan, A. & Stevens, J. W. 2001. A framework for cost-effectiveness analysis from clinical trial data. *Health Econ,* 10, 303-15.

Ofman, J. J., Sullivan, S. D., Neumann, P. J., Chiou, C. F., Henning, J. M., Wade, S. W. & Hay, J. W. 2003. Examining the value and quality of health economic analyses: implications of utilizing the QHES. *J Manag Care Pharm,* 9, 53-61.

Pbca 2008. Guidelines for preparing submissions to the Pharmaceutical Benefits Advisory Committee. *Australian Government - Department of Health and Ageing.,* Camberra, Australia.

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.

Puffer, S., Torgerson, D. & Watson, J. 2003. Evidence for risk of bias in cluster randomised trials: review of recent trials published in three general medical journals. *BMJ,* 327, 785-9.

Puffer, S., Torgerson, D. J. & Watson, J. 2005. Cluster randomized controlled trials. *J Eval Clin Pract,* 11, 479-83.

Raikou, M., Briggs, A., Gray, A. & Mcguire, A. 2000. Centre-specific or average unit costs in multi-centre studies? Some theory and simulation. *Health Economics,* 9, 191-198.

Thompson, S. G. & Barber, J. A. 2000. How should cost data in pragmatic randomised trials be analysed? *British Medical Journal,* 320, 1197-1200.

Timm, N. H. 2002. *Multivariate Analysis,* New York, US, Springer.
Ukoumunne, O. C., Gulliford, M. C., Chinn, S., Sterne, J. A. & Burney, P. G. 1999. Methods for evaluating area-wide and organisation-based interventions in health and health care: a systematic review. *Health Technol Assess,* 3, iii-92.
Weinstein, M. C. & Stason, W. B. 1977. Foundations of Cost-Effectiveness Analysis for Health and Medical Practices. *New England Journal of Medicine,* 296, 716-721.
Willan, A. 2006. Statistical Analysis of cost-effectiveness data from randomised clinical trials. *Expert Revision Pharmacoeconomics Outcomes Research,* 6, 337-346.
Willan, A. & Briggs, A. 2006. *Statistical Analysis of cost-effectiveness data,* Chichester, UK, 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 Econ,* 13, 461-75.
Willke, R. J., Glick, H. A., Polsky, D. & Schulman, K. 1998. Estimating country-specific cost-effectiveness from multinational clinical trials. *Health Econ,* 7, 481-93.
Wooldridge, J. M. 2002. *Econometric analysis of cross section and panel data,* Cambridge, Mass., MIT Press.
Young, T. A. 2005. Estimating mean total costs in the presence of censoring: a comparative assessment of methods. *Pharmacoeconomics,* 23, 1229-42.

# Chapter 2

## Conceptual review of statistical methods for CEA that use CRTs

## 2.1 Introduction

Statistical methods for CEA that use data from RCTs have received considerable attention in the last twenty years. Conceptual ideas from biostatistics, medical statistics and econometrics have stimulated the development of methods for CEA (Gold, 1996; Drummond et al., 2005; Willan and Briggs, 2006; Glick et al., 2007). Despite this methodological progress, commentators have identified areas where these studies still use inappropriate methods, and encouraged further methods development (Willan, 2006; Glick et al., 2007). They highlighted that methodological flaws in applied studies can hinder their usefulness for policy making (Rennie and Luft, 2000; Willan, 2006). It is therefore essential that methods can address the key methodological issues faced by CEA when evaluating the cost-effectiveness of health interventions.

One area where the lack of work on methods has been recognised is in CEA that use CRTs (Klar and Donner, 2001; Flynn and Peters, 2005a; Willan, 2006). A few studies (Flynn and Peters, 2004; Bachmann et al., 2007; Grieve et al., 2010) have considered statistical methods for CEA that use CRTs such as the two-stage bootstrap (TSB) and multilevel models (MLMs), and illustrated their use in practice. However, these studies failed to tackle a number of fundamental questions, which this chapter seeks to address: Can these proposed methods address the key methodological challenges faced by CEA that use CRTs? Are the assumptions underlying these methods satisfied across different circumstances typically observed in CEA that use CRTs? Are additional potentially appropriate methods for CEA that use CRTs available? What are the anticipated relative merits of alternative methods for CEA that use CRTs?

The overall aim of this chapter is, therefore, to identify the key methodological issues faced by CEA that use CRTs and appropriate statistical methods that can address these concerns. The specific objectives of the chapter are:

1. To describe the fundamental statistical issues that can arise in CEA that use CRTs
2. To identify appropriate statistical methods for CEA that use CRTs in settings balanced covariates and systematic covariate imbalance
3. To formulate hypotheses about the relative performance of alternative methods across a range of realistic circumstances in CEA that use CRTs.

To address these objectives, a conceptual review was conducted to gather general methodological insights from medical statistics and health economics literature. The review covered relevant papers concerning the analysis of CRTs and economic evaluation alongside clinical trials available from 1995 to 2010. A broad search of Medline, Scopus, EconLit and Web of Science databases was conducted by combining general search terms such as '*analysis*', '*methods*' and '*models*' with '*cluster randomised trials*' and '*group randomised trials*'. In addition, the citations included in these studies were examined to identify further relevant methodological publications. Working papers databases such as RePec (Research Papers in Economics) and CSSS (Centre for Statistics and Social Sciences) were also considered to cover non-published literature.

This review focused on fundamental issues for statistical analysis in CEA that use CRTs. Other aspects of CEA that use cluster trials such as those pertaining to study design, for example, sample size calculations, were not reviewed here but were considered in the critical appraisal of the applied literature (research paper 1).

The next section of this chapter identifies the main statistical issues that need to be addressed in CEA that use CRTs. Section 2.3 examines the appropriateness of potential statistical methods against the criteria developed in section 2.2. For those methods that are judged appropriate, i.e. can meet all criteria, this section critically assesses their underlying assumptions across different circumstances in CEA that use CRTs. Section 2.4 critically reviews previous evidence on alternative methods considered in the context of CEA that use data from CRTs. The last section discusses the findings of the conceptual review and implications for the empirical investigation.

## 2.2 Fundamental statistical issues in CEA that use CRTs

The first objective of the conceptual review was to help identify key methodological concerns in CEA that use CRTs. The review combined general insights from biostatistics and medical statistics literature together with conceptual ideas from health economics and econometrics, and identified four key issues for statistical analysis in CEA that use CRTs: the clustering of individuals within clusters; the correlation between costs and health outcomes at individual and cluster-level; distributional assumptions for cost and outcome data; and systematic imbalances in baseline covariates. These are discussed in greater detail below.

Firstly, methodological guidelines for the analysis of CRTs (Donner, 1998; Murray et al., 1998; Donner and Klar, 2000; Hayes and Moulton, 2009) highlighted the tendency for patients to be more similar in their characteristics and the care they receive within clusters than between clusters. This means that individual costs and outcomes within a cluster are anticipated to be more homogenous than those in different clusters. In addition, economic theory emphasised that economic factors such as unit costs and resource use tend to be relatively heterogeneous across

clusters (Ukoumunne et al., 1999; Klar and Donner, 2001; Flynn and Peters, 2004). For example, unit costs in teaching hospitals are typically higher than non-teaching hospitals. Hence, costs often exhibit high intra-cluster correlations (ICCs)[4] (Campbell et al., 2005; Flynn and Peters, 2005a).

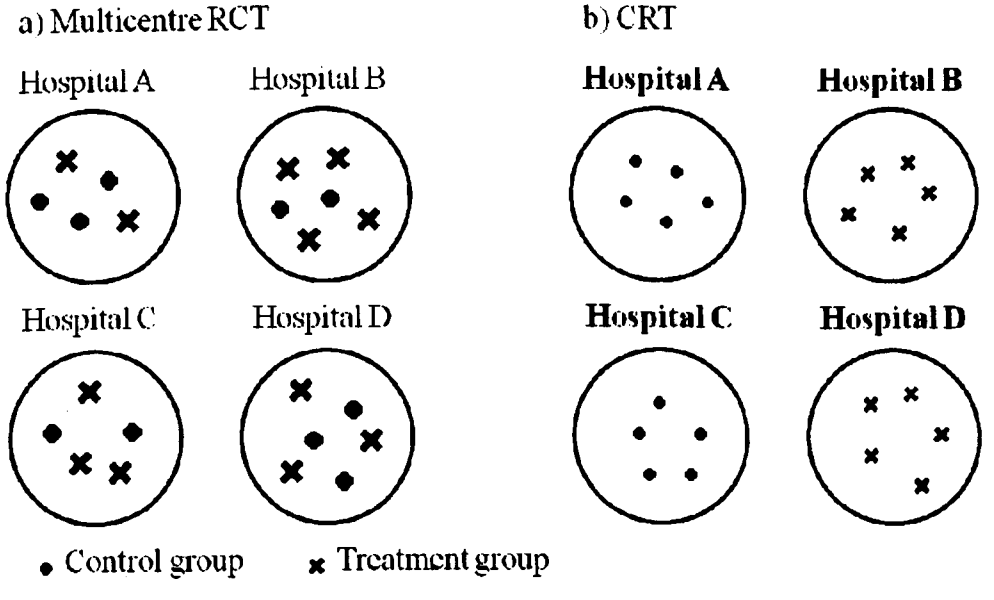
Another important conceptual point that studies need to recognise is that the clustering inherent in CRTs is distinct from that in individually-randomised multicentre trials (Donner, 1998; Murray et al., 1998). In multicentre RCTs, the unit of randomisation is the individual and within each centre patients receive different treatments. On the other hand, when clusters are randomised, individuals within the cluster are allocated to the same treatment group, as described in Figure 2.1 below. Analytical methods are required that allow for this specific type of clustering, otherwise they can underestimate the statistical uncertainty and produce incorrect inferences (Feng et al., 1996; Donner, 1998; Omar and Thompson, 2000; Spiegelhalter, 2001). In addition, when cluster size is correlated with the endpoints, methods that fail to account for clustering may provide biased results (Panageas et al. 2007).

Secondly, methodological guidance for economic evaluation highlighted that methods developed for analysing clinical outcomes in CRTs may not be directly applicable to CEA that use CRTs (Klar and Donner, 2001; Flynn and Peters, 2005a), which tend to have additional complexities. A key requirement identified in the review was that methods for CEA need to recognise the correlation between costs and health outcomes (Briggs et al., 1999; Hoch et al., 2002; Willan et al., 2004; Nixon and Thompson, 2005). For example, patients who respond better to treatment may have shorter hospital lengths of stay and lower costs, which implies a negative correlation

[4] The degree of clustering is commonly summarised by the intra-cluster correlation coefficient, which indicates the proportion of total variance that is at the cluster level.

between individual costs and health outcomes. In CEA that use CRTs, correlation between costs and outcomes may also occur at the cluster-level[5]. For example, teaching hospitals are often associated with higher quality care and better health outcomes, but also higher mean costs. It is therefore important that methods can simultaneously allow for the clustering and correlation between endpoints, which poses specific requirements for the choice of analytical method (Turner et al., 2006).

**Figure 2.1: Clustering inherent to 2-arm multicentre RCTs and CRTs. The unit of randomisation (in bold) is the individual in multicentre RCTs and the cluster (e.g. hospital) in CRTs**




[5] The correlation (corr) between costs and outcomes at the cluster-level can be calculated using the covariance (cov) and variance (var) of cluster-level mean costs (C) and outcomes (E): $corr(C,E) = cov(C,E)/\sqrt{var(C)var(E)}$.

Thirdly, the review found that a general concern in CEA is to make appropriate distributional assumptions about cost and outcome data (Briggs and Gray, 1998; Briggs et al., 2005; Manning et al., 2005; Thompson and Nixon, 2005). Often statistical analyses of clinical outcomes are conducted based on the assumption that data are Normally distributed (Lumley et al., 2002; Mihaylova et al., 2011). However, as Nester (1996) emphasised in his Applied Statistician's Creed:

"*Many methods assume normality (...) simply assert that such assumption is always false. (...) No data are normally distributed*".

In CEA in particular, a plethora of studies has urged analysts to recognise that cost data obtained for individual patients in health care interventions are typically highly skewed (Barber and Thompson, 2000; O'Hagan and Stevens, 2003; Nixon and Thompson, 2004; Briggs et al., 2005; Manning, 2006; Mihaylova et al., 2011). This happens because often a substantial fraction of patients are associated with low or zero costs, while a few patients incur very high costs. Hence, it is important that methods for CEA that use CRTs make appropriate assumptions about the distribution of the data while accounting for both the clustering, and the correlation between costs and outcomes.

Fourthly, while CRTs may be preferred for many public health interventions, commentators have drawn attention to the fact that cluster designs can be vulnerable to systematic imbalances in both individual-level and cluster-level baseline covariates (Donner and Klar, 2000; Puffer et al., 2005; Carter, 2010). Those identifying and recruiting individuals into clusters often have foreknowledge of both the treatment allocation and patient characteristics, and this may lead to systematic imbalance in baseline characteristics (Puffer et al., 2003; Eldridge et al., 2008). For

example, patients with poor prognostic attributes are more likely to enter the control group (Hahn et al., 2005).

Unblinded CRTs can, therefore, cause systematic imbalances in important prognostic factors, potentially leading to biased results. This systematic imbalance is distinct from imbalance due simply to chance, which can arise, for example, when small numbers of clusters are randomised. Therefore, statistical methods for CEA that use CRTs are required that appropriately adjust for any anticipated systematic differences in baseline covariates. Even in settings without systematic covariate imbalance, adjusting for important prognostic factors is expected to correct for any potential chance imbalances and improve the precision of the estimates by explaining some of the sample variability (Senn, 1994; Pocock et al., 2002). In addition, adjusting for covariates can allow the examination of potential subgroup cost-effectiveness effects, which are often of prime interest for policy makers (Sculpher, 2008).

In summary, the conceptual review identified four important statistical issues faced by CEA that use CRTs. While it is recognised that analysts may be concerned with other methodological issues such as censoring or missing data, these were judged to be beyond the scope of this review. This review also focused on CEA that use individual patient data. The findings will help assess the appropriateness of different statistical methods for CEA that use CRTs, as presented in the next section. In addition, these criteria will inform the development of a checklist for critical appraisal of the methods used in practice (research paper 1).

## 2.3 Review of prospective statistical methods for CEA that use CRTs

Using the criteria developed in the previous section, the conceptual review examined the methods literature to identify prospective methods for CEA that use CRTs (for further details on the review, please refer to section 2.1). The focus of the review was on statistical methods for analysing continuous endpoints, for example costs and QALYs, as these are typically of interest for health policy makers (Gold, 1996; Drummond et al., 2005).

### *2.3.1 Hypothesis tests and cluster-level methods*

Hypothesis tests are often used for analysing clinical outcomes from CRTs (Donner and Klar, 2000; Hayes and Moulton, 2009). Conventional methods such as the Wilcoxon rank sum test, two-sample $t$-test and $\chi^2$-tests have been extended to account for the clustering inherent to CRTs (Donner, 1998; Murray et al., 1998; Ukoumunne et al., 1999). However, these standard methods exhibit a number of limitations which make them unlikely to meet all the key criteria for CEA that use CRTs. For example, they lack the flexibility to recognise the correlation between costs and outcomes or to perform covariate adjustment. In addition, previous studies highlighted the fact that hypothesis tests often require particular distributional assumptions such as Normality or equal variances, which are unlikely to be met, for example in the analysis of cost data (Briggs and Gray, 1998; Barber and Thompson, 2000; Mihaylova et al., 2011).

Cluster-level methods such as summary statistics combining data from different clusters and regression analysis at the cluster-level, can also account for the clustering in CRTs since the units of randomisation and analysis are the same. While these approaches are frequently used in

practice because they are relatively simple to implement, they do not appear appropriate for CEA that use CRTs. The key disadvantage of these methods is the loss of information from collapsing all individual observations within the cluster into a single measure (Donner and Klar, 2000). This poses specific limitations to the methods, for example in adjusting for individual-level covariates or accounting for the correlation between individual costs and outcomes. In addition, cluster-level analyses are only appropriate for CRTs with many clusters and equal numbers per cluster. CRTs with these characteristics are not typically found in practice (Ukoumunne et al., 1999; Eldridge et al., 2004; Campbell et al., 2007).

### *2.3.2 Net-benefit regression*

One of the main limitations of the methods discussed in the previous section was the lack of flexibility to take into account the correlation between costs and outcomes. A simple approach to address this correlation is to collapse the endpoints into a single scale before conducting the estimation. For example, costs and health outcomes can be combined into a univariate measure such as the net monetary benefit (Stinnett and Mullahy, 1998). Then, an estimation method, for example linear regression, can be applied to this measure (Hoch et al., 2002). Let $c_{ij}$ and $e_{ij}$ represent the costs and outcomes for the $i$th individual in the $j$th cluster. The individual net monetary benefit is determined as $NB_{ij} = e_{ij}.\lambda - c_{ij}$, where $\lambda$ is the willingness-to-pay for an additional unit of outcome. A net-benefit regression framework can be described as follows (Hoch et al., 2002):

$$NB_{ij} \sim dist(\mu_{ij}^{NB}, \sigma_{NB}) \qquad \mu_{ij}^{NB} = \beta_0^{NB} + \beta_1^{NB} t_j + u_j^{NB} \qquad (1)$$

where $t_j$ is the treatment indicator ($t_j$=0 for control and 1 for treatment group). The net-benefit is assumed to follow a particular distribution with mean $\mu_{ij}^{NB}$ and standard deviation $\sigma_{NB}$. This net-benefit regression approach can allow for the clustering by incorporating a cluster-level random-effect ($u_j^{NB}$), which accounts for the between-cluster variation. While different distributions can be chosen for the net-benefits, Model (1) does not provide sufficient flexibility to make different distributional assumptions for costs versus outcomes. This is an important limitation because costs are typically right skewed whereas outcomes, such as QALYs, may have distributions that are left skewed or Normal (Basu and Manca, 2011). Making inappropriate assumptions about the parametric form of the data may lead to incorrect inferences (Briggs et al., 2005; Thompson and Nixon, 2005; Nixon et al., 2010). In addition, while Model (1) can include covariates, it requires the set of variables for cost and outcomes to be the same (Willan et al., 2004). In fact, a particular covariate may be expected to be an important prognostic factor for either costs or outcomes, but not necessarily for net-benefits. Furthermore, because the net-benefit measure is dependent on the ceiling ratio, separate analyses are required for alternative threshold values.

### *2.3.3 Seemingly Unrelated Regression (SUR)*

The limitations of net-benefit regression suggest that a univariate framework is unlikely to satisfy the key criteria for CEA. A more flexible approach to allow for the correlation between costs and outcomes is to consider bivariate modelling (Timm, 2002; Greene, 2003). For example, methods guidance often proposes the use of a system of seemingly unrelated regression equations (SUR) that allows the error terms to be correlated (Zellner, 1962). The joint estimation of the equations makes full use of the available information and can lead to gains in statistical efficiency when compared with equation-by-equation estimation (Greene, 2003). By allowing

the equations to be linked by their error terms, SUR provides a flexible framework for CEA because it recognises the correlation between individual costs and outcomes as described in Model (2.1):

$$\begin{aligned} c_{ij} &= \beta_0^c + \beta_1^c t_j + \varepsilon_{ij}^c \\ e_{ij} &= \beta_0^e + \beta_1^e t_j + \varepsilon_{ij}^e \end{aligned} \qquad \begin{pmatrix} \varepsilon_{ij}^c \\ \varepsilon_{ij}^e \end{pmatrix} \sim BVN\left( \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_c^2 & \rho\sigma_c\sigma_e \\ & \sigma_e^2 \end{pmatrix} \right) \qquad (2.1)$$

This SUR assumes that the individual error terms $(\varepsilon_{ij}^c, \varepsilon_{ij}^e)$ follow a bivariate Normal distribution (BVN) with mean zero and variances $\sigma_c^2$ and $\sigma_e^2$. Correlation between costs and outcomes is recognised through the parameter $\rho$. The parameters of interest, incremental costs ($\beta_1^c$) and incremental outcomes ($\beta_1^e$), can be estimated by ordinary least squares (OLS). SUR can also be estimated by generalised least squares (GLS), which provides identical estimates to OLS when the same covariates are included for costs and outcomes (Greene, 2003: page 343-344; Willan et al., 2004). When different covariates are included for costs and outcomes, SUR estimation by GLS can improve statistical efficiency (precision) compared to OLS.

To accommodate the clustering, SUR can be extended to include random effects (Singh and Ullah, 1974), but this is not readily available in conventional software packages. A practical alternative way of addressing the clustering in uncertainty estimates is to report robust standard errors (Wooldridge, 2002) (Appendix 2.1 provides further details on the robust variance estimator). The main purpose of robust methods is to produce estimators that are not markedly affected by departures from the key assumptions of classical statistical methods (Huber, 2004). However, a potential concern with using robust methods in CEA that use CRTs is that the asymptotic assumptions underlying the robust variance estimation may not be satisfied in CRTs with few clusters, particularly when there are unequal numbers per cluster (Donner, 1998;

Murray et al., 1998; Omar and Thompson, 2000). Furthermore, another concern with SUR is whether estimates are still unbiased and precise when the model is misspecified, for example, by assuming the error terms to be Normally distributed when costs are highly skewed.

When systematic imbalances are anticipated, covariate adjustment can be easily incorporated in Model (2.1). Unlike net-benefit regression, SUR can allow for the set of covariates to differ for costs and outcomes, but in Model (2.2) below the same individual ($x_{ij}$) and cluster-level ($z_j$) covariates are included for each endpoint:

$$\begin{aligned} c_{ij} &= \beta_0^c + \beta_1^c t_j + \beta_2^c x_{ij} + \beta_3^c z_j + \varepsilon_{ij}^c \\ e_{ij} &= \beta_0^e + \beta_1^e t_j + \beta_2^e x_{ij} + \beta_3^e z_j + \varepsilon_{ij}^e \end{aligned} \qquad \begin{pmatrix} \varepsilon_{ij}^c \\ \varepsilon_{ij}^e \end{pmatrix} \sim BVN\left( \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_c^2 & \rho\sigma_c\sigma_e \\ & \sigma_e^2 \end{pmatrix} \right) \quad (2.2)$$

Model (2.2) can also incorporate interaction terms, for example, of treatment with a continuous individual-level covariate ($x_{ij}$). The covariate $x_{ij}$ can be centred on the mean so that $\beta_1^c$ and $\beta_1^e$ are the incremental costs and outcomes, at the covariate mean.

### *2.3.4 Generalised Estimating Equations (GEEs)*

An alternative approach commonly advocated to acknowledge the clustering is to use population-averaged GEEs (Liang and Zeger, 1986). GEEs offer a flexible extension to likelihood-based generalised linear models for analysing correlated data, and have been generally used in the analysis of CRTs (Donner and Klar, 2000; Hardin and Hilbe, 2003; Hayes and Moulton, 2009). GEEs can account for clustering by incorporating a working correlation matrix which treats the elements that define the within-cluster correlation structure as nuisance parameters (Liang and Zeger, 1986). Unlike SUR, GEEs take a marginal rather than a

conditional approach, i.e. they estimate marginal effects averaged over the population. Commentators argue, however, that for continuous outcomes, marginal and conditional analyses provide the same point estimates (Lee and Nelder, 2004).

Multivariate GEEs have been developed to recognise potential correlation between two or more endpoints (Lipsitz et al., 2009). A bivariate GEE for CEA that use CRTs, with an exchangeable correlation matrix can be written as:

$$\begin{aligned} c_{ij} &= \beta_0^c + \beta_1^c t_j + \varepsilon_{ij}^c \\ e_{ij} &= \beta_0^e + \beta_1^e t_j + \varepsilon_{ij}^e \end{aligned} \qquad V_j = \begin{bmatrix} \Sigma & & & & \\ \boldsymbol{\alpha} & \Sigma & & & \\ \boldsymbol{\alpha} & \boldsymbol{\alpha} & \Sigma & & \\ \vdots & & & \ddots & \\ \boldsymbol{\alpha} & \boldsymbol{\alpha} & \boldsymbol{\alpha} & \cdots & \Sigma \end{bmatrix} \tag{3.1}$$

Where $V_{j\,[n_j \times n_j]}$ is a symmetrical variance-covariance matrix for the $j$th cluster, with the dimension of the cluster size ($n_j$). $\Sigma = \begin{bmatrix} \sigma_c^2 & \sigma_c\sigma_e \\ \sigma_e\sigma_c & \sigma_e^2 \end{bmatrix}$ is the standard variance matrix for costs and outcomes for the $i$th individual, and $\boldsymbol{\alpha} = \begin{bmatrix} \alpha_{c_i,c_{i'}} & \alpha_{c_i,e_{i'}} \\ \alpha_{e_i,c_{i'}} & \alpha_{e_i,e_{i'}} \end{bmatrix}$, with $\alpha$ being the covariance between the individuals $i$ and $i'$ ($i \neq i'$), for each endpoint. Model (3.1) considers an exchangeable correlation matrix, which assumes that the level of correlation between the different observations within the cluster is the same. However, more complex correlation structures could be assumed such as an unstructured matrix which allows for correlations amongst different individuals within each cluster to differ.

While model (3.1) provides a flexible framework to account for the clustering and the correlation between costs and outcomes, its implementation can be complex and it is not currently available

in standard software packages. A simpler alternative is to consider a GEE model with independent estimating equations which stacks costs and outcomes into a single vector but still allowing separate, independent estimates of incremental costs and outcomes (Hardin and Hilbe, 2001; Hardin and Hilbe, 2003). A bivariate GEE model with independent estimating equations can be described as,

$$\begin{aligned} c_{ij} &= \beta_0^c + \beta_1^c t_j + \varepsilon_{ij}^c \\ e_{ij} &= \beta_0^e + \beta_1^e t_j + \varepsilon_{ij}^e \end{aligned} \qquad \varepsilon_{ij}^c \perp \varepsilon_{ij}^e, \quad \varepsilon_{ij}^c \sim N(0,\sigma_c^2),\ \varepsilon_{ij}^e \sim N(0,\sigma_e^2) \tag{3.2}$$

This model structure relies on a general property of population-averaged GEEs, which is that the regression parameter estimates are asymptotically consistent, even if the working correlation matrix is misspecified (Hardin and Hilbe, 2003). This holds as long as the model, i.e. the relationship between the marginal mean and the linear predictor, is correct (Wang and Carey, 2003). Model (3.2) also provides a flexible framework for covariate adjustment, and can include both individual and cluster-level covariates and interaction terms, as described previously for SUR.

Parameter estimates can be obtained by maximum likelihood, assuming that the error terms follow Normal distributions, and provide the same point estimates as OLS estimation. Similar to SUR, a robust estimator for the variance can be considered to allow for clustering when reporting uncertainty (see Appendix 2.1 for further details). A key distinction between the SUR (2.1) and GEEs (3.2) is that the former accounts for correlation between costs and outcomes within the estimation of the parameters of interest. With GEEs, it is recommended that the correlation is acknowledged after the parameter estimation in the robust variance estimator (Williams, 2000; Hardin and Hilbe, 2003: page 30-31).

However, these GEEs share some common concerns with SUR: Firstly, parameter estimates are obtained without acknowledging the clustering. Secondly, the asymptotic properties required for the robust variance estimation may also not be satisfied when there are few clusters (Feng et al., 1996; Bellamy et al., 2000; Ukoumunne and Thompson, 2001). Thirdly, correlation between costs and outcomes at the individual and cluster levels are not separately identified. Fourthly, GEEs considered here also assume that both costs and outcomes are Normally distributed.

### *2.3.5 Multilevel models (MLMs)*

The hierarchical nature of multilevel models (MLMs) provides a suitable approach for analysing clustered data (Goldstein, 2003), and they have been recommended for the analysis of CRTs (Omar and Thompson, 2000; Spiegelhalter, 2001; Turner et al., 2001). Unlike the two previous approaches, MLMs explicitly take into account the clustering in the parameter estimation, by including additional cluster-level random effects $(u_j^c, u_j^e)$ as illustrated in Model (4)[6]:

$$\begin{aligned} c_{ij} &\sim dist(\mu_{ij}^c, \sigma_c^2) \qquad \mu_{ij}^c = \beta_0^c + \beta_1^c t_j + u_j^c \\ e_{ij} &\sim dist(\mu_{ij}^e, \sigma_e^2) \qquad \mu_{ij}^e = \beta_0^e + \beta_1^e t_j + u_j^e + \psi(c_{ij} - \mu_{ij}^c) \end{aligned} \tag{4}$$

---

[6] When costs and outcomes are assumed to follow a bivariate Normal distribution, model (4) naturally extends the SUR model (2.1) as:

$$\begin{pmatrix} \varepsilon_{ij}^c \\ \varepsilon_{ij}^e \end{pmatrix} \sim BVN\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_c^2 & \rho\sigma_c\sigma_e \\ & \sigma_e^2 \end{pmatrix}\right) \qquad \begin{pmatrix} u_j^c \\ u_j^e \end{pmatrix} \sim BVN\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \tau_c^2 & \psi\tau_c\tau_e \\ & \tau_e^2 \end{pmatrix}\right)$$

These random effects are typically assumed to have a joint Normal distribution, with mean zero and cluster-level variances $\tau_c^2$ and $\tau_e^2$ (Spiegelhalter, 2001; Turner et al., 2007), but different specifications can be assumed (Pinheiro et al., 2001). An advantage of these bivariate MLMs over applying a univariate MLM in net-benefit regression is that they can allow for distinct distributions for costs versus outcomes, with means $\mu_{ij}^c, \mu_{ij}^e$ and variances $\sigma_c^2, \sigma_e^2$. For example, it may be reasonable to assume that QALYs follow a Normal distribution while allowing for more skewed distributions for costs. Unlike SUR and GEEs, individual-level ($\rho$) and cluster-level ($\psi$) correlation between costs and outcomes are separately identified and incorporated in the parameter estimation. In model (4) the coefficients $\beta_1^c$ and $\beta_1^e$ still represent incremental costs and outcomes, after allowing for clustering. As with SUR and GEEs, MLMs can include individual and cluster-level covariates, and handle a different set of covariates for cost versus outcome equations.

MLMs can be estimated and interpreted from a frequentist perspective, generally implemented with maximum likelihood (ML), or within a Bayesian approach typically using Markov Chain Monte Carlo (MCMC) methods. Current software options for MCMC estimation afford a wider choice of distributional assumptions (O'Hagan et al., 2001; Nixon and Thompson, 2005). This can be an important advantage when, for example, cost data are highly skewed, and hence incorporating more complex distributions, such as Gamma and Lognormal, may be more plausible. Another strength of the Bayesian framework is that it readily enables more complex MLMs to be implemented, for instance, model (4) could be extended to consider different variances across clusters. It is, however, unclear whether MLMs, using either a ML or MCMC approach, converge when CRTs have few individuals per cluster (Rodriguez and Goldman,

1995; Austin, 2010a). An additional concern is that although MLMs do not require the same set of asymptotic assumptions to be met as in previous studies, the estimation of the variance-covariance structure does still rely on asymptotic properties (Leyland and Goldstein, 2001).

### *2.3.6 The non-parametric two-stage bootstrap (TSB)*

By allowing for a range of alternative parametric distributions of the data, MLMs can potentially make appropriate distributional assumptions. However, the usefulness of assuming a specific parametric distribution relies on the assumption that the chosen form is the true distribution of the data, otherwise results can be misleading (O'Hagan and Stevens, 2003; Nixon and Thompson, 2004; Briggs et al., 2005; Thompson and Nixon, 2005). Non-parametric methods such as the bootstrap can avoid making distributional assumptions, and hence can offer great potential in this context (Barber and Thompson, 2000; Flynn and Peters, 2004; Nixon et al., 2010). The conceptual review identified two non-parametric bootstrap approaches potentially appropriate for CEA that use CRTs. The first procedure is the cluster bootstrap, where only clusters are resampled and individuals within the cluster are kept intact (Davison and Hinkley, 1997: page 100). Although this approach can recognise the clustering, it is associated with similar limitations to other cluster-level methods described in section 2.3.1. A more promising non-parametric bootstrap approach proposed for clustered data involves resampling clusters as well as individuals within the resampled clusters. This non-parametric *two-stage bootstrap* (TSB) (Davison and Hinkley, 1997: page 100) can accommodate the correlation between costs and outcomes by sampling these endpoints in pairs. Davison and Hinkley proposed two distinct ways of implementing the TSB, which are discussed below.

*TSB without shrinkage correction*

One TSB approach proposed by Davison and Hinkley (1997) requires resampling clusters, and then individuals within each resampled cluster, both with replacement. The resultant bootstrap samples are then used to calculate the statistics of interest, for example incremental net benefits (INB) and corresponding confidence intervals (CIs) (Nixon et al., 2010). This approach has been considered for CEA that use CRTs (Bachmann et al., 2007), and the full routine is given in Algorithm 1 below.

***Algorithm 1: Original TSB routine without the shrinkage correction (adapted from Davison and Hinkley, 1997: page 100-101)***

Suppose we have $M_k$ clusters randomised to treatment ($k$=2) and control ($k$=1) groups, with $n_j$ individuals within each cluster $j$.

1. For $i$ in 1 to $n_j$ (individuals in cluster $j$).
2. For $j$ in 1 to $M_k$ (clusters in treatment group $k$).
3. For $k$ in 1 to 2 (treatment groups).
4. Randomly sample (with replacement) $M_k$ clusters in treatment group $k$.
5. Within each of the resampled clusters, randomly select (with replacement) $n_j$ pairs of individual costs and outcomes to preserve the correlation between them.
6. Compute the parameter of interest, $\text{INB} = \lambda * \Delta\text{outcome} - \Delta\text{cost}$, where $\Delta\text{cost} = \bar{c}_{treatment} - \bar{c}_{control}$, and likewise for outcomes.
7. Replicate steps 4 to 6 $R$ times to obtain an estimate of the bootstrap distribution of the parameter of interest.
8. Compute the bias-corrected and accelerated CIs around the mean INB.

However, unless the CRT has many clusters and individuals per cluster, this routine can overestimate the variance. Resampling at the second stage is likely to double-count the within-

cluster variance because the estimated cluster means from resampling at the first stage already incorporate both within and between-cluster variability (Davison and Hinkley, 1997: page 101; Flynn and Peters, 2005a). For example, resampling hospital-level mean costs already recognises the within-hospital homogeneity and further resampling of any deviations from the hospital-level mean costs (individual costs) will overestimate the within-hospital variance.

*TSB with shrinkage correction*

To provide an accurate estimation of the variance, Davison and Hinkley recommend a 'shrinkage estimator' (Davison and Hinkley, 1997: page 102). This procedure requires that shrunken cluster means and standardised individual residuals are calculated before any resampling, as described in steps 4 and 5 of Algorithm 2. Two-stage resampling (with replacement) is then performed by firstly resampling the shrunken cluster means, and secondly, resampling the standardised individual level residuals across all clusters (steps 6 and 7). Shrunken cluster mean costs and outcomes as well as standardised residual costs and outcomes, are resampled in pairs to preserve the individual and cluster-level correlation between costs and outcomes. Bootstrap data sets are constructed by combining the resampled shrunken cluster means and individual level residuals (step 8). As above, the parameter of interest (e.g. the INB) can be taken as average across the bootstrap samples and uncertainty can be reported by calculating bias-corrected and accelerated 95% CIs.

***Algorithm 2: Original TSB routine with the shrinkage correction (adapted from Davison and Hinkley, 1997: page 102)***

Suppose we have $M_k$ clusters randomised to treatment ($k$=2) and control ($k$=1) groups, with $n_j$ individuals within each cluster $j$.

1. For $i$ in 1 to $n_j$ (individuals in cluster $j$).
2. For $j$ in 1 to $M_k$ (clusters in treatment $k$).
3. For $k$ in 1 to 2 (treatments).
4. Calculate shrunken cluster means, ($\hat{x}_j^c$ and $\hat{x}_j^e$), for costs and outcomes[7].
5. Calculate standardized individual-level residuals, ($\hat{z}_{cost,ji}$ and $\hat{z}_{outcome,ji}$), for costs and outcomes[8].
6. Randomly sample (with replacement) $M_k$ pairs of cluster means, ($x^*_{cost,j'}$ and $x^*_{outcome,j'}$), from the shrunken cluster means calculated in step 4.
7. Randomly sample (with replacement) $\sum_{j'=1}^{M_k} n_{j'}$ pairs of residuals, ($z^*_{cost,i'}$ and $z^*_{outcome,i'}$), where $i'$=1... $\sum_{j'=1}^{M_k} n_{j'}$, from the standardized residuals calculated in step 5. Note that the hierarchical structure is ignored in this step.
8. Re-construct the sample, ($y^*_{cost,j'i'}$ , $y^*_{outcome,j'i'}$), by adding the shrunken cluster means from step 6 and the standardized residuals from step 7. For example, $y^*_{cost,j'i'} = x^*_{cost,j'} + z^*_{cost,i'}$ where $i' = 1 \ldots n_{j'}$ and likewise for outcomes; call it a "synthetic" sample.
9. Repeat steps 4 to 8 for each stratum (treatment) and stack these 'synthetic' samples into a single bootstrap sample.
10. Compute the parameter of interest, for example, INB = Δcost × λ - Δoutcome, where Δcost = $\bar{y}^*_{cost,treatment} - \bar{y}^*_{cost,control}$, and likewise for Δoutcome.
11. Replicate steps 6 to 10 $R$ times to form a bootstrap distribution of INB, i.e. a distribution constructed by $R$ replicates of INB.
12. Compute the bias-corrected and accelerated CIs around the mean INB.

---

[7] $\hat{x}_j^c = c\bar{y}_{..}^c + (1-c)\bar{y}_j^c$ where c is given by $(1-c)^2 = \frac{M_k}{M_k - 1} - \frac{SS_W}{b(b-1)SS_B}$ ; $SS_w$= within-sum of squares and $SS_B$ = between-sums of squares, b = average cluster size (a formulation akin to the harmonic mean is used here; see page 412 in Smeeth and Ng (2002). These are similarly calculated for effect and separately so for the two strata (treatments). Note that $j'$ is the new cluster identifier (=1 to $M_k$) which may contain repeats of the old cluster identifier, $j$. All these calculations take place prior to sampling.

[8] $\hat{z}_{cost,ji} = \frac{y_{cost,ji} - \bar{y}_{cost,j.}}{\sqrt{1-b^{-1}}}$, where $y_{cost,ji}$ is the observed cost for the $i$-th individual in cluster $j$. These are similarly calculated for effect and separately for the two strata (treatments). Again, all these calculations take place prior to sampling.

A common concern with any of the bootstrap routines described above is that they still rely on asymptotic properties and it is unclear whether these are satisfied with few clusters, particularly when data are not Normally distributed (O'Hagan and Stevens, 2003; Thompson and Nixon, 2005). Moreover, Davison and Hinkley's original TSB routines were only proposed for clusters with equal sizes (Davison and Hinkley, 1997; Flynn and Peters, 2005a), which may make the method inappropriate for CEA that use CRTs with unequal numbers per cluster.

Algorithm 2 extends the original TSB routine to recognise the variability in cluster size by considering a 'harmonic' mean of the cluster size distribution (see further details in step 4). This measure has been shown to provide a more accurate estimate of the sample mean than the arithmetic mean or median, when cluster sizes are highly variable (Donner and Koval, 1980; Donner and Klar, 2000: page 9).

*TSB combined with SUR for covariate adjustment*

When covariate adjustment is required, the TSB routine described in Algorithm 2 is insufficient. Davison and Hinkley's original resampling approach of combining each shrunken cluster mean with individual residuals drawn across all clusters (steps 7 and 8), does not preserve a relationship between the cluster mean and the covariate information within the cluster. To avoid this problem, Algorithm 2 needs to be modified so that the bootstrap samples respect the cluster membership.

*Algorithm 3: New routine combing the extended TSB with SUR for covariate adjustment*

Suppose we have $M_k$ clusters randomised to treatment ($k$=2) and control ($k$=1) groups, with $n_j$ individuals within each cluster $j$.

1. For $i$ in 1 to $n_j$ (individuals in cluster $j$).
2. For $j$ in 1 to $M_k$ (clusters in treatment $k$).
3. For $k$ in 1 to 2 (treatments).
4. Calculate shrunken cluster means, ($\hat{x}_j^c$ and $\hat{x}_j^e$), for costs and outcomes[9].
5. Calculate standardized individual-level residuals, ($\hat{z}_{cost,ji}$ and $\hat{z}_{outcome,ji}$), for costs and outcomes[10].
6. Randomly sample (with replacement) $M_k$ pairs of cluster means, ($x^*_{cost,j'}$ and $x^*_{outcome,j'}$), from the shrunken cluster means calculated in step 4.
7. Within each resampled cluster, randomly sample (with replacement) $\sum_{j'=1}^{M_k} n_{j'}$ pairs of residuals, ($z^*_{cost,i'}$ and $z^*_{outcome,i'}$), where $i'$=1... $\sum_{j'=1}^{M_k} n_{j'}$, from the standardized residuals calculated in step 5.
8. Re-construct the sample, ($y^*_{cost,j'i'}$ , $y^*_{outcome,j'i'}$), by adding the shrunken cluster means from step 6 and the standardized residuals from step 7. For example, for costs $y^*_{cost,j'i'} = x^*_{cost,j'} + z^*_{cost,i'}$ where $i' = 1 \ldots n_{j'}$, and likewise for outcomes; call it a "synthetic" sample.
9. Incorporate the covariate $w_{j'i'}$ into each synthetic sample as follows: ($y^*_{cost,j'i'} + w_{j'i'}$ , $y^*_{outcome,j'i'} + w_{j'i'}$). Note that the set of covariates can differ for cost and outcomes.
10. Repeat steps 4 to 9 for each treatment arm and stack these 'synthetic' samples into a single bootstrap sample.
11. Replicate steps 6 to 10 $R$ times to construct $R$ bootstrap samples.
12. Apply SUR without robust SE to each bootstrap sample generated in step 11, to estimate mean and SE of incremental costs(ΔC), incremental outcomes (ΔE) and the covariance (ΔC, ΔE), adjusted for potential confounders.
13. Calculate the parameter of interest, e.g. INB, by averaging SUR estimates across the $R$ replications: $\widehat{INB} = (\sum_{r=1}^{R} \widehat{\Delta E_r} * \lambda - \widehat{\Delta C_r})/R$, where $\lambda$ is the willingness-to-pay for a QALY gain.
14. Applying the CLT, CIs for INB can be constructed as $\widehat{INB} \pm 1.96 SE(\widehat{INB})$ where,

$$SE(\widehat{INB}) = \sqrt{[\sum_{r=1}^{R} \mathrm{SE}(\widehat{\Delta \mathrm{E}}_r)^2 \lambda^2 + \mathrm{SE}(\widehat{\Delta \mathrm{C}}_r)^2 - 2\lambda \mathrm{cov}(\widehat{\Delta \mathrm{E}}_r, \widehat{\Delta \mathrm{E}}_r)]/\mathrm{R}}$$

[9] $\hat{x}_j^c = c\bar{y}_{..}^c + (1-c)\bar{y}_j^c$ where c is given by $(1-c)^2 = \frac{M_k}{M_k - 1} - \frac{SS_W}{b(b-1)SS_B}$ ; $SS_w$= within-sum of squares and $SS_B$ = between-sums of squares, b = average cluster size (a formulation akin to the harmonic mean is used here; see page 412 in Smeeth and Ng (2002). These are similarly calculated for effect and separately so for the two strata (treatments). Note that $j'$ is the new cluster identifier (=1 to $M_k$) which may contain repeats of the old cluster identifier, $j$. All these calculations take place prior to sampling.

[10] $\hat{z}_{cost,ji} = \frac{y_{cost,ji} - \bar{y}_{cost,j.}}{\sqrt{1-b^{-1}}}$, where $y_{cost,ji}$ is the observed cost for the $i$-th individual in cluster $j$. These are similarly calculated for effect and separately for the two strata (treatments). Again, all these calculations take place prior to sampling.

For example, patient-level prognostic factors may determine the cluster-level mean outcome, and hence, individual residuals within a cluster need to be combined with the corresponding cluster-level mean. In the modified algorithm (see Algorithm 3), shrunken cluster means and standardised residuals are calculated as before, but each cluster mean is now combined with individual residuals drawn from that same cluster (see steps 7 and 8). Covariate adjustment can be conducted by applying, for example, the SUR model (2.2)[11] to each bootstrap sample. Adjusted incremental costs and outcomes and INBs can then be averaged across the bootstrap replicates (steps 12-14).

The SUR reports SEs for each incremental measure, without applying the robust estimator, because any clustering is expected to be recognised by the TSB routine. The standard errors are also averaged across the bootstrap replicates, to report 95% CIs. A potential concern is that while TSB avoids distributional assumptions, the SUR adjustment assumes that cost and outcome data in the bootstrap replicates are from bivariate Normal distributions.

### *2.3.7 Summary*

The conceptual review identified four potentially appropriate methods for CEA that use CRTs: SUR and GEEs, both with robust standard errors, MLMs and the TSB. Each of these methods can address the main methodological challenges in CEA that use CRTs, as summarised in Table 2.1. This means that a method that accounts only for clustering in univariate analyses of costs and outcomes (e.g. hypothesis tests), or addresses correlation but not clustering (e.g. SUR without robust SE), will be insufficient. Methods are required that allow for both clustering and correlation in the estimation of incremental cost-effectiveness.

---

[11] GEEs or MLMs could also be combined with TSB to adjust for the covariates.

Methods also need to make appropriate distributional assumptions and address the covariate imbalance. For example, TSB alone is not able to adjust for potential confounders when these are anticipated. These criteria will provide an important basis for developing a checklist for critically appraising CEA that use data from cluster trials (Research paper 1).

### *2.3.8 Generating hypotheses about alternative appropriate methods for CEA that use CRTs*

After having identified prospective statistical methods for CEA that use CRTs, the review critically assessed the assumptions underlying these methods to pose hypotheses for the empirical investigation (Chapters 4 and 5).

**Table 2.1: The ability of each method to address the main statistical issues in CEA that use CRTs**

| | Account for the clustering of individuals within clusters? | Recognise the correlation between costs and outcomes? | Flexibility to allow different distributions for costs and outcomes? | Address systematic imbalances in baseline covariates? |
|---|---|---|---|---|
| **Hypothesis tests & cluster-level methods** | Yes | No | No | No |
| **Net-benefit regression** | Yes | Yes | No | No |
| **SUR**[†] | Yes | Yes | Yes | Yes |
| **GEEs**[†] | Yes | Yes | Yes | Yes |
| **MLMs** | Yes | Yes | Yes | Yes |
| **TSB** | Yes | Yes | Yes | No |
| **TSB + SUR*** | Yes | Yes | Yes | Yes |

[†]with robust SE; *other parametric methods such as GEEs and MLMs could also be combined with TSB.

In ideal settings, for example, with large numbers of clusters and individuals per cluster, equal cluster sizes, Normally distributed data and balanced baseline covariates, each method is anticipated to perform well. However, across more realistic circumstances typically found in CRTs, such as few clusters, unequal numbers per cluster and skewed data, differences in performance across methods may be expected. Table 2.2 summarises the anticipated appropriateness of the alternative methods across different circumstances in CEA that use CRTs. The ensuing empirical investigations will test the hypotheses raised for each of the settings, and consider combinations of settings judged *a priori* to differentiate the performance amongst methods.

One of the key features of CRTs is that they tend to have a small number of clusters (Ukoumunne et al., 1999; Eldridge et al., 2004). In these circumstances, the GEEs considered may be less appropriate because the asymptotics required for the robust estimation of the variance may not be satisfied (Feng et al., 1996; Bellamy et al., 2000; Ukoumunne and Thompson, 2001). Although less evidence is available on SUR using robust standard errors, this method is anticipated to have similar limitations (Wooldridge, 2002: pages 149-152). While TSB make a different set of asymptotic assumptions to the SUR and GEEs with robust variance estimators, it still works asymptotically (Davison and Hinkley, 1997: pages 100-102). That is, the bootstrap distribution approximates the true sampling distribution as the original data increases, which may be less conceivable with few clusters (O'Hagan and Stevens, 2003).

When there are few individuals per cluster, previous studies suggested that MLMs may fail to converge because there is little within-cluster information for the estimation of the cluster-level random effects (Rodriguez and Goldman, 1995). When the MLMs are estimated by MCMC, the parameter estimates may be sensitive to prior information on the random effects (Browne and Draper, 2006; Austin, 2010b).

The non-parametric TSB identified in this review has been considered for CRTs with equal cluster sizes (Flynn and Peters, 2005b; Bachmann et al., 2007). It is unclear how this method performs in circumstances where the CRTs have unequal cluster sizes.

In settings where costs are highly skewed, the non-parametric TSB may be preferred to parametric approaches as it avoids distributional assumptions (Barber and Thompson, 2000; Flynn and Peters, 2004). SUR and GEEs assume that data are Normally distributed which may not be plausible under these circumstances. MLMs can allow for skewed distributions such as Gamma and Lognormal, and have been illustrated for CEA usually within a Bayesian framework (Grieve et al., 2005; Nixon and Thompson, 2005). However, the true shape of cost data is still unknown and misspecification of its parametric form may lead to misleading results (Nixon and Thompson, 2004; Briggs et al., 2005; Thompson and Nixon, 2005).

ICCs can be high, particularly for cost data (Campbell et al., 2005), and this may challenge SUR and GEEs with robust SE. In general, robust estimation is expected to provide unbiased, imprecise estimates when the deviation from the classical assumption (i.e., that error terms are identical and independently distributed) is relatively small (Huber, 2004). This is unlikely to be the case with high ICCs. An additional concern with both SUR and GEEs considered here is that the correlation between costs and outcomes at the individual and cluster levels are not separately identified.

In CEA that use CRTs where covariate imbalance is anticipated, the methods described in Table 2.2 should be able to adjust for potential confounders, otherwise they may provide biased estimates. The conceptual review suggested that parametric approaches, for example MLMs or SUR, may have more potential for addressing the systematic imbalances in CEA that use CRT as they can adjust for covariates that are anticipated to be confounders (Willan et al., 2004; Nixon and Thompson, 2005).

**Table 2.2: Anticipated appropriateness of methods proposed for CEA that use CRTs across typical circumstances**

| Characteristics of the CRT | SUR | GEEs | MLMs | TSB |
|---|---|---|---|---|
| **Small number of clusters** | L | L | H | M |
| **Few individuals per cluster** | H | H | M | H |
| **Unequal cluster sizes** | H | H | H | M |
| **Highly skewed costs** | M | M | M | H |
| **High ICCs** | M | M | H | H |
| **Correlation between costs and outcomes** | M | M | H | H |

**Note: H**: Highly appropriate; **M**: Moderately appropriate; **L**: Less appropriate

Non-parametric bootstrap methods such as the TSB have typically less appeal for covariate adjustment (Barber and Thompson, 2000; Dinh and Zhou, 2006; Nixon et al., 2010). One way of addressing this is to combine TSB with one of the parametric methods reviewed in Table 2.2 to adjust for the covariates. While all adjusted methods are expected to provide unbiased results, they may differ in the estimation of the uncertainty across the circumstances presented in Table 2.2. For example, covariate-adjusted SUR and GEEs (with robust SE) are still anticipated to be less appropriate in settings with few clusters. Also, the combination of TSB with a parametric method, for example SUR, means that this method will not be distribution-free. However, the combination of the TSB with SUR may improve the precision of the estimates, compared with SUR alone, if there are few clusters.

## 2.4 Current evidence on methods proposed for CEA that use CRTs

Informed by the criteria developed above, this section critically reviews current evidence on alternative methods proposed for CEA that use CRTs. The focus of the review was methodological, and did not include applied studies. The search strategy was similar to that described in section 2.1. The results are summarised in Table 2.3.

The only study that used simulations, compared two alternative non-parametric bootstrap approaches for CEA that use CRTs (Flynn and Peters, 2005b). This paper assessed the performance of the cluster bootstrap and the TSB across different numbers of clusters, individuals per cluster and levels of correlation between costs and outcomes. The study shows that the cluster bootstrap performs poorly across most scenarios and seems an inappropriate method for CEA that use CRTs. The TSB provides better CI coverage than the cluster bootstrap, but the results vary according to the number of clusters and level of correlation between costs and outcomes. This study has an important limitation in that the bootstrap approaches considered here were only appropriate for CRTs with equal cluster sizes, a feature that is not typical in practice (Eldridge et al., 2004; Eldridge et al., 2006; Carter, 2010). In addition, this paper did not compare the performance of bootstrapping with that of alternative methods.

Another study compared the TSB with MLMs and net-benefit regression, but did not draw on simulations (Bachmann et al., 2007). Using a single case-study with relatively ideal characteristics such as moderate numbers of clusters, equal cluster sizes and low ICCs, Bachmann and colleagues compared these methods to estimate incremental cost-effectiveness ratios (ICER). However, the study found little differences across methods and no general methodological insights could be provided across more realistic circumstances. Another

limitation of the study was that the TSB considered did not apply the recommended shrinkage estimator (Davison and Hinkley, 1997: page 102).

Grieve and others (2010) proposed a range of MLMs for CEA that use CRTs. Firstly, they compared these MLMs with a method that did not account for clustering, simple OLS regression. The authors used a case study to illustrate that failing to take clustering into account can lead to different cost-effectiveness results. This study concluded that MLMs can provide a flexible framework for CEA that use CRTs but did not compare them to other potentially appropriate methods.

**Table 2.3: Evidence comparing appropriate statistical methods for CEA that use CRTs**

| | **Flynn and Peters 2005** | **Bachmann et al. 2007** | **Grieve et al. 2010** |
|---|---|---|---|
| **Methods considered** | Cluster bootstrap and TSB | MLMs, net-benefit regression and TSB | MLMs and OLS |
| **Type of study** | Simulations | Case study | Case study |
| **Parameter of interest** | ICER | ICER | INB |
| **Main CRT features considered** | Clusters: 12 and 24 Cluster sizes: 25 and 50 ICCs of costs and outcomes: 0.01 and 0.1 Costs: moderate skew Covariate imbalance: not considered | Clusters: 40 Cluster sizes: 50 ICCs of costs and outcomes: 0.01 Costs: moderate skew Covariate imbalance: not assessed | Clusters: 70 Cluster sizes: 1 to 77 ICCs of costs and outcomes: 0.01 and 0.18 Costs: moderate skew Covariate imbalance: not assessed |
| **Key insights** | Cluster bootstrap performs worse than TSB, particularly with few clusters | Methods provide similar cost-effectiveness results | Bayesian MLMs seems to provide a flexible framework for CEA that use CRTs |
| **Major limitations** | Only compares bootstrap methods; considers equal cluster sizes | Case-study with ideal characteristics not representative of typical CEA that use CRTs | Does not compare MLMs to other potentially appropriate methods |

This review of current evidence on alternative methods in the context of CEA that use CRTs identifies important gaps in the literature that this thesis aims to address. Firstly, there is no evidence about the methodological quality of published CEA that use CRTs. Secondly, further methodological work is required to assess the relative merits of alternative methods that are potentially appropriate for CEA that use CRTs. This requires simulation work that can test the performance of the methods across a wide range of circumstances, combined with case-studies to compare the methods in practice. Simulation studies will be grounded in the findings of the conceptual review and the results of the applied literature review. Thirdly, none of the studies reviewed in Table 2.3 considered SUR or GEEs, identified in the conceptual review as potentially appropriate methods for CEA that use CRTs. Fourthly, no attention has been devoted to statistical methods that can address systematic covariate imbalance in CEA that use CRTs and an assessment of alternative approaches is therefore warranted.

## 2.5 Discussion

The purpose of this chapter is threefold: to describe the fundamental statistical issues that can arise in CEA that use CRTs; to identify appropriate statistical methods for CEA that use CRTs in settings both where baseline covariates are balanced and where there is systematic covariate imbalance; and to formulate hypotheses about the relative performance of alternative methods across a range of realistic circumstances in CEA that use CRTs. The conceptual review identified four key requirements that statistical methods need to meet in CEA that use CRTs: to account for the clustering inherent in CRTs; to recognise the correlation between costs and outcomes; to make appropriate assumptions about the distribution of cost and outcome data; and to adjust for potential systematic imbalance in

baseline covariates. The second stage of the review found four statistical methods that were judged to satisfy these criteria for CEA that use CRTs: SUR and GEEs, both with robust standard errors; MLMs and a non-parametric TSB. For simplicity, all these methods were described for CEA that use CRTs with two treatment arms but the methods extend to evaluations with more than two randomised groups. SUR, GEEs and MLMs took the common approach of assuming linear additive effects for both cost and outcomes (O'Hagan and Stevens, 2001; Willan et al., 2004; Nixon and Thompson, 2005).

Under ideal circumstances, all these prospective methods are anticipated to be highly appropriate and perform well. However, the relative performance of alternative methods may differ across the more realistic circumstances seen in practice. The review highlighted that the proposed robust variance estimators for the SUR and GEEs rely on asymptotic properties which may not be satisfied in CRTs with few clusters (Feng et al., 1996; Bellamy et al., 2000; Ukoumunne and Thompson, 2001). Another important assumption made by these methods is that residuals are Normally distributed, which may not be reasonable for skewed cost data (Nixon and Thompson, 2005; Thompson and Nixon, 2005). MLMs are expected to perform better with few clusters (Omar and Thompson, 2000) but they still make asymptotic assumptions (Leyland and Goldstein, 2001). In addition, these methods may experience some convergence issues with few individuals per cluster, and also make distributional assumptions. The TSB is expected to perform reasonably well with few clusters (Flynn and Peters, 2004) and avoids parametric assumptions, and therefore, is appealing in settings with highly skewed costs (Barber and Thompson, 2000). However, TSB routines have previously only been proposed for CRTs with equal numbers per cluster (Davison and Hinkley, 1997) and it is unclear how they perform with unequal cluster sizes. Moreover, this method seems less appealing for covariate adjustment, and a combination with a parametric method is required in these circumstances.

Although the review was intended to be general and comprehensive, it had some limitations. Firstly, it focused on statistical methods for continuous endpoints as these are often more informative for policy making (Neumann et al., 2009). Health outcomes may assume other forms such as binary or count, but these were judged to be outside the scope of this review. Secondly, the review did not consider all potential statistical issues that can arise in CEA that use CRTs. For example, methods were not compared across circumstances where costs or outcomes are subject to censoring or missingness. Even when censoring and missing data are completely at random, i.e. not associated with any variables, ignoring these issues may lead to inconsistent results (Willan et al., 2002; Briggs et al., 2003). However, methods proposed here could be extended to address censoring or missing data. For example, when data are censored, Kaplan Meier (Lin et al., 1997) and inverse probability weighting (Bang and Tsiatis, 2000) estimators could be used with, say MLMs and SUR (Lin, 2003; Willan et al., 2005; Liu et al., 2007). With missing data, it is usually assumed that the data are missing at random, i.e. conditional on observed variables. Here, a common approach is to use multiple imputation to handle the missing values (Rubin, 1987). Then, different approaches for imputation can be combined with any of the methods identified in the review (Briggs et al., 2003; Diggle et al., 2007; Lambert et al., 2008).

Thirdly, the review identified other approaches that can help address systematic covariate imbalance in CEA that use CRTs when this is anticipated to be of concern. For example, to help balance observed baseline characteristics between treatment groups, propensity score matching could be applied (Rosenbaum and Rubin, 1983). Then, any of the methods identified above for covariate adjustment such as MLMs and SUR could be used after the propensity score matching to correct for any residual bias (Ho et al., 2007; Abadie and Imbens, 2011) and to address clustering and correlation. In circumstances where unobserved confounders are anticipated, methods such as instrumental variables estimation (Angrist et al.,

1996; Polsky and Basu, 2006) may be required to fully adjust for both observed (overt) and unobserved (hidden) confounding. However, these approaches rely on the availability of a plausible instrument, which may not be realistic in practice.

This conceptual review provides important methodological underpinnings for the empirical work that will be reported in the ensuing chapters. Firstly, the review provides important conceptual insights to help construct specific criteria for critical appraisal of the methodological quality of CEAs that use CRTs (research paper 1). These criteria will supplement existing generic checklists that cover more general aspects of the design and interpretation of CEAs. Secondly, the review identified appropriate statistical methods for CEA that use CRT to be considered for the empirical investigations (research papers 2 and 3). The empirical investigation will consider simulations and case studies to assess the relative performance of alternative methods. This can help provide more general insights on the use of different methods across different circumstances. Thirdly, the review raised a number of hypotheses about the anticipated relative merits of alternative statistical methods across different settings. *A priori* reasoning will help identify key scenarios in which the performances of alternative method are expected to differ and where the empirical investigation should focus.

This review concludes that CEA that use CRTs can place a number of specific requirements on statistical methods. Analytical methods are available to address these challenges but there is little evidence about their relative merits across the complex settings typically observed in CEA that use CRTs. The subsequent chapters provide methods for CEA that use cluster trials to help address the gaps in the literature identified in this chapter.

## References


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

Angrist, J. D., Imbens, G. W. & Rubin, D. B. 1996. Identification of causal effects using instrumental variables. *Journal of the American Statistical Association,* 91, 444-455.

Austin, P. C. 2010a. A Comparison of the Statistical Power of Different Methods for the Analysis of Repeated Cross-Sectional Cluster Randomization Trials with Binary Outcomes. *Int J Biostat,* 6.

Austin, P. C. 2010b. Estimating multilevel logistic regression models when the number of clusters is low: a comparison of different statistical software procedures. *Int J Biostat,* 6, Article 16.

Bachmann, M. O., Fairall, L., Clark, A. & Mugford, M. 2007. Methods for analyzing cost effectiveness data from cluster randomized trials. *Cost Eff Resour Alloc,* 5, 12.

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

Barber, J. A. & Thompson, S. G. 2000. Analysis of cost data in randomized trials: an application of the non-parametric bootstrap. *Statistics in Medicine,* 19, 3219-3236.

Basu, A. & Manca, A. 2011. Regression Estimators for Generic Health-Related Quality of Life and Quality-Adjusted Life Years. *Med Decis Making,* (in press).

Bellamy, S. L., Gibberd, R., Hancock, L., Howley, P., Kennedy, B., Klar, N., Lipsitz, S. & Ryan, L. 2000. Analysis of dichotomous outcome data for community intervention studies. *Stat Methods Med Res,* 9, 135-59.

Briggs, A., Clark, T., Wolstenholme, J. & Clarke, P. 2003. Missing... presumed at random: cost-analysis of incomplete data. *Health Econ,* 12, 377-92.

Briggs, A. & Gray, A. 1998. The distribution of health care costs and their statistical analysis for economic evaluation. *J Health Serv Res Policy,* 3, 233-45.

Briggs, A., Nixon, R., Dixon, S. & Thompson, S. 2005. Parametric modelling of cost data: some simulation evidence. *Health Econ,* 14, 421-8.

Briggs, A. H., Mooney, C. Z. & Wonderling, D. E. 1999. Constructing confidence intervals for cost-effectiveness ratios: an evaluation of parametric and non-parametric techniques using Monte Carlo simulation. *Stat Med,* 18, 3245-62.

Browne, W. J. & Draper, D. 2006. A comparison of Bayesian and likelihood-based methods for fitting multilevel models. *Bayesian Analysis,* 1, 473-514.

Campbell, M. J., Donner, A. & Klar, N. 2007. Developments in cluster randomized trials and Statistics in Medicine. *Stat Med,* 26, 2-19.

Campbell, M. K., Fayers, P. M. & Grimshaw, J. M. 2005. Determinants of the intracluster correlation coefficient in cluster randomized trials: the case of implementation research. *Clinical Trials,* 2, 99-107.

Carter, B. 2010. Cluster size variability and imbalance in cluster randomized controlled trials. *Stat Med,* 29, 2984-93.

Davidson, R. & Mackinnon, J. G. 1993. *Estimation and inference in econometrics,* New York, Oxford University Press.

Davison, A. C. & Hinkley, D. V. 1997. *Bootstrap methods and their application,* Cambridge, UK, Cambridge University Press.

Diggle, P., Farewell, D. & Henderson, R. 2007. Analysis of longitudinal data with drop-out: objectives, assumptions and a proposal. *Journal of the Royal Statistical Society Series C-Applied Statistics,* 56, 499-529.

Dinh, P. & Zhou, X. H. 2006. Nonparametric statistical methods for cost-effectiveness analyses. *Biometrics,* 62, 576-588.

Donner, A. 1998. Some aspects of the design and analysis of cluster randomization trials. *Applied Statistics,* 47, 95-113.

Donner, A. & Klar, N. 2000. *Design and analysis of cluster randomization trials in health research,* London, UK, Hodder Arnold Publishers.

Donner, A. & Koval, J. J. 1980. Estimation of Intra-Class Correlation in the Analysis of Family Data. *Biometrics,* 36, 19-25.

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

Eldridge, S., Ashby, D., Bennett, C., Wakelin, M. & Feder, G. 2008. Internal and external validity of cluster randomised trials: systematic review of recent trials. *British Medical Journal,* 336, 876-880.

Eldridge, S. M., Ashby, D., Feder, G. S., Rudnicka, A. R. & Ukoumunne, O. C. 2004. Lessons for cluster randomized trials in the twenty-first century: a systematic review of trials in primary care. *Clin Trials,* 1, 80-90.

Eldridge, S. M., Ashby, D. & Kerry, S. 2006. Sample size for cluster randomized trials: effect of coefficient of variation of cluster size and analysis method. *Int J Epidemiol,* 35, 1292-300.

Feng, Z. D., Mclerran, D. & Grizzle, J. 1996. A comparison of statistical methods for clustered data analysis with Gaussian error. *Statistics in Medicine,* 15, 1793-1806.

Flynn, T. & Peters, T. 2005a. Conceptual issues in the analysis of cost data within cluster randomized trials. *J Health Serv Res Policy,* 10, 97-102.

Flynn, T. N. & Peters, T. J. 2004. Use of the bootstrap in analysing cost data from cluster randomised trials: some simulation results. *Bmc Health Services Research,* 4, 33-43.

Flynn, T. N. & Peters, T. J. 2005b. Cluster randomized trials: Another problem for cost-effectiveness ratios. *International Journal of Technology Assessment in Health Care,* 21, 403-409.

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

Gold, M. R. 1996. *Cost-effectiveness in health and medicine,* New York, Oxford University Press.

Goldstein, H. 2003. *Multilevel Statistical Models,* Oxford, UK, Oxford University Press.

Greene, W. H. 2003. *Econometric analysis,* Upper Saddle River, N.J., Great Britain, Prentice Hall.

Grieve, R., Nixon, R. & Thompson, S. G. 2010. Bayesian hierarchical models for cost-effectiveness analyses that use data from cluster randomized trials. *Med Decis Making,* 30, 163-75.

Grieve, R., Nixon, R., Thompson, S. G. & Normand, C. 2005. Using multilevel models for assessing the variability of multinational resource use and cost data. *Health Econ,* 14, 185-96.

Hahn, S., Puffer, S., Torgerson, D. J. & Watson, J. 2005. Methodological bias in cluster randomised trials. *BMC Med Res Methodol,* 5, 10.

Hardin, J. W. & Hilbe, J. 2001. *Generalized linear models and extensions,* College Station, Tex., Stata Press.

Hardin, J. W. & Hilbe, J. M. 2003. *Generalized Estimating Equations,* Boca Raton, Florida. US, Chapman & Hall/CRC.

Hayes, R. & Moulton, L. 2009. *Cluster randomised trials,* Boca Raton - Florida, US, CRC Press, Taylor & Francis Group.

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 Econ,* 11, 415-30.

Huber, P. J. 2004. *Robust statistics,* New York ; Chichester, Wiley.

Klar, N. & Donner, A. 2001. Current and future challenges in the design and analysis of cluster randomization trials. *Stat Med,* 20, 3729-40.

Lambert, P. C., Billingham, L. J., Cooper, N. J., Sutton, A. J. & Abrams, K. R. 2008. Estimating the cost-effectiveness of an intervention in a clinical trial when partial cost information is available: a Bayesian approach. *Health Economics,* 17, 67-81.

Lee, Y. & Nelder, J. A. 2004. Conditional and marginal models: another view. *Statistical Science,* 19, 219-238.

Leyland, A. & Goldstein, H. 2001. *Multilevel Modelling of Health Statistics,* Chichester, UK, John Wiley & Sons, Ltd.

Liang, K. Y. & Zeger, S. 1986. Longitudinal data analysis using generalized linear models. *Biometrika,* 73, 13-22.

Lin, D. Y. 2003. Regression analysis of incomplete medical cost data. *Stat Med,* 22, 1181-200.

Lin, D. Y., Feuer, E. J., Etzioni, R. & Wax, Y. 1997. Estimating medical costs from incomplete follow-up data. *Biometrics,* 53, 419-34.

Lipsitz, S., Fitzmaurice, G., Ibrahim, J., Sinha, D., Parzen, M. & Lipshultz, S. 2009. Joint generalized estimating equations for multivariate longitudinal binary outcomes with missing data: an application to acquired immune defficiency syndrome data. *Journal of Royal Statistical Society, Series A,* 172, 3-20.

Liu, L., Wolfe, R. A. & Kalbfleisch, J. D. 2007. A shared random effects model for censored medical costs and mortality. *Statistics in Medicine,* 26, 139-155.

Lumley, T., Diehr, P., Emerson, S. & Chen, L. 2002. The importance of the normality assumption in large public health data sets. *Annu Rev Public Health,* 23, 151-69.

Manning, W. 2006. Dealing with skewed data on costs and expenditures. *In:* JONES, A. (ed.) *The Elgar Companion to Health Economics.* Cheltenham, UK: Edward Elgar.

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. G. 2011. Review of statistical methods for analysing healthcare resources and costs. *Health Econ,* 20, 897-916.

Murray, D. M., Hannan, P. J., Wolfinger, R. D., Baker, W. L. & Dwyer, J. H. 1998. Analysis of data from group-randomized trials with repeat observations on the same groups. *Stat Med,* 17, 1581-600.

Nester, M. R. 1996. An applied statistician's creed. *Applied Statistics-Journal of the Royal Statistical Society Series C,* 45, 401-410.

Neumann, P. J., Fang, C. H. & Cohen, J. T. 2009. 30 Years of Pharmaceutical Cost-Utility Analyses Growth, Diversity and Methodological Improvement. *Pharmacoeconomics,* 27, 861-872.

Nixon, R. M. & Thompson, S. G. 2004. Parametric modelling of cost data in medical studies. *Stat Med,* 23, 1311-31.

Nixon, R. M. & Thompson, S. G. 2005. Methods for incorporating covariate adjustment, subgroup analysis and between-centre differences into cost-effectiveness evaluations. *Health Econ,* 14, 1217-29.

Nixon, R. M., Wonderling, D. & Grieve, R. D. 2010. Non-parametric methods for cost-effectiveness analysis: the central limit theorem and the bootstrap compared. *Health Econ,* 19, 316-33.

O'hagan, A. & Stevens, J. W. 2001. A framework for cost-effectiveness analysis from clinical trial data. *Health Econ,* 10, 303-15.

O'hagan, A. & Stevens, J. W. 2003. Assessing and comparing costs: how robust are the bootstrap and methods based on asymptotic normality? *Health Economics,* 12, 33-49.

O'hagan, A., Stevens, J. W. & Montmartin, J. 2001. Bayesian cost-effectiveness analysis from clinical trial data. *Stat Med,* 20, 733-53.

Omar, R. Z. & Thompson, S. G. 2000. Analysis of a cluster randomized trial with binary outcome data using a multi-level model. *Stat Med,* 19, 2675-88.

Pinheiro, J. C., Liu, C. H. & Wu, Y. N. 2001. Efficient algorithms for robust estimation in linear mixed-effects models using the multivariate t distribution. *Journal of Computational and Graphical Statistics,* 10, 249-276.

Pocock, S. J., Assmann, S. E., Enos, L. E. & Kasten, L. E. 2002. Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: current practice and problems. *Stat Med,* 21, 2917-30.

Polsky, D. & Basu, A. 2006. Selection bias in observational data. *In:* JONES, A. (ed.) *The Elgar Companion to Health Economics.* Cheltenham, UK: Edward Elgar.

Puffer, S., Torgerson, D. & Watson, J. 2003. Evidence for risk of bias in cluster randomised trials: review of recent trials published in three general medical journals. *BMJ,* 327, 785-9.

Puffer, S., Torgerson, D. J. & Watson, J. 2005. Cluster randomized controlled trials. *J Eval Clin Pract,* 11, 479-83.

Rennie, D. & Luft, H. S. 2000. Pharmacoeconomic analyses: making them transparent, making them credible. *JAMA,* 283, 2158-60.

Rodriguez, G. & Goldman, M. 1995. An assessment of estimation procedures for multilevel models with binary responses. *Journal of the royal Statistical Society, Series A,* 158, 73-89.

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

Rubin, D. 1987. *Multiple imputation for nonresponse in surveys,* New York, US, Wiley.

Sculpher, M. 2008. Subgroups and heterogeneity in cost-effectiveness analysis. *Pharmacoeconomics,* 26, 799-806.

Senn, S. 1994. Testing for baseline balance in clinical trials. *Stat Med,* 13, 1715-26.

Singh, B. & Ullah, A. 1974. Estimation of seemingly unrelated regressions with random coefficients. *Journal of the American Statistical Association,* 69, 191-195.

Skrondal, A. & Rabe-Hesketh, S. 2004. *Generalized Latent Variable Modeling: Multilevel, Longitudinal and Structural Equation Models,* Boca Raton, FL, Chapman & Hall/CRC.

Smeeth, L. & Ng, E. S. 2002. Intraclass correlation coefficients for cluster randomized trials in primary care: data from the MRC Trial of the Assessment and Management of Older People in the Community. *Control Clin Trials,* 23, 409-21.

Spiegelhalter, D. J. 2001. Bayesian methods for cluster randomized trials with continuous responses. *Stat Med,* 20, 435-52.

Stinnett, A. A. & Mullahy, J. 1998. Net health benefits: a new framework for the analysis of uncertainty in cost-effectiveness analysis. *Med Decis Making,* 18, S68-80.

Thompson, S. G. & Nixon, R. M. 2005. How sensitive are cost-effectiveness analyses to choice of parametric distributions? *Med Decis Making,* 25, 416-23.

Timm, N. H. 2002. *Multivariate analysis,* New York, US, Springer.

Turner, R. M., Omar, R. Z. & Thompson, S. G. 2001. Bayesian methods of analysis for cluster randomized trials with binary outcome data. *Stat Med,* 20, 453-72.

Turner, R. M., Omar, R. Z. & Thompson, S. G. 2006. Modelling multivariate outcomes in hierarchical data, with application to cluster randomised trials. *Biom J,* 48, 333-45.

Turner, R. M., White, I. R. & Croudace, T. 2007. Analysis of cluster randomized cross-over trial data: a comparison of methods. *Stat Med,* 26, 274-89.

Ukoumunne, O. C., Gulliford, M. C., Chinn, S., Sterne, J. A. & Burney, P. G. 1999. Methods for evaluating area-wide and organisation-based interventions in health and health care: a systematic review. *Health Technol Assess,* 3, iii-92.

Ukoumunne, O. C. & Thompson, S. G. 2001. Analysis of cluster randomized trials with repeated cross-sectional binary measurements. *Stat Med,* 20, 417-33.

Wang, Y. G. & Carey, V. 2003. Working correlation structure misspecification, estimation and covariate design: Implications for generalised estimating equations performance. *Biometrika,* 90, 29-41.

Willan, A. 2006. Statistical Analysis of cost-effectiveness data from randomised clinical trials. *Expert Revision Pharmacoeconomics Outcomes Research,* 6, 337-346.

Willan, A. & Briggs, A. 2006. *Statistical Analysis of cost-effectiveness data,* Chichester, UK, 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 Econ,* 13, 461-75.

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. *Stat Med,* 24, 131-45.

Williams, R. L. 2000. A note on robust variance estimation for cluster-correlated data. *Biometrics,* 56, 645-6.

Wooldridge, J. M. 2002. *Econometric analysis of cross section and panel data,* Cambridge, Mass., MIT Press.

Zeger, S. L., Liang, K. Y. & Albert, P. S. 1988. Models for longitudinal data: a generalized estimating equation approach. *Biometrics,* 44, 1049-60.

Zellner, A. 1962. An efficient method of estimating seemingly unrelated regressions and tests of aggregation bias. *Journal of the American Statistical Association,* 57, 348-368.

## Appendix 2.1: Robust estimators of the variance for SUR and GEEs

### *SUR*

Let the Model (*A*.1) below be a generalisation of the SUR model (2.1), described in section 2.3.3 above, for costs ($c_{ij}$) and outcomes ($e_{ij}$) for the *i*th individual in the *j*th cluster.

$$\begin{aligned} c_{ij} &= \Delta_c t_j + \omega_0 + \omega_1 z_{1,ij} + \omega_2 z_{2,ij} + ...\omega_k z_{k,ij} + \varepsilon_{ij}^c \\ e_{ij} &= \Delta_e t_j + \theta_0 + \theta_1 w_{1,ij} + \theta_2 w_{2,ij} + ...\theta_k w_{k,ij} + \varepsilon_{ij}^e \end{aligned} \qquad (A.1)$$

where $t_j$ is the treatment indicator ($t_j$=0 for control and 1 for treatment group) as above. $z_{1,ij}, z_{2,ij}, ...z_{k,ij}$ and $w_{1,ij}, w_{2,ij}, ...w_{k,ij}$ are the k covariates for costs and outcomes, respectively.

The error terms $\begin{pmatrix} \varepsilon_{ij}^c \\ \varepsilon_{ij}^e \end{pmatrix}$ are independent and identically distributed (i.i.d.) with mean $\begin{pmatrix} 0 \\ 0 \end{pmatrix}$ and covariance $\Sigma = \begin{pmatrix} \sigma_c^2 & \sigma_{ec} \\ \sigma_{ce} & \sigma_e^2 \end{pmatrix}$.

Model (A.1) for *N* individual costs and outcomes can be written as a system of seemingly unrelated regression equations (Greene, 2003; Willan et al., 2004) as

$$\mathbf{y} = X\beta + \varepsilon, \qquad \mathbf{y} = \begin{pmatrix} \mathbf{c} \\ \mathbf{e} \end{pmatrix} \qquad X = \begin{pmatrix} Z & 0 \\ 0 & W \end{pmatrix} \qquad \beta = \begin{pmatrix} \omega \\ \theta \end{pmatrix} \qquad \varepsilon = \begin{pmatrix} \varepsilon_c \\ \varepsilon_e \end{pmatrix} \qquad (A.2)$$

where $\mathbf{c} = (c_1, ...c_N)'$, the *i*th row of $Z = t_j,\ 1, z_{1,ij}, ...z_{k,ij}$, $\omega = (\Delta_c, \omega_0, \omega_1, ...\omega_k)'$,

$\varepsilon_c = (\varepsilon_{c1}, ...\varepsilon_{cN})'$, and $\mathbf{e} = (e_1, ...e_N)'$, the *i*th row $W = t_j,\ 1, w_{1,ij}, ...w_{k,ij}$, $\theta = (\Delta_e, \theta_0, \theta_1, ...\theta_k)'$,

$\varepsilon_e = (\varepsilon_{e1}, ...\varepsilon_{eN})'$.

The vector of parameters of interest ($\beta$) can be estimated by feasible GLS (Greene, 2003: chap. 14) as, $\hat{\beta}_{GLS} = (X'(I \otimes \hat{\Sigma}^{-1})X)^{-1} X'(I \otimes \hat{\Sigma}^{-1})\mathbf{y}$, where $I$ is the identity matrix of dimension $N$. The standard Huber-White sandwich estimator of the variance, which allows for heteroskedastic (non-identical) error terms (Wooldridge, 2002: chap. 7), can be determined as

$$V(\hat{\beta}_{GLS}) = \left( \sum_{i=1}^{N} X_i'(I \otimes \hat{\Sigma}^{-1})X_i \right)^{-1} \sum_{i=1}^{N} X_i'(I \otimes \hat{\Sigma}^{-1})\hat{\varepsilon}'\hat{\varepsilon}\ (I \otimes \hat{\Sigma}^{-1})X_i \left( \sum_{i=1}^{N} X_i'(I \otimes \hat{\Sigma}^{-1})X_i \right)^{-1} \quad (A.3)$$

However, this estimator assumes that the $N$ individual observations are independent. In CRTs, individuals are nested within clusters, and individual observations are no longer independent. A modified (cluster-robust) sandwich estimator of the variance ($V_c$) that allows for the clustering have been proposed as follows (Davidson and MacKinnon, 1993, chap. 6):

$$V_c(\hat{\beta}_{GLS}) = \left( \sum_{i=1}^{N} X_i'\hat{\Sigma}^{-1}X_i \right)^{-1} \sum_{j=1}^{N_c} e_j'e_j \left( \sum_{i=1}^{N} X_i'\hat{\Sigma}^{-1}X_i \right)^{-1} \quad (A.4)$$

Where $N_c$ is the total number of clusters and $e_j = \sum_{i=1}^{n_j} X_i'\hat{\Sigma}^{-1}\hat{\varepsilon}_i'$, with $n_j$ being the total number of individuals in the $j$th cluster. This modified sandwich estimator relies on the assumption that the clusters are independent.

***GEEs***

Let Model ($A$.1) also be a generalisation of GEE Model (3.2) described in the main methods section. Similarly to SUR, let Model ($A$.1) be written as a system of estimating equations as above ($A$.2). The log-likelihood function can be described as $l(\beta, \mathbf{y} \mid X) = \ln f(\mathbf{y} \mid X; \beta)$. The parameters of interest ($\beta$) can be obtained by maximizing the likelihood function, i.e.

satisfying the following condition $\hat{\beta}_{ML} = \partial l(\beta, \mathbf{y} \mid X)/\partial\beta = 0$. The standard sandwich estimator in this context can be written as (Liang and Zeger, 1986; Zeger et al., 1988):

$$V(\hat{\beta}_{ML}) = \left(\sum_{i=1}^{N} X_i' X_i\right)^{-1} \sum_{i=1}^{N} \nabla_i' \nabla_i \left(\sum_{i=1}^{N} X_i' X_i\right)^{-1} \qquad (A.5)$$

where $\nabla_i = \partial l(.)/\partial\beta$ is the score statistic for the *i*th individual. As with SUR, this generic sandwich estimator of the variance needs to be modified to allow for the fact that individuals may be correlated within clusters, and hence the individual-level scores ($\nabla_{ij}$) are no longer independent. A cluster-robust sandwich estimator for GEEs using independent estimation equations can be given by:

$$V_c(\hat{\beta}_{ML}) = \left(\sum_{i=1}^{N} X_i' X_i\right)^{-1} \sum_{j=1}^{N_c} \left(\sum_{i=1}^{n_j} \nabla_{ij}\right)' \left(\sum_{i=1}^{n_j} \nabla_{ij}\right) \left(\sum_{i=1}^{N} X_i' X_i\right)^{-1} \qquad (A.6)$$

Here, $\nabla_{ij}$ is the score statistic for the *i*th individual in the *j*th cluster, $N_c$ is the total number of clusters, and $n_j$ is the total number of individuals in the *j*th cluster. This estimation of the variance addresses the potential dependence structure of the observations within clusters and it has been shown to be an unbiased estimator for data correlated within clusters in a general setting (Williams, 2000; Hardin and Hilbe, 2003: pag. 30-31; Skrondal and Rabe-Hesketh, 2004: pag. 260). In addition, this cluster-robust estimator of the variance can be scaled by $(N_c/N_c\text{-}1)$ for use with small samples (Hardin and Hilbe, 2003, pag. 31).

# Chapter 3

## Checklist for critical appraisal of CEA that use CRTs

## 3.1 Preamble to research paper 1

In the previous chapter, the conceptual review identified key methodological issues that need to be addressed in CEA that use CRTs. Using these criteria, the review then assessed the appropriateness of potential methods for CEA that use CRTs. While a number of statistical methods identified in the literature were judged potentially appropriate for CEA that use CRTs, it is unclear whether these methods are implemented in practice. General checklists and methodological guidelines have been proposed to appraise the methodological standards of CEA studies (Drummond et al., 2005, Evers et al., 2005, Ramsey et al., 2005b). However, these generic checklists do not include sufficient criteria to assess the quality of the methods used in applied CEAs that use CRTs.

Research paper 1 aims to fill these gaps in the literature by developing a new checklist for critical appraisal of CEA that use CRTs. The development of this checklist is informed by the conceptual review (Chapter 2) and includes criteria on key methodological issues in CEA that use CRTs, not covered in more general checklists. Methodological considerations from a panel of experts of different areas such as medical statistics, health economics and epidemiology are also integrated in the development of the checklist. For example, in addition to the key requirements identified in the conceptual review, it was judged important to consider whether the studies have recognised the need for sample size calculations to incorporate any clustering anticipated in costs and health outcomes (Campbell et al., 2005, Donner, 1998, Ukoumunne et al., 1999, Murray et al., 2004). Conversely, the checklist does not include one element raised in the conceptual review which is on whether CEA that use CRTs address any systematic imbalance in baseline covariates.

The checklist is applied in a review of applied literature, which follows the key recommendations of recently published guidelines for good quality systematic reviews (Moher et al., 2009b). The methodological quality of the studies identified in the review is

critically appraised using both the new checklist and a more general checklist for CEA (Drummond et al., 2005: page 30). To help the reviewer judge whether or not a paper meets the criteria of the new checklist, this paper provides a methodological guideline (Appendix 3.1). Findings from the review will also help inform the subsequent empirical investigations comparing the performance of alternative methods for CEA that use CRTs (chapters 4 and 5).

## 3.2. Research paper 1

**Statistical methods for cost-effectiveness analyses that use data from cluster randomised trials: a systematic review and checklist for critical appraisal.**

**Manuel Gomes MSc[1], Richard Grieve PhD[1], Richard Nixon PhD[2], W. J. Edmunds PhD[3]**

[1]Department of Health Services Research and Policy, London School of Hygiene and Tropical Medicine, London, UK.
[2] Modeling and Simulation Group, Novartis Pharma AG, Basel, Switzerland.
[3] Department of Infectious Disease Epidemiology, London School of Hygiene and Tropical Medicine, London, UK.

**Status:** Published in Medical Decision Making (2012), 32 (1): 209-220.

**Contributions:** The candidate designed the research question and developed the checklist in collaboration with RG. The candidate conducted the systematic literature review and applied the checklist to the studies reviewed. RN contributed to the development of the checklist and the accompanying methodological guidance. JE also contributed to the design of the checklist and interpretation of the results. The candidate wrote the first draft of the manuscript. He managed each round of comments and suggestions from 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 | The supervisor

## Abstract

**Introduction:** The best data for cost-effectiveness analysis (CEA) of group-level interventions often come from cluster randomised trials (CRTs), where randomisation is by cluster, for example the hospital attended, not by individual. Methods for these CEA need to recognise both the correlation between costs and outcomes, and that these data may be dependent on the cluster. General checklists and methodological guidance for critically appraising CEA fail to address these issues. This paper develops a new checklist and applies it in a systematic review of CEAs that use CRTs.

**Methods:** We developed a checklist for CEA that use CRTs, informed by a conceptual review of statistical methods. This checklist included criteria such as whether the analysis allowed for both clustering and the correlation between individuals' costs and outcomes. We undertook a systematic literature review of full economic evaluations that used CRTs. The quality of studies was assessed with the new checklist and by the 'Drummond checklist'.

**Results:** We identified 62 papers that met the inclusion criteria. On average, studies satisfied nine out of the ten criteria for the Drummond checklist, but only 20% of criteria for the new checklist. More than 40% of papers adopted statistical methods that completely ignored clustering, and 75% disregarded any correlation between costs and outcomes. Only four studies employed appropriate statistical methods that allowed for both clustering and correlation.

**Conclusions:** Most economic evaluations that use data from CRTs ignored clustering or correlation. Statistical methods that address these issues are available, and their use should be encouraged. The new checklist can supplement generic CEA guidelines and highlight where research practice can be improved.

## Introduction

Policy-makers worldwide require economic evaluations to help decide which health care technologies to provide (NICE, 2008, CADTH, 2006, PBCA, 2008, IQWIG, 2009). Economic evaluations are now also used to help identify which public health interventions are priorities (NICE, 2009), and to evaluate different ways of organising health services (Grieve et al. 2007, Hutchings et al . 2009) . Methods for the economic evaluation of health care programmes are relatively well established and encourage the use of data from randomised controlled trials (RCTs), where patients are individually randomised to alternative interventions (Drummond et al., 2005, Glick et al., 2007, Gold, 1996, Willan and Briggs, 2006). However, for the evaluation of group-level interventions, a cluster randomised trial (CRT) may be preferred. Here the unit of randomisation is the 'cluster', for example the hospital or primary-care physician, not the individual. A cluster design may be chosen because the intervention operates at a group rather than an individual level (e.g. changing incentives for providers), or if there is a high risk of ''contamination'' amongst the individuals within clusters (e.g. evaluating different advertising strategies to encourage smoking cessation).

A fundamental issue in CRTs is that individuals within a cluster are likely to be relatively similar in their characteristics and the care they receive, compared to individuals in different clusters. Individual outcomes or costs within each cluster may therefore be more similar to each other than outcomes or costs from a different cluster. General methodological guidance for CRTs strongly encourages researchers to recognise this clustering in both the design and the analysis of these studies (Donner and Klar, 2000, Donner and Klar, 2004, Hayes and Moulton, 2009, Klar and Donner, 2001, Murray et al., 2004, Ukoumunne et al., 1999). Methods that accommodate clustering in the analysis of clinical outcomes are relatively well-

established in the medical statistics literature (Campbell et al., 2007, Omar and Thompson, 2000, Spiegelhalter, 2001, Turner et al., 2001).

Cost-effectiveness analysis (CEA) that use data from CRTs can face particular methodological challenges. These studies require methods that address clustering in both costs and outcomes, recognise the correlation between individual and cluster-level costs and outcomes (Briggs and O'Brien, 2001, Hoch et al., 2002, Nixon and Thompson, 2005, Willan et al., 2004), and make appropriate assumptions about the distribution of these endpoints (Mihaylova et al., 2010, Nixon and Thompson, 2004, Thompson and Nixon, 2005). Statistical methods for CEA that use CRTs have been proposed (Bachmann et al., 2007, Flynn and Peters, 2005, Grieve et al., 2010), but it is unknown whether suitable methods are used in practice; no previous paper has reviewed the methodological quality of these studies. This is a potential concern, because as a reanalysis of a CRT-based CEA demonstrated (Bachmann et al., 2007, Flynn and Peters, 2005, Grieve et al., 2010), use of methods that ignore clustering will underestimate statistical uncertainty and can lead to inaccurate point estimates. These methodological concerns face all CEA that use CRTs, whether they use data from a single CRT or combine that data with other evidence in a decision model (Philips et al., 2006).

General concerns about methodological standards in CEA have encouraged a plethora of methodological guidelines and critical appraisal criteria which aim to improve methods and reporting transparency (Drummond et al., 2005, Puffer et al., 2005, Graves et al., 2002, Hjelmgren et al., 2001, Ofman et al., 2003, Philips et al., 2006, Ramsey et al., 2005a). However, these generic guidelines do not include sufficiently detailed criteria for CEA based on CRT and a more specific tool for appraising the quality of these studies is warranted.

This paper aims to develop a new checklist for improving methods in CEA that use CRTs, and applies these criteria in a systematic review of previous studies. This checklist is not intended to replace a generic checklist such as that reported in Drummond et al. (Drummond et al., 2005) that covers important, general issues about the overall design and interpretation of a health economic evaluation. Instead, this checklist is proposed for use alongside such general checklists. It covers fundamental statistical issues that arise in CEA that use data from CRTs. The next section presents the key concepts, the new checklist, and the methods used in the systematic literature review. We then present the results of the review, and discuss the findings and the implications for future research.

**Methods**

We undertook a conceptual review that had two main purposes. Firstly, to develop criteria for assessing the methodological quality of economic evaluations that use data from CRTs, and secondly to identify appropriate methods. The review covered relevant methodological guidance for CRTs (Campbell et al., 2007, Donner and Klar, 2000, Donner and Klar, 2004, Hayes and Moulton, 2009, Murray et al., 2004), and statistical methods for CEA (Drummond et al., 2005, Glick et al., 2007, Gold, 1996, Willan, 2006, Willan and Briggs, 2006) and included methodological studies published from 1997 to 2009.

The findings from this conceptual review highlighted that the form of clustering in CRTs is distinct from that in multicentre RCTs where patients are individually randomised, and hence alternative methods are required (Kim, 2010). In multicentre RCTs, although data may be clustered, individuals within each centre are randomised to different treatments; in a CRT all individuals within a cluster receive the same treatment. This specific form of clustering needs

to be anticipated, and accounted for in the sample size calculation, otherwise the study will be underpowered, i.e. it will have less than the nominal power against the study's alternative hypothesis (Donner, 1998, Hayes and Bennett, 1999). The statistical analysis must also recognise any clustering, otherwise type I errors will be higher than anticipated and inference will underestimate the statistical uncertainty (Feng et al., 1996, Omar and Thompson, 2000, Ukoumunne et al., 1999, Ukoumunne and Thompson, 2001). If the CRT has unequal numbers per cluster (imbalance) and a relationship between cluster size and the mean endpoints in each cluster, methods that ignore clustering can provide biased estimates (Panageas et al., 2007).

Methods developed for analysing clinical outcomes in CRTs may not be directly applicable to CEA, which tend to have additional complexities (Glick et al., 2007, Willan and Briggs, 2006). Firstly, CEA would ideally use CRTs with sample sizes calculated according to variances of both costs and effects, with both variances inflated to anticipate clustering (Al et al., 1998, Briggs, 2000). However, in CRTs (like RCTs) sample sizes are usually only calculated to detect differences between treatment groups in clinical endpoints rather than costs, where variation relative to the mean tends to be relatively large (Al et al., 1998, Briggs, 2000, Drummond and O'Brien, 1993, Williamson et al., 2003). However, a relevant recommendation for CEA is that rather than basing power calculations on a single primary endpoint, studies should ideally present several sample size calculations and anticipated measures of uncertainty for incremental effectiveness, cost and measures of cost-effectiveness such as incremental net-benefit (INB) (Bland, 2009).

Secondly, cost function theory and previous evidence suggests that resource use, unit costs and efficiency may vary widely across clusters leading to potentially higher intra-cluster

correlation coefficients (ICCs)[12] for costs than for outcomes (Campbell et al., 2005, Grieve et al., 2005, Thompson et al., 2006). Thirdly, CEA methods need to recognise correlation between costs and outcomes at both the individual and cluster level (Hoch et al., 2002, Nixon and Thompson, 2005, O'Hagan et al., 2001, Willan et al., 2004). The size or direction of the correlation may differ according to whether it is at the individual or the cluster-level. For example, within clusters individuals with lower health status may incur higher costs, i.e. at the individual level there is a strong negative correlation. By contrast, clusters (e.g. hospitals) that have higher mean costs per patient may have on average better outcomes. Fourthly, allowing for clustering in separate, univariate estimation of incremental costs and effectiveness is insufficient for correct inferences; methods need to simultaneously recognise correlation and clustering when reporting incremental cost-effectiveness. For example, using a simple non-parametric bootstrap approach that recognises correlation but ignores clustering would be inadequate.

Finally, statistical methods including those that acknowledge clustering and correlation should make plausible assumptions about the distributions of both costs and outcomes (Briggs et al., 2005, Manning et al., 2005, Hoffman et al., 2001, Nixon and Thompson, 2004, Thompson and Nixon, 2005). For CEA that use individual-patient data (IPD) from a CRT it is important that they carefully consider whether the results are sensitive to alternative assumptions about the distribution of the data. Likewise, for CEA that use endpoints from CRTs as parameters in a decision model, any probabilistic sensitivity analysis should carefully justify the distributional assumptions of these input parameters (Briggs et al., 2005, Mihaylova et al., 2010).

---

[12] The ICC reports the proportion of the total variation that is at the cluster rather than the individual level.

To summarise, statistical methods in CEA that use CRT should recognise clustering, correlation and make plausible distributional assumptions. Studies can be categorised according to whether the methods can accommodate clustering and correlation. We define a Type D study as one that either completely ignores clustering and correlation, or only allows for clustering in one of the univariate measures (e.g. incremental outcome). Type C refers to studies that account for correlation between cost and outcomes but ignore clustering using, for instance, methods such as seemingly unrelated regression (Willan et al., 2004) without robust standard errors. By contrast, studies may account for clustering in both costs and outcomes, but assume they are uncorrelated, for example by estimating incremental costs and outcomes with separate multilevel models (type B). Ideally, studies should use a statistical method that simultaneously accommodates both clustering and correlation (type A) (Grieve et al., 2007, Nixon and Thompson, 2005).

The results of CEA based on a CRT can differ according to these methodological choices. A re-analysis (Bachmann et al., 2007, Flynn and Peters, 2005, Grieve et al., 2010) of a CRT (Morrell et al., 2009) reported that if the analytical method recognised both clustering and correlation (type A study) the probability that the intervention was cost-effective was 0.52, but when the analysis ignored the clustering and correlation, the corresponding probability was 0.80 (type D study) (Table 3.1).

For analyses that acknowledged clustering but ignored correlation (type B), or accommodated correlation but not clustering (type C) the probabilities of the intervention being cost-effective were 0.51 and 0.79, respectively. This example had low levels of correlation between individual costs and outcomes (rho=0.05), but moderate to high levels of clustering (ICCs of 0.05 for outcomes and 0.18 for costs). Here methods that ignored clustering led to underestimation of uncertainty and also provided different point estimates. More generally, it

is unclear a priori whether the choice of method matters and hence analytical methods that can accommodate both clustering and correlation are required. We describe below several methods that can meet the criteria for type A studies.

**Table 3.1: Results from a CEA of a CRT (PoNDER), reanalysed according to whether the statistical methods accounted for clustering and correlation**

| | Clustering and Correlation? | | | |
|---|---|---|---|---|
| | **Neither (Type D)** | **Correlation (Type C)** | **Clustering (Type B)** | **Both (Type A)** |
| **Incremental cost (£)** | -72.0 | -72.2 | -21.4 | -22.31 |
| **(SE)** | (12.73) | (12.64) | (28.83) | (29.85) |
| **Incremental QALY** | 0.00192 | 0.00189 | 0.00196 | 0.00177 |
| **(SE)** | (0.0015) | (0.0015) | (0.0018) | (0.0017) |
| **Incremental cost per QALY (£)** | -37 510 | -38 175 | -10 715 | -12 605 |
| **INB (λ=£20 000)** | 110.3 | 110.0 | 61.46 | 57.61 |
| **(SE)** | (31.93) | (32.41) | (44.46) | (45.83) |
| **P (Cost-effective)** | 0.80 | 0.79 | 0.51 | 0.52 |

**Appropriate statistical methods for CEA that use CRTs**

Methods that recognise simultaneously the clustering and correlation between costs and outcomes in estimating incremental cost-effectiveness can either estimate incremental costs and effectiveness on their original scales (bivariate approaches) (Nixon and Thompson, 2005) or calculate a composite measure of net benefit for each individual (univariate approaches) (Hoch et al., 2002). While either approach can accommodate clustering, the bivariate approaches are generally more flexible (Willan et al., 2004). For example, they can make

appropriate distributional assumptions about the distributions of costs as distinct from outcomes. The conceptual review identified three main groups of bivariate methods able to handle clustering and correlation and make appropriate distributional assumptions: multilevel models (MLMs) (Goldstein, 2003, Leyland and Goldstein, 2001), generalised estimating equations (GEEs) (Hardin and Hilbe, 2003) and the two-stage non-parametric bootstrap (TSB) (Davison and Hinkley, 1997). While bivariate GEEs (Lipsitz et al., 2009) are a recent development, the bootstrap method and bivariate MLMs have been available in the literature for some time.

*Multilevel models (MLMs)*

MLMs can accommodate the hierarchical structure of cost-effectiveness data from CRTs (Bachmann et al., 2007, Grieve et al., 2010) (model 1). Suppose that the costs (c) and effects (e) for the ith individual, within the jth cluster, in the kth trial arm, follows a certain distribution characterised by its mean $(\mu_{ij})$ and variance $(\sigma^2)$. The clustering is explicitly recognised by including parameters $(u_j^c, u_j^e)$ to account for the cluster-specific random-effects. The correlation between the individual costs and effects is introduced through the parameter $\psi$. This bivariate model can then report incremental costs $(\beta_1^c)$ and effects $(\beta_1^e)$ after allowing for the clustering and correlation.

$$\begin{aligned} c_{ij} &\sim dist(\mu_{ij}^c, \sigma_c^2) & \mu_{ij}^c &= \beta_0^c + \beta_1^c t_j + u_j^c \\ e_{ij} &\sim dist(\mu_{ij}^e, \sigma_e^2) & \mu_{ij}^e &= \beta_0^e + \beta_1^e t_j + u_j^e + \psi(c_{ij} - \mu_{ij}^c) \end{aligned} \qquad (1)$$

Model (1) can be estimated and interpreted with a frequentist approach, generally implemented by maximum likelihood (Grieve et al., 2007), or from a Bayesian perspective typically implemented with Markov Chain Monte Carlo (MCMC) methods (Nixon and Thompson, 2005). Fitting MLMs by MCMC in WinBUGS offers a particularly flexible alternative since the wide range of parametric distributions available can help the study make more plausible distributional assumptions than estimating MLMs by maximum likelihood (Grieve et al., 2010).

*Generalised Estimating Equations (GEEs)*

GEEs are a flexible extension of likelihood-based generalised linear models (GLMs) that can accommodate clustered data (Hardin and Hilbe, 2003, Liang and Zeger, 1986), and are often used to analyse clinical outcomes in CRTs (Austin, 2007, Campbell et al., 2007, Turner et al., 2007, Ukoumunne and Thompson, 2001). GEEs take a marginal rather than a conditional approach, i.e. they estimate marginal effects averaged over the population of individuals. Then, the estimated coefficients (e.g. treatment effect) report how the population-averaged outcome, rather than one individual's outcome, responds to the covariate (e.g. treatment indicator).[13] Bivariate GEEs can recognise correlations between dependent endpoints and are a potential alternative for CEA that use CRT data (Lipsitz et al., 2009). These GEEs rely on the general property of population-averaged GEE models in that they provide asymptotically consistent parameter estimates even if the working correlation matrix is misspecified as long as the model, the relationship between the marginal mean and the linear predictor, is correct.

[13] For (clustered) continuous outcomes, marginal and conditional analyses provide the same estimates (Lee and Nelder, 2004).

When the parametric assumptions underlying MLMs are not satisfied, GEEs can be relatively statistically efficient, i.e. they may report smaller variances. By contrast a GEEs may be less efficient than MLMs that make plausible parametric assumptions. A further concern with GEEs is that the asymptotic assumptions rely on the study having sufficient clusters. Methodological guidelines generally recommend that for the GEE to provide reliable estimates the CRT should have at least 20 clusters (Austin, 2007, Feng et al., 1996, Omar and Thompson, 2000, Ukoumunne and Thompson, 2001).

*The non-parametric two-stage bootstrap (TSB)*

The TSB proposed by Davison and Hinkley (Davison and Hinkley, 1997) involves re-sampling clusters and then individuals (both with replacement). This two-stage process accounts for clustering by recognising that the sample variance is partitioned within and between clusters. Mean endpoints are then calculated arithmetically across the bootstrap re-samples. The TSB can account for the correlation between costs and effects by re-sampling them in pairs (Briggs et al., 1999). This algorithm was proposed for balanced clusters (equal numbers per cluster) and it is unknown from the methodological literature, whether the TSB performs well when the clusters are highly imbalanced. Similarly, it is unclear how the TSB performs when the number of clusters is small, in particular when costs are highly skewed (Flynn and Peters, 2005, Nixon et al., 2010, O'Hagan and Stevens, 2003).

Rather than keeping costs and outcomes on their original scales, net benefits (NB) can be calculated for each individual as either net monetary benefits (NMB) or net health benefits (NHB) (Hoch et al., 2002). Clustering can then be recognised in univariate versions of any of

the bivariate methods listed above. For example, the MLM in equation (1) could be re-written as:

$$NB_{ij} \sim dist(\mu_{ij}^{NB}, \sigma_{NB}) \qquad \mu_{ij}^{NB} = \beta_0^{NB} + \beta_1^{NB} t_j + u_j^{NB} \qquad (2)$$

The TSB algorithm described could also be implemented by re-sampling individual net benefits rather than pairs of costs and outcomes. A univariate GEE with net benefits as the dependent variable could be applied to estimate INBs. However, while these univariate models can allow for correlation and clustering, they are more restrictive (Willan, 2006, Willan et al., 2004); for example, these methods do not allow for different covariates to be included in the estimation of incremental costs versus effectiveness (Nixon and Thompson, 2005, Willan et al., 2004).

*Cluster-level summaries and statistical tests*

Individual-level analyses using parametric or non-parametric statistical tests adjusted for clustering (e.g. adjusted two-sample t-test or adjusted $\chi^2$-test), or cluster-level summary statistics (e.g. two-sample t-test, Wilcoxon rank sum test) are simple to implement and can be appropriate for the analysis of clinical outcomes in CRTs (Donner and Klar, 2000, Hayes and Moulton, 2009). However, for CEA these methods lack the flexibility required to address key statistical issues such as the correlation between individual costs and outcomes.

**New checklist for economic evaluations that use data from CRTs**

Based on the conceptual review, we developed specific criteria for assessing how well CEA that use CRT address key methodological issues not covered in generic checklists. Provisional versions of the checklist were reviewed by a panel with relevant expertise (medical statisticians, health economists and epidemiologists). In addition, three researchers not involved in developing the checklist, piloted the tool on 15 papers. The final version of the checklist is reported in Table 3.2. The checklist gives 'credit' to those CEAs that met recommended practice and both used appropriate statistical methods and reported them transparently in the paper (Drummond et al., 2005, NICE, 2008, Philips et al., 2006, Ramsey et al., 2005a). To ascertain whether a study uses appropriate methods but fails to report them in the main CEA paper, the checklist can also be applied using information from additional published sources such as the main CRT paper, previous CEA or reports such as those published by the National Institute of Health Research Health Technology Assessment Programme.

Question 1 in the checklist assesses whether sample size calculations have incorporated any clustering anticipated in outcomes and costs (Al et al., 1998, Briggs, 2000, Donner, 1998, Hayes and Bennett, 1999). Question 2 appraises whether clustering has been recognised in the univariate analysis of incremental costs and outcomes (Campbell et al., 2007, Donner and Klar, 2000, Flynn and Peters, 2004, Hayes and Moulton, 2009, Omar and Thompson, 2000, Spiegelhalter, 2001). Question 3 assesses whether the statistical methods accounted for the correlation between individual costs and effects (Briggs et al., 1999, Hoch et al., 2002, Nixon and Thompson, 2005, Willan et al., 2004). Even if a study has allowed for clustering in the univariate endpoints (Question 2) or correlation between costs and effects (Question 3), it may fail to recognise both clustering and correlation in the joint estimation of costs and

effects or in the estimation of INBs. Question 4 considers whether methods allowed simultaneously for clustering and correlation when estimating incremental cost-effectiveness (Bachmann et al., 2007, Flynn and Peters, 2005, Grieve et al., 2010). A study that uses methods that meet the criterion for Question 4, would also be assumed to satisfy the criteria for Questions 2 and 3. Question 5 considers whether the study has used statistical methods which made explicit, appropriate assumptions about the distribution of costs and outcomes (Briggs et al., 2005, Manning et al., 2005, Mihaylova et al., 2010, Nixon and Thompson, 2004, Thompson and Nixon, 2005).

The checklist can be scored to give a total score for each paper. Each paper can be credited with one point for each criterion met (0.5 points for each sub-question), otherwise zero, with the scores then summed across the criteria to give a total score out of five. Appendix 3.1 offers guidance on how to judge whether each criterion is met.

**Systematic Review of CEAs that use CRTs**

We conducted a systematic literature review of economic evaluations that used data from CRTs. The review included full economic evaluations (Drummond et al., 2005) as they provide information on the relative tradeoffs between the effect of the intervention on costs and outcomes, and hence the most relevant information for health care decision-making (Briggs and O'Brien, 2001). The review satisfied the requirements of a systematic literature review, according to the updated PRISMA statement (Moher et al., 2009a).

**Table 3.2 – Proposed checklist for CEA that use CRTs**

| Criteria | |
|---|---|
| 1. Was the cluster design recognised in the sample size calculation for (44-47, 50): | |
| a) outcomes? | Yes ☐ No ☐ |
| b) costs? | Yes ☐ No ☐ |
| 2. Was clustering account for in the univariate analysis of (12, 14, 18-20, 71): | |
| a) outcomes? | Yes ☐ No ☐ |
| b) costs? | Yes ☐ No ☐ |
| 3. Did the statistical analysis account for the correlation between costs and outcomes (23-25, 68)? | Yes ☐ No ☐ |
| 4. Did the study account for clustering and correlation in the estimation of incremental cost-effectiveness (29-31)? | Yes ☐ No ☐ |
| 5. Did the study explicitly make appropriate assumptions about the distribution of (26-28, 55, 56): | |
| a) outcomes? | Yes ☐ No ☐ |
| b) costs? | Yes ☐ No ☐ |

**Search strategy**

A consistent and transparent literature search was undertaken over a wide range of databases in health economics, public health and medicine used in previous systematic reviews in economic evaluation (Briggs, 1999, Philips et al., 2004, Sculpher et al., 2004). The databases included Health Economic Evaluations Database (HEED), NHS Economic Evaluations Database (NEED), EconLit, EMBASE, PubMed, MedLine, Scopus and Web of Science. In addition, non-published literature was also searched in working papers databases (e.g. Ideas, NetEc, EconWPA) and Conference Papers Index (CPI). The search strategy combined two

sets of free-text terms related to: 1) "economic evaluations", where terminology used in previous search strategies (Sculpher et al., 2004, Puffer et al., 2003) was extended; 2) 'cluster randomised trials', where the terms 'cluster', 'group', 'community', 'clinic', 'centre' or 'area' were used to identify CRTs that included any of these terms in the title or abstract. Appendix 3.2 describes the search strategy for MedLine, which was slightly modified for the other databases.

**Inclusion criteria**

To minimise the risk of missing potentially relevant economic evaluations, the inclusion criteria were broad, and the search was conducted on all the available evidence up to the end of 2009. Titles and abstracts were screened to check whether the study met the following inclusion criteria:

1 - Study must be undertaken alongside a CRT.

2 - Paper must compare both cost and outcomes of alternative interventions.

3 - Results must be reported on an incremental basis.

4 - The paper must be a cost-effectiveness, cost-utility or cost-benefit analysis, but not a cost minimisation or cost-consequence analysis.

5 - Paper published in any language but with an abstract in English.

6 - Several papers that use data from the same CRT can be included provided each of them reports results not published elsewhere.

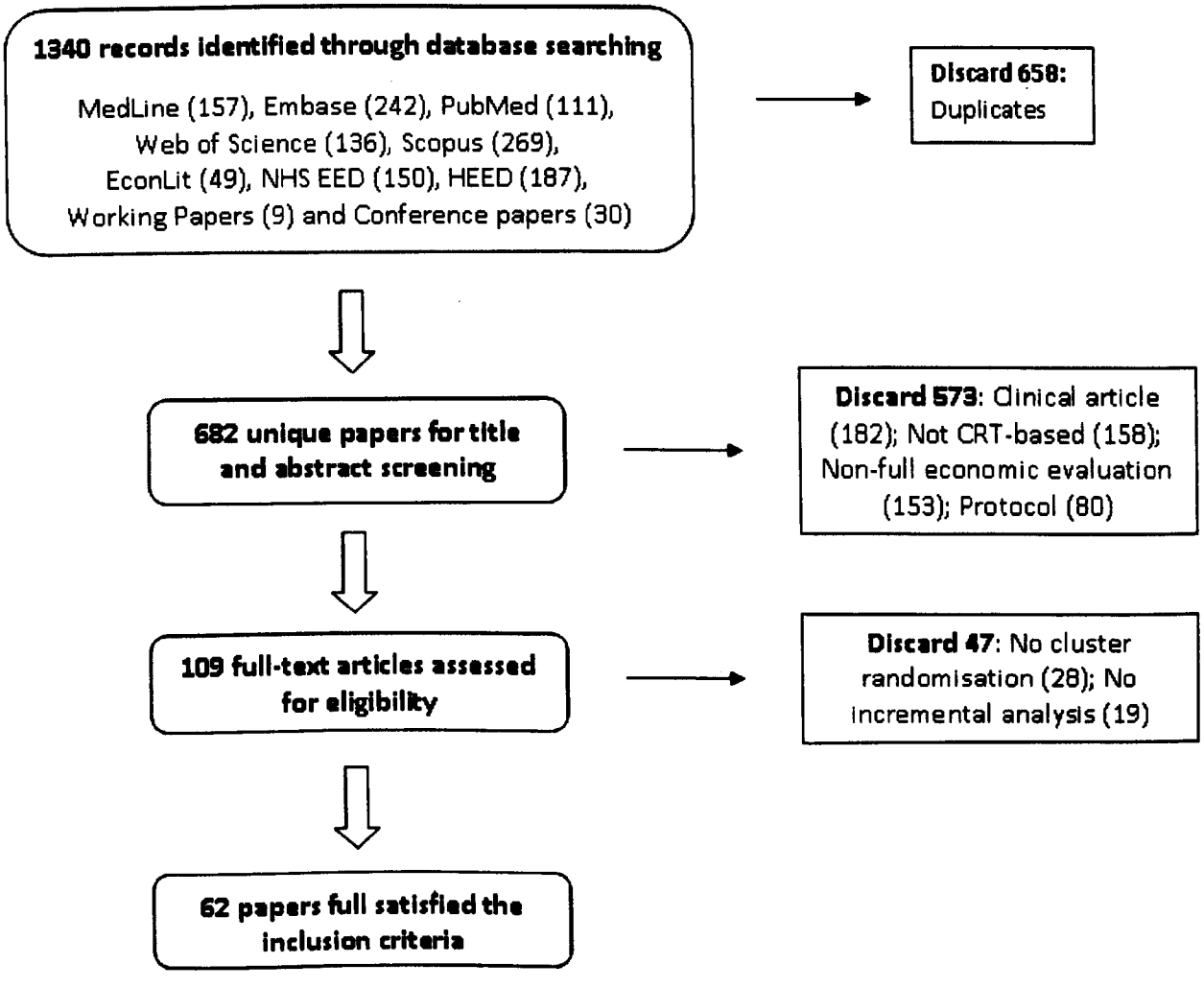
The title and abstract screening were conducted by one reviewer. To check the reliability of the way the inclusion criteria were applied at the abstract screening stage, the abstracts for those articles excluded by the first reviewer were re-screened by a second reviewer. At the abstract and title screening stage both reviewers were blinded to the authorship. There were no disagreements in the articles excluded. Full-text review and data extraction were conducted by the first reviewer, who at this stage was not blinded to the authorship. The selected articles were critically appraised with the Drummond checklist (Drummond et al., 2005), chosen because it has been commonly used to critically appraise CEA. The economic evaluation papers were critically appraised by both reviewers independently, using the cluster-specific checklist. The reliability of the checklist was good, the reviewers only disagreed on whether a certain criterion was met in 10 studies (kappa > 0.9). These disagreements were resolved in consultation with a third reviewer. Finally, to consider the possibility that studies that may have used appropriate methods but failed to report them in the main CEA paper, we re-applied the cluster-specific checklist using additional information from further sources, such as the accompanying clinical paper or previous economic evaluations of the same study.

The review reported the overall results of critically appraising the papers with both checklists. In addition, several prior hypotheses were considered concerning the context of the studies. To assess whether studies had improved since the publication of several key papers on relevant statistical methods, results were compared over time (post 2005 versus 2005 or earlier). Other pre-specified hypotheses were: whether the results differed according to the type of journal in which the CEA was published (medical versus other), and according to the overall study design (CEA that used IPD versus alternatives such as decision-models that used summary inputs from CRTs).

## Results

The database search strategy yielded 682 unique articles, 573 of which were excluded after screening the title and abstract and a further 47 after full text review. The most common reason for exclusion was that studies were either not based on CRTs or were not full economic evaluations (Figure 3.1). A total of 62 papers (54 CRTs) satisfied the inclusion criteria, and were included in the review (see Appendix 3.3 for a full list). For 45 of the 62 papers we identified a relevant accompanying article for review such as the main clinical paper on the CRT.

**Figure 3.1: Study selection flow diagram**



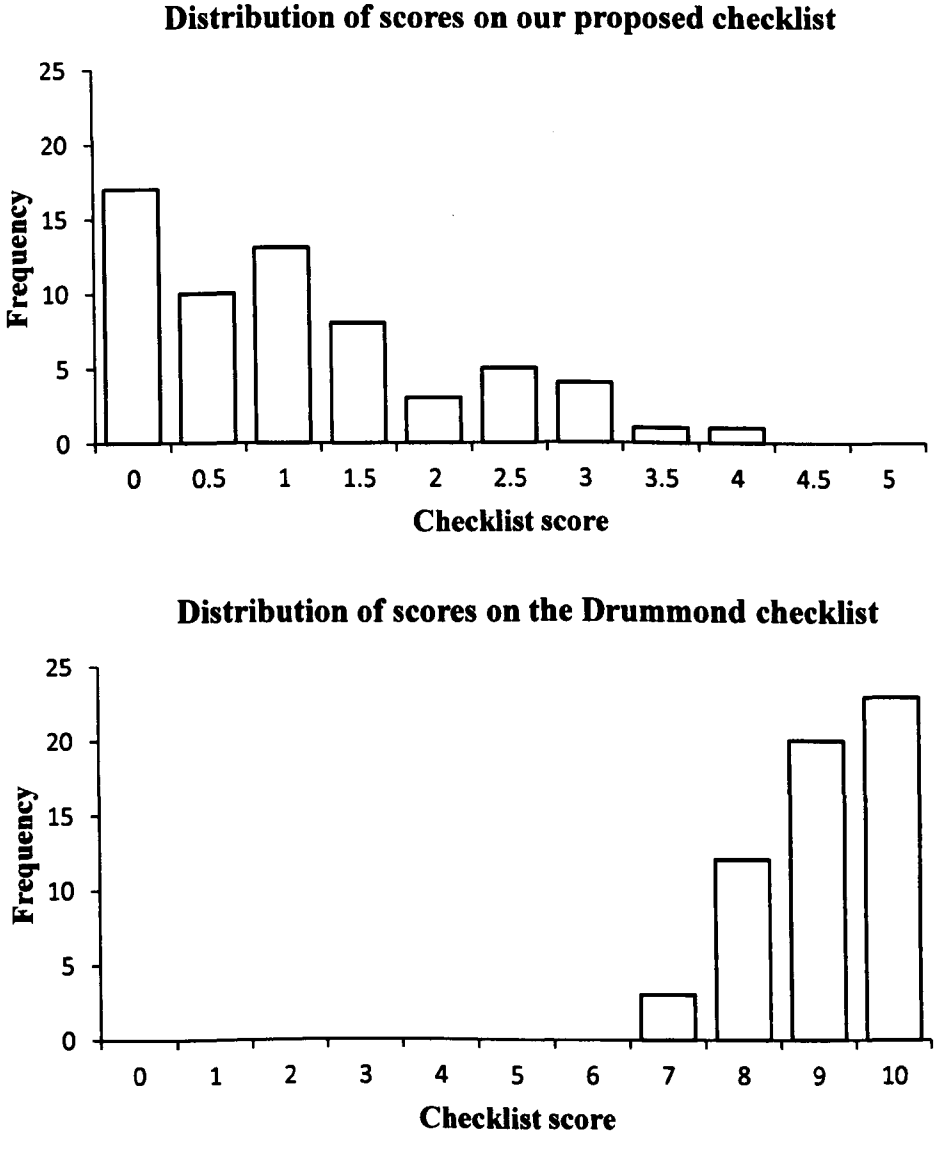
The selected papers were published from 1997 to 2009 (inclusive), mainly in medical and public health journals (Table 3.3). The economic evaluations covered a wide range of group-level interventions: alternative ways of organising health care services (e.g. cost-sharing incentives programmes), different disease management programs, screening, health care promotion strategies, and evaluations of clinical guidelines. More than two thirds of papers were based on IPD from the CRT, the remaining studies used summary data in decision-analytic models or simply reported aggregated, deterministic measures (Appendix 3.4 reports additional characteristics of the reviewed studies).

**Table 3.3: Characteristics of the studies included in the review (n=62)**

| Characteristic | N (%) |
|---|---|
| **Year** | |
| 2009 | 12 (19.4%) |
| 2008 | 5 (8.1%) |
| 2007 | 9 (14.5%) |
| 2006 | 10 (16.1%) |
| 2005 | 7 (11.3%) |
| Before 2004 | 9 (14.5%) |
| **Journal** | |
| Medical | 36 (58.1%) |
| Health Economics | 11 (17.7%) |
| Public Health | 14 (22.6%) |
| Statistics | 1 (1.6%) |
| **Intervention type** | |
| Health services | 21 (33.9%) |
| Disease management | 16 (25.8%) |
| Screening | 9 (17.7%) |
| Prevention | 11 (14.5%) |
| Guidelines | 6 (8.1%) |
| **Study design** | |
| IPD | 42 (67.7%) |
| Decision model | 8 (12.9%) |
| Aggregate analysis | 12 (19.4%) |

On average the papers met 90% of the Drummond checklist criteria. When the studies were assessed using the cluster-specific checklist, 20% of the criteria were met, the median total quality score was 1 out of 5. Figure 3.2 describes the distributions of the criteria met for each checklist. The distribution of the criteria met is heavily left skewed for the Drummond checklist and right skewed for the cluster-specific checklist.

**Figure 3.2: Methodological quality of the selected papers using the Drummond checklist and our proposed checklist**



**Results from applying the checklists**

Table 3.4 presents disaggregated results according to whether the study met each criterion in the checklist, firstly according to the information reported in the main CEA paper and secondly according to additional information available from other sources.

**Table 3.4: Results from applying the CRT checklist to a) the economic evaluation paper, and b) the economic evaluation and supplementary sources. N(%) of studies that met each criterion and total score (n=62)**

| Question | Economic evaluation | Economic evaluation & supplementary papers |
|---|---|---|
| 1. Clustering recognised in the sample size calculation for: | | |
| a) outcomes | 12 (19.4%) | 38 (61.3%) |
| b) costs | 0 (0%) | 7 (11.3%) |
| 2. Clustering accounted for in univariate analysis of | | |
| a) outcomes | 32 (51.6%) | 36 (58.1%) |
| b) costs | 20 (32.3%) | 20 (32.3%) |
| 3. Accounted for correlation between costs and outcomes | 16 (25.8%) | 16 (25.8%) |
| 4. Accounted for clustering and correlation for cost-effectiveness | 4 (6.5%) | 4 (6.5%) |
| 5. Appropriate assumptions about the distribution of: | | |
| a) outcomes | 10 (16.1%) | 14 (22.6%) |
| b) costs | 17 (27.4%) | 20 (32.3%) |
| Median (IQR) total score | 1 (1.5) | 1.25 (1.5) |
| Mean (sd) total score | 1.1 (1.0) | 1.41 (1.2) |

The results, in general, were similar across these two sets of sources apart from for the criterion on the sample size for outcomes; more than 80% of the economic evaluations did not report that clustering was recognised in sample size calculations whereas, once supplementary papers such as the corresponding clinical paper were considered, less than 40% of studies failed this criterion.

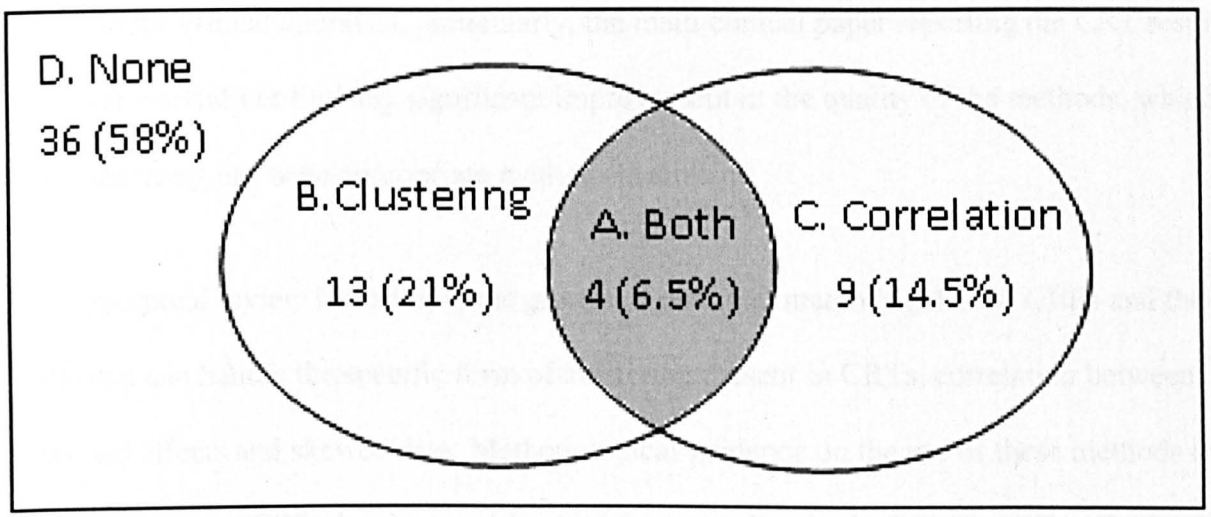
More than 40% of papers completely ignored clustering in the analyses (37% when supplementary information was also considered). Almost 70% ignored clustering in the univariate analysis of costs and 50% in the univariate analysis of outcomes. Those studies that allowed for clustering tended to use multilevel models or GEEs. A total of 37 papers reported ICCs for outcomes, 5% suggested that levels of clustering were 'high' (ICC>0.1), 65%, that they were 'moderate' (0.01<ICC<=0.1) and 30% that they were 'low' (ICC<=0.01).

More than 70% of papers neglected the correlation between costs and outcomes (Table 3.4). We found only four studies (6.5%) that used statistical methods (bivariate MLMs and TSB) that allowed for both clustering and correlation in the estimation of incremental costs and outcomes, of which three did not make appropriate distributional assumptions. Overall more than 60% of studies failed to make appropriate assumptions about the distribution of the costs or outcomes.

Figure 3.3 summarises the main findings. The majority of papers (58%) were defined as 'type D' studies as they either ignored both clustering and correlation or neglected correlation and only accounted for clustering in one of the outcomes (costs or effects). One fifth of papers reported accounting for clustering in the univariate analyses of both costs and effects but did not allow for correlation (type B), and 15% recognised correlation but failed to correctly

acknowledge the clustering (type C). Only four studies appropriately accounted for both clustering and correlation in the estimation of cost and effects (type A).

**Figure 3.3: The proportions of papers that allow for clustering and correlation (n=62)**




**Discussion**

This study makes two important contributions to the methodological literature on CEA. Firstly, this paper presents a tool to assess and help improve the quality of economic evaluations that use data from CRTs. Secondly, our systematic review finds that although the methodological quality of papers was generally good when judged against the Drummond checklist, it was poor when assessed against the cluster-specific checklist. The main purpose of the cluster-specific checklist is to summarise how well studies address the key statistical issues at a particular point in time. This purpose is reflected in the way summary scores can be calculated with each item given equal weight (Ofman et al., 2003). The checklist could be used to assess whether studies improve over time, both by comparing the overall score but also specific components, such as the proportion of studies that allow for both clustering and correlation (type A studies).

This review followed previous recommendations for critical appraisal and judged study quality not just according to whether studies used appropriate methods but also whether they reported them transparently. It is conceivable that the studies used sound methods but did not report them in the main economic evaluation paper. We therefore considered additional papers in the critical appraisal, particularly, the main clinical paper reporting the CRT results. However, we did not find any significant improvement in the quality of the methods, which raises the question: were appropriate methods available?

The conceptual review identified three groups of potential methods (MLMs, GEEs and the TSB) that can handle the specific form of clustering present in CRTs, correlation between costs and effects and skewed data. Methodological guidance on the use of these methods to analyse data from CRTs has been established for some time in the literature (Campbell et al., 2007, Davison and Hinkley, 1997, Donner and Klar, 2000, Hayes and Moulton, 2009, Turner et al., 2006). However, these methods have yet to permeate applied health economic evaluations that use data from CRTs. Indeed we found no evidence that the use of these potentially appropriate methods improved over time. The most flexible way to implement these methods is in bivariate approaches that jointly estimate costs and effects. Alternatively, net benefits can be calculated for each individual, and then any of the above methods can be applied to allow for clustering in the net benefit estimates. While the latter approach lacks flexibility, allowing for clustering using any of the approaches outlined would improve on the status quo.

The finding that, in practice most CEAs fail to use or report appropriate statistical methods is consistent with previous reviews on design and analysis of CRTs and statistical methods in CEA based on RCTs (Puffer et al., 2005, Briggs and Gray, 1999, Eldridge et al., 2004, Ukoumunne et al., 1999). The poor quality of CEAs based on CRTs may reflect the relative

lack of attention given to statistical methods for CEA that use CRTs. Only one study (Bachmann et al., 2007) has attempted to compare alternative statistical methods for CEA from CRTs, but the study used a CRT with relatively 'ideal' characteristics, not representative of the majority of studies we reviewed. That CRT had many balanced clusters (n=50), and so the methodological comparison offered limited generalisability; For example, more than 70% of studies in our review had imbalanced clusters. Likewise, Flynn and Peters (Flynn and Peters, 2005) only considered the TSB for circumstances when CRTs have balanced clusters. A recent study proposed MLMs for economic evaluations of CRTs and suggested that this approach led to different cost-effectiveness results compared to methods that ignored clustering (Grieve et al., 2010). However, this study did not consider alternative approaches (e.g. GEEs or TSB). Research is currently underway to investigate the relative merits of alternative statistical methods for CEA from CRTs across the range of circumstances representative of the studies reviewed.

Although this paper has developed criteria for improving methods for economic evaluations that use CRTs and carefully applied them in a systematic literature review, it does have some limitations. Firstly, the checklist does not attempt to cover all the statistical issues that can arise when designing or analysing an economic evaluation alongside CRTs. For example, it does not include questions on whether the study accounted for missing or censored data (Manca and Palmer, 2005, Willan et al., 2005). Indeed the cluster-specific criteria developed are intended to complement generic checklists and guidelines for statistical methods in CEA. Secondly, although a careful search strategy was undertaken to try and capture all published studies that used CRTs, if the article did not include appropriate index terms that enabled us to identify a CEA that used data from CRTs, a relevant study could have been omitted. However, to minimise this problem the search strategy did include many alternative search terms for CEA and CRT. Thirdly, when the reviewers applied the checklists they were not

blinded to the paper's title and authorship, and so the possibility for investigator bias cannot be ruled out. Fourthly, it is plausible that studies may have used appropriate methods but did not report them. Distinguishing between inappropriate methods and inadequate reporting of appropriate methods is a general challenge facing researchers who conduct critical appraisals (Puffer et al., 2005, Philips et al., 2006). Our findings suggest that there is room for improvement in both the methods used and the reporting of those methods in CEA that use CRTs.

In conclusion, economic evaluations that use CRTs frequently ignore clustering in the data or correlation between costs and outcomes. Statistical methods that address these issues are available and their use should be encouraged to help these studies provide a sound basis for policy-making. Methodological guidelines for the evaluation of public health interventions (NICE, 2009) could incorporate the additional criteria developed. Our proposed checklist can help raise awareness of poor research practice, and provide a starting point for improving the quality of economic evaluations that use CRTs. The checklist can be updated to recognise future methodological developments.

**Acknowledgements**

We thank Edmond Ng, Simon Thompson, James Carpenter, Andrew Hutchings, Zia Sadique, Catherine Pitt and Noemi Kreif for their help and comments. We thank Jane Morrell (PI) and Simon Dixon for permission to use, and for providing access to, the PoNDER data. We also thank participants at the Health Economists' Study Group meeting, at the London School of Economics, January 2010, where this paper was presented. In particular, we thank Helen Dakin for her thoughtful discussion.

**References**


Al, M. J., Van Hout, B. A., Michel, B. C. & Rutten, F. F. 1998. Sample size calculation in economic evaluations. *Health Econ,* 7, 327-35.

Austin, P. C. 2007. A comparison of the statistical power of different methods for the analysis of cluster randomization trials with binary outcomes. *Stat Med,* 26, 3550-65.

Bachmann, M. O., Fairall, L., Clark, A. & Mugford, M. 2007. Methods for analyzing cost effectiveness data from cluster randomized trials. *Cost Eff Resour Alloc,* 5, 12.

Bland, J. M. 2009. The tyranny of power: is there a better way to calculate sample size? *BMJ,* 339, b3985.

Briggs, A. 2000. Economic evaluation and clinical trials: size matters. *BMJ,* 321, 1362-3.

Briggs, A. & Gray, A. 1998. The distribution of health care costs and their statistical analysis for economic evaluation. *J Health Serv Res Policy,* 3, 233-45.

Briggs, A., Nixon, R., Dixon, S. & Thompson, S. 2005. Parametric modelling of cost data: some simulation evidence. *Health Econ,* 14, 421-8.

Briggs, A. H. 1999. A Bayesian approach to stochastic cost-effectiveness analysis. *Health Econ,* 8, 257-61.

Briggs, A. H. & Gray, A. M. 1999. Handling uncertainty when performing economic evaluation of healthcare interventions. *Health Technol Assess,* 3, 1-134.

Briggs, A. H., Mooney, C. Z. & Wonderling, D. E. 1999. Constructing confidence intervals for cost-effectiveness ratios: an evaluation of parametric and non-parametric techniques using Monte Carlo simulation. *Stat Med,* 18, 3245-62.

Briggs, A. H. & O'brien, B. J. 2001. The death of cost-minimization analysis? *Health Economics,* 10, 179-84.

Cadth 2006. Guidelines for the Economic Evaluation of Health Tecnologies: Canada. 3rd Ed. *Canadian Agency for Drugs and Technologies in Health.* , Ottawa, Canada.

Campbell, M. J., Donner, A. & Klar, N. 2007. Developments in cluster randomized trials and Statistics in Medicine. *Stat Med,* 26, 2-19.

Campbell, M. K., Fayers, P. M. & Grimshaw, J. M. 2005. Determinants of the intracluster correlation coefficient in cluster randomized trials: the case of implementation research. *Clinical Trials,* 2, 99-107.

Davison, A. C. & Hinkley, D. V. 1997. *Bootstrap methods and their application,* Cambridge, UK, Cambridge University Press.

Donner, A. 1998. Some aspects of the design and analysis of cluster randomization trials. *Applied Statistics,* 47, 95-113.

Donner, A. & Klar, N. 1994. Methods for comparing event rates in intervention studies when the unit of allocation is a cluster. *Am J Epidemiol,* 140, 279-89; discussion 300-1.

Donner, A. & Klar, N. 2000. *Design and analysis of cluster randomization trials in health research,* London, UK, Hodder Arnold Publishers.

Donner, A. & Klar, N. 2004. Pitfalls of and controversies in cluster randomization trials. *Am J Public Health,* 94, 416-22.

Drummond, M. & O'brien, B. 1993. Clinical importance, statistical significance and the assessment of economic and quality-of-life outcomes. *Health Econ,* 2, 205-12.

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

Efron, B. & Tibshirani, R. 1993. *An introduction to Bootstrap,* New York, US, Chapman and Hall.

Eldridge, S. M., Ashby, D., Feder, G. S., Rudnicka, A. R. & Ukoumunne, O. C. 2004. Lessons for cluster randomized trials in the twenty-first century: a systematic review of trials in primary care. *Clin Trials,* 1, 80-90.

Evers, S., Goossens, M., De Vet, H., Van Tulder, M. & Ament, A. 2005. Criteria list for assessment of methodological quality of economic evaluations: Consensus on Health Economic Criteria. *Int J Technol Assess Health Care,* 21, 240-5.

Feng, Z., Diehr, P., Peterson, A. & Mclerran, D. 2001. Selected statistical issues in group randomized trials. *Annu Rev Public Health,* 22, 167-87.

Feng, Z. D., Mclerran, D. & Grizzle, J. 1996. A comparison of statistical methods for clustered data analysis with Gaussian error. *Statistics in Medicine,* 15, 1793-1806.

Flynn, T. N. & Peters, T. J. 2004. Use of the bootstrap in analysing cost data from cluster randomised trials: some simulation results. *Bmc Health Services Research,* 4, 33-43.

Flynn, T. N. & Peters, T. J. 2005. Cluster randomized trials: Another problem for cost-effectiveness ratios. *International Journal of Technology Assessment in Health Care,* 21, 403-409.

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

Gold, M. R. 1996. *Cost-effectiveness in health and medicine,* New York, Oxford University Press.

Goldstein, H. 2003. *Multilevel Statistical Models,* Oxford, UK, 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,* (in press). DOI:10.1177/0272989X11407341.

Graves, N., Walker, D., Raine, R., Hutchings, A. & Roberts, J. A. 2002. Cost data for individual patients included in clinical studies: no amount of statistical analysis can compensate for inadequate costing methods. *Health Econ,* 11, 735-9.

Grieve, R., Nixon, R. & Thompson, S. G. 2010. Bayesian hierarchical models for cost-effectiveness analyses that use data from cluster randomized trials. *Med Decis Making,* 30, 163-75.

Grieve, R., Nixon, R., Thompson, S. G. & Cairns, J. 2007. Multilevel models for estimating incremental net benefits in multinational studies. *Health Econ,* 16, 815-26.

Grieve, R., Nixon, R., Thompson, S. G. & Normand, C. 2005. Using multilevel models for assessing the variability of multinational resource use and cost data. *Health Econ,* 14, 185-96.

Hardin, J. W. & Hilbe, J. M. 2003. *Generalized Estimating Equations,* Boca Raton, Florida. US, Chapman & Hall/CRC.

Hayes, R. & Moulton, L. 2009. *Cluster randomised trials,* Boca Raton - Florida, US, CRC Press, Taylor & Francis Group.

Hayes, R. J. & Bennett, S. 1999. Simple sample size calculation for cluster-randomized trials. *Int J Epidemiol,* 28, 319-26.

Hjelmgren, J., Berggren, F. & Andersson, F. 2001. Health economic guidelines similarities, differences and some implications. *Value Health,* 4, 225-50.

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 Econ,* 11, 415-30.

Hoffman, E., Sen, P. & Weinberg, C. 2001. Within-cluster resampling. *Biometrika,* 88, 1121-1134.

Hutchings, A., Durand, M. A., Grieve, R., Harrison, D., Rowan, K., Green, J., Cairns, J. & Black, N. 2009. Evaluation of modernisation of adult critical care services in England: time series and cost effectiveness analysis. *BMJ,* 339, b4353.

Iqwig 2009. Methods for assessment of the relation of Benefits to Costs in the German Statutory Health Care System. *Institute for Quality and Efficiency in Health Care.,* Cologne, Germany.

Kim, Y. J. 2010. Regression analysis of clustered interval-censored data with informative cluster size. *Stat Med.*

Klar, N. & Donner, A. 2001. Current and future challenges in the design and analysis of cluster randomization trials. *Stat Med,* 20, 3729-40.

Lee, A. Y. & Nelder, J. 2004. Conditional and Marginal Models: Another View. *Statistical Science,* 19, 219-238.

Leyland, A. & Goldstein, H. 2001. *Multilevel Modelling of Health Statistics,* Chichester, UK, John Wiley & Sons, Ltd.

Liang, K. Y. & Zeger, S. 1986. Longitudinal data analysis using generalized linear models. *Biometrika,* 73, 13-22.

Lipsitz, S., Fitzmaurice, G., Ibrahim, J., Sinha, D., Parzen, M. & Lipshultz, S. 2009. Joint generalized estimating equations for multivariate longitudinal binary outcomes with missing data: an application to acquired immune defficiency syndrome data. *Journal of Royal Statistical Society, Series A,* 172, 3-20.

Manca, A., Lambert, P. C., Sculpher, M. & Rice, N. 2007. Cost-effectiveness analysis using data from multinational trials: the use of bivariate hierarchical modeling. *Med Decis Making,* 27, 471-90.

Manca, A. & Palmer, S. 2005. Handling missing data in patient-level cost-effectiveness analysis alongside randomised clinical trials. *Appl Health Econ Health Policy,* 4, 65-75.

Manning, W. G. 1998. The logged dependent variable, heteroscedasticity, and the retransformation problem. *J Health Econ,* 17, 283-95.

Manning, W. G., Basu, A. & Mullahy, J. 2005. Generalized modeling approaches to risk adjustment of skewed outcomes data. *J Health Econ,* 24, 465-88.

Mihaylova, B., Briggs, A., O'hagan, A. & Thompson, S. G. 2010. Review of statistical methods for analysing healthcare resources and costs. *Health Econ,* DOI: 10.1002/hec.1653.

Moher, D., Liberati, A., Tetzlaff, J. & Altman, D. G. 2009a. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *BMJ,* 339, b2535.

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G. & Prisma Group 2009b. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *British Medical Journal,* 339, b2700.

Morrell, C. J., Warner, R., Slade, P., Dixon, S., Walters, S., Paley, G. & Brugha, T. 2009. Psychological interventions for postnatal depression: cluster randomised trial and economic evaluation. The PoNDER trial. *Health Technol Assess,* 13, iii-iv, xi-xiii, 1-153.

Murray, D. M., Varnell, S. P. & Blitstein, J. L. 2004. Design and analysis of group-randomized trials: a review of recent methodological developments. *Am J Public Health,* 94, 423-32.

Nice 2008. Methods for Technology Appraisal. *National Institute for Health and Clinical Excellence,* London, UK.

Nice 2009. Methods for development of NICE public health guidance. *National Institute for Health and Clinical Excellence.,* London, UK.

Nixon, R. M. & Thompson, S. G. 2004. Parametric modelling of cost data in medical studies. *Stat Med,* 23, 1311-31.

Nixon, R. M. & Thompson, S. G. 2005. Methods for incorporating covariate adjustment, subgroup analysis and between-centre differences into cost-effectiveness evaluations. *Health Econ,* 14, 1217-29.

Nixon, R. M., Wonderling, D. & Grieve, R. D. 2010. Non-parametric methods for cost-effectiveness analysis: the central limit theorem and the bootstrap compared. *Health Econ,* 19, 316-33.

O'hagan, A. & Stevens, J. W. 2003. Assessing and comparing costs: how robust are the bootstrap and methods based on asymptotic normality? *Health Economics,* 12, 33-49.

O'hagan, A., Stevens, J. W. & Montmartin, J. 2001. Bayesian cost-effectiveness analysis from clinical trial data. *Stat Med,* 20, 733-53.

Ofman, J. J., Sullivan, S. D., Neumann, P. J., Chiou, C. F., Henning, J. M., Wade, S. W. & Hay, J. W. 2003. Examining the value and quality of health economic analyses: implications of utilizing the QHES. *J Manag Care Pharm,* 9, 53-61.

Omar, R. Z. & Thompson, S. G. 2000. Analysis of a cluster randomized trial with binary outcome data using a multi-level model. *Stat Med,* 19, 2675-88.

Panageas, K. S., Schrag, D., Russell Localio, A., Venkatraman, E. S. & Begg, C. B. 2007. Properties of analysis methods that account for clustering in volume-outcome studies when the primary predictor is cluster size. *Stat Med,* 26, 2017-35.

Pbca 2008. Guidelines for preparing submissions to the Pharmaceutical Benefits Advisory Committee. *Australian Government - Department of Health and Ageing.,* Camberra, Australia.

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.

Philips, Z., Ginnelly, L., Sculpher, M., Claxton, K., Golder, S., Riemsma, R., Woolacoot, N. & Glanville, J. 2004. Review of guidelines for good practice in decision-analytic modelling in health technology assessment. *Health Technol Assess,* 8, iii-iv, ix-xi, 1-158.

Puffer, S., Torgerson, D. & Watson, J. 2003. Evidence for risk of bias in cluster randomised trials: review of recent trials published in three general medical journals. *BMJ,* 327, 785-9.

Puffer, S., Torgerson, D. J. & Watson, J. 2005. Cluster randomized controlled trials. *J Eval Clin Pract,* 11, 479-83.

Ramsey, S., Willke, R., Briggs, A., Brown, R., Buxton, M., Chawla, A., Cook, J., Glick, H., Liljas, B., Petitti, D. & Reed, S. 2005a. Good research practices for cost-effectiveness analysis alongside clinical trials: the ISPOR RCT-CEA Task Force report. *Value Health,* 8, 521-33.

Ramsey, S., Willke, R., Briggs, A., Brown, R., Buxton, M., Chawla, A., Cook, J., Glick, H., Liljas, B., Petitti, D. & Reed, S. 2005b. Good research practices for cost-effectiveness analysis alongside clinical trials: The ISPOR RCT-CEA task force report. *Value in Health,* 8, 521-533.

Sculpher, M. J., Pang, F. S., Manca, A., Drummond, M. F., Golder, S., Urdahl, H., Davies, L. M. & Eastwood, A. 2004. Generalisability in economic evaluation studies in healthcare: a review and case studies. *Health Technol Assess,* 8, iii-iv, 1-192.

Spiegelhalter, D. J. 2001. Bayesian methods for cluster randomized trials with continuous responses. *Stat Med,* 20, 435-52.

Thompson, S. G. & Nixon, R. M. 2005. How sensitive are cost-effectiveness analyses to choice of parametric distributions? *Med Decis Making,* 25, 416-23.

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.

Turner, R. M., Omar, R. Z. & Thompson, S. G. 2001. Bayesian methods of analysis for cluster randomized trials with binary outcome data. *Stat Med,* 20, 453-72.

Turner, R. M., Omar, R. Z. & Thompson, S. G. 2006. Modelling multivariate outcomes in hierarchical data, with application to cluster randomised trials. *Biom J,* 48, 333-45.

Turner, R. M., White, I. R. & Croudace, T. 2007. Analysis of cluster randomized cross-over trial data: a comparison of methods. *Stat Med,* 26, 274-89.

Ukoumunne, O. C., Gulliford, M. C., Chinn, S., Sterne, J. A. & Burney, P. G. 1999. Methods for evaluating area-wide and organisation-based interventions in health and health care: a systematic review. *Health Technol Assess,* 3, iii-92.

Ukoumunne, O. C. & Thompson, S. G. 2001. Analysis of cluster randomized trials with repeated cross-sectional binary measurements. *Stat Med,* 20, 417-33.

Willan, A. 2006. Statistical Analysis of cost-effectiveness data from randomised clinical trials. *Expert Revision Pharmacoeconomics Outcomes Research,* 6, 337-346.

Willan, A. & Briggs, A. 2006. *Statistical Analysis of cost-effectiveness data,* Chichester, UK, 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 Econ,* 13, 461-75.

Willan, A. R., Lin, D. Y. & Manca, A. 2005. Regression methods for cost-effectiveness analysis with censored data. *Stat Med,* 24, 131-45.

Williamson, J. M., Datta, S. & Satten, G. A. 2003. Marginal analyses of clustered data when cluster size is informative. *Biometrics,* 59, 36-42.

## Appendix 3.1: Guidance on the methods considered appropriate for a paper to meet each criterion of the proposed checklist

The aim of this guidance is to help the reviewer to judge whether or not a paper meets each criterion of the cluster-specific checklist. Although the guidance aims to provide a broad set of examples of how each criterion can be met, it does not aim to provide an exhaustive list of potential methods. Moreover, this guidance, like the checklist, will need to be updated in response to methodological developments.

**Question 1 – Was the cluster design recognised in the sample size calculation of:**
**a) outcomes?** **Yes ☐ No ☐**
**b) costs?** **Yes ☐ No ☐**

**a)** The paper would be defined to meet these criteria if it states that the sample size calculations were inflated *a priori* using anticipated measures of the intra-cluster correlation coefficient (ICC) for each endpoint (Al et al., 1998, Briggs, 2000, Drummond and O'Brien, 1993, Williamson et al., 2003). Or if the sample size calculations have inflated the anticipated CI widths for each endpoint to allow for the potential ICCs (Bland, 2009).
**b)** The ICCs were anticipated to be zero, and adequate justification was provided, for example from a pilot study, or previous literature.

**Question 2 - Was clustering accounted for in the univariate analyses of**
**a) outcomes?** **Yes☐ No ☐**
**b) costs?** **Yes☐ No ☐**

Clustering should be accounted for in the univariate analysis of each endpoint. Methods that would be defined as recognising the clustering in the univariate (separate) analyses of incremental costs and outcomes are:

**a)** Hierarchical, multilevel or random effects models (Goldstein, 2003, Leyland and Goldstein, 2001).
**b)** Generalised estimating equations (Hardin and Hilbe, 2003).
**c)** The non-parametric bootstrap when this is conducted in two-stages (first sample clusters and then individuals within the sampled clusters) (Davison and Hinkley, 1997).
**d)** Cluster-level summary statistics (e.g. two-sample *t*-test or Wilcoxon rank sum test), unless the numbers per cluster are very small ($\leq 5$) and the clusters are highly imbalanced (Donner and Klar, 1994, Donner and Klar, 2000, Hayes and Moulton, 2009, Ukoumunne and Thompson, 2001).
**e)** Parametric tests adjusted for clustering (e.g. adjusted two sample *t*-test) when data are normally distributed (Donner and Klar, 1994, Donner and Klar, 2000, Feng et al., 2001).
**f)** For clinical outcomes, but not costs or QALYs, non-parametric tests adjusted for clustering equivalents (e.g. adjusted $\chi^2$ - test) when there are a sufficient large number of clusters (>10) (Donner and Klar, 1994, Donner and Klar, 2000, Feng et al., 2001).
**g)** Robust standard errors (sandwich estimator) as long as the clusters are not highly imbalanced.
**h)** Method does not allow for clustering but ICC is calculated and reported as zero.
**i)** Paper uses method that allows for clustering and correlation (Question 4) would also be considered to have satisfied the criterion for this question.

**Question 3 – Did the statistical analysis account for the correlation between costs and outcomes?** **Yes ☐ No ☐**

The paper is defined as having met this criterion if it is clear that the statistical method accommodated any correlation between individual costs and outcomes. The following approaches would be anticipated to meet this criterion:

**a)** Bivariate regression analysis, which assumes that costs and outcomes are drawn from a bivariate distribution. For example, costs and effects can be modelled jointly in a set of regression equations, where the individual-level correlation is introduced parametrically (Nixon and Thompson, 2005, Willan et al., 2004).
**b)** Bivariate generalised estimating equations (Lipsitz et al., 2009).
**c)** Non-parametric bootstrap, if it is clearly stated that costs and outcomes were re-sampled in pairs, which ensures that the endpoints were drawn from a joint distribution (Efron and Tibshirani, 1993). The bootstrap is not considered to allow for correlation in circumstances when it is used only to calculate CIs around the ICER (Flynn and Peters, 2005).
**d)** A univariate measure of net benefit is calculated for each individual, and then for example a regression approach is taken to estimate incremental net benefits (net benefit regression) (Hoch et al., 2002).
**e)** If costs and outcomes are analysed separately after clearly justifying that the correlation between costs and outcomes is indeed zero (for example the correlation coefficient is calculated and reported as zero).
**f)** In a decision model where information from the CRT on the correlation between costs and effects is introduced as an aggregate input parameter to the model.
**g)** If the paper used methods that meet the criterion for Question 4, it would also be considered to satisfy the criterion for this question.

**Question 4 – Did the study account for clustering and correlation in the estimation of incremental costs-effectiveness?** **Yes☐ No ☐**

To meet this criterion, statistical methods should simultaneously recognise the clustered nature of both costs and outcomes, and the correlation between the endpoints. The following methods would be anticipated to meet this criterion.

**a)** Bivariate multi-level models (MLMs) (also called random-effects models or hierarchical models), which assume that costs and outcomes are drawn from a bivariate distribution and accommodate the hierarchical nature of the data through a random effect parameter (Grieve et al., 2007, Manca et al., 2007, Nixon and Thompson, 2005).
**b)** Non-parametric bootstrap, if conducted in two stages (re-sampling clusters and then individuals within clusters) (TSB), provided it is clearly stated that costs and outcomes were re-sampled in pairs to allow for correlation between costs and outcomes (Bachmann et al., 2007, Flynn and Peters, 2005).
**c)** Bivariate Generalised Estimating Equations (GEEs) (Lipsitz et al., 2009).
**d)** Univariate MLMs, the TSB or univariate GEEs applied to a univariate measure of net benefit for each individual (Grieve et al., 2007).
**e)** Robust standard errors (sandwich estimator) for joint models such as seemingly unrelated regression.

**Question 5 – Did the statistical analysis make explicit, appropriate assumptions about the distribution of:**

**a) outcomes?** **Yes ☐ No ☐**

**b) costs?** **Yes ☐ No ☐**

The paper would be defined to meet these criteria if:

**a)** The study assumed that the costs or outcomes data are drawn from normal distributions (e.g. MLMs assuming normality) and provided adequate justification (e.g. by presenting a histogram or Quantile-Quantile plot) (Briggs and Gray, 1998).

**b)** A parametric model (e.g. GLMs, two-part models, mixture models) was used in the analysis, which assumed that individual level errors were drawn from a distribution other than the normal (e.g. GLM family), with adequate justification (for example relevant distributional plots) (Briggs et al., 2005, Thompson and Nixon, 2005).

**c)** The data were transformed in order to achieve normality with appropriate justification (e.g. a histogram was provided to justify that the data were lognormal). The data were then appropriately back transformed to give the arithmetic means on the original scale (Manning, 1998, Hoffman et al., 2001, O'Hagan and Stevens, 2003).

**d)** Non-parametric methods were used and an adequate justification provided. For example, if the study invoked the Central Limit Theorem (CLT), then this could not be considered appropriate if $n<30$ and the data were skewed (Nixon et al., 2010, O'Hagan and Stevens, 2003). Similarly, if the non parametric Bootstrap is used in the analysis, to meet the criteria the study would have to acknowledged that the method does rely on asymptotic properties, and offer justification that these were satisfied. In addition, if non parametric tests were chosen for statistical inference, then to be considered appropriate for incremental costs, then they should be based on means, rather than medians (e.g. Mann-Whitney tests) (Donner and Klar, 2000, Hayes and Moulton, 2009).

**e)** For papers that report clinical outcomes that are binary, count, or survival then appropriate (non-linear) models should be chosen (for example, use logistic, Poisson or Cox regression models) and justification provided.

**f)** If the study uses mean endpoints from the CRT as inputs to a decision model, then a probabilistic sensitivity analysis should provide an adequate justification for the distributional assumptions made. For example, reasonable justifications for assuming that mean costs have a gamma rather than normal distribution, is that the mean costs in each resample of the PSA should always take positive values (Briggs et al., 2005, Mihaylova et al., 2010).

## Appendix 3.2: Search strategy for the database MedLine (March 1, 2010)

| Set | Search |
|---|---|
| #1 | (cluster randomi$ adj2 trial*).af |
| #2 | cluster RCT*.af |
| #3 | (group- adj1 randomi$).af |
| #4 | (community- adj1 randomi$).af |
| #5 | (cent$2- adj1 randomi$).af |
| #6 | (area- adj1 randomi$).af |
| #7 | (practice- adj1 randomi$).af |
| #8 | Cost-effectiveness analy$.af |
| #9 | Cost-benefit analy$.af |
| #10 | Cost-utility analy$.af |
| #11 | Economic evaluation*.af |
| #12 | value for money.af |
| #13 | #1 or #2 or #3 or #4 or #5 or #6 or #7 |
| #14 | #8 or #9 or #10 or #11 or #12 |
| #15 | #13 and #14 |

## Appendix 3.3: List of the papers that satisfied the inclusion criteria


1. Subramanian S, Sankaranarayanan R, Bapat B, Somanathan T, Thomas G, Mathew B, et al. Cost-effectiveness of oral cancer screening: Results from a cluster randomized controlled trial in India. Bulletin of the World Health Organization. 2009;87(3):200-6.
2. Salize HJ, Merkel S, Reinhard I, Twardella D, Mann K, Brenner H. Cost-effective primary care-based strategies to improve smoking cessation. Archives of Internal Medicine. 2009;169(3):230-5.
3. Morrell CJ, Warner R, Slade P, Dixon S, Walters S, Paley G, et al. Psychological interventions for postnatal depression: Cluster randomised trial and economic evaluation. The PoNDER trial. Health Technology Assessment. 2009;13(30):1-153.
4. McKenna C, Bojke L, Manca A, Adebajo A, Dickson J, Helliwell P, et al. Shoulder acute pain in primary health care: is retraining GPs effective? The SAPPHIRE randomized trial: a cost-effectiveness analysis. Rheumatology. 2009;48(5):558-63.
5. Lewin RJ, Coulton S, Frizelle DJ, Kaye G, Cox H. A brief cognitive behavioural preimplantation and rehabilitation programme for patients receiving an implantable cardioverter-defibrillator improves physical health and reduces psychological morbidity and unplanned readmissions. Heart. 2009;95(1):63-9.
6. König HH, Born A, Heider D, Matschinger H, Heinrich S, Riedel-Heller SG, et al. Cost-effectiveness of a primary care model for anxiety disorders. British Journal of Psychiatry. 2009;195(4):308-17.
7. Hammar T, Rissanen P, Perälä ML. The cost-effectiveness of integrated home care and discharge practice for home care patients. Health Policy. 2009;92(1):10-20.
8. Grieve R, Nixon R, Thompson SG. Bayesian Hierarchical Models for Cost-Effectiveness Analyses that Use Data from Cluster Randomized Trials. Med Decis Making (in press).
9. Graves N, Barnett AG, Halton KA, Veerman JL, Winkler E, Owen N, et al. Cost-effectiveness of a telephone-delivered intervention for physical activity and diet. PLoS ONE. 2009;4(9).
10. Fairall L, Bachmann MO, Zwarenstein M, Bateman ED, Niessen LW, Lombard C, et al. Cost-effectiveness of educational outreach to primary care nurses to increase tuberculosis case detection and improve respiratory care: economic evaluation alongside a randomised trial. Trop Med Int Health. 2009.
11. Dixon S, Mason S, Knowles E, Colwell B, Wardrope J, Snooks H, et al. Is it cost effective to introduce paramedic practitioners for older people to the ambulance service? Results of a cluster randomised controlled trial. Emergency Medicine Journal. 2009;26(6):446-51.
12. Cleveringa FG, Welsing PM, van den Donk M, Gorter KJ, Niessen LW, Rutten GE, et al. Cost-effectiveness of the diabetes care protocol, a multifaceted computerized decision support diabetes management intervention that reduces cardiovascular risk. Diabetes Care. 2009;33(2):258-63.
13. Welton NJ, Ades AE, Caldwell DM, Peters TJ. Research prioritization based on expected value of partial perfect information: A case-study on interventions to increase uptake of breast cancer screening. Journal of the Royal Statistical Society Series A: Statistics in Society. 2008;171(4):807-34.
14. Turner DA, Paul S, Stone MA, Juarez-Garcia A, Squire I, Khunti K. Cost-effectiveness of a disease management programme for secondary prevention of coronary heart disease and heart failure in primary care. Heart. 2008;94(12):1601-6.

15. Temperley M, Mueller DH, Njagi JK, Akhwale W, Clarke SE, Jukes MC, et al. Costs and cost-effectiveness of delivering intermittent preventive treatment through schools in western Kenya. Malar J. 2008;7:196.
16. Oluboyede Y, Goodacre S, Wailoo A. Cost effectiveness of chest pain unit care in the NHS. BMC Health Serv Res. 2008;8:174.
17. Bellary S, O'Hare J, Raymond N, Gumber A, Mughal S, Szczepura A, et al. Enhanced diabetes care to patients of south Asian ethnic origin (the United Kingdom Asian Diabetes Study): a cluster randomised controlled trial. The Lancet. 2008;371(9626):1769-76.
18. Wright J, Bibby J, Eastham J, Harrison S, McGeorge M, Patterson C, et al. Multifaceted implementation of stroke prevention guidelines in primary care: Cluster-randomised evaluation of clinical and cost effectiveness. Quality and Safety in Health Care. 2007;16(1):51-9.
19. Wells KB, Schoenbaum M, Duan N, Miranda J, Tang L, Sherbourne C. Cost-effectiveness of quality improvement programs for patients with subthreshold depression or depressive disorder. Psychiatr Serv. 2007;58(10):1269-78.
20. Stanback J, Griffey S, Lynam P, Ruto C, Cummings S. Improving adherence to family planning guidelines in Kenya: an experiment. International journal for quality in health care. 2007;19(2):68-73.
21. Pinget C, Martin E, Wasserfallen JB, Humair JP, Cornuz J. Cost-effectiveness analysis of a European primary-care physician training in smoking cessation counseling. European Journal of Cardiovascular Prevention and Rehabilitation. 2007;14(3):451-5.
22. Low N, McCarthy A, Macleod J, Salisbury C, Campbell R, Roberts TE, et al. Epidemiological, social, diagnostic and economic evaluation of population screening for genital chlamydial infection. Health Technol Assess. 2007;11(8):1-165.
23. Jellema P, Van Der Roer N, Van Der Windt DAWM, Van Tulder MW, Van Der Horst HE, Stalman WAB, et al. Low back pain in general practice: Cost-effectiveness of a minimal psychosocial intervention versus usual care. European Spine Journal. 2007;16(11):1812-21.
24. Hurley MV, Walsh NE, Mitchell HL, Pimm TJ, Williamson E, Jones RH, et al. Economic evaluation of a rehabilitation program integrating exercise, self-management, and active coping strategies for chronic knee pain. Arthritis and Rheumatism. 2007;57(7):1220-1229.
25. Franzini L, Boom J, Nelson C. Cost-Effectiveness Analysis of a Practice-Based Immunization Education Intervention. Ambulatory Pediatrics. 2007;7(2):167-75.
26. Bachmann M O FLCAMM. Methods for analyzing cost effectiveness data from cluster randomized trials. Cost Effectiveness and Resource Allocation. 2007;5:12.
27. Wolters R, Grol R, Schermer T, Akkermans R, Hermens R, Wensing M. Improving initial management of lower urinary tract symptoms in primary care: Costs and patient outcomes. Scandinavian Journal of Urology and Nephrology. 2006;40(4):300-6.
28. Richardson G, Sculpher M, Kennedy A, Nelson E, Reeves D, Roberts C, et al. Is self-care a cost-effective use of resources: evidence from a randomized trial in inflammatory bowel disease. Journal of Health Services Research and Policy. 2006;11(4):225-230.
29. Kronborg C, Vass M, Lauridsen J, Avlund K. Cost Effectiveness of Preventive Home Visits to the Elderly: Economic Evaluation Alongside Randomized Controlled Study. European Journal of Health Economics. 2006;7(4):238-46.
30. Jarbol DE, Bech M, Kragstrup J, Havelund T, Schaffalitzky de Muckadell OB. Economic evaluation of empirical antisecretory therapy versus Helicobacter pylori test for management of dyspepsia: a randomized trial in primary care. Int J Technol Assess Health Care. 2006;22(3):362-71.

31. Dijkstra RF, Niessen LW, Braspenning JCC, Adang E, Grol RTPM. Patient-centred and professional-directed implementation strategies for diabetes guidelines: A cluster-randomized trial-based cost-effectiveness analysis. Diabetic Medicine. 2006;23(2):164-70.
32. Colvin M, Bachmann MO, Homan RK, Nsibande D, Nkwanyana NM, Connolly C, et al. Effectiveness and cost effectiveness of syndromic sexually transmitted infection packages in South African primary care: Cluster randomised trial. Sexually Transmitted Infections. 2006;82(4):290-4.
33. Burls A, Jordan R, Barton P, Olowokure B, Wake B, Albon E, et al. Vaccinating healthcare workers against influenza to protect the vulnerable: is it a good use of healthcare resources? A systematic review of the evidence and an economic evaluation. Vaccine. 2006;19: 4212-4221
34. Brown J. A Bayesian Approach to Analysing the Cost-Effectiveness of Two Primary Care Interventions Aimed at Improving Attendance for Breast Screening. Health Economics. 2006;15(5):435-45.
35. Bosmans J, De Bruijne M, Van Hout H, Van Marwijk H, Beekman A, Bouter L, et al. Cost-effectiveness of a disease management program for major depression in elderly primary care patients. Journal of General Internal Medicine. 2006;21(10):1020-6.
36. Tilley C, McIntosh E, Bahrami M, Clarkson J, Deery C, Pitts N. An economic analysis of implementing the SIGN third molar guideline: implications for the design and analysis of implementation studies. Journal of Health Services Research and Policy. 2005;3:143-149.
37. Sullivan SD, Lee TA, Blough DK, Finkelstein JA, Lozano P, Inui TS, et al. A multisite randomized trial of the effects of physician education and organizational change in chronic asthma care: Cost-effectiveness analysis of the Pediatric Asthma Care Patient Outcomes Research Team II (PAC-PORT II). Archives of Pediatrics and Adolescent Medicine. 2005;159(5):428-34.
38. Schoenbaum M, Sherbourne C, Wells K. Gender patterns in cost effectiveness of quality improvement for depression: results of a randomized, controlled trial. Journal of Affective Disorders. 2005;2-3:319-325.
39. Rost K, Pyne JM, Dickinson LM, LoSasso AT. Cost-effectiveness of enhancing primary care depression management on an ongoing basis. Annals of Family Medicine. 2005;3(1):7-14.
40. Plaza V, Çobos A, Ignacio-Garcia JM, Molina J, Bergonon S, Garcia-Alonso F, et al. Cost-effectiveness of an intervention based on the Global INitiative for Asthma (GINA) recommendations using a computerized clinical decision support system: a physicians randomized trial. Medicina Clinica. 2005;124(6):201-6.
41. Meyer G, Wegscheider K, Kersten JF, Icks A, Muhlhauser I. Increased use of hip protectors in nursing homes: Economic analysis of a cluster randomized, controlled trial. Journal of the American Geriatrics Society. 2005;53(12):2153-8.
42. Legood R, Gray AM, Mahé C, Wolstenholme J, Jayant K, Nene BM, et al. Screening for cervical cancer in India: How much will it cost? A trial based analysis of the cost per case detected. International Journal of Cancer. 2005;117(6):981-7.
43. Ginnelly L, Sculpher M, Bojke C, Roberts I, Wade A, Diguseppi C. Determining the cost effectiveness of a smoke alarm give-away program using data from a randomized controlled trial. European Journal of Public Health. 2005;15(5):448-53.
44. Feldman PH, Murtaugh CM, Pezzin LE, McDonald MV, Peng TR. Just-in-time evidence-based e-mail "reminders" in home health care: impact on patient outcomes. Health Services Research. 2005;3:865-885.
45. Borghi J, Thapa B, Osrin D, Jan S, Morrison J, Tamang S, et al. Economic assessment of a women's group intervention to improve birth outcomes in rural Nepal. Lancet. 2005;366(9500):1882-4.

46. Schoenbaum M, et al. Cost-Effectiveness of Interventions for Depressed Latinos. Journal of Mental Health Policy and Economics. 2004;7(2):69-76.
47. Poulos C, Bahl R, Whittington D, Bhan MK, Clemens JD, Acosta CJ. A cost-benefit analysis of typhoid fever immunization programmes in an Indian urban slum community. J Health Popul Nutr. 2004;22(3):311-21.
48. Munro JF, Nicholl JP, Brazier JE, Davey R, Cochrane T. Cost effectiveness of a community based exercise programme in over 65 year olds: Cluster randomised trial. Journal of Epidemiology and Community Health. 2004;58(12):1004-10.
49. Morgan K, Dixon S, Mathers N, Thompson J, Tomeny M. Psychological treatment for insomnia in the regulation of long-term hypnotic drug use. Health Technology Assessment. 2004;8(8).
50. Goodacre S, Nicholl J, Dixon S, Cross E, Angelini K, Arnold J, et al. Randomised controlled trial and economic evaluation of a chest pain observation unit compared with routine care. British Medical Journal. 2004;328(7434):254-7.
51. Elley CR, Kerse N, Arroll B, Swinburn B, Ashton T, Robinson E. Cost-effectiveness of physical activity counselling in general practice. New Zealand Medical Journal. 2004;117(1207).
52. Chirikos T N, K. CL, S. H, G. RR. Cost-effectiveness of an intervention to increase cancer screening in primary care settings. Preventive Medicine. 2004;39(2):230-238.
53. Bhatia MR, Fox-Rushby J, Mills A. Cost-effectiveness of malaria control interventions when malaria mortality is low: Insecticide-treated nets versus in-house residual spraying in India. Social Science and Medicine. 2004;59(3):525-39.
54. Johnston K, Gray A, Moher M, Yudkin P, Wright L, Mant D. Reporting the cost-effectiveness of interventions with nonsignificant effect differences: example from a trial of secondary prevention of coronary heart disease. International Journal of Technology Assessment in Health Care. 2003;3:476-489.
55. Loisel P, Lemaire J, Poitras S, Durand MJ, Champagne F, Stock S, et al. Cost-benefit and cost-effectiveness analysis of a disability prevention model for back pain management: a six year follow up study. Occupational and Environmental Medicine. 2002;12:807-815.
56. Kovacs FM, Llobera J, Abraira V, Lázaro P, Pozo F, Kleinbaum D. Effectiveness and cost-effectiveness analysis of neuroreflexotherapy for subacute and chronic low back pain in routine general practice: A cluster randomized, controlled trial. Spine. 2002;27(11):1149-59.
57. Finkelstein EA, Troped PJ, Will JC, Palombo R. Cost-effectiveness of a cardiovascular disease risk reduction program aimed at financially vulnerable women: the Massachusetts WISEWOMAN project. J Womens Health Gend Based Med. 2002;11(6):519-26.
58. Andersen MR, Hager M, Su C, Urban N. Analysis of the cost-effectiveness of mammography promotion by volunteers in rural communities. Health Education and Behavior. 2002;6:755-760.
59. Schoenbaum M, Unutzer J, Sherbourne C, Duan N, Rubenstein LV, Miranda J, et al. Cost-effectiveness of practice-initiated quality improvement for depression: results of a randomized controlled trial. Journal of the American Medical Association. 2001;286(11):1325-1330.
60. Richards SH, Bankhead C, Peters TJ, Austoker J, Hobbs FD, Brown J, et al. Cluster randomised controlled trial comparing the effectiveness and cost-effectiveness of two primary care interventions aimed at improving attendance for breast screening. Journal of Medical Screening. 2001;8(2):91-8.
61. Salkeld G, Phongsavan P, Oldenburg B, Johannesson M, Convery P, Graham-Clarke P, et al. The cost-effectiveness of a cardiovascular risk reduction program in general practice. Health Policy. 1997;41:105-119.

62. Gilson L, Mkanje R, Grosskurth H, Mosha F, Picard J, Gavyole A, et al. Cost-effectiveness of improved treatment services for sexually transmitted diseases in preventing HIV-1 infection in Mwanza Region, Tanzania. Lancet. 1997;350(9094):1805-9.

## Appendix 3.4: Additional characteristics of the reviewed studies (N=62)

| Characteristic | N (%) |
|---|---|
| **Number of clusters per arm** | |
| < 10 | 15 (24.3%) |
| ≥ 10 and < 20 | 27 (43.5%) |
| > 20 | 20 (32.2%) |
| **Cluster size** | |
| Equal | 11 (17.7%) |
| Unequal | 51 (82.3%) |
| **ICC** (in outcomes) | |
| < 0.01 | 11 (17.7%) |
| ≥ 0.01 and < 0.1 | 24 (38.7%) |
| ≥ 0.1 | 2 (3.3%) |
| Not reported | 25 (40.3%) |
| **Baseline covariates** | |
| Balanced | 11 (17.7%) |
| Imbalanced | 16 (25.8%) |
| Not reported | 35 (56.5%) |

# Chapter 4

## Assessment of the relative performance of alternative statistical methods for CEA that use CRTs in settings with balanced covariates

## 4.1 Preamble to research paper 2

The conceptual review developed criteria for identifying appropriate statistical methods for CEA that use CRTs, and found several methods which could meet these criteria: SUR and GEEs, both with robust standard errors; MLMs; and a non-parametric TSB (chapter 2). The checklist developed in the previous chapter found that most applied CEAs that use cluster trials fail to use any of these methods in practice. In addition, the review highlighted the very limited amount of evidence about which methods are most appropriate across typical circumstances faced by CEA that use CRTs. To help address these concerns, this paper compares the relative performance of these methods judged appropriate for CEA that use CRTs across a range of realistic scenarios. The focus of the paper is on settings where baseline covariates are balanced.

This study firstly considers simulations to allow the methods to be tested across a wide range of scenarios reflecting realistic situations observed in practice. The choice of scenarios investigated is made *a priori*, informed by hypotheses raised in the conceptual review (chapter 2). For example, characteristics such as the number of clusters, levels of cluster size variation and cost skewness were anticipated to influence the relative merits of the alternative methods. Similarly, the choice of parameter values is informed by the previous review of the applied literature.

The paper also compares the methods in a case-study to assess the implications of the choice of methods practice. This paper provides general insights into which methods perform best across different circumstances, and makes specific recommendations for future studies. To encourage the dissemination of appropriate methods, this paper provides software code to assist future researchers.

## 4.2 Research paper 2

**Developing appropriate methods for cost-effectiveness analysis of cluster randomized trials**

**Manuel Gomes MSc[1], Edmond SW. Ng MSc[1], Richard Grieve PhD[1], Richard Nixon PhD[2], James Carpenter PhD[3] and Simon G. Thompson DSc[4]**

[1]Department of Health Services Research and Policy, London School of Hygiene and Tropical Medicine, London, UK.
[2]Modeling and Simulation Group, Novartis Pharma AG, Basel, Switzerland.
[3]Department of Medical Statistics, London School of Hygiene and Tropical Medicine, London, UK.
[4]MRC Biostatistics Unit, Cambridge, UK.

**Status:** Published in Medical Decision Making, Oct 2011, DOI: 10.1177/0272989X11418372

**Contributions:** The research question for this paper was linked to the MRC project and identified by RG, principal investigator of the project. The candidate led on the design of the Monte Carlo simulations while visiting the Modelling and Simulation Group at Novartis Pharma (Switzerland), and guided by RN, researcher at Novartis and collaborator in the project. EN, a statistician working in the project, developed the bivariate GEEs with robust standard errors in collaboration with JC. EN helped the candidate write code for implementing the TSB and running the simulations on the LSHTM computing cluster. The candidate conducted the analysis of the case-study, and interpreted all the results in consultation with RG. The candidate wrote the first draft of the manuscript. He managed each round of comments and suggestions from co-authors in collaboration with RG. JC and ST, collaborators in the project, contributed to the analysis and interpretation of the results. All authors read and approved the final draft prior to journal submission and inclusion in the dissertation.

The candidate_ The supervisor_

## Abstract

**Aim:** Cost-effectiveness analysis (CEA) may use data from cluster randomized trials (CRTs) where the unit of randomization is the cluster not the individual. However, most studies use analytical methods that ignore clustering. This paper compares alternative statistical methods for accommodating clustering in CEA that use CRTs.

**Methods**: Our simulation study compared the performance of statistical methods for CEA that use CRTs with two treatment arms. The study considered a method that ignored clustering: seemingly unrelated regression (SUR) without a robust standard error (SE), and four methods that recognized clustering: SUR and generalized estimating equations (GEE) both with robust SE, a 'two-stage' non-parametric bootstrap (TSB) with shrinkage correction, and a multilevel models (MLMs). The base-case assumed CRTs with moderate numbers of balanced clusters (20 per arm) and normally distributed costs. Other scenarios included CRTs with few clusters, imbalanced cluster sizes and skewed costs. Performance was reported as bias, root mean squared error (rMSE) and confidence interval (CI) coverage for estimating incremental net benefits (INB). We also compared the methods in a case-study.

**Results**: Each method reported low levels of bias. Without the robust SE, SUR gave poor CI coverage (base-case 0.89 vs. nominal level 0.95). The MLMs and TSB performed well in each scenario (CI coverage 0.92 to 0.95). With few clusters, the GEEs and SUR (with robust SE) had coverage below 0.90. In the case-study, the mean INB were similar across all methods but ignoring clustering underestimated statistical uncertainty and the value of further research.

**Conclusions**: MLMs and the TSB are appropriate analytical methods for CEA that use CRTs with the characteristics described. SUR and GEEs are not recommended for studies with few clusters.

## Introduction

Cost-effectiveness analysis (CEA) of group-based interventions often use data from cluster randomized trials (CRTs) (Gomes et al., 2011). A cluster design may be preferred in evaluations of interventions which operate at a group level (for example alternative incentives for health providers) or where there is a high risk of 'contamination' amongst the individuals within clusters (for example vaccination programs) (Donner and Klar, 2000, Hayes and Moulton, 2009). Agencies such as National Institute for Health and Clinical Excellence may use these CEAs especially when recommending which public health interventions should be provided (Williamson et al., 2003). For these studies to provide a sound basis for decision-making, appropriate statistical methods need to be developed and used (Eckermann and Willan, 2007, Williamson et al., 2007). CEA based on randomized controlled trials (RCTs) where individual patients are randomized, have well-established methods (Glick et al., 2007, Gold, 1996, Willan and Briggs, 2006). However, statistical methods for CEA that use CRTs have received limited attention (Willan, 2006). A review found that less than 10% of published CEAs that use CRTs adopted appropriate statistical methods (Gomes et al., 2011). A distinct feature of CRTs is that the unit of randomization is the cluster (for example the hospital), not the patient. Each patient within a cluster is randomized to receive the same treatment, and so the form of clustering differs from multicentre RCTs, where patients within a centre are randomized to different treatments. In CRTs, individuals within a cluster are likely to be somewhat similar in their characteristics and the care they receive, and therefore, individual outcomes or costs within the same cluster tend to be more homogeneous than those in different clusters. The extent of such clustering can be summarized by the intra-cluster correlation coefficient (ICC), which reports the proportion of the overall variation that is at the cluster level. For the analysis of clinical outcomes it is recognized that ignoring clustering underestimates statistical uncertainty (Campbell et al., 2007, Donner and Klar, 2000, Hayes

and Moulton, 2009), encourages incorrect inferences (Austin, 2007, Feng et al., 1996, Nixon and Thompson, 2003, Ukoumunne and Thompson, 2001), and can also lead to bias (Middleton, 2008, Panageas et al., 2007). Appropriate methods for handling clustering in clinical outcomes are well-developed and can include multilevel models (MLMs) and generalized estimating equations (GEEs) (Omar and Thompson, 2000).

CEA that use CRTs raise additional challenges for statistical methods. Here, methods are required that not only allow for clustering but also acknowledge the correlation between individual costs and outcomes (Briggs et al., 1999, Nixon and Thompson, 2005, Willan et al., 2004) and make plausible assumptions about the distribution of costs and outcomes. Based on a conceptual review, we identified four main groups of statistical methods that may be appropriate for CEA that use CRTs: seemingly unrelated regression (SUR) (Willan et al., 2004); GEEs (Lipsitz et al., 2009); the non-parametric two-stage bootstrap (TSB) (Davison and Hinkley, 1997) and MLMs (Nixon and Thompson, 2005). Each of these methods can accommodate both clustering and correlation in a bivariate approach. We did not consider univariate net benefit regression analysis, as this method has less flexibility: for example, it does not allow for separate distributional assumptions to be made for costs (which tend to be highly skewed) as opposed to outcomes.

There is limited evidence comparing these alternative statistical methods for CEA that use CRTs. The TSB (Flynn and Peters, 2005) and MLMs (Grieve et al., 2010) have been proposed for CEA that use CRTs, but the only study (Bachmann et al., 2007) to compare these methods used data from a single CRT. A simulation study (Flynn and Peters, 2005) assessed the performance of the TSB but did not compare it to MLMs or GEEs, and assumed balanced clusters (equal numbers per cluster). It is therefore unclear which method performs best across the range of circumstances faced in CEA that use CRTs.

The aim of this paper is to assess the relative performance of alternative statistical methods for CEA of two-arm CRTs. We address this by conducting an extensive simulation study and illustrate the practical use of the methods in a case-study. In the next section we describe each analytical method, the design of the Monte Carlo simulations and the case-study. We then present the results of the simulations and case-study. The last section discusses the key findings and outlines an agenda for further research.

## Methods

### Statistical methods for CEA that use CRTs

We consider four methods for CEA that use CRT data. We use the following notation: let $c_{ij}$ and $e_{ij}$ represent the costs and outcomes for the $i$th individual in the $j$th cluster. For simplicity the models and the simulation study are described for CEA with two alternative treatments but the models extend to evaluations with more than two randomized treatments. Each method takes the common approach of assuming linear additive treatment effects for both costs and outcomes (Nixon and Thompson, 2005, O'Hagan and Stevens, 2001, Willan and Briggs, 2006, Willan et al., 2004).

### Seemingly Unrelated Regression (SUR)

SUR consists of a system of regression equations which can recognize the correlation between individual costs and outcomes (Willan, 2006, Willan and Briggs, 2006, Willan et al., 2004). SUR model (1) allows the individual-level error terms ($\varepsilon$) to be correlated through the parameter $\rho$ :

$$\begin{aligned} c_{ij} &= \beta_0^c + \beta_1^c t_j + \varepsilon_{ij}^c \\ e_{ij} &= \beta_0^e + \beta_1^e t_j + \varepsilon_{ij}^e \end{aligned} \qquad \begin{pmatrix} \varepsilon_{ij}^c \\ \varepsilon_{ij}^e \end{pmatrix} \sim BVN\left( \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_c^2 & \rho\sigma_c\sigma_e \\ & \sigma_e^2 \end{pmatrix} \right) \tag{1}$$

where $t_j$ is the treatment indicator ($t_j$=0 for control and 1 for treatment group). The parameters of interest, the incremental costs ($\beta_1^c$) and outcomes ($\beta_1^e$), can be estimated by ordinary least squares (OLS). SUR assumes the individual error terms ($\varepsilon$) have a bivariate Normal distribution (BVN), with variances $\sigma_c^2$ and $\sigma_e^2$. Conceptually SUR can be extended to accommodate clustering by including random effects (Singh and Ullah, 1974), but this cannot be readily implemented in conventional software packages. A practical way of allowing the uncertainty estimates to reflect clustering is to report robust SE by iterative feasible generalized nonlinear least squares (IFGNLS) (nlsur package, STATA 11). Estimates are identical to OLS when the same covariates are included for costs and outcomes (Willan et al., 2004)[14].

A limitation of SUR is that its implementation in most standard statistical packages assumes the errors are normally distributed, which may not be plausible in the context of CEA that use CRTs. In addition, it is unclear whether the robust SE recognize the correlation at the cluster-level, i.e. between cluster-level mean costs and mean outcomes (Zellner and Ando, 2010b, Zellner and Ando, 2010a). Finally, the asymptotic assumptions underlying the robust variance estimator may not hold in CRT with few clusters per treatment arm (Smeeth and Ng, 2002). The problem can be exacerbated by skewed outcomes (or costs) or imbalanced cluster sizes (Omar and Thompson, 2000). More details on the robust variance estimator are given in Appendix 4.1.

**Generalized Estimating Equations (GEEs)**

A similar approach for handling clustering is to use a GEE model with robust SE. In general GEEs offer a flexible extension of likelihood-based generalized linear models, and are commonly used to analyze clinical outcomes in CRTs (Donner and Klar, 2000, Hardin and Hilbe, 2003, Hayes and Moulton, 2009). While multivariate GEEs have been developed to

[14] Where different covariates are included for costs and outcomes, SUR estimation by IFGNLS can improve statistical efficiency (precision) compared to OLS.

recognize potential correlation between binary endpoints (Lipsitz et al., 2009), they are complex to implement and have not been extended to continuous endpoints. As a practical alternative we used a GEE model with independent estimating equations, stacking costs and outcomes, into a single vector but still allowing separate, independent estimates of incremental costs and outcomes. A bivariate GEE model with independent estimating equations can be written as,

$$\begin{aligned} c_{ij} &= \beta_0^c + \beta_1^c t_k + \varepsilon_{ij}^c \\ e_{ij} &= \beta_0^e + \beta_1^e t_k + \varepsilon_{ij}^e \end{aligned} \qquad \varepsilon_{ij}^c \perp \varepsilon_{ij}^e; \quad \varepsilon_{ij}^c \sim N(0, \sigma_c^2),\ \varepsilon_{ij}^e \sim N(0, \sigma_e^2) \tag{2}$$

This structure relies on a general property of population-averaged GEEs ensuring asymptotically consistent regression parameter estimates, even if the working correlation matrix is misspecified. This holds as long as the model, i.e. the relationship between the marginal mean and the linear predictor, is correct. However, if the working correlation matrix is misspecified, the parameter estimates may be less statistically efficient.

Parameter estimates can be obtained by maximum likelihood assuming that the errors have Normal distributions, and can provide the same point estimates to OLS estimation. As with SUR, we assumed that the error terms have a bivariate Normal distribution, although the model could be extended to allow for other distributions. We have used a robust estimator for the variance to allow for clustering when reporting uncertainty: see Appendix 4.1 for further details. However, the asymptotic properties required may not hold when there are few clusters (Bellamy et al., 2000, Feng et al., 1996, Ukoumunne and Thompson, 2001).

**The non-parametric two-stage bootstrap (TSB)**

Non-parametric bootstrap methods can avoid parametric assumptions and are easy to apply in simple settings (for example, RCTs) (Briggs et al., 1999). However, the conventional non-parametric bootstrap that resamples individuals has to be extended to recognize the clustering

inherent in CRTs. Davison and Hinkley (Davison and Hinkley, 1997) propose a two-stage routine for CRTs which resamples clusters as well as individuals, and this approach has been considered for CEA (Bachmann et al., 2007, Flynn and Peters, 2005, Flynn and Peters, 2004). The TSB can recognize the individual-level correlation between costs and outcomes by *bivariate resampling*, and the resampling can also stratify by treatment group (Flynn and Peters, 2005).

*TSB without shrinkage correction*

One proposed TSB algorithm requires resampling clusters, and then individuals within each resampled cluster (both with replacement) (Davison and Hinkley, 1997). The resultant datasets are used to calculate the statistics of interest, for example incremental net benefits (INB) and confidence intervals (CI). However, unless the CRT has many clusters and individuals per cluster, this routine can overestimate the variance. Resampling at the second stage is likely to double-count the within-cluster variance because the estimated cluster means from resampling at the first stage already incorporate both within and between-cluster variability (Davison and Hinkley, 1997, Flynn and Peters, 2005, Flynn and Peters, 2004).

*TSB with shrinkage correction*

Davison and Hinkley (Davison and Hinkley, 1997) recommend a 'shrinkage estimator' to correct for possible overestimation of the variance. Here before any resampling, cluster means are calculated with a shrinkage correction and individual level residuals estimated from the cluster means. Two-stage resampling (with replacement) is then performed by firstly resampling the shrunken cluster means, and secondly resampling the standardized individual level residuals across all clusters. Bootstrap data sets are compiled by combining the resampled shrunken cluster means and individual level residuals. Unlike the previous routine

where clusters and individuals are resampled from the original data, this routine resamples the shrunken means and residuals: see Appendix 4.2 for more details about the algorithms. Both bootstrap routines rely on asymptotic assumptions and it is unclear whether they are satisfied with few clusters, particularly if data are not Normal (O'Hagan and Stevens, 2003, Thompson and Nixon, 2005). Furthermore, the TSB routines described above were only proposed for balanced clusters (Davison and Hinkley, 1997, Flynn and Peters, 2005, Flynn and Peters, 2004), which may make the method inappropriate for CEA that use CRTs with imbalanced clusters (Gomes et al., 2011). Our implementation therefore extends Davison and Hinkley's original algorithms to allow for imbalanced clusters (see Appendix 4.2).

**Multilevel models (MLMs)**

MLMs can allow for the correlation between costs and outcomes, and recognize clustering (Grieve et al., 2010). Unlike SUR, MLMs can explicitly recognize clustering by including additional random terms, $u_j^c, u_j^e$ which in model (3) below represent the differences in the cluster mean costs and outcomes from the overall means in each treatment group. These random effects are assumed to follow a bivariate Normal distribution, with variances $\tau_c^2$ and $\tau_e^2$. MLMs acknowledge individual and cluster-level correlation between costs and outcomes through the parameters $\rho$ and $\psi$. The coefficients $\beta_1^c$ and $\beta_1^e$ still represent incremental costs and outcomes after allowing for clustering. Like the SUR model, this particular MLM assumes that the individual error terms ($\varepsilon$) are normally distributed but more generally, alternative distribution assumptions can be made for costs, outcomes or both.

$$
\begin{aligned}
c_{ij} &= \beta_0^c + \beta_1^c t_j + u_j^c + \varepsilon_{ij}^c \\
e_{ij} &= \beta_0^e + \beta_1^e t_j + u_j^e + \varepsilon_{ij}^e
\end{aligned}
\qquad
\begin{aligned}
\begin{pmatrix} \varepsilon_{ij}^c \\ \varepsilon_{ij}^e \end{pmatrix} &\sim BVN\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_c^2 & \rho\sigma_c\sigma_e \\ & \sigma_e^2 \end{pmatrix}\right) \\
\begin{pmatrix} u_j^c \\ u_j^e \end{pmatrix} &\sim BVN\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \tau_c^2 & \psi\tau_c\tau_e \\ & \tau_e^2 \end{pmatrix}\right)
\end{aligned}
\qquad (3)
$$

MLMs can be estimated and interpreted from a frequentist perspective, generally implemented with maximum likelihood, or with a Bayesian approach typically using Markov Chain Monte Carlo (MCMC) methods. Current software options for MCMC estimation afford a wide choice of distributional assumptions (Nixon and Thompson, 2005). A concern with either approach is that MLMs may fail to converge if the CRT has few individuals per cluster (Austin, 2010, Rodriguez and Goldman, 1995).

**Monte Carlo simulations**

*Data generating process*

The simulation study was designed to test the methods across a wide range of circumstances typically found in CEA that use CRTs. Our conceptual review suggested it was important to allow the following to differ: number of clusters per treatment arm; number of individuals per cluster; level of cluster size imbalance; ICCs; skewness in the cost distribution; and correlation between costs and outcomes at both the individual and cluster level (see rationale in Table 4.1). To consider this range of settings required a flexible data generating process. Data were constructed to reflect the specific form of clustering found in CRTs (Austin, 2007, Austin, 2010, Turner et al., 2007). The design allowed for a wide range of parameters to be varied and could accommodate different parametric distributions for costs and outcomes. As in previous simulation studies in economic evaluation, we assumed a linear additive treatment effect throughout (Flynn and Peters, 2005, Pinto et al., 2005, Willan et al., 2004). We simulated cost ($c$) and outcome data ($e$) from CRTs with $M$ clusters per arm and $n_m$ ($m$ =

1…*M*) individuals per cluster. Data were generated firstly at the cluster-level, and then at the individual-level according to model (4) below.

$$\begin{aligned} &\text{Cluster-level means:} && \phi_j^c \sim dist(\beta_0^c + \beta_1^c t_j \ , \ \tau_c) \\ & && \phi_j^e \sim dist(\beta_0^e + \beta_1^e t_j + \gamma(\phi_j^c - (\beta_0^c + \beta_1^c t_j)) \ , \ \tau_e) \\ &\text{Individual-level data:} && c_{ij} \sim dist(\phi_j^c, \ \sigma_c) \\ & && e_{ij} \sim dist(\phi_j^e + \theta(c_{ij} - \phi_j^c) \ , \ \sigma_e) \end{aligned} \qquad (4)$$

Cluster-level mean costs ($\phi_j^c$) and outcomes ($\phi_j^e$) were simulated for the *j*th cluster. These were assumed to follow a certain distribution characterized by the cluster means for the control ($\beta_0^c, \beta_0^e + \gamma(\phi_j^c - \beta_0^c)$) and treatment ($\beta_0^c + \beta_1^c, \ \beta_0^e + \beta_1^e + \gamma(\varphi_j^c - (\beta_0^c + \beta_1^c t_j))$) groups, and the corresponding cluster-level standard deviations ($\tau_c, \tau_e$). This mechanism allowed costs and outcomes to be correlated at the cluster level through the parameter $\gamma$, where $\gamma = \psi(\tau_e / \tau_c)$. Costs ($c_{ij}$) and outcomes ($e_{ij}$) for the *i*th individual were simulated from distributions centered at the previously simulated cluster-level means, and with the corresponding individual-level standard deviations ($\sigma_c, \sigma_e$). Costs and outcomes were also allowed to be correlated at the individual level through the term $\theta$, where $\theta = \rho(\sigma_e / \sigma_c)$. ICCs were set to recognize the proportion of the total variance at the cluster level, where for example for costs $ICC_c = \tau_c^2 / (\sigma_c^2 + \tau_c^2)$. The size of the clusters was assumed to follow a Gamma distribution according to a mean and a coefficient of variation ($cv_{imb}$), which is obtained by dividing the SD of cluster size by its mean; so for balanced (equal) cluster sizes $cv_{imb} = 0$.

*Definition of scenarios*

The simulation study initially considered a base-case scenario, then one-way and multi-way sensitivity analyses, and finished with a final 'most realistic' scenario. Under the base-case scenario parameter values were chosen not only to be plausible, but also to represent circumstances where each method was anticipated to perform well. This scenario provided a benchmark for the subsequent sensitivity analyses (Table 4.1). The choices of which parameters to vary in the sensitivity analyses, and which scenarios to combine in the multi-way sensitivity analyses, were informed by general insights from the methods literature. For each parameter, the range of values chosen was grounded in a systematic literature review of 62 studies (Gomes et al., 2011), previous methods papers and simulation studies (Briggs et al., 2005, Eldridge et al., 2006, Flynn and Peters, 2005, Nixon et al., 2010), and eight case studies (Cheyne et al., 2008, Davies et al., 2008, Fairall et al., 2005, Morgan et al., 2003, Morrell et al., 2009, Munro et al., 2004, Murphy et al., 2009, Oluboyede et al., 2008). In the final scenario each parameter was set to its 'most realistic' value, taking median values from the literature review and case studies. For example, costs followed a Normal distribution in the base-case, but increasingly skewed Gamma distributions in the sensitivity analyses with coefficient of variation ($cv_{cost}$) ranging from 0.25 to 3.0 (final case 0.5).

Table 4.1 lists the parameters changed across the scenarios; other parameters such as the true incremental costs and outcomes (QALYs) were held constant throughout. For example, the 'true' incremental costs, incremental QALYs and INB (assuming a threshold of £20 000 per QALY) were £500, 0.075 and £1 000, respectively.

*Implementation*

The performance of the different estimation methods was assessed according to mean (SE) bias, root mean squared error (rMSE), CI coverage, the error rate for lower and upper CI limits, and CI width (see Appendix 4.3 for definitions). We used 2000 simulations for each scenario[15]. The performance of each method in estimating incremental costs, incremental QALYs and INB was reported.

MLMs, GEEs and TSB were implemented in R (R, 2011) and SUR in STATA (Kim, 2010). The SUR was estimated by IFGNLS, without and with a robust standard error. The GEEs was estimated with a robust SE, and the TSB before and then after shrinkage correction. The MLMs were estimated by maximum likelihood across all scenarios. For selected scenarios (base-case, 3 clusters per treatment and the final case) estimation was also carried out via MCMC by calling WinBUGS from R (Berger and Weinstein, 2004). The MCMC estimation consisted of 5000 iterations, 3 parallel chains with different starting values, and assuming diffuse priors (Kass and Wasserman, 1996).

**Case-study**

To consider the potential implications of the choice of methods in practice, we compared the methods in a case-study of a CEA alongside a CRT. This approach extends the simulation study as, for example, the cost and outcome data do not follow specified distributions; this allows for a more pragmatic comparison of the methods.

[15] This was judged to be sufficient to report CI coverage with a margin of MC error of less than 1%, i.e. for true coverage of 0.95, 2000 simulations would be 95% certain to give coverage rates of 0.94 to 0.96.

**Table 4.1: Description, rationale and evidence for the parameter values allowed to vary across the different scenarios**

| Parameter | Rationale | Base case | SA* | Final case | Justification for parameter levels |
|---|---|---|---|---|---|
| **No. clusters per arm** | GEEs, SUR and TSB all rely on asymptotic assumptions | 20 | 3 to 30 | 15 | **Base-case:** 20 clusters per arm suggested for asymptotics to hold (Donner, 1998, Donner and Klar, 2000). **SA:** takes lower, upper quartiles from lit review. **Final case:** median no. clusters from lit review |
| **No. individuals per cluster** | MLMs may have convergence issues with few cases per cluster | 50 | 10 to 80 | 30 | **Base-case:** within the range of values from lit review. **SA:** the lower, upper quartiles from the lit review. **Final case:** median no. per cluster from the lit review |
| **Level of imbalance ($cv_{imb}$) of cluster size** | GEEs, SUR and TSB have not been assessed with imbalanced clusters | 0 | 0 to 1 | 0.5 | **Base-case:** Previous methods papers (Davison and Hinkley, 1997, Flynn and Peters, 2005, Flynn and Peters, 2004) **SA:** Cluster size imbalance informed by range of values reported across case studies and previous study (Eldridge et al., 2006). **Final case:** median from the case studies |
| **ICC for costs** | To assess if methods can handle high levels of clustering | 0.01 | 0 to 0.3 | 0.05 | **Base-case:** Start with low ICC as per previous methods papers (Flynn and Peters, 2005, Flynn and Peters, 2004). **SA:** range of ICCs from case studies and previous study (Campbell et al., 2005). **Final case:** median from case studies |
| **ICC for outcomes** | As above | 0.01 | 0 to 0.3 | 0.02 | **Base-case:** 30% of studies from lit review have ICCs for outcomes <=0.01 **SA:** range from lit review and previous methods studies (Eldridge et al., 2006). **Final case:** median from lit review |
| **Coefficient of variation ($cv_{cost}$) of cost distribution** | SUR, our MLMs and GEEs assume errors follow a Normal distribution | 0.2 | 0.25 to 3 | 0.5 | **Base-case:** start with Normal distribution, no skewness as per previous simulation studies (Nixon and Thompson, 2004, Thompson and Nixon, 2005). **SA:** Gamma distribution, range for $cv_{cost}$ from previous simulation studies (Briggs et al., 2005, Nixon et al., 2010). **Final case:** Gamma distribution, median $cv_{cost}$ from case studies. |
| **Individual level correlation of costs and effects** | GEEs assumes costs and outcomes are independent | 0.2 | -0.5 to +0.5 | -0.2 | **Base-case:** plausible level of individual level correlation (Flynn and Peters, 2005, Nixon et al., 2010). **SA:** based on the range from case studies. **Final case:** median from the case studies |
| **Cluster level correlation of costs and effects** | GEEs as above. SUR ignores cluster-level correlation. | 0 | -0.5 to +0.5 | 0.1 | **Base-case:** Conservative value assuming no correlation at the cluster level (Flynn and Peters, 2005). **SA** based on the range from case studies. **Final case:** median from case studies |

* **SA** - sensitivity analyses

We compare estimates of both relative cost-effectiveness and potential value of further research across the methods. The potential value of further research is the gain from resolving decision uncertainty, given the current state of knowledge. In other words, the expected value of perfect information (EVPI) is the increase in net benefits from taking the optimal decision after resolving current uncertainty (Claxton, 1999).

The case-study consists of a CRT that evaluates an educational intervention intended to improve the management of lung disease in adults attending outpatient clinics in South Africa (Fairall et al., 2005). The CRT included 40 balanced clusters (clinics) randomized to intervention or control. This re-analysis used complete data for 1851 patients. For each patient the study measured health service costs for three months consisting mainly of the costs of the educational intervention clinic, outpatient visits and drugs. EQ-5D data were recorded at three months follow-up and we calculated QALYs assuming that there was no mortality. The ICCs for costs and outcomes were both low (around 0.01). While the outcome data were approximately normally distributed, the costs were moderately skewed ($cv_{cost} = 1.6$). Hence the characteristics of this study were fairly similar to those in the base-case scenario in the simulation.

Each of the above statistical methods were used to report incremental costs, QALYs and INB, calculated at realistic levels of the ceiling ratio for the local South African context. We then used these estimates across the alternative methods to compare the EVPI per patient, as reported in other trial-based CEAs (Fenwick et al., 2008, Hoffman et al., 2001). EVPI was calculated assuming that the INB was normally distributed (Claxton, 1999).

## Results

### Simulation study

*Base-case*

In the base-case each method reported low bias and similar rMSE for the INB (Table 4.2). The method that ignored clustering, SUR without the robust standard error (SE), performed poorly with CI coverage below 0.9.

**Table 4.2: Bias, rMSE, CI coverage and width of the mean INB for the base-case (true INB=£1 000)**

| | SUR | | GEEs | TSB | | MLMs |
|---|---|---|---|---|---|---|
| | Without robust SE | With robust SE | With robust SE | Without shrinkage correction | With shrinkage correction | ML* |
| **Mean bias (SE)** | -1.999 (2.45) | -1.999 (2.45) | -1.999 (2.45) | -2.108 (2.45) | -2.041 (2.45) | -1.999 (2.45) |
| **rMSE** | 109.45 | 109.45 | 109.45 | 109.52 | 109.52 | 109.45 |
| **CI coverage** | 0.891 | 0.940 | 0.933 | 0.981 | 0.943 | 0.950 |
| **Mean CI width** | 353.6 | 423.7 | 417.7 | 539.1 | 427.5 | 440.7 |
| **Lower tail error rate** | 0.048 | 0.030 | 0.033 | 0.009 | 0.028 | 0.024 |
| **Upper tail error rate** | 0.051 | 0.029 | 0.035 | 0.011 | 0.030 | 0.026 |

* **ML**-Maximum likelihood. MLM estimated by MCMC in WinBUGS produced similar results.

The TSB without shrinkage correction reported wide 95% CIs and coverage above the nominal level, but with correction, coverage was similar to the other methods that recognized clustering. The MLMs had coverage close to the nominal level whether estimated by maximum likelihood (Table 4.2) or MCMC (CI coverage of 0.94). The relative performance across methods was similar for incremental QALYs, incremental cost and INB calculated with alternative levels of the ceiling ratio.

*One-way sensitivity analysis*

The bias was low across all scenarios; for example in the scenario with 3 clusters per treatment arm the mean (SE) biases for the estimated INB were -9.98 (6.05) for SUR, -6.93 (6.28) for the MLMs and GEEs, and -7.15 (6.29) for the TSB with shrinkage correction (true INB=£1 000). The rMSE differed across scenarios but was similar for each method. For example, with 3 clusters per arm, rMSE was about 280 for all methods.

Table 4.3 reports CI coverage for the one-way sensitivity analyses. The bootstrap without correction reported CI coverage above the nominal level for most scenarios, but the other methods generally reported good coverage, unless there were few clusters. Here CI coverage remained good for the MLMs and TSB (following correction), but the SUR and GEEs, both with robust standard errors, reported poor coverage.

**Table 4.3: CI coverage of the mean INB (nominal level is 0.95) for the one-way sensitivity analysis**

| | | SUR | GEEs | TSB | | MLMs |
|---|---|---|---|---|---|---|
| | | With robust SE | With robust SE | Without shrinkage correction | With shrinkage correction | ML |
| **Base-case** | | 0.940 | 0.933 | 0.981 | 0.943 | 0.950 |
| **Few clusters per arm** | (M=3) | 0.856 | 0.841 | 0.962 | 0.941 | 0.933 |
| **Few individuals per cluster** | ($n_m$=10) | 0.937 | 0.945 | 0.991 | 0.961 | 0.958 |
| **Highly imbalanced cluster size** | ($cv_{imb}$=1) | 0.919 | 0.916 | 0.981 | 0.960 | 0.951 |
| **High ICC for costs** | (ICCc=0.3) | 0.936 | 0.935 | 0.980 | 0.944 | 0.953 |
| **High ICC for outcomes** | (ICCe=0.3) | 0.941 | 0.941 | 0.941 | 0.943 | 0.945 |
| **Highly skewed Gamma costs** | ($cv_{cost}$=3) | 0.941 | 0.941 | 0.982 | 0.942 | 0.952 |

ML-Maximum likelihood.

With high levels of cluster size imbalance, coverage levels for these latter two methods were also low. All the methods (except the TSB without correction) performed well in scenarios with few individuals per cluster, high ICCs, and highly skewed costs (Table 4.3). CI coverage

also remained close to the nominal level with high levels of correlation at the individual or cluster level.

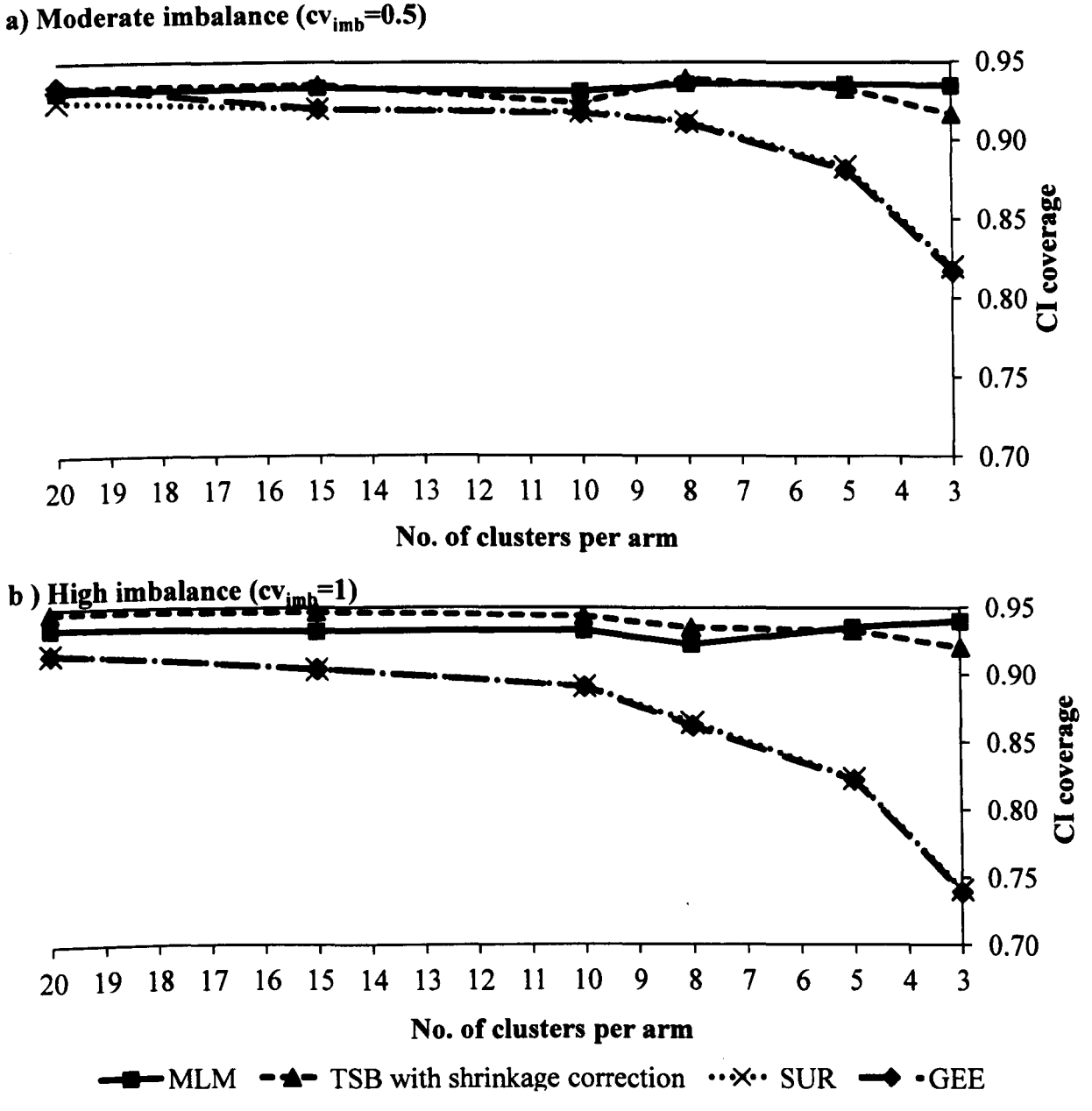
*Multi-way sensitivity analysis*

The multi-way sensitivity analyses combined variation in the number of clusters, levels of cluster size imbalance and cost skewness. Bias remained low (between -5 and 5) across all multi-way sensitivity analyses. While rMSE increased when fewer clusters were combined with high levels of imbalance, the differences between methods were small.

Figure 4.1 reports CI coverage for CRTs with decreasing number of clusters (20, 15, 10, 8, 5, and 3 clusters per treatment arm), moderate and high cluster-size imbalance ($cv_{imb}$ of 0.5 and 1) combined with highly skewed costs ($cv_{cost}$=3).

In CRTs with moderate levels of imbalance, the performance of SUR and GEEs is worse than for the MLMs and TSB if there are eight or fewer clusters per treatment arm (Figure 4.1a). With high levels of cluster size imbalance, the coverage levels for the SUR and GEEs are poor with fewer than 10 clusters per arm (Figure 4.1b). For the MLMs and TSB (with shrinkage correction), the CI coverage remains relatively good even when the study has few highly imbalanced clusters and highly skewed costs. In further scenarios that combined variation in cluster size imbalance and number of clusters with other parameters, such as different levels of individual and cluster level correlation, all methods performed well except in scenarios with few clusters, where SUR and GEEs reported poor coverage.

**Figure 4.1: CI coverage (nominal level is 0.95) for multi-way sensitivity analyses: high skewness of costs ($cv_{cost}$ = 3), decreasing number of clusters combined with a) moderate and b) high cluster size imbalance[16]**




*Final 'most realistic' scenario*

In the final scenario with all parameters set to their 'most realistic' levels (Table 4.1), bias and rMSE were again similar across methods (Table 4.4). The SUR without a robust SE, and the

[16] The CI coverage is very similar for the GEE and SUR and hence their lines show considerable overlap.

TSB without correction reported levels of CI coverage that diverged from the nominal, but the MLMs and TSB with correction both had good levels of CI coverage.

**Table 4.4: Bias, rMSE, CI coverage and width of the mean INB for the final case (true INB=£1 000)**

| | SUR | | GEEs | TSB | | MLMs |
|---|---|---|---|---|---|---|
| | Without robust SE | With robust SE | With robust SE | Without shrinkage correction | With shrinkage correction | ML |
| **Mean Bias** | 6.63 | 6.63 | 6.63 | 7.10 | 6.85 | 7.95 |
| **(SE)** | (4.40) | (4.41) | (4.40) | (4.38) | (4.38) | (4.33) |
| **rMSE** | 197 | 197 | 197 | 196 | 196 | 194 |
| **CI coverage** | 0.858 | 0.921 | 0.920 | 0.978 | 0.944 | 0.938 |
| **Mean CI width** | 583 | 726 | 724 | 924 | 778 | 754 |
| **Lower tail error rate** | 0.072 | 0.041 | 0.041 | 0.014 | 0.029 | 0.033 |
| **Upper tail error rate** | 0.120 | 0.038 | 0.039 | 0.010 | 0.028 | 0.030 |

ML-Maximum likelihood.

**Case-study**

Table 4.5 presents cost-effectiveness results from applying the alternative methods to the case-study. Each method reported that the intervention had positive incremental costs, negative incremental QALYs and negative INB. While the means were similar across methods, applying SUR without allowing for clustering led to standard errors that were substantially smaller than for the other methods. For SUR without the robust errors the EVPI (per patient) was more than 50% lower when compared to methods that accommodate clustering.

**Table 4.5: Case-study results: incremental cost, incremental QALY, INB (threshold of R100 000 per QALY)[17], and individual EVPI**

| | SUR | | GEEs | TSB | | MLMs |
|---|---|---|---|---|---|---|
| | Without Robust SE | With Robust SE | With Robust SE | Without shrinkage correction | With shrinkage correction | ML |
| **Incremental cost [ZAR]** | 14.16 | 14.16 | 14.16 | 13.73 | 15.45 | 14.78 |
| **(SE)** | (15.84) | (19.49) | (19.47) | (24.67) | (18.94) | (19.27) |
| **Incremental QALY** | -0.057 | -0.057 | -0.057 | -0.061 | -0.059 | -0.058 |
| **(SE)** | (0.020) | (0.046) | (0.046) | (0.051) | (0.045) | (0.046) |
| **INB [ZAR]** | -5762 | -5762 | -5762 | -6073 | -5926 | -5793 |
| **(SE)** | (2003) | (4651) | (4647) | (5127) | (4529) | (4583) |
| **INB [GBP]** | -824 | -824 | -824 | -869 | -848 | -829 |
| **(SE)** | (286) | (665) | (664) | (733) | (648) | (656) |
| **EVPI [GBP]** | 114 | 266 | 265 | 293 | 280 | 262 |

ML-Maximum likelihood.

## Discussion

This study compares the relative merits of alternative statistical methods for CEA that use CRTs. The simulation study finds that each method reports low bias and similar MSE across the settings considered, with the MLMs and TSB (with correction) providing good levels of CI coverage throughout. The simulation study highlights that robust methods (SUR and GEEs), which rely on asymptotic assumptions, can perform poorly for studies with few clusters. Both the simulation study and the case-study illustrate that methods that ignore clustering (for example, SUR without a robust SE) can seriously underestimate statistical uncertainty. As our empirical example illustrates, ignoring clustering can therefore understate the expected value of further research. Future studies should not attempt to justify statistical methods that ignore clustering on the basis of low estimated ICCs.

[17] One Pound (GBP) corresponded to approximately 6.99 Rands (ZAR) in terms of purchasing power parity (OECD, 2010).

This is the first paper to compare a range of statistical methods for CEA that use CRTs. Previous simulation studies (Flynn and Peters, 2005, Flynn and Peters, 2004) did not consider MLMs or GEEs, and other studies just compared the methods using a single case-study (Bachmann et al., 2007). The design of the simulation study is sufficiently general to consider the methods across common circumstances faced by CEA that use CRTs. In particular, the simulation includes scenarios with few clusters, unequal numbers per cluster (imbalance) and highly skewed costs. The choice of scenarios and parameters values are grounded in a previous review of methods and of published CEA that use CRTs (Gomes et al., 2011). These features help ensure that the simulation study provides relevant insights on the choice of analytical method for future CEAs. While for illustrative purposes we consider two-armed CRTs, the findings extend directly to CRTs with three or more randomized treatments.

The simulation study finds the TSB performs as well as the MLMs across the circumstances considered, once the shrinkage correction factor proposed by Davison and Hinkley is applied. A previous CEA used the TSB, but did not apply the shrinkage correction, and reported wide CIs compared to a MLM (26). We find that without the shrinkage correction the TSB overstates the uncertainty, but once the correction is applied the method gives good CI coverage. This finding contrasts with those of a previous simulation study (Flynn and Peters, 2005) that only considered balanced clusters but reported relatively poor performance for the TSB (even after correction). We extended the implementation to recognize cluster size imbalance and find that the method still performs well. To help improve the translation of appropriate methods into practice we are developing user-friendly software for implementing the TSB. Sample code for the TSB, GEEs and MLMs is included in Appendix 4.4.

This paper considers GEEs for the first time in this context. We develop a robust variance estimator to account for the clustering that also allows for the joint distribution of individual

costs and outcomes. A general concern for such a robust variance estimator is that it relies on asymptotic assumptions which, in these circumstances, pertain to the number of clusters per treatment arm. Our work provides specific guidance for CEA that use CRTs on the number of clusters per treatment arm required for asymptotic assumptions to hold. Our findings suggest that between 8 and 15 clusters per arm are required, depending on the other features of the study; in particular more clusters are required when the cluster sizes are highly imbalanced. This is pertinent for CEA where about 40% of such studies have fewer than 15 clusters per treatment arm, and 15% less than 8 (Gomes et al., 2011). The general literature on GEEs has reported similar sample size requirements for asymptotic assumptions to hold (Feng et al., 1996, Omar and Thompson, 2000, Ukoumunne and Thompson, 2001), and the same requirements apply to the robust estimator for SUR. The simulation study also finds that the performance of these methods does not improve in CRTs with more individuals per cluster. Grieve and others (Grieve et al., 2010) proposed a flexible Bayesian hierarchical model to tackle the main statistical issues faced by CEA that use CRTs. However, such models are complex to implement and other more accessible MLMs may be required to improve practice. Our simulation showed that a MLM estimated by maximum likelihood, assuming a bivariate Normal distribution for costs and outcomes, can perform well even when costs are highly skewed. Although in a different context, this corroborates previous findings which suggest methods assuming normality may be quite robust to skewed cost data (Briggs et al., 2005, Nixon et al., 2010, Pinto et al., 2005, Willan et al., 2004). However, it would be worth investigating whether MLMs which better accommodate skewed costs would lead to gains in precision.

This study has several limitations. While the simulation considers a wide range of circumstances and the case-study provides a useful illustration, in practice some CEAs face further complications. If for example there are baseline imbalances between the treatment

groups, or cost-effectiveness estimates are required for particular subgroups, the methods would need to consider covariates. The effects of baseline covariates, and indeed treatment group on costs and outcomes may be multiplicative, not additive (Thompson et al., 2006). Also CEA may have more complex variance structures than those considered (Grieve et al., 2010, Turner et al., 2001). These methods have not been tested under such circumstances, but MLMs may have more scope for adaptation to these broader settings than the other methods (Nixon and Thompson, 2005, Nixon et al., 2010, Omar and Thompson, 2000, Willan et al., 2004). In addition, we have not considered censored or missing data, or combining CRT data with evidence from other sources in decision models. These are all avenues for further research.

In conclusion, CEA that use CRTs may inform recommendations on the provision of area-level or public health interventions. This study finds that MLMs and TSB (with correction) are appropriate analytical methods for CEA that use CRTs across a wide range of circumstances. While methods that use a robust variance estimator such as SUR and the GEE model considered here are simple to implement, they are not recommended for CEA that use CRTs with few (less than 10) clusters in each treatment arm.

## Acknowledgments

The authors are grateful to Simon Dixon and John Cairns for helpful comments and Max Bachmann for providing access to the Outreach data. We also thank participants including the discussant Joshua Pink at the Health Economists' Study Group meeting, University of York, January 2011, where an earlier version of this paper was presented. Finally, we would like to thank the collaborators on the study- Graham Scotland, Patrick Gillespie, Allan Clark, and Mike Gillett for useful discussions.

## References


Austin, P. C. 2007. A comparison of the statistical power of different methods for the analysis of cluster randomization trials with binary outcomes. *Stat Med,* 26, 3550-65.

Austin, P. C. 2010. A Comparison of the Statistical Power of Different Methods for the Analysis of Repeated Cross-Sectional Cluster Randomization Trials with Binary Outcomes. *Int J Biostat,* 6.

Bachmann, M. O., Fairall, L., Clark, A. & Mugford, M. 2007. Methods for analyzing cost effectiveness data from cluster randomized trials. *Cost Eff Resour Alloc,* 5, 12.

Bellamy, S. L., Gibberd, R., Hancock, L., Howley, P., Kennedy, B., Klar, N., Lipsitz, S. & Ryan, L. 2000. Analysis of dichotomous outcome data for community intervention studies. *Stat Methods Med Res,* 9, 135-59.

Berger, V. 2005. *Selection bias and covariate imbalances in randomised clinical trials,* New York, Wiley.

Berger, V. W. & Weinstein, S. 2004. Ensuring the comparability of comparison groups: is randomization enough? *Control Clin Trials,* 25, 515-24.

Briggs, A., Nixon, R., Dixon, S. & Thompson, S. 2005. Parametric modelling of cost data: some simulation evidence. *Health Econ,* 14, 421-8.

Briggs, A. H., Mooney, C. Z. & Wonderling, D. E. 1999. Constructing confidence intervals for cost-effectiveness ratios: an evaluation of parametric and non-parametric techniques using Monte Carlo simulation. *Stat Med,* 18, 3245-62.

Campbell, M. J., Donner, A. & Klar, N. 2007. Developments in cluster randomized trials and Statistics in Medicine. *Stat Med,* 26, 2-19.

Campbell, M. K., Fayers, P. M. & Grimshaw, J. M. 2005. Determinants of the intracluster correlation coefficient in cluster randomized trials: the case of implementation research. *Clinical Trials,* 2, 99-107.

Carpenter, J. & Bithell, J. 2000. Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians. *Statistics in Medicine,* 19, 1141-1164.

Cheyne, H., Hundley, V., Dowding, D., Bland, J. M., Mcnamee, P., Greer, I., Styles, M., Barnett, C. A., Scotland, G. & Niven, C. 2008. Effects of algorithm for diagnosis of active labour: cluster randomised trial. *BMJ,* 337, a2396.

Claxton, K. 1999. The irrelevance of inference: a decision-making approach to the stochastic evaluation of health care technologies. *J Health Econ,* 18, 341-64.

Davies, M. J., Heller, S., Skinner, T. C., Campbell, M. J., Carey, M. E., Cradock, S., Dallosso, H. M., Daly, H., Doherty, Y., Eaton, S., Fox, C., Oliver, L., Rantell, K., Rayman, G. & Khunti, K. 2008. Effectiveness of the diabetes education and self management for ongoing and newly diagnosed (DESMOND) programme for people with newly diagnosed type 2 diabetes: cluster randomised controlled trial. *BMJ,* 336, 491-5.

Davison, A. C. & Hinkley, D. V. 1997. *Bootstrap methods and their application,* Cambridge, UK, Cambridge University Press.

Donner, A. 1998. Some aspects of the design and analysis of cluster randomization trials. *Applied Statistics,* 47, 95-113.

Donner, A. & Klar, N. 2000. *Design and analysis of cluster randomization trials in health research,* London, UK, Hodder Arnold Publishers.

Eckermann, S. & Willan, A. R. 2007. Expected value of information and decision making in HTA. *Health Econ,* 16, 195-209.

Eldridge, S. M., Ashby, D. & Kerry, S. 2006. Sample size for cluster randomized trials: effect of coefficient of variation of cluster size and analysis method. *Int J Epidemiol,* 35, 1292-300.

Fairall, L. R., Zwarenstein, M., Bateman, E. D., Bachmann, M., Lombard, C., Majara, B. P., Joubert, G., English, R. G., Bheekie, A., Van Rensburg, D., Mayers, P., Peters, A. C. & Chapman, R. D. 2005. Effect of educational outreach to nurses on tuberculosis case detection and primary care of respiratory illness: pragmatic cluster randomised controlled trial. *BMJ,* 331, 750-4.

Feng, Z. D., Mclerran, D. & Grizzle, J. 1996. A comparison of statistical methods for clustered data analysis with Gaussian error. *Statistics in Medicine,* 15, 1793-1806.

Fenwick, E., Marshall, D. A., Blackhouse, G., Vidaillet, H., Slee, A., Shemanski, L. & Levy, A. R. 2008. Assessing the impact of censoring of costs and effects on health-care decision-making: an example using the Atrial Fibrillation Follow-up Investigation of Rhythm Management (AFFIRM) study. *Value Health,* 11, 365-75.

Flynn, T. & Peters, T. 2005. Conceptual issues in the analysis of cost data within cluster randomized trials. *J Health Serv Res Policy,* 10, 97-102.

Flynn, T. N. & Peters, T. J. 2004. Use of the bootstrap in analysing cost data from cluster randomised trials: some simulation results. *Bmc Health Services Research,* 4, 33-43.

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

Gold, M. R. 1996. *Cost-effectiveness in health and medicine,* New York, 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,* (in press). DOI:10.1177/0272989X11407341.

Grieve, R., Nixon, R. & Thompson, S. G. 2010. Bayesian hierarchical models for cost-effectiveness analyses that use data from cluster randomized trials. *Med Decis Making,* 30, 163-75.

Hardin, J. W. & Hilbe, J. M. 2003. *Generalized Estimating Equations,* Boca Raton, Florida. US, Chapman & Hall/CRC.

Hayes, R. & Moulton, L. 2009. *Cluster randomised trials,* Boca Raton - Florida, US, CRC Press, Taylor & Francis Group.

Hoffman, E., Sen, P. & Weinberg, C. 2001. Within-cluster resampling. *Biometrika,* 88, 1121-1134.

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

Kass, R. & Wasserman, L. 1996. The selection of prior distributions by formal rules. *Journal of the American Statistical Association,* 91, 1343-1370.

Kim, Y. J. 2010. Regression analysis of clustered interval-censored data with informative cluster size. *Stat Med.*

Lipsitz, S., Fitzmaurice, G., Ibrahim, J., Sinha, D., Parzen, M. & Lipshultz, S. 2009. Joint generalized estimating equations for multivariate longitudinal binary outcomes with missing data: an application to acquired immune defficiency syndrome data. *Journal of Royal Statistical Society, Series A,* 172, 3-20.

Middleton, J. 2008. Bias of the regression estimator for experiments using clustered random assignment. *Statistics and Probability Letters,* 78, 2654-2659.

Morgan, K., Thompson, J., Dixon, S., Tomeny, M. & Mathers, N. 2003. Predicting longer-term outcomes following psychological treatment for hypnotic-dependent chronic insomnia. *J Psychosom Res,* 54, 21-9.

Morrell, C. J., Slade, P., Warner, R., Paley, G., Dixon, S., Walters, S. J., Brugha, T., Barkham, M., Parry, G. J. & Nicholl, J. 2009. Clinical effectiveness of health visitor training in psychologically informed approaches for depression in postnatal women: pragmatic cluster randomised trial in primary care. *BMJ,* 338, a3045.

Munro, J. F., Nicholl, J. P., Brazier, J. E., Davey, R. & Cochrane, T. 2004. Cost effectiveness of a community based exercise programme in over 65 year olds: cluster randomised trial. *J Epidemiol Community Health,* 58, 1004-10.

Murphy, A. W., Cupples, M. E., Smith, S. M., Byrne, M., Byrne, M. C. & Newell, J. 2009. Effect of tailored practice and patient care plans on secondary prevention of heart disease in general practice: cluster randomised controlled trial. *BMJ,* 339, b4220.

Ng, E. S. W. 2005. *A review of mixed-effects models in S-plus (version 6.2). Centre for Multilevel Modelling* [Online]. University of Bristol, Bristol. Available: http://www.bristol.ac.uk/cmm/learning/mmsoftware/reviewsplus.pdf [Accessed].

Nixon, R. M. & Thompson, S. G. 2003. Baseline adjustments for binary data in repeated cross-sectional cluster randomized trials. *Stat Med,* 22, 2673-92.

Nixon, R. M. & Thompson, S. G. 2004. Parametric modelling of cost data in medical studies. *Stat Med,* 23, 1311-31.

Nixon, R. M. & Thompson, S. G. 2005. Methods for incorporating covariate adjustment, subgroup analysis and between-centre differences into cost-effectiveness evaluations. *Health Econ,* 14, 1217-29.

Nixon, R. M., Wonderling, D. & Grieve, R. D. 2010. Non-parametric methods for cost-effectiveness analysis: the central limit theorem and the bootstrap compared. *Health Econ,* 19, 316-33.

O'hagan, A. & Stevens, J. W. 2001. A framework for cost-effectiveness analysis from clinical trial data. *Health Econ,* 10, 303-15.

O'hagan, A. & Stevens, J. W. 2003. Assessing and comparing costs: how robust are the bootstrap and methods based on asymptotic normality? *Health Economics,* 12, 33-49.

OECD 2010. *Country Statistical Profiles: South Africa.* [Online]. Available: http://stats.oecd.org/index.aspx?queryid=23123 [Accessed].

Oluboyede, Y., Goodacre, S. & Wailoo, A. 2008. Cost effectiveness of chest pain unit care in the NHS. *BMC Health Serv Res,* 8, 174.

Omar, R. Z. & Thompson, S. G. 2000. Analysis of a cluster randomized trial with binary outcome data using a multi-level model. *Stat Med,* 19, 2675-88.

Panageas, K. S., Schrag, D., Russell Localio, A., Venkatraman, E. S. & Begg, C. B. 2007. Properties of analysis methods that account for clustering in volume-outcome studies when the primary predictor is cluster size. *Stat Med,* 26, 2017-35.

Pinto, E. M., Willan, A. R. & O'brien, B. J. 2005. Cost-effectiveness analysis for multinational clinical trials. *Stat Med,* 24, 1965-82.

R 2011. The R project for statistical computing. http://www.r-project.org/.

Rodriguez, G. & Goldman, M. 1995. An assessment of estimation procedures for multilevel models with binary responses. *Journal of the royal Statistical Society, Series A,* 158, 73-89.

Singh, B. & Ullah, A. 1974. Estimation of seemingly unrelated regressions with random coefficients. *Journal of the American Statistical Association,* 69, 191-195.

Smeeth, L. & Ng, E. S. 2002. Intraclass correlation coefficients for cluster randomized trials in primary care: data from the MRC Trial of the Assessment and Management of Older People in the Community. *Control Clin Trials,* 23, 409-21.

Thompson, S. G. & Nixon, R. M. 2005. How sensitive are cost-effectiveness analyses to choice of parametric distributions? *Med Decis Making,* 25, 416-23.

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.

Turner, R. M., Omar, R. Z. & Thompson, S. G. 2001. Bayesian methods of analysis for cluster randomized trials with binary outcome data. *Stat Med,* 20, 453-72.

Turner, R. M., White, I. R. & Croudace, T. 2007. Analysis of cluster randomized cross-over trial data: a comparison of methods. *Stat Med,* 26, 274-89.

Ukoumunne, O. C. & Thompson, S. G. 2001. Analysis of cluster randomized trials with repeated cross-sectional binary measurements. *Stat Med,* 20, 417-33.

White, H. 1980. A Heteroskedasticity-Consistent Covariance-Matrix Estimator and a Direct Test for Heteroskedasticity. *Econometrica,* 48, 817-838.

Willan, A. 2006. Statistical Analysis of cost-effectiveness data from randomised clinical trials. *Expert Revision Pharmacoeconomics Outcomes Research,* 6, 337-346.

Willan, A. & Briggs, A. 2006. *Statistical Analysis of cost-effectiveness data,* Chichester, UK, 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 Econ,* 13, 461-75.

Williams, R. L. 2000. A note on robust variance for cluster-correlated data. *Biometrics,* 56, 645-646.

Williamson, J. M., Datta, S. & Satten, G. A. 2003. Marginal analyses of clustered data when cluster size is informative. *Biometrics,* 59, 36-42.

Williamson, J. M., Kim, H. Y. & Warner, L. 2007. Weighting condom use data to account for nonignorable cluster size. *Ann Epidemiol,* 17, 603-7.

Zellner, A. & Ando, T. 2010a. Bayesian and non-Bayesian analysis of the seemingly unrelated regression model with Student-t errors, and its application for forecasting. *International Journal of Forecasting,* 26, 413-434.

Zellner, A. & Ando, T. 2010b. A direct Monte Carlo approach for Bayesian analysis of the seemingly unrelated regression model. *Journal of Econometrics,* 159, 33-45.

## Appendix 4.1: Robust variance estimator

The robust variance estimator proposed by White (White, 1980) and Huber (Huber, 1967) allows for heteroskedasticity where data can be independent but not necessarily identically distributed (i.n.i.d) such as the stacked responses of individual costs ($\sigma_c^2$) and outcomes ($\sigma_e^2$). In other words, the data are assumed to be independent but with non-constant variances, $\sigma_i^2$, for the *i*th observation. It is also known as the sandwich estimator due to its mathematical form of **DMD** where D is the conventional covariance estimator and M in the middle is a correction term. For independent observations, it can be written as

$$V(\hat{\beta}) = D\left(\frac{n}{n-1}\sum_{i=1}^{n}\Delta_i'\Delta_i\right)D \qquad (5)$$

where $\Delta_i$ is the score statistic ($=\frac{dlnL_i}{d\beta}_{(1\times p')}$), $lnL_i$ is the log-likelihood for the *i*th observation, $\beta = (\beta_o^c, \beta_o^e, \beta_1^c, \beta_1^e))$ and $D = (X^{*'}X^*)^{-1}{}_{(p'\times p')}$ is the traditional covariance estimate. $X^*{}_{(n\times p')}$ is the covariate matrix, $p'$ the total number of parameters in the cost and outcome models, $n$ the total number of observations (= two responses × number of individuals).

For grouped data such as costs and outcomes from individuals nested within general practices in CRTs, the observation-level scores are no longer independent and a simple modification of the robust variance estimator that relies on the assumption of independent clusters is required. The modification allows for an arbitrary dependence structure of the observations within clusters (Williams, 2000) and has been described in ((Hardin and Hilbe, 2003), page 30-31), ((Kim, 2010), page 362) and ((Berger, 2005), page 260). Williams (Williams, 2000) presents a proof that the robust variance estimator is unbiased for data correlated within clusters in a general setting.

The modified robust variance estimator is given by

$V_M(\hat{\beta}) = D\left\{\frac{n_c}{n_c-1}\sum_{j=1}^{n_c}\left(\sum_{i=1}^{C_j}\Delta_{ij}\right)'\left(\sum_{i=1}^{C_j}\Delta_{ij}\right)\right\}D$ where . $\Delta_{ij}$ is the score statistic for the *i*th individual in the *j*th cluster, , $n_c$ the total number of clusters and $C_j$ the total number of individuals in cluster *j*. The rest are similarly defined as in (5). Since clusters are assumed to be independent, observation-level scores within clusters can be summed to form independent cluster-level scores and the same formula can be applied on the cluster-level scores for the robust variance estimator. Here the modification works by simply replacing the observation-level scores, $\Delta_i$, in (5) by the sum of the scores from cluster *j*, $(\sum_{i=1}^{C_j}\Delta_{ij})$, to form cluster-level scores. The *k* subscript has been omitted in the calculation for simplicity sake. The outer summation of $\Sigma_{j=1}^{n_c}$ goes from first to the last cluster irrespective to treatment assignment and the estimate is scaled by $(\frac{n_c}{n_c-1})$ for use with small samples.

## Appendix 4.2: Algorithms for the non-parametric two-stage bootstrap

Suppose we have $M_k$ clusters randomized to treatment ($k$=2) and control ($k$=1) groups, with $n_j$ individuals within each cluster $j$.

**Algorithm 1** – Routine without the shrinkage correction

1. For $i$ in 1 to $n_j$ individuals in cluster $j$.
2. For $j$ in 1 to $M_k$ clusters in treatment group $k$.
3. For $k$ in 1 to 2 treatment groups.
4. Randomly sample (with replacement) $M_k$ clusters in treatment group $k$.
5. Within each of the resampled clusters, randomly select (with replacement) $n_j$ pairs of individual costs and effects to preserve the correlation between them.
6. Compute the parameter of interest, INB = Δeffect × λ - Δcost where Δcost = $\bar{y}^c_{treatment} - \bar{y}^c_{control}$ and likewise for Δeffect.
7. Replicate steps 4 to 6 $R$ times to obtain an estimate of the bootstrap distribution of the parameter of interest.
8. Compute the bias-corrected and accelerated CIs around the mean INB.

**Algorithm 2** – Routine with the shrinkage correction

1. For $i$ in 1 to $n_j$ individuals in cluster $j$.
2. For $j$ in 1 to $M_k$ clusters in treatment $k$.
3. For $k$ in 1 to 2 treatments.
4. Calculate shrunken cluster means, $\hat{x}^c_j$ and $\hat{x}^e_j$, for cost and effect[18].
5. Calculate standardized individual-level residuals, $\hat{z}_{cost,ji}$ and $\hat{z}_{effect,ji}$, for cost and effect[19].

---

[18] $\hat{x}^c_j = c\bar{y}^c_{..} + (1-c)\bar{y}^c_j$ where c is given by $(1-c)^2 = \frac{M_k}{M_k - 1} - \frac{SS_W}{b(b-1)SS_B}$ ; $SS_w$= within-sum of squares and $SS_B$ = between-sums of squares, b = average cluster size (a formulation akin to the harmonic mean is used here; see page 412 in Smeeth and Ng (Smeeth and Ng, 2002). These are similarly calculated for effect and separately so for the two strata (treatments). Note that $j'$ is the new cluster identifier (=1 to $M_k$) which may contain repeats of the old cluster identifier, $j$. All these calculations take place before sampling.

[19] $\hat{z}_{cost,ji} = \frac{y_{cost,ji} - \bar{y}_{cost,j.}}{\sqrt{1-b^{-1}}}$, where $y_{cost,ji}$ is the observed cost for the $i$-th individual in cluster $j$. These are similarly calculated for effect and separately for the two strata (treatments). Again, all these calculations take place before sampling.

6. Randomly sample (with replacement) $M_k$ pairs of cluster means, $x^*_{cost,j'}$ and $x^*_{effect,j'}$, from the shrunken cluster means calculated in step 4.
7. Randomly sample (with replacement) $\sum_{j'=1}^{M_k} n_{j'}$ pairs of residuals, $z^*_{cost,i'}$ and $z^*_{effect,i'}$, where $i'$=1... $\sum_{j'=1}^{M_k} n_{j'}$ , from the standardized residuals calculated in step 5. Note that the hierarchical structure is ignored in this step.
8. Re-construct the sample ($y^*_{cost,j'i'}$ , $y^*_{effect,j'i'}$) by adding the shrunken cluster means from step 6 and the standardized residuals from step 7, i.e. $y^*_{cost,j'i'} = x^*_{cost,j'} + z^*_{cost,i'}$ where $i' = 1 \ldots n_{j'}$ and likewise for effects; call it a "synthetic" sample.
9. Repeat steps 4 to 8 for each stratum (treatment) and stack these 'synthetic' samples into a single bootstrap sample.
10. Compute the parameter of interest, INB, by INB = Δeffect × λ - Δcost where Δcost = $\bar{y}^*_{cost,treatment} - \bar{y}^*_{cost,control}$ and likewise for Δeffect.
11. Replicate steps 6 to 10 $R$ times to form a bootstrap distribution of INB, i.e. a distribution constructed by $R$ replicates of INB.
12. Compute the bias-corrected and accelerated CIs around the mean INB.

**Appendix 4.3: Definition of performance measures for a given parameter of interest (θ)**

| Measure | Definition | Best performance |
|---|---|---|
| **Bias** | $\frac{\sum_{k=1}^{n.sim}(\hat{\theta}_k - \theta)}{n.sim}$ | Bias = 0 |
| **SE of bias** | $SE(bias) = SD(\hat{\theta})/\sqrt{n.sim}$ where $SD(\hat{\theta}) = \sqrt{\sum_{k=1}^{n.sim}(\hat{\theta}_k - \bar{\hat{\theta}})^2 /(n.sim - 1)}$ | Lowest SE |
| **rMSE** | $\sqrt{\frac{\sum_{k=1}^{n.sim}(\hat{\theta}_k - \theta)^2}{n.sim}}$ | Lowest rMSE |
| **CI coverage** | $\frac{\sum_{k=1}^{n.sim} \mathbf{1}[LL(\hat{\theta}_k) \leq \theta \leq UL(\hat{\theta}_k)]}{n.sim}$ | Nominal level (0.95) |
| **LL error rate** | $\frac{\sum_{k=1}^{n.sim} \mathbf{1}[\theta < LL(\hat{\theta}_k)]}{n.sim}$ | Nominal LL error rate (0.025) |
| **UL error rate** | $\frac{\sum_{k=1}^{n.sim} \mathbf{1}[\theta > UL(\hat{\theta}_k)]}{n.sim}$ | Nominal UL error rate (0.025) |
| **Mean CI width** | $\frac{\sum_{k=1}^{n.sim}[UL(\hat{\theta}_k) - LL(\hat{\theta}_k)]}{n.sim}$ | Smallest Mean CI width |
| **Median CI width** | Median value of $[UL(\hat{\theta}_k) - LL(\hat{\theta}_k)]$ for k=1...n.sim | Smallest Median CI width |

**Note:** $\theta$ = true parameter; $\hat{\theta}$ = estimator for $\theta$; n.sim = total number of simulations; SE = standard error; SD = standard deviation; rMSE = root mean square error; CI = confidence interval; UL = CI upper limit; LL = CI lower limit; $\mathbf{1}[\ ]$ an indicator function for the event in brackets

## Appendix 4.4: R code for implementing GEEs, MLMs and TSB

These following sets of R code were used to obtain the results shown in Tables 4.2 to 4.5 and Figure 4.1 in our manuscript.

### 1. GEEs with independent estimating equations and robust standard errors

The following excerpt shows the derivation of the regression parameter estimates and the robust standard errors for the GEE models that we described in the GEEs section of our manuscript.

```
**********************************************************************
## Excerpts of R code for GEEs with independent estimating equations with (modified)
## robust standard errors (see page 30-31 of Hardin and Hilbe, 2003).
##
## Descriptions of objects:
## While the number of responses (R) in our manuscript is 2, the code has been    ##
generalised to the multivariate case where there can be more than 2 responses.
##
## Xlw: response-specific covariate matrix with dimension, (R*N,tot.P).
##      where R = no. of responses
##            N = total no. of individuals
##        tot.P = no. of unique regression parameters in the multivariate model
## yl: stacked response (e.g. cost and outcome) vector of dimension, (R*N,1)
## Nc: Total no. of clusters
## beta: regression parameters
## sandwich_varcov_b: robust variance estimator
## sandwich_se_b: robust standard errors
## fit multi-variate model ##
mvglm <- glm.fit(Xlw,yl,family=gaussian(),intercept=F)
beta <- coef(mvglm)
# Save fitted values
fvl <- Xlw %*% beta # fvl - fitted values held in long format
## Calc u's for each subject (record for each subject is split in R rows)
ul <- matrix(NA,R*N,tot.P)
o.res <- yl - fvl  # o.res = observed residuals
for (i in 1:(tot.P)) {
      ul[,i] <- o.res * Xlw[,i]
}
# Collapse rows for each subject into one
u <- matrix(NA,N,tot.P)
ii <- 1
for (i in 1:N) {
      u[i,] <- colSums( ul[ii:(ii+R-1),] )
      ii <- R*i+1
}
# Calculate scores within clusters and total cluster scores
J <- tapply(rep(1,N),cid,sum)
u.k <- rep(NA,tot.P)
S <- matrix(0,tot.P,tot.P)
ii <- 1
for (k in 1:Nc) {
      if (J[k]==1) {
            u.k <- u[ii:cumsum(J)[k],]
      } else {
            u.k <- colSums(u[ii:cumsum(J)[k],]) # sum of subject's scores
      }
      S <- S + u.k %*% t(u.k) # rolling sum of products of cluster scores
      ii <- cumsum(J)[k]+1
```

```
}
# Finally, put the sandwich estimator together
D <- solve(t(Xlw)%*%Xlw); dim(D)
sandwich_varcov_b <- Nc/(Nc-1)*D%*%S%*% D
sandwich_se_b <- sqrt(diag(sandwich_varcov_b))
****************************** End of GEE ********************************
```

## 2. Multilevel models (MLMs)

The following excerpt shows how to use the function, *lme*, in the package, *nlme*, for fitting a bivariate Normal multilevel model (Model 3).

```
**********************************************************************
## Excerpt of R code for fitting MLM using lme from the package, nlme.
## See section 4.3 in Ng (2005) for technical details on fitting
## multivariate Normal response models using lme in R and Splus (Ng, 2005)
##
## Descriptions of objects:
## datalong: A dataframe containing observed data matrix. Observed data
##     (including responses, covariates, cluster and individual ids) are
##     re-structured into a single matrix such that the two responses (cost
##     and outcome) are stacked to form the first (leftmost) column, yl,
##     in datalong. The rightmost 2 columns are cluster and individual
##     identifiers, cidl and indl. The columns in between are those for
##     the response-specific covariates. For the example given below,
##     the four response-specific covariates are cons.1, treat.1
##     (constant and treatment terms for cost), cons.2 and
##     treat.2 (likewise for outcome).
## cidl: Cluster identifier (long formatted)
## indl: Individual identifier (long formatted)
## beta: beta
## varcov_beta: variance-covariance matrix for beta
## se_beta: standard errors for beta
##
library(nlme)
# control setting for lme
lmc<-lmeControl(
 msTol=1e-7,tolerance=1e-6,msMaxIter=3000, msMaxEval=3000,
 opt="optim",optimMethod="SANN", # NOTE: non-default optim method used!
 msVerbose = TRUE)
mlm <- lme(yl~-1+cons.1+treat.1+cons.2+treat.2,
       random=~-1+cons.1+cons.2|cidl,
       weights=varIdent(form=~1|cons.1),
       corr=corCompSymm(form=~1|cidl/indl),
       data=datalong,
       control=lmc)
beta <- mlm$coeff$fixed
varcov_beta <- mlm$varFix
se_beta <- sqrt(diag(mlm$varFix))
****************************** End of MLM ********************************
```

## 3. Two-stage bootstrap (TSB) with shrinkage correction

The following code can be used to implement the routine described in Algorithm 2, Appendix 4.2. The function *tsbshrink* performs a two-stage bootstrap sampling routine with shrinkage

correction as described in Davison and Hinkley (Davison and Hinkley, 1997). The statistic of interest is supplied to *tsbshrink* through *user.fun* (see *cestats.r*). The options *corrbystrat* and *tssampling* are experimental and should be kept at their default values. Finally, bias-corrected and accelerated confidence intervals are calculated on the bootstrap sample of the statistic of interest using *npci1.3.r* (see below).

```
##########################################################################
## tsbshrink - function to perform two-stage bootstrap with shrinkage
## correction (Davison and Hinkley, 1997 page 100-102).
tsbshrink <-
function(cost=cost,qaly=qaly,cid=cluster,strata=treat,user.fun,unbalclus="donner",corr
bystrata=T,tssampling="varystratumsize",warning=T,seed.value){
### OPTIONS ##############################################
# 'unbalclus' - estimator to use for average cluster size
# 'corrbystrata' - shrinkage correction performed by strata
##########################################################
### sampling, shrinkage correction, standardising deviations ###
 count <- 0
 n.strata <- length(unique(strata))
 # stop if !=2 strata
 if (n.strata!=2){ stop("procedure designed for 2 strata only.") } else {}
 data<-data.frame(cost,qaly,cid,strata)
 shrunk.data <- c()
# use predetermined seed if specified; else do nothing
if (!missing(seed.value)){
 set.seed(seed.value)
} else { }
# Option for performing correction by strata or not
if (corrbystrata){
 } else {
 n.strata<-1
}
while (count<n.strata){
 count <- count+1
 if (corrbystrata){
  data1 <- data.frame(data[data$strata==unique(data$strata)[count],])
 } else {
  data1 <- data.fram(data)
 }
 clus.size <- table(data1$cid)
 # calc cluster means
 cost.x <- tapply(data1$cost,data1$cid,mean)
 qaly.x <- tapply(data1$qaly,data1$cid,mean)
 # STANDARDIZE Z: calc b for standardizing z
 a <- length(unique(data1$cid))
 if (var(clus.size)==0){
  b <- unique(clus.size)
  } else {
  if (unbalclus=="donner"){
   ifelse(warning,print("'average' clus size = Donner"),NA)
   n <- sum(clus.size)
   b <- (n-(sum(clus.size^2)/n))/(a-1)
  } else if (unbalclus=="median"){
   ifelse(warning,print("'average' clus size = median"),NA)
   b <- median(clus.size)
  } else if (unbalclus=="mean"){
   ifelse(warning,print("'average' clus size = mean"),NA)
   b <- mean(clus.size)
  } else {}
```

```
} # End of 'else'
# standardize z using cluser means (dfm = deviation from cluster mean)
cost.dfm <- data1$cost-rep(cost.x,times=clus.size)
qaly.dfm <- data1$qaly-rep(qaly.x,times=clus.size)
cost.z <- (cost.dfm)/sqrt(1-1/b)
qaly.z <- (qaly.dfm)/sqrt(1-1/b)
# SHRINKAGE: calc c for shrinking x
cost.ssw <- sum(cost.dfm^2); qaly.ssw <- sum(qaly.dfm^2)
cost.ssb <- sum((cost.x-mean(cost.x))^2); qaly.ssb <- sum((qaly.x-mean(qaly.x))^2)
cost.rhs <- a/(a-1) - cost.ssw/(b*(b-1)*cost.ssb)
qaly.rhs <- a/(a-1) - qaly.ssw/(b*(b-1)*qaly.ssb)
ifelse(cost.rhs<0, cost.c<-1, cost.c<-1-sqrt(cost.rhs))
ifelse(qaly.rhs<0, qaly.c<-1, qaly.c<-1-sqrt(qaly.rhs))
## re-calc x
cost.x <- cost.c*mean(data1$cost) + (1-cost.c)*cost.x
qaly.x <- qaly.c*mean(data1$qaly) + (1-qaly.c)*qaly.x
# TWO-STAGE SAMPLING & RE-CONSTRUCT OBS WITH SHRUNKEN MEANS AND STANDARDIZED
RESIDUALS
# gen random clus (order) id with replacement
sampled.x.cid <- sample(1:length(unique(data1$cid)),replace=T)
if (tssampling=="varystratumsize"){
 sampled.z.iid <- sample(1:length(cost.z),sum(clus.size[sampled.x.cid]),replace=T) #
chosen ind ids for varying stratum sizes
 sampled.cost <-
rep(cost.x[sampled.x.cid],times=clus.size[sampled.x.cid])+cost.z[sampled.z.iid]
 sampled.qaly <-
rep(qaly.x[sampled.x.cid],times=clus.size[sampled.x.cid])+qaly.z[sampled.z.iid]
 # bind data from multiple strata together
 shrunk.data <- as.data.frame(rbind(shrunk.data,cbind(sampled.cost,sampled.qaly,
rep(unique(data1$cid)[sampled.x.cid],times=clus.size[sampled.x.cid]),
rep(unique(data$strata)[count],times=sum(clus.size[sampled.x.cid])))))
 } else if (tssampling=="fixedstratumsize") {
 sampled.z.iid <- sample(1:length(data1$cid),replace=T) # chosen ind ids for fixed
stratum size (=original data)
 sampled.cost <- rep(cost.x[sampled.x.cid],times=clus.size)+cost.z[sampled.z.iid]
 sampled.qaly <- rep(qaly.x[sampled.x.cid],times=clus.size)+qaly.z[sampled.z.iid]
 # bind data from multiple strata together
 shrunk.data <-
as.data.frame(rbind(shrunk.data,cbind(sampled.cost,sampled.qaly,data1$cid,data1$strata
)))
 }
} # end of while
colnames(shrunk.data) <- c("cost","qaly","cid","treat")
# apply user-provided function
tsb.shrunk.inb<-user.fun(shrunk.data,warning=F)
return(tsb.shrunk.inb)
}
######################## End of tsbshrink ###################################
```

**Excerpt of "cestats.r" for calculating incremental net benefit (INB)**

The following excerpt shows the function, *calcinb*, within *cestata.r,* for calculating the parameter of interest, the INB. *calcinb* is called by *tsbshrink* above for obtaining a bootstrap distribution of INB using the two-stage bootstrap method.

```
#############################################################################
### calcinb - function to calculate INB with a single input argument (data) ###
calcinb <- function(data,lambda=20000,warning=T){
### Data checking ###
# 1. check if expected vars are all present in data
cea.names <-c("cost","qaly","treat")
count<-0
```

```
for (name in cea.names){
 count <- count+any(names(data)==name)
}
if (count!=3){ stop("Not all of the expected variable names (cost, qaly and treat)
present in data.") }
# 2. is 'data' a data.frame
stopifnot(is.data.frame(data))
# 3. Non-0/1 values in treatment variable
if (warning){
 if ((length(unique(data$treat))!=2)||(min(data$treat)!=0|max(data$treat)!=1)){
  cat("\n",
  "Warning: Either values in 'treat' are not 0/1 or number of unique values in 'treat'
is not two.","\n",
  "         Function assumes the lower of the two values in 'treat' to represent the
control ","\n",
  "         group and the higher the treatment group.","\n")
 } else {}
} else {}
### End of data checking ###
### Calc INB ###
inc.qaly <- by(data$qaly,data$treat,mean)[2]-by(data$qaly,data$treat,mean)[1]
inc.cost <- by(data$cost,data$treat,mean)[2]-by(data$cost,data$treat,mean)[1]
inb <- inc.qaly*lambda - inc.cost
return(inb)
}
############################## End of 'calcinb'################################
```

## npci1.3.r – constructing non-parametric confidence intervals

This function can be used to construct the bias-corrected and accelerated (BCa) confidence intervals, using the observed data and the bootstrap distribution of the statistic of interest.

```
#############################################################################
## npciX.X.r - Non-Parametric Confidence Intervals
## Use this to calculate the upper and lower limits of non-parametric confidence
## intervals using the percentile ("perc"), bias-corrected ("bcor" and
## bias-corrected and accelerated ("bca") methods described by Carpenter and
## Bithell (Carpenter and Bithell, 2000).
#############################################################################
### npci - Function to return lower and upper limits of BCa confidence interval ###
npci <- function(orig.data,bsample,user.fun,twosided.alpha=0.05,type,interpolate=T){
 half.alpha <- twosided.alpha/2
 ul <-
ulnpci(orig.data,bsample,user.fun,upp.alpha=half.alpha,type=type,interpolate=interpola
te)
 ll <- ulnpci(orig.data,bsample,user.fun,upp.alpha=1-
half.alpha,type=type,interpolate=interpolate)
 return(c(ll,ul))
}
### End of 'npci' ###
#############################################################################
### ulnpci - function to calc upper limit of non-parametric CI ###
ulnpci <- function(orig.data,bsample,user.fun,upp.alpha=0.025,type,interpolate=T){
### 1. Data checking ###
# 1. Missing 'type'
if (missing(type)){
 stop("Warning: 'type' of confidence interval not specified.")
} else {}
# 2. Missing 'orig.data' if type!='perc'
if (type!="perc" && missing(orig.data)){
 stop("Warning: original data not specified.")
} else {}
### End of data checking ###
############################
```

```
if (type!="perc"){
 # calc statistic of interest
 theta.obs <- user.fun(orig.data)
 ### calc b ###
 p <- sum(bsample<theta.obs)
 B <- length(bsample)
 b <- qnorm(p/B)
 ### calc a ###
 theta.jack <- c()
 for (i in 1:dim(orig.data)[1]){
  theta.jack[i] <- user.fun(orig.data[-i,]) # jackknife estimate of user-supplied
function
 }
 theta.tilda <- mean(theta.jack)
 a.top <- sum((theta.tilda-theta.jack)^3)
 a.btm <- 6*(sum((theta.tilda-theta.jack)^2)^1.5)
 if (type=="bca"){
  a <- a.top/a.btm
  } else if (type=="bcor"){
   a <- 0
  }
 ### calc q ###
 z.upp.alpha <- qnorm(upp.alpha)
 q <-(B+1)*pnorm(b-(z.upp.alpha-b)/(1+a*(z.upp.alpha-b)))
 q.tilda <- floor(q)
 q.a <- floor(q)
 q.b <- q.a+1
 ### Handling extreme values of q.tilda ###
 if (q.a==0){   # extreme low value of q.tilda
  warning("Integer part of Q_tilda=0! CI limit replaced by lowest value of bootstrap
sample.
    Interpolation not carried out.")
  return(min(bsample))
 } else {}
 if (q.a==B){   # extreme low value of q.tilda
  warning("Integer part of Q_tilda=size of bootstrap sample! CI limit replaced by
largest value of bootstrap sample.
    Interpolation not carried out.")
  return(max(bsample))
 } else {}
 ### End of handling extreme q.tilda values ###
 theta.a <- sort(bsample)[q.a]
 theta.b <- sort(bsample)[q.b]
 if (interpolate){
  q1<-qnorm(q/(B+1))
  q2<-qnorm(q.b/(B+1))
  q3<-qnorm(q.a/(B+1))
  theta.q = theta.a + ((q1-q3)/(q2-q3))*(theta.b-theta.a)
  return(theta.q)
  } else {
  return(theta.a)
 }
### percentile CI ###
} else if (type=="perc"){
 B <- length(bsample)
 theta.q <- sort(bsample)[(1-upp.alpha)*(B+1)]
 return(theta.q)
}}
### End of 'ulnpci' ###
########################## End of npci1.3.r ###################################
```

# Chapter 5

## Comparison of alternative methods for covariate adjustment in CEA that use CRTs.

## 5.1 Preamble to research paper 3

The conceptual review (chapter 2) highlighted the fact that the cluster design can encourage systematic imbalances in both individual and cluster-level baseline covariates between the treatment groups (Eldridge et al., 2008, Hahn et al., 2005, Morgan et al., 2003, Puffer et al., 2003). In those circumstances, the methods considered in research paper 2 are insufficient to allow for the potential confounding arising from the systematic covariate imbalance. No previous work has considered methods to address this issue in the context of CEA that use CRTs. This study extends research paper 2 by comparing alternative methods for CEA that use cluster trials across settings with covariate imbalance. This paper considers the following methods: SUR with robust standard errors, MLMs, and a method combining the non-parametric TSB with SUR to adjust for the potential confounding.

This paper firstly considers an empirical application (the PoNDER study) with covariate imbalance to illustrate the implications of the choice of method for covariate adjustment. The study followed general methodological guidance on covariate adjustment (Altman, 2005, Imai et al., 2008) and limited the adjustment to those covariates that were anticipated *a priori* to be important prognostic factors. The case-study also illustrates circumstances where different prognostic relationships between treatment groups are anticipated, a concern often raised more generally (Assmann et al., 2000, Gelman and Pardoe, 2007, Pocock et al., 2002).

This paper considers new simulations motivated by the case study, and grounded in the conceptual review. For example, it was judged important to allow for scenarios with different levels of baseline imbalances and levels of correlation between the covariates and the endpoints, and to allow for prognostic relationships to differ between treatment groups. As in research paper

2, methods were compared under other common settings, where differences amongst methods were anticipated, such as in CRTs with few clusters and unequal numbers per cluster.

## 5.2 Research paper 3

## Methods for covariate adjustment in cost-effectiveness analyses of cluster randomised trials

**Manuel Gomes MSc[1], Richard Grieve PhD[1], Richard Nixon PhD[2], Edmond SW. Ng MSc[1], James Carpenter PhD[3], Simon G. Thompson DSc[4]**

[1]Department of Health Services Research and Policy, London School of Hygiene & Tropical Medicine, London.
[2]Modeling and Simulation Group, Novartis Pharma AG, Basel, Switzerland.
[3]Department of Medical Statistics, London School of Hygiene & Tropical Medicine, London.
[4]Department of Public Health and Primary Care, University of Cambridge, Cambridge.

**Status:** Submitted to Health Economics (under review). The paper was presented at the 20th European Workshop on Econometrics and Health Economics, York, September 2011.

**Contributions:** The candidate led on the conception of the research question in collaboration with RG. The candidate carried out a secondary analysis of the PoNDER study, and interpreted the findings with RG. The candidate was responsible for designing the simulations, writing additional code for implementing the statistical methods and conducting the simulations. RG and RN were involved in the design of the simulations and interpretation of the results. EN, JC and ST also contributed to the analyses and interpretation of the findings. The candidate wrote the first draft of the manuscript and managed each round of comments and suggestions from co-authors with RG. All authors read and approved the final draft prior to journal submission and inclusion in the dissertation.

The candidate. The supervisor

**Abstract**

Statistical methods have been developed for cost-effectiveness analysis (CEA) of cluster randomised trials (CRTs) where baseline covariates are balanced. However, CRTs may show systematic differences in individual and cluster-level covariates between the treatment groups. This paper presents three methods to adjust for imbalances in observed covariates: seemingly unrelated regression (SUR) with a robust standard error, a 'two-stage' bootstrap (TSB) approach combined with SUR, and multilevel models (MLMs). We consider the methods in a CEA of a CRT with covariate imbalance, unequal cluster sizes and a prognostic relationship that varied by treatment group. The cost-effectiveness results differed according to the approach for covariate adjustment.

A simulation study then assessed the relative performance of methods for addressing systematic imbalance in baseline covariates. The simulations extended the case study and considered scenarios with: different levels of confounding, cluster size variation and few clusters. Performance was reported as bias, root mean squared error and confidence interval (CI) coverage of the incremental net benefit. Even with low levels of confounding, unadjusted methods were biased, but all adjusted methods were unbiased. MLMs performed well across all settings, and unlike the other methods, reported CI coverage close to nominal levels even with few clusters of unequal sizes.

## 1. Introduction

Econometric evaluation often uses observational data to estimate 'average treatment effects' (ATEs). In non-randomised studies, baseline characteristics may be correlated with both treatment choice and the endpoints of interest, i.e. the distribution of potential confounders (both observed and unobserved) can differ across treatment groups. Approaches such as regression, instrumental variables estimation, matching and inverse probability weighting have been advocated for reducing selection bias in observational studies (Basu and Rathouz, 2005, Sekhon and Grieve, 2011, Jones and Rice, 2011). In cost-effectiveness analysis (CEA), many studies use data from clinical trials where individual patients are randomised. Here, if the randomisation is properly conducted, systematic differences in baseline characteristics between the treatment groups can be avoided, and the resultant estimates will be unbiased (Imai et al., 2008, Senn, 1989). For CEA of clinical trials, regression approaches have been proposed for the purposes of improving precision or conducting pre-specified subgroup analyses, (Barber and Thompson, 2004, Briggs, 2006, Hoch et al., 2002, Manca et al., 2005, Nixon and Thompson, 2005, Willan and Briggs, 2006, Willan et al., 2004).

For CEA of interventions that operate at a group rather than an individual-level (e.g. changing incentives for providers), or where there is a high risk of contamination amongst individuals within a geographical setting (e.g. alternative strategies for containing an infectious disease), a cluster randomised trial (CRT) may be preferred. Here the unit of randomisation is the cluster, for example the primary care physician, not the patient. The CRT can be designed to try and avoid selection bias, for example by concealing treatment allocation, and also recruiting individuals before cluster randomisation.

A general concern with CRTs is that studies tend to be unblinded, with individuals recruited after treatment allocation is known (Donner, 1998, Donner and Klar, 2000, Puffer et al., 2005). Those recruiting individuals into clusters often know both the treatment allocation and patients' characteristics prior to their inclusion. CRTs with this design are prone to differences between the treatment groups in patient and cluster-level baseline characteristics that are systematic, rather than due simply to chance (Eldridge et al., 2008, Puffer et al., 2003). For example, potential participants with poor prognostic characteristics may be more likely to enter the control group once assignment is known. Hence, the CRT design can yield systematic imbalances in baseline characteristics, which if associated with endpoints, can lead to biased results (Eldridge et al., 2008, Hahn et al., 2005). An additional concern is that CRTs typically have clusters of unequal size, for example due to different recruitment rates (Carter, 2010). If cluster size is correlated with an endpoint, such as costs, for example due to (dis)economies of scale, then this can lead to biased estimates (Panageas et al., 2007). Furthermore, baseline covariates may have prognostic relationships that differ by treatment group (Gelman and Pardoe, 2007, Liu and Gustafson, 2008); this may occur if for example, the study protocol is less rigid for the control than the treatment group.

Hence, for CEA that use CRTs to provide unbiased estimates, analytical methods are required to adjust appropriately for systematic differences in observed baseline covariates. This raises the issue of which covariates to include and how best to undertake the adjustment (Austin et al., 2010). Methodological guidance emphasises that covariate adjustment should be limited to those variables anticipated to be strongly associated with the endpoints of interest (Altman, 2005, Imai et al., 2008). Consideration should also be given to non-linear terms and covariate by treatment interactions if these are anticipated to be important (Assmann et al., 2000, Gelman and Pardoe,

2007). Hence, the choice of covariates for adjustment should not simply be according to whether or not there are statistically significant baseline differences between the treatment groups (Imai et al., 2008).

In CEA that use CRTs, little attention has been given to analytical methods (Gomes et al., 2011a). A recent paper presented methods that allow for clustering and the correlation between costs and outcomes: these were seemingly unrelated regressions (SUR) and generalised estimating equations (GEEs) both with a robust variance estimator, multilevel models (MLMs) and a two-stage non-parametric bootstrap (TSB) (Gomes et al., 2011b). The study assumed that baseline covariates were balanced between the treatment groups. Indeed, the potential for selection bias seems to be generally ignored in CEA that use CRTs. Our review (Gomes et al., 2011a) found that of 62 published CEAs that use CRTs, about 60% did not report an assessment of covariate balance, and of the 27 studies reporting baseline information, only 16 adjusted for any baseline imbalances. The remaining 11 studies justified reporting unadjusted results by the lack of any statistically significant baseline differences.

The aim of this paper is to assess the relative performance of alternative methods for CEA that use CRTs when there are systematic imbalances in individual and cluster-level baseline covariates. This paper considers alternative approaches for CEA that use CRT in an empirical application and an extensive simulation study. We consider regression-based methods such as MLMs and SUR, and extend a non-parametric TSB to handle covariate adjustment. We do not consider net benefit regression because the approach lacks flexibility (Nixon and Thompson, 2005, Willan et al., 2004), nor GEEs as these performed poorly in studies with few clusters (Gomes et al., 2011b). We estimate ATEs, as these are of prime interest for policy makers (Claxton, 1999, Imbens and Wooldridge, 2009, Jones and Rice, 2011). In the next section, we

outline the methods under comparison. Section 3 presents the motivating example. Sections 4 and 5 report the design and results of the simulation study. The last section discusses the findings and suggests areas for further research.

## 2. Statistical methods for covariate adjustment in CEA that use CRTs

In CEA that use CRTs, statistical methods are required that adjust for covariate imbalances while accounting for the clustering and the correlation between costs and health outcomes. We consider three methods: SUR with robust standard errors (SE), MLMs, and an approach that combines the TSB with SUR (TSB+SUR).

We use the following notation: let $c_{ij}$ and $e_{ij}$ represent the costs and outcomes for the $i$th individual in the $j$th cluster. For simplicity the models and the simulation study are described for CEA with two alternative treatments but the models extend to evaluations with more than two randomised groups. Each method is illustrated assuming linear additive effects for treatment and covariates (Nixon and Thompson, 2005, Willan and Briggs, 2006). For simplicity, we illustrate adjustment for one individual-level ($x_{ij}$) and one cluster-level ($z_j$) covariate.

*Seemingly unrelated regressions (SUR)*

SUR consists of a system of regression equations with residuals that are allowed to be correlated (Zellner, 1962). The set of covariates can differ for each endpoint, but in Model (1) below we use the same individual ($x_{ij}$) and cluster-level ($z_j$) covariates for each endpoint.

$$c_{ij} = \beta_0^c + \beta_1^c t_j + \beta_2^c x_{ij} + \beta_3^c z_j + \varepsilon_{ij}^c$$
$$e_{ij} = \beta_0^e + \beta_1^e t_j + \beta_2^e x_{ij} + \beta_3^e z_j + \varepsilon_{ij}^e$$
$$\begin{pmatrix} \varepsilon_{ij}^c \\ \varepsilon_{ij}^e \end{pmatrix} \sim BVN\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_c^2 & \rho\sigma_c\sigma_e \\ & \sigma_e^2 \end{pmatrix}\right) \quad (1)$$

where $t_j$ is the treatment indicator ($t_j$=0 for control and 1 for treatment group). The incremental costs ($\beta_1^c$) and outcomes ($\beta_1^e$), can be estimated by ordinary least squares (OLS). SUR can also assume that the individual error terms ($\varepsilon$) follow a bivariate Normal distribution (BVN), with mean zero and variances $\sigma_c^2$ and $\sigma_e^2$. The correlation between costs and outcomes, conditional on covariates, is recognised through the parameter $\rho$. Model 1 can also include covariate by treatment interaction terms. The covariates can be centred on their means so that $\beta_1^c$ and $\beta_1^e$ are the incremental costs and outcomes, at the mean level of each covariate. The uncertainty estimates can account for clustering with robust SE (Wooldridge, 2002). However, potential concerns with SUR are: (i) parameters estimates are obtained without acknowledging the clustering; (ii) correlation between costs and outcomes at individual and cluster levels are not separately identified; (iii) the asymptotic assumptions required for the robust variance estimation may not be satisfied in CRTs with few clusters, particularly when there are unequal numbers per cluster (Gomes et al., 2011b).

*Multilevel models (MLMs)*

Unlike SUR, MLMs can explicitly recognise clustering in the parameter estimation by incorporating the cluster-level random effects ($u_j^c, u_j^e$), while adjusting for cluster and

individual-level covariates (Nixon and Thompson, 2005). For example, an MLM that includes one individual-level covariate ($x_{ij}$) and one cluster-level ($z_j$) and can be described as:

$$\begin{aligned} c_{ij} &= \beta_0^c + \beta_1^c t_j + \beta_2^c x_{ij} + \beta_3^c z_j + u_j^c + \varepsilon_{ij}^c \\ e_{ij} &= \beta_0^e + \beta_1^e t_j + \beta_2^e x_{ij} + \beta_3^e z_j + u_j^e + \varepsilon_{ij}^e \end{aligned} \qquad \begin{aligned} \begin{pmatrix} \varepsilon_{ij}^c \\ \varepsilon_{ij}^e \end{pmatrix} &\sim BVN\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_c^2 & \rho\sigma_c\sigma_e \\ & \sigma_e^2 \end{pmatrix}\right) \\ \begin{pmatrix} u_j^c \\ u_j^e \end{pmatrix} &\sim BVN\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \tau_c^2 & \psi\tau_c\tau_e \\ & \tau_e^2 \end{pmatrix}\right) \end{aligned} \tag{2}$$

which as above can be extended to include treatment by covariate interactions. Model (2) acknowledges separately individual and cluster-level correlations between costs and outcomes, conditional on the covariates, through the parameters $\rho$ and $\psi$. This particular MLM (2) assumes the error terms are normally distributed but alternative distributions such as a Gamma distribution for costs could be chosen (Grieve et al., 2007, Nixon and Thompson, 2005). A general concern with MLMs or SUR is whether estimates are unbiased and precise if the distribution model is misspecified, by for example, assuming that the individual-level residuals are normally distributed when cost data are highly skewed. Unlike SUR, MLMs do not require the same set of asymptotic assumptions to be met, but the estimation of the variance-covariance structure still relies on asymptotic properties (Leyland and Goldstein, 2001).

*Two-stage bootstrap (TSB)*

We also considered a non-parametric TSB, which can accommodate clustering and the correlation between costs and outcomes, but avoids making distributional assumptions. We

provide an overview below, and define the steps taken in the algorithm (Appendix 5.1) but for full details of the TSB approach readers are referred elsewhere to (Gomes et al., 2011b)

A simple TSB resamples clusters and then individuals within each resampled cluster. However, to provide an accurate estimation of the variance, Davison and Hinkley advocate a 'shrinkage correction'. This procedure requires that shrunken cluster means and standardised individual residuals are calculated before any resampling. Bootstrap datasets are then constructed by combining resampled shrunken means with resampled individual-level residuals. The ATE of interest, for example the incremental net benefit (INB), can be taken as the mean of the INBs across the bootstrap replicates. Uncertainty can be reported by calculating bias-corrected and accelerated 95% CIs (Nixon et al., 2010). This approach can provide unbiased estimates of the INB and good CI coverage, even with few clusters of unequal size, if baseline covariates are balanced (Gomes et al., 2011b). We use this approach for the TSB without covariate adjustment.

When systematic imbalances are anticipated and covariate adjustment is required, the TSB described above may be insufficient. The previous resampling approach of combining each shrunken cluster mean with individual residuals drawn across all clusters, does not preserve a relationship between the cluster mean and the covariate information within the cluster. To avoid this problem we modify Davison and Hinkley's original resampling routine so that the bootstrap datasets respect the cluster membership. In the modified algorithm, shrunken cluster means and standardised residuals are calculated as before, but each cluster mean is now combined with individual residuals drawn from that same cluster (see Appendix 5.1 for further details).

We then adjust for covariate imbalances by applying SUR (model 1) to each bootstrap resample, to report adjusted incremental costs and outcomes and INBs, which are then averaged across the

bootstrap replicates. The SUR reports SEs for each incremental measure, without applying the robust estimator, because any clustering is recognised by the bootstrap routine. The SEs are then also averaged across the bootstrap replicates, to report 95% CIs. A potential concern is that while the TSB avoids distributional assumptions, the SUR adjustment assumes that the cost and outcome data in the bootstrap replicates are from bivariate Normal distributions.

## 3. Motivating example

*Design and description*

This CEA of a CRT evaluated alternative interventions for preventing postnatal depression (PoNDER) (Morrell et al., 2009). The CRT included 2659 patients attending 101 GP practices (clusters), and as is typical (Gomes et al., 2011a), the number of patients per cluster varied widely (from 1 to 77). Intra-cluster correlation coefficients (ICCs) were moderate for quality-adjusted life years (QALYs) ($ICCe$=0.04), but high for costs ($ICCc$=0.17). While QALYs were approximately Normally distributed, costs were moderately skewed.

In PoNDER, prior to patient recruitment, clusters were randomly allocated to usual care (control) or a psychological intervention delivered by a health visitor (treatment). The intervention consisted of health visitor training to identify and manage patients with postnatal depression. Baseline measurements were recorded for variables anticipated *a priori* to be potential confounders (Morrell et al., 2009). Previous studies suggest that cluster size, the number of patients randomised in each cluster, may be a confounder (Campbell et al., 2000, Omar and Thompson, 2000). In PoNDER, because clinical protocols were less restrictive in the control

than treatment group, it was anticipated that any relationship between the cluster size and the endpoints would be stronger in the control group. Hence, models were considered *a priori* that included an interaction of treatment with cluster size. This analysis used baseline and 6 month endpoints for 1,732 patients (70 clusters) with complete information.

Table 5.1 describes covariate balance between treatment arms, reported as percent standardised mean differences, which allows comparison across different types of variables (e.g. continuous, binary) and is invariant to sample size (Austin, 2009). For a continuous covariate ($x$), the standardised mean difference is calculated as $d_x = (\bar{x}_1 - \bar{x}_0)/\sqrt{(\text{var}_x^1 + \text{var}_x^0)/2} * 100$, with $\bar{x}_1, \bar{x}_0$ and $\text{var}_x^1, \text{var}_x^0$ the means and variances for each group. There is no consensus on the level of imbalance that is of concern, but if a standardised difference exceeds 10% this has been judged meaningful (Austin, 2009, Rosenbaum and Rubin, 1985).

In PoNDER, a cluster-level covariate, cluster size, and some individual-level covariates were relatively imbalanced (Table 5.1). Cluster size was strongly correlated with costs and QALYs but only for the control group. When the full data set was considered rather than the subset with complete information, covariate imbalance was similar.

We compare the analytical approaches described above, in pre-specified analyses: i) without covariate adjustment ii) with adjustment for main covariate effects and iii) with adjustment that includes main effects and a treatment by cluster size interaction. SUR was estimated in STATA by iterative feasible generalized least squares with a robust SE. The bivariate Normal MLM was implemented by maximum likelihood (in R).

**Table 5.1: The PoNDER case-study. Covariate balance for baseline characteristics, and correlation of those covariates with endpoints.**

| Covariates | Control group (n=495) | Treatment group (n=1237) | Standardised difference (%) | Correlated with endpoints | |
|---|---|---|---|---|---|
| **Cluster-level** | | | | | |
| Cluster size | 35.2 (21.08) | 39.8 (19.71) | 26.3 | $r_0^{cost} = 0.46$<br>$r_0^{qaly} = 0.29$ | $r_1^{cost} = -0.03$<br>$r_1^{qaly} = 0.05$ |
| **Individual-level** | | | | | |
| Age | 32.0 (5.12) | 31.3 (5.03) | 13.8 | $r_0^{cost} = 0.03$<br>$r_0^{qaly} = -0.04$ | $r_1^{cost} = -0.04$<br>$r_1^{qaly} = -0.02$ |
| Baseline QALY | 0.256 (0.035) | 0.259 (0.034) | 7.4 | $r_0^{cost} = -0.12$<br>$r_0^{qaly} = 0.77$ | $r_1^{cost} = -0.19$<br>$r_1^{qaly} = 0.76$ |
| Depression score | 6.85 (4.95) | 6.57 (4.81) | 5.7 | $r_0^{cost} = 0.10$<br>$r_0^{qaly} = -0.56$ | $r_1^{cost} = 0.30$<br>$r_1^{qaly} = -0.54$ |
| Economic status | 345 (69.8%) | 876 (70.8%) | 2.3 | $r_0^{cost} = 0.03$<br>$r_0^{qaly} = 0.08$ | $r_1^{cost} = 0.02$<br>$r_1^{qaly} = 0.01$ |
| Major life events | 202 (40.8%) | 492 (39.8%) | 2.1 | $r_0^{cost} = -0.02$<br>$r_0^{qaly} = -0.16$ | $r_1^{cost} = 0.05$<br>$r_1^{qaly} = -0.17$ |
| Previous depression | 40 (8.1%) | 107 (8.6%) | 2.1 | $r_0^{cost} = -0.02$<br>$r_0^{qaly} = -0.09$ | $r_1^{cost} = 0.11$<br>$r_1^{qaly} = -0.16$ |
| Living alone | 22 (4.4%) | 44 (3.6%) | 4.5 | $r_0^{cost} = -0.06$<br>$r_0^{qaly} = -0.05$ | $r_1^{cost} = 0.04$<br>$r_1^{qaly} = -0.13$ |

**Note:** continuous covariates reported as Mean (SD) and binary covariates as N (%), $r$ – correlation between the covariate and endpoint.

An MLM that allowed costs to take a Gamma distribution was fitted using Markov Chain Monte Carlo Methods (MCMC) by calling WinBUGS from R (Spiegehalter et al., 2003). The MCMC estimation was with 5000 iterations, three parallel chains with different starting values and assuming diffuse, vague priors (Lambert et al., 2005).

The unadjusted TSB was implemented with Davison and Hinkley's shrinkage correction (Davison and Hinkley, 1997). For covariate adjustment after the TSB, we combined our new TSB routine with SUR, but without a robust SE. Bootstrap methods were implemented in R, with 1000 replicates. We reported mean (SE) incremental costs, QALYs and INBs (at a ceiling ratio of £20 000 per QALY), and accompanying Akaike Information Criteria (AIC)[20].

*Case study results*

The treatment group had lower mean costs, higher mean QALYs, a positive INB and a high probability of being cost-effective (above 0.9) (Table 5.2). Without covariate adjustment, the MLMs reported a less negative incremental cost than the other methods; the MLMs gave relatively high weight to smaller clusters which in the control group had relatively low costs; hence the mean cost for the control group was lower for the MLMs versus SUR (£272 *vs* £303). After each model adjusted for main covariate effects, the estimated INBs were about 50% lower, with substantially smaller SEs, and the AICs were much reduced. Once the models included the treatment by cluster size interaction, SUR and the MLMs gave similar estimates, and lower AICs. When the MLMs were specified with Gamma rather than Normal costs, the estimated INB was similar, but model fit improved further. The TSB combined with SUR provided relatively similar point estimates to the other methods but with substantially smaller SEs. The differences across methods motivates the simulation study described in the next two sections, which aims to

[20] For SUR the AIC is computed from the least squares statistics and does not take into account the robust estimation. For TSB+SUR, the AIC is also taken from the same least squares statistics and averaged over the bootstrap samples.

provide generalisable conclusions about which methods are most appropriate across different circumstances.

## 4. Monte Carlo simulations

*Data generating process (DGP)*

The simulation study was designed to test the methods across a range of settings where systematic imbalances in baseline covariates may be anticipated in CEA that use CRTs. The choice of scenarios was based on the PoNDER case-study, a systematic review of published CEAs that use CRTs (Gomes et al., 2011a) and previous methodological studies (Campbell et al., 2005, Eldridge et al., 2006, Flynn and Peters, 2005, Pocock et al., 2002, Senn, 1994, Turner et al., 2007). It was judged important to allow the following to differ: the level of covariate imbalance, the correlation of each covariate (individual and cluster-level) with cost and QALY endpoints, the ICCs, the variation in cluster size and the number of clusters per treatment arm.

We designed a flexible DGP that incorporated baseline imbalances and correlations between covariates and endpoints, while recognising clustering, and correlation between costs and health outcomes. Briefly, costs and outcomes were simulated from a bivariate distribution in two stages, at the cluster then the individual level, to reflect the clustering inherent in CRTs. The DGP allowed for a wide range of parameters to be varied, and for each endpoint to have different parametric distributions. The DGP considered linear additive effects for both treatment and covariates (Turner et al., 2007).

**Table 5.2: PoNDER case-study. Mean (SE) incremental cost (£), incremental QALY, INB (λ=£20 000) for models without and with covariate. adjustment.**

| | SUR | | | MLM | | | TSB | | |
|---|---|---|---|---|---|---|---|---|---|
| | **No adjustment[1]** | **Adjusted for key covariates[2]** | **With interaction[3]** | **No adjustment[1]** | **Adjusted for key covariates[2]** | **With interaction[3]** | **No adjustment[1]** | **Adjusted for key covariates[2]** | **With interaction[3]** |
| Incremental cost | -63.4 (50.2) | -67.5 (45.0) | -86.4 (29.1) | -21.4 (25.3) | -19.9 (25.2) | -78.4 (29.7) | -61.7 (45.7) | -37.2 (10.1) | -43.0 (10.4) |
| Incremental QALY | 0.0043 (0.0020) | 0.0019 (0.0012) | 0.0021 (0.0013) | 0.0044 (0.0021) | 0.0019 (0.0013) | 0.0021 (0.0013) | 0.0042 (0.0024) | 0.0027 (0.0011) | 0.0028 (0.0012) |
| INB | 149.4 (70.1) | 105.5 (57.9) | 127.8 (47.8) | 109.0 (50.0) | 58.1 (36.8) | 119.7 (42.4) | 146.1 (65.3) | 91.7 (25.5) | 99.6 (28.8) |
| AIC | 16 886 | 15 110 | 14 808 | 16 630 | 14 936 | 14 742 | 16 894 | 15 090 | 14 840 |

[1]Model without covariates; [2]Model adjusted for cluster size, socio-economic status, age and other key clinical factors (Morrell et al., 2009); [3]Model with previous covariates plus a treatment interaction with cluster size, results reported at the mean cluster size.

We illustrate below a simple DGP with one continuous cluster-level covariate[21] and one continuous individual-level covariate (equations 3.1 and 3.2). We simulated cost ($c$) and outcome ($e$) data from a potential CRT with $M$ clusters per arm and $n_m$ ($m = 1,\ldots M$) individuals per cluster. We firstly generated cluster-level mean costs and outcomes ($\phi_j^c, \phi_j^e$) that followed distributions with means ($\mu_c, \mu_e$) and cluster-level standard deviations ($\tau_c, \tau_e$). Then, individual-level data ($c_{ij}, e_{ij}$) were simulated from distributions centred at the cluster-level means, and with individual-level standard deviations ($\sigma_c, \sigma_e$). Costs and outcomes were allowed to be correlated at both the cluster ($\psi$) and individual-level ($\rho$). The level of clustering was defined by the ICCs; for example for costs $ICC_c = \tau_c^2/(\sigma_c^2 + \tau_c^2)$. The number of individuals per cluster was drawn from a Gamma distribution defined by a mean and coefficient of variation, which ensured cluster size remained positive (Eldridge et al., 2006).

Cluster-level means:

$$\phi_j^c \sim dist(\mu_c,\ \tau_c), \quad \phi_j^e \sim dist(\mu_e,\ \tau_e) \qquad (3.1)$$
$$\mu_c = \beta_0^c + \beta_1^c t_j + \beta_3^c z_j, \quad \mu_e = \beta_0^e + \beta_1^e t_j + \beta_3^e z_j + \psi(\phi_j^c - \mu_c)$$

Individual-level data:

$$c_{ij} \sim dist(\phi_j^c + \beta_2^c x_{ij},\ \sigma_c) \qquad (3.2)$$
$$e_{ij} \sim dist(\phi_j^e + \beta_2^e x_{ij} + \rho(c_{ij} - (\phi_j^c + \beta_2^c x_{ij}))\ ,\ \sigma_e)$$

---

[21] In PoNDER, the imbalanced cluster level covariate was cluster size. To afford more flexibility in the simulation study, a different cluster-level characteristic was assumed imbalanced between the treatment groups.

We incorporated the cluster-level covariate ($z_j$) when simulating the cluster-level mean costs and outcomes, and the individual-level covariate ($x_{ij}$) when simulating individual-level data[22]. Both cluster and individual-level covariates were assumed to be continuous and drawn from Normal distributions, $z_j \sim N(\mu_z, \sigma_z)$ and $x_{ij} \sim N(\mu_x, \sigma_x)$.

The DGP introduced systematic baseline imbalances by allowing the covariate means to differ across treatment arms set according to standardised mean differences (Austin, 2009)[23]. For the individual ($\beta_2^c, \beta_2^e$) and cluster-level ($\beta_3^c, \beta_3^e$) covariates, coefficients were simulated as a function of the correlation coefficient ($r$) between each covariate and the corresponding endpoint (Turner et al., 2007). For instance, the coefficient of the individual-level covariate (Normal) on health outcomes (Normal) was determined as $\beta_2^e = \frac{\sigma_e}{\sigma_x}\sqrt{r_e^2 / (1 - r_e^2)}$, and the corresponding coefficient on costs (Gamma) as $\beta_2^c = \frac{\mu_c}{\sigma_x}\sqrt{(1/shape_c)r_c^2 / (1 - r_c^2)}$. The DGP easily extends to allow the prognostic strength of a covariate to differ by treatment group, by including treatment by covariate interaction terms.

*Definition of scenarios*

Table 5.3 lists parameters allowed to vary across the scenarios. Other parameters, such as the level of correlation between costs and health outcomes (0.2), mean cluster size (50) and true INB

[22] As individuals within a cluster tend to be relatively similar, the covariate was allowed to be clustered.
[23] The standardised mean differences assumed constant variance across treatment arms.

(£1 000; ceiling ratio £20 000 per QALY), were held constant across scenarios. Covariates $x_{ij}$ and $z_j$ were assumed to follow Normal distributions (mean 50 and SD 20) throughout.

The first group of scenarios (Table 5.3, S1-S5), considered different levels of imbalance for an individual-level covariate, and confounding just for health outcomes. In the initial scenario, baseline imbalance and the correlation between the covariate and health outcome were both set to zero (S1). We then simulated scenarios with increasing levels of baseline imbalance and correlation with health outcomes (S2-S5). For these scenarios, we reported the performance for each method before and after adjustment. The scenario, S5, characterised by high levels of imbalance and confounding, was taken as the base case for subsequent scenarios.

The second group of scenarios, considered the choice of adjustment method across a broader set of circumstances (Table 5.3, S6-S11). These scenarios allowed for confounding in the cost endpoint, assumed to follow a Gamma distribution (S6). Subsequent scenarios allowed: for imbalance in a cluster-level covariate, assumed correlated with both endpoints (S7); high ICCs (S8); unequal cluster sizes (S9); and few clusters (S10). In addition to the change described, each scenario incorporated the characteristics of the preceding setting. The final scenario (S11), motivated by PoNDER, and anticipated in CRTs more generally (Campbell et al., 2000), allowed the prognostic relationship of a cluster-level covariate to differ by treatment arm.

*Implementation*

For each scenario, each method estimated INBs before and after covariate adjustment. MLMs and the TSB were implemented in R (R, 2011) and SUR in STATA (STATA, 2009). SUR was

estimated by iterative feasible generalized least squares with a robust SE, and the bivariate Normal MLMs by maximum likelihood. The TSB was implemented before, and after adjustment with SUR (no robust SE) as in the case study. We conducted 2000 simulations for each scenario[24]. The relative performance of the alternative methods was assessed according to mean (SE) bias, root mean squared error (rMSE), variance, confidence interval (CI) coverage, and CI width of the INB (ceiling ratio of £20 000 per QALY). We reported performance before and after adjustment (S1-6, S11), and across the adjusted methods (S6-10).

**Table 5.3: Description of the main parameter values allowed to vary across the different scenarios in the simulation study**

| Scenario | Individual-level covariate | | | Cluster-level covariate | | | Costs | ICCs | $cv_{imb}$ | $M$ |
|---|---|---|---|---|---|---|---|---|---|---|
| | $d$ | $r_e$ | $r_c$ | $d$ | $r_e$ | $r_c$ | | | | |
| **S1** | 0 | 0 | 0 | 0 | 0 | 0 | Normal | 0.01 | 0 | 20 |
| **S2** | 0 | 0.1 | 0 | 0 | 0 | 0 | Normal | 0.01 | 0 | 20 |
| **S3** | 5 | 0.1 | 0 | 0 | 0 | 0 | Normal | 0.01 | 0 | 20 |
| **S4** | 5 | 0.3 | 0 | 0 | 0 | 0 | Normal | 0.01 | 0 | 20 |
| **S5** | 20 | 0.3 | 0 | 0 | 0 | 0 | Normal | 0.01 | 0 | 20 |
| **S6** | 20 | 0.3 | -0.3 | 0 | 0 | 0 | Gamma | 0.01 | 0 | 20 |
| **S7** | 20 | 0.3 | -0.3 | 20 | 0.3 | -0.3 | Gamma | 0.01 | 0 | 20 |
| **S8** | 20 | 0.3 | -0.3 | 20 | 0.3 | -0.3 | Gamma | 0.2 | 0 | 20 |
| **S9** | 20 | 0.3 | -0.3 | 20 | 0.3 | -0.3 | Gamma | 0.2 | 1 | 20 |
| **S10** | 20 | 0.3 | -0.3 | 20 | 0.3 | -0.3 | Gamma | 0.2 | 1 | 3 |
| **S11** | 20 | 0.3 | -0.3 | 20 | $0.3^{\dagger}$ | $-0.3^{\dagger}$ | Gamma | 0.2 | 1 | 20 |

**Notes:** $d$- standardised difference; $r_e$ – correlation between covariate and outcomes; $r_c$ – correlation between covariate and costs; $cv_{imb}$ - coefficient of variation of the cluster size; $M$ – no. of clusters per arm; $^{\dagger}$correlation was 50% higher for treatment arm (differential prognostic strength).
The choice of parameter values was informed by previous systematic and conceptual reviews (Gomes et al., 2011a), and from data extracted from eight case studies (Gomes et al., 2011b).

[24]2000 simulations provide coverage rates of 0.94 to 0.96 (for true coverage of 0.95) with 95% confidence.

## 5. Simulation results

Table 5.4 reports the results for the first set of scenarios where an individual-level baseline covariate had different levels of imbalance and correlation with health outcome. Even with low levels of baseline imbalance and correlation (S3), methods without adjustment produced slightly biased results. At increased levels of imbalance and correlation (S5), the unadjusted approaches reported high bias (>10%) and low CI coverage (below 0.9 for a nominal level of 0.95).

All adjusted approaches reported unbiased estimates of the INB, including the new TSB routine combined with SUR[25]. However, the CI coverage for the TSB combined with SUR was lower than for the other methods (0.91 *vs* 0.94) across all scenarios.

In the scenario without imbalance and confounding (S1) covariate adjustment increased the variance of the INB (after covariate adjustment with the MLMs, the average variance was 12 125 vs 12 027 before adjustment). By contrast, if the covariate was balanced but correlated with outcome (S2), the corresponding variance was slightly smaller after adjustment (12 122 *vs* 12 227).

For the scenarios with confounding on costs (S6), an imbalanced cluster-level covariate correlated with both endpoints (S7), high ICCs (S8), cluster size variation (S9) and few clusters (S10) all unadjusted methods reported biased estimates and low CI coverage (below 0.9). Following covariate adjustment, each method provided unbiased estimates of the INB (Appendix 5.2).

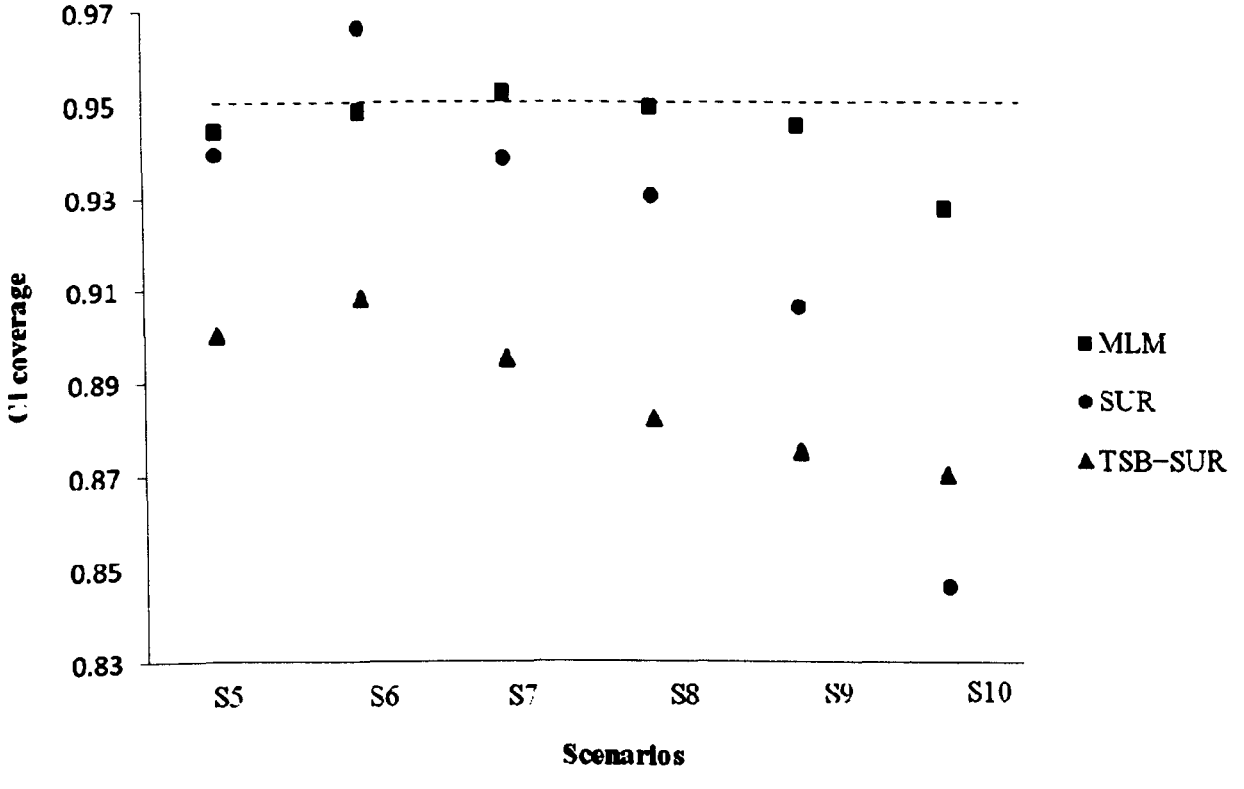
[25]Using Davison and Hinkley's original TSB routine, combined with SUR provided biased results; for example for S5 the mean (SE) bias was 23.6 (2.56).

**Table 5.4: Bias (SE) of the INB for a set of scenarios (S1-S5) which allow for increasing levels of baseline imbalance for an individual-level covariate, and increasing levels of correlation of that covariate with health outcome (QALYs, true INB=£1 000)**

| | | | SUR | | MLM | | TSB | |
|---|---|---|---|---|---|---|---|---|
| **Scenario** | **Baseline imbalance** | **Correlation between covariate and outcome** | Without covariate adjustment | With covariate adjustment | Without covariate adjustment | With covariate adjustment | Without covariate adjustment | With covariate adjustment |
| **S1** | **None** | **None** | 0.14 (2.46) | 0.56 (2.50) | 0.14 (2.46) | 0.56 (2.50) | 0.13 (2.47) | 0.43 (2.47) |
| **S2** | **None** | **Low (0.1)** | 0.26 (2.47) | 0.11 (2.46) | 0.26 (2.47) | 0.11 (2.46) | 0.24 (2.48) | 0.11 (2.46) |
| **S3** | **Low (5)** | **Low (0.1)** | 9.79 (2.47) | 0.07 (2.46) | 9.79 (2.47) | 0.07 (2.46) | 9.81 (2.48) | 0.04 (2.46) |
| **S4** | **Low (5)** | **High (0.3)** | 30.9 (2.58) | 0.08 (2.46) | 30.9 (2.58) | 0.08 (2.46) | 31.0 (2.58) | 0.02 (2.46) |
| **S5** | **High (20)** | **High (0.3)** | 125.3 (2.58) | 0.01 (2.47) | 125.3 (2.58) | 0.01 (2.47) | 125.3 (2.58) | 0.03 (2.47) |

However, as Figure 5.1 shows, CI coverage differed across methods. The combination of the TSB with SUR gives poor CI coverage (0.91 or less) under each scenario. The CI coverage with SUR is lower than for the MLMs, when the numbers per cluster vary[26] (S9) and there are few clusters (S10). For these scenarios, MLMs also reports lower variance and rMSE than SUR (see Appendix 5.2 for further details). For scenario S10, characterised by imbalanced individual and cluster-level covariates correlated with endpoints, high ICCs, few clusters (8 per arm) and cluster size variation, the adjusted MLMs still gives reasonable coverage (0.93).

**Figure 5.1: CI coverage of the INB (nominal level is 0.95) for adjusted methods for the following scenarios: base case (S5); confounding on costs (S6); imbalanced cluster-level covariate (S7); high ICCs (S8); high cluster size variation (S9); few clusters (S10)***




*Each scenario includes the other characteristics of the preceding scenario.

[26] Here, for cluster size we assumed a coefficient of variation of 1. Even with a coefficient of variation of 0.5, SUR reports variance and rMSE that are 20% higher than for the MLM.

**Table 5.5: Bias, variance, rMSE CI coverage and width of the INB for a scenario (S11) with a cluster-level prognostic relationship that differs by treatment arm (true INB=£1 000)**

| | SUR | | | MLM | | | TSB | | |
|---|---|---|---|---|---|---|---|---|---|
| | Without covariate adjustment | Adjust for main effect only | Adjust for interaction* | Without covariate adjustment | Adjust for main effect only | Adjust for interaction | Without covariate adjustment | Adjust for main effect only | Adjust for interaction |
| **Mean (SE) bias** | 421.9 (28.9) | 167.9 (9.4) | 3.93 (28.0) | 422.3 (27.9) | 167.5 (7.6) | 3.51 (22.3) | 423.9 (29.1) | 168.2 (9.0) | 4.71 (26.2) |
| **variance** | 1 673 434 | 176 655 | 183 833 | 1 555 618 | 116 477 | 112 697 | 1 695 850 | 162 577 | 158 030 |
| **rMSE** | 1 361 | 453 | 438 | 1 317 | 380 | 367 | 1 369 | 437 | 425 |
| **CI coverage** | 0.808 | 0.879 | 0.885 | 0.790 | 0.919 | 0.947 | 0.809 | 0.875 | 0.881 |
| **Mean CI width** | 1 742 | 1 472 | 1 352 | 1 711 | 1 343 | 1 194 | 1 749 | 1 482 | 1 401 |

* ATE is reported at the covariate mean. This scenario is characterised by high ICCs (0.2), unequal numbers per cluster, and 20 clusters per treatment arm

Table 5.5 reports the results for the last scenario (S11), where the prognostic relationship for a cluster-level covariate differed by treatment arm, there were unequal numbers per cluster, high ICC (0.2), but moderate numbers of clusters (20 per arm)[27]. The results show that unless the treatment by covariate interaction is incorporated, each method reported biased estimates of the INB and low CI coverage. After including the interaction term, each method provided unbiased estimates, lower rMSE and improved CI coverage. The MLMs with the interaction term reported the lowest rMSE and was the only approach that reported CI coverage close to the nominal level.

## 6. Discussion

This study presents alternative methods for CEA that use CRT where baseline covariates differ between treatment groups. These adjusted methods address systematic imbalances in both individual and cluster-level covariates. The case study illustrates that in CEA that use CRT, cost-effectiveness estimates can differ according to method. The simulation extends the case study, and shows that without adjustment, CEA can report biased estimates even with low levels of confounding.

By contrast, each adjustment method provides unbiased estimates. Of the alternative methods, the MLMs report CI coverage close to nominal levels across all the circumstances considered (CI coverage of 0.93 to 0.95). In settings with unequal numbers per cluster and few clusters, SUR with a robust variance estimator, reports low CI coverage and high rMSE compared to the MLMs. The TSB and SUR approach proposed gives low CI coverage in each setting considered.

[27] We also considered a scenario where the interaction of treatment is with an individual-level rather than a cluster-level covariate, but the results were similar to those presented for S11.

This is the first paper to consider analytical methods for addressing systematic covariate imbalance in CEA that use CRT. A previous simulation study (Gomes et al., 2011b) suggested that MLMs or a TSB approach were appropriate for CEA that use CRTs, but only considered circumstances with balanced covariates. Our paper shows that where the CRT has systematic baseline differences between the treatment groups, methods that assume covariate balance are insufficient. We consider a simple approach to adjusting for systematic imbalances in patient or cluster-level covariates, which is to apply SUR with a robust SE. Previous work reported that SUR performed well for CEA that use CRTs unless the number of clusters was small (Gomes et al., 2011b). By contrast, our paper shows that when there are unequal numbers per cluster, adjusted SUR can report poor coverage even with a moderate number of clusters (20 per treatment arm). This is an important concern, as a previous review reported that 75% of studies have uneven numbers per cluster, and of these about 50% have fewer than 20 clusters per arm (Gomes et al., 2011a). While improved robust estimators have been proposed for setting with few clusters (Pan and Wall, 2002, Skene and Kenward, 2010), it is unclear how they would perform with unequal cluster sizes.

Rather than relying on the asymptotics required for robust variance estimation, or the distributional assumptions made by MLMs, we extend a previous TSB algorithm and combine it with SUR. While this new approach performs well in terms of bias and rMSE, it provides too narrow CIs, as observed in the case study. Hence, TSB appears less appealing for CEA when covariate adjustment is required. While one alternative would be to combine the TSB with a SUR or GEE that has a robust variance estimator, as our results show the asymptotic assumptions required are unlikely to be satisfied by the numbers of clusters commonly in CRTs.

An alternative approach to avoiding distributional assumptions about the endpoints, would be to bootstrap individual and cluster-level residuals from adjusted MLMs (Carpenter et al., 2003).

The MLMs proposed have more general appeal for CEA that use CRTs. The MLMs that assume bivariate Normality, perform relatively well even with highly skewed costs; this corroborates previous findings suggesting that methods that assume Normality may be reasonably robust to skewed cost data (Nixon et al., 2010, Willan et al., 2004). In the case study, the MLM extended to assume a Gamma distribution for costs, and as in previous studies, this slightly improved the precision of the estimates (Grieve et al., 2010). The MLMs presented here can be easily extended to report multiplicative treatment effects (Thompson et al., 2006) or ATEs for each subgroup of policy-interest (Vaness and Mullahy, 2006).

In addressing systematic imbalances, issues beyond the choice of estimation method warrant careful consideration. In particular, pre-specified analysis plans for CEA should consider *a priori* what form the potential confounding may take, informed by theory, previous literature and expert opinion. In our case-study, as may be present more generally in CEA, adjusting for main effects was judged insufficient. Here, it was important that each method recognised that a prognostic relationship can differ by treatment group. Indeed, the simulation highlighted that ignoring a more complex prognostic relationship can bias the overall cost-effectiveness estimates.

This research does have some limitations. The methods proposed allow for systematic differences in potential confounders that were observed. The CRT design may also lead to systematic imbalances in unobserved characteristics. Hence methods such as instrumental variable estimation that can address unobserved differences also warrant careful consideration

(Basu et al., 2007, Polsky and Basu, 2006). Some CRTs may be designed so that the only baseline imbalances are by chance; our study does not apply to these circumstances. The MLMs proposed performed well across a range of settings including skewed cost data, but the simulation study did not consider some complexities that can arise including variances that differ across clusters, or non-Normal distributions for cluster-level residuals, In principle, the MLMs presented could be extended to allow for such complexities, but previous research suggest the improvements in inference may be relatively small (Grieve et al., 2010).

This paper opens up several areas for further research. In particular, it would be useful to extend the methods to handle nonlinear relationships between covariates and endpoints, missing and censored data. A complementary approach, which can offer protection against misspecification of the covariate adjustment model would be to extend the MLMs to doubly robust estimation (Bang and Robins, 2005). Here, a model for treatment choice, a propensity score, could be estimated including covariates anticipated to be potential confounders, with the MLMs weighted according to the inverse probability of treatment (Imbens, 2004). Such doubly robust estimators are consistent as long as either the treatment or the endpoint model is correctly specified (Bang and Robins, 2005). Another area for further research is to consider circumstances where missing data are a concern. In the context of CEA that use CRTs, the use of multilevel multiple imputation may be warranted (Carpenter et al., 2011). This approach extends conventional multiple imputation (Rubin, 1987) to reflect the multilevel structure of the original data.

This paper extends the literature examining the relative merits of hierarchical models (Cameron and Trivedi, 2005, Jones, 2009), robust variance estimation (Greene, 2003, Wooldridge, 2010), and non-parametric bootstrap approaches for covariate adjustment. In a context where adjustment methods are required to address systematic differences between treatment groups as

well as accommodate clustering and the correlation of costs with health outcomes, we find that MLMs perform well. While any of the adjustment methods proposed reports unbiased estimates, the MLMs can provide more precise estimates with better CI coverage than the other approaches.

**Acknowledgments**

The authors are grateful to John Cairns and Simon Dixon for helpful comments and Jane Morrell for providing full access to the PoNDER data. We also thank participants at the Twentieth European Workshop on Econometrics and Health Economics (York, 2011), where this paper was presented.

**References**


Altman, D. G. 2005. Adjustment for covariate imbalance. *In:* ARMITAGE, P. & COLTON, T. (eds.) *Encyclopedia of Biostatistics.* Chichester, UK: John Wiley.

Assmann, S. F., Pocock, S. J., Enos, L. E. & Kasten, L. E. 2000. Subgroup analysis and other (mis)uses of baseline data in clinical trials. *Lancet,* 355, 1064-9.

Austin, P. C. 2009. Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. *Stat Med,* 28, 3083-107.

Austin, P. C., Manca, A., Zwarenstein, M., Juurlink, D. N. & Stanbrook, M. B. 2010. A substantial and confusing variation exists in handling of baseline covariates in randomized controlled trials: a review of trials published in leading medical journals. *Journal of Clinical Epidemiology,* 63, 142-153.

Bang, H. & Robins, J. M. 2005. Doubly robust estimation in missing data and causal inference models. *Biometrics,* 61, 962-73.

Barber, J. & Thompson, S. 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. & 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. 2006. Statistical Methods for cost-effectiveness analysis alongside clinical trials. *In:* JONES, A. (ed.) *The Elgar Companion to Health Economics.* Cheltenham, UK: Edward Elgar Publishing.

Cameron, A. C. & Trivedi, P. K. 2005. *Microeconometrics : methods and applications,* Cambridge ; New York, Cambridge University Press.

Campbell, M. K., Fayers, P. M. & Grimshaw, J. M. 2005. Determinants of the intracluster correlation coefficient in cluster randomized trials: the case of implementation research. *Clinical Trials,* 2, 99-107.

Campbell, M. K., Mollison, J., Steen, N., Grimshaw, J. M. & Eccles, M. 2000. Analysis of cluster randomized trials in primary care: a practical approach. *Family Practice,* 17, 192-196.

Carpenter, J., Goldstein, H. & Kenward, M. 2011. REALCOM-IMPUTE software for multilevel multiple imputation with mixed response types. *Journal of Statistical Software* (in press).

Carpenter, J. R., Goldstein, H. & Rasbash, J. 2003. A novel bootstrap procedure for assessing the relationship between class size and achievement. *Journal of the Royal Statistical Society Series C-Applied Statistics,* 52, 431-443.

Carter, B. 2010. Cluster size variability and imbalance in cluster randomized controlled trials. *Stat Med,* 29, 2984-93.

Claxton, K. 1999. The irrelevance of inference: a decision-making approach to the stochastic evaluation of health care technologies. *J Health Econ,* 18, 341-64.

Davison, A. C. & Hinkley, D. V. 1997. *Bootstrap methods and their application,* Cambridge, UK, Cambridge University Press.

Donner, A. 1998. Some aspects of the design and analysis of cluster randomization trials. *Applied Statistics,* 47, 95-113.

Donner, A. & Klar, N. 2000. *Design and analysis of cluster randomization trials in health research,* London, UK, Hodder Arnold Publishers.

Eldridge, S., Ashby, D., Bennett, C., Wakelin, M. & Feder, G. 2008. Internal and external validity of cluster randomised trials: systematic review of recent trials. *British Medical Journal,* 336, 876-880.

Eldridge, S. M., Ashby, D. & Kerry, S. 2006. Sample size for cluster randomized trials: effect of coefficient of variation of cluster size and analysis method. *Int J Epidemiol,* 35, 1292-300.

Flynn, T. N. & Peters, T. J. 2005. Cluster randomized trials: Another problem for cost-effectiveness ratios. *International Journal of Technology Assessment in Health Care,* 21, 403-409.

Gelman, A. & Pardoe, I. 2007. Average Predictive Comparisons for Models with Nonlinearity, Interactions, and Variance Components. *Sociological Methodology 2007, Vol 37,* 37, 23-51.

Gomes, M., Grieve, R., Edmunds, J. & Nixon, R. 2011a. Statistical methods for cost-effectiveness analyses that use data from cluster randomised trials: a systematic review and checklist for critical appraisal. *Medical Decision Making,* (in press). DOI:10.1177/0272989X11407341.

Gomes, M., Ng, E. S., Grieve, R., Nixon, R., Carpenter, J. & Thompson, S. 2011b. Developing appropriate analytical methods for cost-effectiveness analyses that use cluster randomized trials. *Medical Decision Making,* Submitted (March 2011).

Greene, W. H. 2003. *Econometric analysis,* Upper Saddle River, N.J., Great Britain, Prentice Hall.

Grieve, R., Nixon, R. & Thompson, S. G. 2010. Bayesian hierarchical models for cost-effectiveness analyses that use data from cluster randomized trials. *Med Decis Making,* 30, 163-75.

Grieve, R., Nixon, R., Thompson, S. G. & Cairns, J. 2007. Multilevel models for estimating incremental net benefits in multinational studies. *Health Econ,* 16, 815-26.

Hahn, S., Puffer, S., Torgerson, D. J. & Watson, J. 2005. Methodological bias in cluster randomised trials. *BMC Med Res Methodol,* 5, 10.

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 Econ,* 11, 415-30.

Imai, K., King, G. & Stuart, E. A. 2008. Misunderstandings between experimentalists and observationalists about causal inference. *Journal of the Royal Statistical Society Series a-Statistics in Society,* 171, 481-502.

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

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

Jones, A. 2009. Panel data methods and applications to health economics. *In:* MILLS, T. & PATTERSON, K. (eds.) *Palgrave Handbook of Econometrics, Volume II: Applied Econometrics.* Hampshire, UK: Palgrave MacMillan.

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

Lambert, P. C., Sutton, A. J., Burton, P. R., Abrams, K. R. & Jones, D. R. 2005. How vague is vague? A simulation study of the impact of the use of vague prior distributions in MCMC using WinBUGS. *Stat Med,* 24, 2401-28.

Leyland, A. & Goldstein, H. 2001. *Multilevel Modelling of Health Statistics,* Chichester, UK, John Wiley & Sons, Ltd.

Liu, J. X. & Gustafson, P. 2008. On Average Predictive Comparisons and Interactions. *International Statistical Review,* 76, 419-432.

Manca, A., Hawkins, N. & Sculpher, M. J. 2005. Estimating mean QALYs in trial-based cost-effectiveness analysis: the importance of controlling for baseline utility. *Health Econ,* 14, 487-96.

Morgan, K., Thompson, J., Dixon, S., Tomeny, M. & Mathers, N. 2003. Predicting longer-term outcomes following psychological treatment for hypnotic-dependent chronic insomnia. *J Psychosom Res,* 54, 21-9.

Morrell, C. J., Slade, P., Warner, R., Paley, G., Dixon, S., Walters, S. J., Brugha, T., Barkham, M., Parry, G. J. & Nicholl, J. 2009. Clinical effectiveness of health visitor training in psychologically informed approaches for depression in postnatal women: pragmatic cluster randomised trial in primary care. *BMJ,* 338, a3045.

Nixon, R. M. & Thompson, S. G. 2005. Methods for incorporating covariate adjustment, subgroup analysis and between-centre differences into cost-effectiveness evaluations. *Health Econ,* 14, 1217-29.

Nixon, R. M., Wonderling, D. & Grieve, R. D. 2010. Non-parametric methods for cost-effectiveness analysis: the central limit theorem and the bootstrap compared. *Health Econ,* 19, 316-33.

Omar, R. Z. & Thompson, S. G. 2000. Analysis of a cluster randomized trial with binary outcome data using a multi-level model. *Stat Med,* 19, 2675-88.

Pan, W. & Wall, M. M. 2002. Small-sample adjustments in using the sandwich variance estimator in generalized estimating equations. *Statistics in Medicine,* 21, 1429-1441.

Panageas, K. S., Schrag, D., Russell Localio, A., Venkatraman, E. S. & Begg, C. B. 2007. Properties of analysis methods that account for clustering in volume-outcome studies when the primary predictor is cluster size. *Stat Med,* 26, 2017-35.

Pocock, S. J., Assmann, S. E., Enos, L. E. & Kasten, L. E. 2002. Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: current practice and problems. *Stat Med,* 21, 2917-30.

Polsky, D. & Basu, A. 2006. Selection bias in observational data. *In:* JONES, A. (ed.) *The Elgar Companion to Health Economics.* Cheltenham, UK: Edward Elgar.

Puffer, S., Torgerson, D. & Watson, J. 2003. Evidence for risk of bias in cluster randomised trials: review of recent trials published in three general medical journals. *BMJ,* 327, 785-9.

Puffer, S., Torgerson, D. J. & Watson, J. 2005. Cluster randomized controlled trials. *J Eval Clin Pract,* 11, 479-83.

R 2011. The R project for statistical computing. http://www.r-project.org/.

Rosenbaum, P. R. & Rubin, D. B. 1985. Constructing a Control-Group Using Multivariate Matched Sampling Methods That Incorporate the Propensity Score. *American Statistician,* 39, 33-38.

Rubin, D. 1987. *Multiple imputation for nonresponse in surveys,* New York, US, Wiley.

Sekhon, J. S. & Grieve, R. 2011. A matching method for improving covariate balance in cost-effectiveness analyses. *Health Economics,* Accepted (April 2011).

Senn, S. 1994. Testing for baseline balance in clinical trials. *Stat Med,* 13, 1715-26.

Senn, S. J. 1989. Covariate Imbalance and Random Allocation in Clinical-Trials. *Statistics in Medicine,* 8, 467-475.

Skene, S. S. & Kenward, M. G. 2010. The analysis of very small samples of repeated measurements II: a modified Box correction. *Stat Med,* 29, 2838-56.

Smeeth, L. & Ng, E. S. 2002. Intraclass correlation coefficients for cluster randomized trials in primary care: data from the MRC Trial of the Assessment and Management of Older People in the Community. *Control Clin Trials,* 23, 409-21.

Spiegehalter, D., Thomas, A., Best, N. & Lunn, D. 2003. WinBUGS User Manual, version 1.4. MRC Biostatistics Unit. Cambridge, UK. .

Stata 2009. Stata programming reference manual, Version 11. Texas, US: StataCorp.

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.

Turner, R. M., White, I. R. & Croudace, T. 2007. Analysis of cluster randomized cross-over trial data: a comparison of methods. *Stat Med,* 26, 274-89.

Vaness, D. & Mullahy, J. 2006. Perspectives on mean-based evaluation of health care. *In:* JONES, A. (ed.) *The Elgar Companion to Health Economics.* Cheltenham, UK: Edward Elgar.

Willan, A. & Briggs, A. 2006. *Statistical Analysis of cost-effectiveness data,* Chichester, UK, 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 Econ,* 13, 461-75.

Wooldridge, J. M. 2002. *Econometric analysis of cross section and panel data,* Cambridge, Mass., MIT Press.

Wooldridge, J. M. 2010. *Econometric analysis of cross section and panel data,* Cambridge, Mass., MIT Press.

Zellner, A. 1962. An efficient method of estimating seemingly unrelated regressions and tests of aggregation bias. *Journal of the American Statistical Association,* 57, 348-368.

## Appendix 5.1: Algorithm for the non-parametric TSB combined with SUR

Suppose we have $M_k$ clusters randomised to treatment ($k$=2) and control ($k$=1) groups, with $n_j$ individuals within each cluster $j$.

13. For $i$ in 1 to $n_j$ (individuals in cluster $j$)
14. For $j$ in 1 to $M_k$ (clusters in treatment $k$)
15. For $k$ in 1 to 2 (treatments)
16. Calculate shrunken cluster means, $\hat{x}_j^c$ and $\hat{x}_j^e$, for cost and outcome[28].
17. Calculate standardized individual-level residuals, $\hat{z}_{cost,ji}$ and $\hat{z}_{effect,ji}$, for cost and outcome[29].
18. Randomly sample (with replacement) $M_k$ pairs of cluster means, $x^*_{cost,j'}$ and $x^*_{effect,j'}$, from the shrunken cluster means calculated in step 4.
19. Within each resampled cluster, randomly sample (with replacement) $\sum_{j'=1}^{M_k} n_{j'}$ pairs of standardized residuals (step 5), $z^*_{cost,i'}$ and $z^*_{effect,i'}$, where $i'$=1... $\sum_{j'=1}^{M_k} n_{j'}$.
20. Re-construct the sample ($y^*_{cost,j'i'}$, $y^*_{effect,j'i'}$) by adding the shrunken cluster means from step 6 and the standardized residuals from step 7, i.e. $y^*_{cost,j'i'} = x^*_{cost,j'} + z^*_{cost,i'}$ where $i' = 1 \ldots n_{j'}$ and likewise for effects; call it a 'synthetic' sample.
21. Incorporate the covariate set ($w_{j'i'}$) into each synthetic sample: ($y^*_{cost,j'i'} + w_{j'i'}, y^*_{effect,j'i'} + w_{j'i'}$). Covariates can be different for costs versus outcomes.
22. Repeat steps 4 to 9 for each treatment arm and stack these 'synthetic' samples into a single bootstrap sample.
23. Replicate steps 6 to 10 $R$ times to construct $R$ bootstrap samples.
24. Apply SUR without robust standard error to each bootstrap sample generated in step 11, to estimate mean and standard error (SE) of incremental costs (ΔC), incremental outcomes (ΔE) and the covariance (ΔC,ΔE), adjusted for potential confounders.
25. Calculate the parameter of interest, e.g. INB, by averaging SUR estimates across the $R$ replications: $\widehat{INB} = (\sum_{r=1}^{R} \widehat{\Delta E}_r \lambda - \widehat{\Delta C}_r)/R$, where λ is the willingness-to-pay for a QALY.
26. Applying the Central Limit Theorem, CIs for INB can be constructed as $\widehat{INB} \pm 1.96SE(\widehat{INB})$ (Nixon et al., 2010) where,

$$SE(\widehat{INB}) = \sqrt{[\sum_{r=1}^{R} \mathrm{SE}(\widehat{\Delta E}_r)^2 \lambda^2 + \mathrm{SE}(\widehat{\Delta C}_r)^2 - 2\lambda \mathrm{cov}(\widehat{\Delta E}_r, \widehat{\Delta E}_r)]}/R.$$

---

[28] $\hat{x}_j^c = c\bar{y}_{..}^c + (1-c)\bar{y}_j^c$ where c is given by $(1-c)^2 = \frac{M_k}{M_k - 1} - \frac{SS_W}{b(b-1)SS_B}$; $SS_w$= within-sum of squares and $SS_B$ = between-sums of squares, b = average cluster size (a formulation akin to the harmonic mean is used here (Smeeth and Ng, 2002)).

[29] $\hat{z}_{cost,ji} = \frac{y_{cost,ji} - \bar{y}_{cost,j.}}{\sqrt{1-b^{-1}}}$, where $y_{cost,ji}$ is the observed cost for the $i$-th individual in cluster $j$. These are similarly calculated for outcomes, and separately for the two treatment arms.

**Appendix 5.2: Bias (True INB=£1 000), variance and rMSE of the INB for adjusted methods, across scenarios S5-S10***

| | Bias | | | Variance | | |
|---|---|---|---|---|---|---|
| | SUR | MLM | TSB+SUR | SUR | MLM | TSB+SUR |
| **Base-case (S5)** | 0.04 (2.47) | 0.01 (2.47) | 0.26 (2.47) | 12 168 | 12 172 | 12 174 |
| **Confounding on costs (S6)** | 1.77 (2.65) | 1.78 (2.65) | 1.59 (2.65) | 14 092 | 14 092 | 14 073 |
| **Cluster-level covariate (S7)** | 0.06 (2.69) | 0.06 (2.69) | 3.24 (2.67) | 14 475 | 14 468 | 14 258 |
| **High ICCs (S8)** | 7.84 (7.06) | 8.11 (7.05) | 8.25 (7.05) | 99 431 | 99 549 | 99 431 |
| **High cluster size variation (S9)** | 10.3 (9.54) | 2.04 (7.76) | 9.07 (9.22) | 182 142 | 120 300 | 169 880 |
| **Few clusters (S10)** | 0.15 (15.5) | 0.56 (12.8) | 1.48 (14.5) | 478 875 | 329 378 | 422 329 |

* Each scenario includes the other characteristics of the preceding scenario.

# Chapter 6

## Discussion

## 6.1 Introduction

Health policy makers are increasingly using CEA to inform resource allocation decisions. For CEA of many public health interventions, the best cost-effectiveness data come from CRTs. The analysis of patient-level cost-effectiveness data from CRTs raises many challenges which need to be addressed so that studies can provide sound evidence for policy making. Despite methodological progress in CEA (Glick et al., 2007, Willan and Briggs, 2006), the conception of this thesis recognised that little attention had been given to statistical methods for CEA that use CRTs (Flynn and Peters, 2005a, Willan, 2006). This thesis has helped address this gap in the literature.

The overall aim of the thesis was to identify appropriate statistical methods for CEA that use CRTs and assess their relative performance across a wide range of realistic settings. The specific objectives were:

1. To develop criteria for identifying appropriate statistical methods for CEA that use CRTs.
2. To critically appraise the methods used in applied CEAs that use CRTs.
3. To assess the relative performance of alternative statistical methods for CEA that use CRTs in settings where baselines covariates are balanced.
4. To compare alternative methods to adjust for systematic covariate imbalance in CEA that use CRTs.

The next section discusses the overall findings from the thesis. Sections 3 and 4 address the general contributions to the literature. Sections 5 and 6 summarise the limitations and identify areas for future research. Sections 7 and 8 discuss the implications for applied researchers and policy making. The last section provides the conclusion.

## 6.2 Overall findings of the thesis

Firstly, the conceptual review highlighted that statistical methods for CEA that use CRTs were required to address key statistical issues such as the clustering, correlation between costs and outcomes, skewed nature of cost data; and systematic imbalances in baseline covariates (research paper 2). The findings from the review informed a checklist for critical appraisal of the applied literature. This checklist found that most economic evaluations of cluster trials failed to adopt appropriate statistical methods (research paper 1). More specifically, the majority of applied CEAs using cluster trials did not recognise clustering in costs or health outcomes. Studies often justified the use of statistical methods that ignored clustering on the basis of low estimated ICCs. Research paper 2 demonstrated that methods that ignored clustering could underestimate uncertainty, even with low levels of clustering, and therefore that justification is inappropriate. Similarly, most of the studies did not recognise the correlation between costs and health outcomes, by conducting separate analyses of costs versus outcomes. Importantly, only four out of the 62 reviewed studies accounted for both clustering and correlation in the estimation of incremental costs and outcomes.

Secondly, while the review revealed that poor methods were being used in practice, the conceptual review identified four groups of methods that could address the main challenges in CEA that use CRTs: SUR and GEE, both using a robust estimation of the variance, MLMs, and a non-parametric TSB. The methods were first compared across a range of realistic scenarios with balanced covariates (research paper 2). While each method reported low levels of bias, methods differed in terms of CI coverage. MLMs and the TSB with a shrinkage correction performed well with CI coverage close to nominal levels across the scenarios considered. SUR and GEEs reported low CI coverage when the CRT had few clusters.

Thirdly, the methods were compared in circumstances where systematic differences in baseline characteristics were anticipated (research paper 3). The motivating example showed that cost-effectiveness estimates could differ according to the adjustment method. For example, the INB decreased by more than 50% once key anticipated confounders were adjusted for. After adjustment, methods reported similar mean cost-effectiveness estimates, but different uncertainty estimates. Simulations compared SUR, MLMs and the TSB combined with SUR for the covariate adjustment. All covariate-adjusted methods provided unbiased estimates across all scenarios considered. The TSB combined with SUR reported lower CI coverage than the other methods throughout. SUR reported lower rMSE and CI coverage than MLMs, in particular when the CRT had unequal cluster sizes, few clusters or there was a covariate by treatment interaction. MLMs performed best, reporting lower rMSE than the other methods and CI coverage close to nominal levels.

## 6.3 Main contributions of the thesis

### *6.3.1 Developing criteria for identifying appropriate methods for CEA that use CRTs and critical appraisal of applied literature*

This thesis developed a new checklist to critically appraise the methodological quality of CEA that use CRTs. By identifying important methodological flaws, this checklist provided a starting point for improving the methods used in practice. The checklist was accompanied by a methodological guideline to help future reviewers and researchers to judge the appropriateness of the methods adopted. The review also highlighted that there is room for improvement, not only in the methods used in practice, but also in the way potentially appropriate methods are reported.

### *6.3.2 Methodological insights on the relative merits of alternative methods for CEA that use CRTs with balanced covariates*

This thesis provided the first simulation study comparing alternative appropriate methods for CEA that use CRTs. A previous study (Flynn and Peters, 2005b) also conducted simulations in the context of CEA that use cluster trials but only considered bootstrap methods. Only one study has previously attempted to compare different methods for CEA that use CRTs (Bachmann et al., 2007). However, this study compared the methods in a case-study with many clusters, equal numbers per cluster and small ICCs. The simulations conducted in this thesis provided a more comprehensive testing ground for the methods and provided novel results, for example, showing differences across methods in settings with few clusters or unequal cluster sizes (research paper 2).

This thesis examined robust methods for the first time in the context of CEA that use CRTs. Research paper 2 considered SUR and GEEs with robust variance estimators, and showed that although these methods can be simple to implement, they may only be appropriate for CRTs which have at least a moderate number of clusters (15 per arm).

### *6.3.3 Comparative assessment of alternative methods for CEA that use CRTs with systematic imbalance in baseline covariates*

No previous work has addressed systematic covariate imbalance in CEA that use cluster trials. This thesis added to previous work on covariate adjustment for CEA based on RCTs (Nixon and Thompson, 2005, Willan et al., 2004, Hoch et al., 2002). A key contribution of this study is to consider circumstances where bias can arise in a systematic way, not by chance as considered in previous papers (Manca et al., 2005, Nixon and Thompson, 2005, Willan et al., 2004). When imbalance occurs by chance, adjustment is generally

recommended only when the predictor is anticipated to be strongly related to endpoints (Pocock et al., 2002) such as the baseline QALY (Manca et al., 2005). Unlike these findings, research paper 3 demonstrated that methods were required that adjust for systematic covariate imbalance even when the prognostic strength of the confounders is low, otherwise the results may be biased.

## 6.4 Other general methodological contributions emerging from the thesis

Findings from this thesis also contributed to current methodological debate on a number of more general themes in CEA. Specific insights were added to the following areas of knowledge: the use of robust methods in the analysis of hierarchical data; 2) methods that assume Normal distributions in settings with skewed cost data; and 3) non-parametric bootstrap methods for CEA.

### *6.4.1 The use of robust methods in the analysis of hierarchical data*

The plausibility of assumptions underlying robust variance methods for the analysis of hierarchical data depends largely on the sample size (Huber, 2004). In the context of CRTs, the crucial factor is the number of clusters randomised to each treatment arm. Research paper 2 found that in the context of CEA that use cluster trials, the robust methods generally required at least 10 clusters per arm to provide CI coverage above 0.9. Previous studies examining robust variance estimators for analysing clinical outcomes in CRTs, reported that a similar number of clusters were required for asymptotic properties to hold (Feng et al., 1996, Omar and Thompson, 2000). However, this thesis found circumstances where 10 clusters per arm were insufficient. For example, in CRTs with unequal cluster sizes, robust

estimators required at least 15 clusters per arm to report reasonable (above 0.9) CI coverage (research paper 2). This is an important finding as most applied CEAs that use cluster trials have unequal numbers per cluster (research paper 1). Research paper 3 also showed that when covariate adjustment is required, SUR with the robust SE provided poor CI coverage (below 0.9) even with 20 clusters per treatment arm.

### *6.4.2 Methods that assume Normal distributions in settings with skewed cost data*

The conceptual review emphasised that methods which assume data are Normally distributed may not be appropriate for the analysis of skewed costs (Jones, 2000, Mihaylova et al., 2011, Briggs et al., 2005, Manning, 2006). By contrast, this thesis showed that methods which assume a Normal distribution for costs in the context of CEA that use CRTs performed relatively well. For example, research paper 2 and 3 found that a bivariate Normal MLM provided good CI coverage even when costs were highly skewed. Research paper 3 demonstrated that when the bivariate MLMs allowed for a Gamma distribution for costs, it led to little improvement in CI coverage or precision. Similarly, the performances of SUR and GEEs, which have also assumed Normal costs, were fairly similar between scenarios with skewed and Normal costs. This adds to previous findings suggesting that, in CEA, methods that assume data are from a Normal distribution may be quite robust to skewed costs (Thompson and Barber, 2000, Willan et al., 2004).

### *6.4.3 Non-parametric bootstrap methods in CEA*

This thesis provided important methodological insights to the current debate on the appropriateness of non-parametric bootstrap methods for CEA (Barber and Thompson, 2000,

Nixon et al., 2010, O'Hagan and Stevens, 2003, Flynn and Peters, 2005b). Research paper 2 extended seminal work on the non-parametric TSB (Davison and Hinkley, 1997) to allow for circumstances where there was variation in the cluster size. Unlike previous findings (Flynn and Peters, 2005b), this method provided good CI coverage, even when the CRT had few clusters. The TSB proposed here was implemented with a shrinkage correction, recommended more generally to improve the precision of the estimates (Davison and Hinkley, 1997: page 102). The simulations showed that when, as in a previous study (Bachmann et al., 2007), the TSB is applied without the shrinkage estimator, it overestimates the variance.

A rationale for using non-parametric bootstrap methods in CEA is that they can avoid distributional assumptions, which may have particular advantages when the parametric form of the data is unknown, as is typical for costs (Briggs et al., 1999, Chaudhary and Stearns, 1996, Mullahy and Manning, 1994). Research paper 2 showed that the TSB reported good CI coverage, even if costs were highly skewed. Previous studies have considered TSB only for the estimation of confidence intervals to characterise the uncertainty around cost-effectiveness estimates (Flynn and Peters, 2004, Flynn and Peters, 2005b). The TSB considered here is used to estimate both the mean (Barber and Thompson, 2000) and corresponding CIs of the parameter of interest (INB). The simulations demonstrated that the means calculated from the bootstrap samples were unbiased estimates of the mean INB.

Research paper 3 considered, for the first time, a non-parametric bootstrap method for covariate adjustment in CEA. This study showed that when covariate adjustment was required to adjust for systematic covariate imbalance, the non-parametric bootstrap performed poorly. This paper proposed a combination of TSB with a regression-based method, for example SUR to adjust for the confounding. While this improved approach

provided unbiased estimates of the parameter of interest, the CI coverage remained poor across the scenarios considered.

## 6.5 Limitations

While this thesis presented a comprehensive assessment and comparison of alternative methods for CEA that use CRTs, it has some limitations. In this section, I acknowledge general weaknesses of the thesis regarding the criteria for critical appraisal of CEA that use CRTs, breadth of methods, and range of circumstances considered.

### *6.5.1 Criteria for critical appraisal of CEA that use CRTs*

This thesis developed a checklist for critical appraisal of CEA that use CRTs based on a conceptual review of the methods literature conducted at the outset. As with more general methodological guidelines, the checklist should be updated to recognise future methodological developments. For example, it was recognised later in the conceptual review that CEA that use CRTs needed to assess whether there was any anticipated systematic imbalance in baseline covariates. To consider this point, the following criterion could be added to a future update of the checklist developed in research paper 1:

*"Did the study assess the balance of baseline covariates a priori anticipated as potential confounders, and use an appropriate method to correct for anticipated confounding?"*

The conceptual review also identified other specific issues that could arise in the statistical analyses. For example, data may be subject to censoring (e.g. individuals lost to follow-up) or missing (e.g. due to patient non-response). However, the aim of the checklist is to critically

appraise whether central statistical issues arising in CEA that use CRTs have been adequately addressed and did not attempt to cover all aspects of the analyses. The checklist is therefore intended to supplement rather than replace more general methodological guidelines for trial-based CEA (Glick et al., 2007, Willan and Briggs, 2006), where particular issues such as censoring and missing data are addressed.

### *6.5.2 Range of methods considered for assessment*

The thesis identified a comprehensive range of methods for empirical investigation. Each method could address all key statistical issues faced by CEA that use CRTs, and hence was judged potentially appropriate for this context. The review also identified other methods that satisfied only some of the criteria but could still improve current practice. For example, the net benefit regression framework (Hoch et al., 2002) could be implemented with a robust estimation of the variance or with cluster-level random effects to account for the clustering. However, as the conceptual review highlighted, the assumptions underlying this approach are less plausible than the methods considered for CEA that use CRTs.

The non-parametric TSB considered here was one of many possible alternative non-parametric bootstrapping approaches. For example, another bootstrap approach could be to sample individual and cluster-level residuals from the bivariate MLM (Carpenter et al., 2003). However, this approach also requires the use of a shrinkage correction, which can be complex to implement for bivariate models (Carpenter et al., 2003).

In circumstances where covariate adjustment is required, the methods considered in this thesis assumed no unobserved confounding. Methods such as the instrumental variable estimation (Basu et al., 2007, Heckman and Navarro-Lozano, 2004), which could adjust for

systematic imbalances in unobserved baseline characteristics, may warrant future investigation.

### *6.5.3 Range of circumstances considered*

This thesis compared the alternative methods across a wide range of settings typically observed in CEA that use CRTs. However, it did not cover all possible circumstances that could potentially arise in CEA that use CRTs. For example, the design of the simulations considered constant variances for costs and outcomes between treatment groups and across clusters. Costs and outcomes often exhibit systematic variations across clusters that may not be fully captured in the cluster-level random effects (Turner et al., 2001). In these circumstances, allowing the variances to differ by treatment or across clusters may be more appropriate (Grieve et al., 2010). Another source of the heterogeneity across clusters is when the treatment or covariates act multiplicatively on endpoints; for example, the covariate effect multiplies the endpoints by a cluster-specific factor (Barber and Thompson, 2004, Manning and Mullahy, 2001). Unlike the linear additive models considered across all scenarios, multiplicative models might have been more appropriate, for example to address high heterogeneity in skewed cost data (Thompson et al., 2006).

Another limitation in the estimation of costs is that the methods considered in this thesis accounted for the clustering in the overall estimation of the incremental cost. However, clustering may affect each cost component differently. For example, there may be circumstances where although clinical protocols encourage high level of variation between patients, clusters may manage particular groups of patients similarly, and hence, yielding a low ICC for some resource use items. By contrast, there may be different financial incentives between clusters, leading to a high level of variation in unit costs across clusters. Modelling

different health care resource use items before applying unit costs may improve the estimate of individual cost components (Gauthier et al., 2009). Another potential advantage is that it can allow for alternative models for different types of resource use (e.g. continuous and count endpoints) (Conti et al., 2007). The MLMs considered in this thesis could be extended to allow for the estimation of individual cost components, and implemented in a Bayesian framework using WinBUGs. However, the joint estimation of individual cost components and health outcomes can prove difficult. This would require complex Bayesian methods, which can address the correlation across cost components, the correlation between costs and health outcomes and the clustering (Lambert et al 2008). In addition, it has been highlighted that while modelling resource use has potential to improve estimation of individual cost components, it is unclear whether combining these estimates will provide an unbiased estimate of the mean total cost (Gauthier et al 2009).

Both case-studies and simulations in research papers 2 and 3 only compared the methods across scenarios with complete cost and outcome data. However, cost-effectiveness data from CRTs may be incomplete, for example missing due to individual non-response. Comparing methods under complete-case analysis may only be appropriate when data are missing completely at random, i.e. the missing mechanism does not depend on either observed or unobserved factors (Little and Rubin, 1987). However, data are often missing at random, i.e. when the missing mechanism depends on observable variables, and here applying complete case-analysis may lead to incorrect inferences (Briggs et al., 2003). Accounting for missing data in CEA that use CRTs can be difficult because the missing mechanism may differ between clusters or treatment groups, and assuming the same imputation model for the whole sample may be incorrect (Little and Rubin, 1987). In addition, missingness may be associated with both endpoints, and handling the missing data in the joint estimation of costs and outcomes can be complex (Lambert et al., 2008).

Methods were compared in terms of their ability to estimate the overall average treatment effect (ATE), as policy makers often make recommendations about a particular intervention for the overall population (Claxton, 1999, Imbens and Wooldridge, 2009, Jones and Rice, 2011). However, there may be circumstances where heterogeneous subgroups of patients respond differently to treatment, and making the same recommendation the whole population based on the overall ATE may not be appropriate (Vaness and Mullahy, 2006, Sculpher, 2008). Research papers 2 and 3 did not consider settings where treatment effects differed by subgroups of patients. While research paper 3 considered prognostic relationships that differ by treatment group (Gelman and Pardoe, 2007, Liu and Gustafson, 2008), the aim was still to provide an overall ATE rather than effects by subgroups. Nonetheless, any of the methods considered in this thesis could allow for the estimation of different subgroup effects.

## 6.6 Areas of further research

This thesis identified some areas that were potentially worthy of further investigation: comparison of the methods under more complex circumstances; assessment of the impact of method choice on long-term cost-effectiveness using decision models; and development of a general analytical strategy for CEA that use CRTs.

### *6.6.1 Comparison of the methods under more complex circumstances*

The nature of cost-effectiveness data from CRTs can pose further challenges, which may need to be taken into consideration in future research on methods for CEA that use CRTs. For example, comparing methods in circumstances where variance varies across clusters or between treatment arms may be warranted (Omar and Thompson, 2000, Turner et al., 2001).

Similarly, correlation between costs and outcomes may be different between treatment arms. Methods that can better accommodate these more complex variance structures, such as MLMs (Grieve et al., 2010) may provide relative advantages over alternative approaches.

In addition, both costs and outcomes are assumed to be continuous and defined as linear functions of treatment and other covariates, across all scenarios. Investigation of the performance of the methods when endpoints are binary and show non-linear relationships with covariates is warranted. More specifically, when the distribution model form is a concern, more complex Bayesian model-averaging approaches may be preferred to address this structural uncertainty (Conigliani and Tancredi, 2009). In addition, a method which can offer protection against misspecification of the structural model for endpoints would be to extend the MLMs to doubly robust estimation (Bang and Robins, 2005, Imbens, 2004).

Methodological concerns raised by ignoring missing data in CEA that use CRTs, encourage further research on this area. In the context of CEA that use CRTs, the use of multilevel multiple imputation (Carpenter et al., 2011) may be required to take into account the full uncertainty associated with the missing values. Multilevel multiple imputation is an extension of standard multiple imputation (Rubin, 1987) which allows for the variability of the imputed data to reflect the multilevel structure of the data. Another alternative in the context of clustered data is to use weighted estimators, usually implemented with GEEs (Rotnitzky and Robins, 1997).

### *6.6.2 Assessment of the impact of method choice on long-term cost-effectiveness using decision models*

The conceptual review highlighted the fact that the key methodological concerns in CEA that use CRTs affect not only studies that used data from a single CRT, but also studies that

combined that data with other evidence in a decision model. The scope of the empirical investigations was, however, limited to CEA that used data from a single CRT. Differences between alternative statistical methods for analysing patient-level cost-effectiveness data are likely to translate directly into differences in both the mean and uncertainty for input parameters in decision models (Briggs et al., 2006). These differences would also be expected to impact on the expected value of information, as illustrated in the case study of research paper 2. The use of Bayesian methods, which can incorporate prior evidence when estimating the parameters of interest, may be preferred in this context.

### *6.6.3 Development of a general analytical strategy for CEA that use CRTs*

This thesis used two case-studies to examine whether discrepancies between methods identified in the simulations are translated into different cost-effectiveness inferences. The results demonstrated that the choice of method could matter in practice. There is scope to consider further case-studies which may exhibit additional pragmatic circumstances such as complex data distributions or further levels of the hierarchy (e.g. GP practices within cities). Additional studies would offer further insights to help develop an analytical strategy for CEA that use CRTs. Such a strategy should help inform, for example, the circumstances in which more complex methods are more appropriate than simpler alternative approaches.

## 6.7 Recommendations for applied researchers

Grounded in the conceptual review and appraisal of applied studies, research papers 2 and 3 offered an assessment of alternative methods across a wide range of realistic scenarios commonly seen in practice. This broad testing allowed the thesis to provide methodological

insights and practical recommendations for applied analysts on the future use of alternative methods in CEA that use CRTs.

This work provides specific guidance on the use of robust methods in the context of CEA that use CRTs. These are not recommended unless the CRT has at least 10 clusters per treatment arm. However, larger samples may be needed when the CRT has unequal cluster sizes or when systematic covariate imbalance is anticipated. Recent studies have proposed an adjusted robust estimator for small samples (Skene and Kenward, 2010, Pan and Wall, 2002), which can help improve precision.

This thesis also offers some insights about the use of methods based on assuming Normal distributions for both costs and outcomes. It is suggested that the bivariate Normal MLMs perform well for CEA that use cluster trials, even when costs are skewed. This has practical advantages as these methods are relatively simple to implement and are available in a wider range of statistical packages. While the scenarios considered aimed to reflect realistic settings (Bachmann et al., 2007, Grieve et al., 2010), some interventions may have more complex cost structures, for example, with heavier tails. In these circumstances, applied researchers should assess whether more appropriate distributions, such as the Lognormal and inverse Gaussian, can lead to more precise estimates.

To encourage the use of appropriate statistical methods in practice, user-friendly software for implementing alternative methods is provided in research paper 2 (Appendix 4). The SUR was implemented in STATA because the cluster-robust estimator is readily available to use with the method (package *nlsur*). All other methods were implemented in R but they can be used with conventional software. For example, code for implementing TSB in STATA has also been developed. Similarly, bivariate MLMs considered here can be applied in MLWin (Rasbash et al., 2004) or STATA (package *gllamm*). Each method is relatively easy to

implement, and is not computationally expensive. The non-parametric TSB, even with 2000 replications, takes less than 10 minutes to complete the estimation.

## 6.8 Implications for policy making

CEA that use CRTs can provide valuable evidence to help inform decision making about resource allocation in health care. Decision makers should, however, be aware that studies which fail to use appropriate methods may produce misleading results and lead to inappropriate decisions. This thesis provides methods to help improve CEA that use CRTs so that future studies can better inform policy making.

Decision makers such as NICE provide methodological guidelines for the evaluation of health care interventions (NICE, 2008), but these do not offer guidance on how issues raised in CEA that use CRTs should be addressed. Decision makers should incorporate the new criteria developed in the cluster-specific checklist into future methodological guidelines. For example, one area where the criteria developed here can be integrated is in the evaluation of public health interventions (NICE, 2009), where methods guides have received relatively less attention than for health care technologies appraisal (Weatherly et al., 2009, Kelly et al., 2005).

Although the main focus of the thesis was on issues concerning the statistical analyses in CEA that use CRTs, some considerations can be drawn for the design of future CRTs. For example, one of the main challenges for analysts is that CRTs typically have few clusters. Simulation work suggested that the gains from increasing the number of clusters randomised to each treatment arm seem larger than increasing the number of individuals recruited for each cluster. Another important consideration is the variation in the size of the clusters.

Future cluster trials are encouraged to avoid the cluster sizes to be too uneven, particularly when the actual dimension of the cluster is anticipated to be associated with the endpoints.

## 6.7 Conclusion

The overall aim of the thesis was to help address the lack of work on methods for CEA that use CRTs. A primary objective of the thesis was to develop criteria for critical appraisal of economic evaluations that use data from CRTs. The checklist showed that applied studies adopted poor statistical methods. The conceptual review highlighted that unless these methods were improved, these studies would not provide a strong basis for policy making.

The thesis identified potentially appropriate statistical methods for CEA that use CRTs and assessed their relative performance across a wide range of realistic settings. The results from simulations and case studies provide convincing evidence that MLMs are the most appropriate method for CEA that use CRTs. The thesis also offers specific recommendations on when alternative methods such as robust variance methods or bootstrap approaches could perform well. The thesis provides an important contribution to the development of analytical methods for CEA that use CRTs, and identifies potentially fruitful areas for future research.

## References


Bachmann, M. O., Fairall, L., Clark, A. & Mugford, M. 2007. Methods for analyzing cost effectiveness data from cluster randomized trials. *Cost Eff Resour Alloc,* 5, 12.

Bang, H. & Robins, J. M. 2005. Doubly robust estimation in missing data and causal inference models. *Biometrics,* 61, 962-73.

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

Barber, J. A. & Thompson, S. G. 2000. Analysis of cost data in randomized trials: an application of the non-parametric bootstrap. *Statistics in Medicine,* 19, 3219-3236.

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.

Briggs, A., Clark, T., Wolstenholme, J. & Clarke, P. 2003. Missing... presumed at random: cost-analysis of incomplete data. *Health Econ,* 12, 377-92.

Briggs, A., Claxton, K. & Sculpher, M. 2006. *Decision modelling for health economic evaluation,* Oxford, UK, Oxford University Press.

Briggs, A., Nixon, R., Dixon, S. & Thompson, S. 2005. Parametric modelling of cost data: some simulation evidence. *Health Econ,* 14, 421-8.

Briggs, A. H., Mooney, C. Z. & Wonderling, D. E. 1999. Constructing confidence intervals for cost-effectiveness ratios: an evaluation of parametric and non-parametric techniques using Monte Carlo simulation. *Stat Med,* 18, 3245-62.

Carpenter, J., Goldstein, H. & Kenward, M. 2011. REALCOM-IMPUTE software for multilevel multiple imputation with mixed response types. *Journal of Statistical Software* (in press).

Carpenter, J. R., Goldstein, H. & Rasbash, J. 2003. A novel bootstrap procedure for assessing the relationship between class size and achievement. *Journal of the Royal Statistical Society Series C-Applied Statistics,* 52, 431-443.

Chaudhary, M. A. & Stearns, S. C. 1996. Estimating confidence intervals for cost-effectiveness ratios: An example from a randomized trial. *Statistics in Medicine,* 15, 1447-1458.

Claxton, K. 1999. The irrelevance of inference: a decision-making approach to the stochastic evaluation of health care technologies. *J Health Econ,* 18, 341-64.

Conigliani, C. & Tancredi, A. 2009. A Bayesian model averaging approach for cost-effectiveness analyses. *Health Econ,* 18, 807-21.

Conti S, Manca A, Lambert P, et al. Bayesian multivariate modelling of patient level healthcare resource use data in RCTs. International Health Economics Association (iHEA); 2007 Jul 9; Copenhagen.

Davison, A. C. & Hinkley, D. V. 1997. *Bootstrap methods and their application,* Cambridge, UK, Cambridge University Press.

Feng, Z. D., Mclerran, D. & Grizzle, J. 1996. A comparison of statistical methods for clustered data analysis with Gaussian error. *Statistics in Medicine,* 15, 1793-1806.

Flynn, T. & Peters, T. 2005a. Conceptual issues in the analysis of cost data within cluster randomized trials. *J Health Serv Res Policy,* 10, 97-102.

Flynn, T. N. & Peters, T. J. 2004. Use of the bootstrap in analysing cost data from cluster randomised trials: some simulation results. *Bmc Health Services Research,* 4, 33-43.

Flynn, T. N. & Peters, T. J. 2005b. Cluster randomized trials: Another problem for cost-effectiveness ratios. *International Journal of Technology Assessment in Health Care,* 21, 403-409.

Gauthier, A., Manca, A. & Anton S. Bayesian Modelling of healthcare resource use in multinational randomized clinical trials. *Pharmacoeconomics,* 27 (17): 1017: 1029.

Gelman, A. & Pardoe, I. 2007. Average Predictive Comparisons for Models with Nonlinearity, Interactions, and Variance Components. *Sociological Methodology 2007, Vol 37,* 37, 23-51.

Glick, H. A., Doshi, J. A., Sonnad, S. S. & Polsky, D. 2007. *Economic evaluation in clinical trials,* Oxford, UK, 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. *Med Decis Making,* 30, 163-75.

Heckman, J. & Navarro-Lozano, S. 2004. Using matching, instrumental variables, and control functions to estimate economic choice models. *Review of Economics and Statistics,* 86, 30-57.

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 Econ,* 11, 415-30.

Huber, P. J. 2004. *Robust statistics,* New York ; Chichester, Wiley.

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

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

Jones, A. 2000. Health Econometrics. *In:* CULYER, A. & NEWHOUSE, J. (eds.) *Handbook of Health Economics.* Amsterdam: Elsevier.

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

Kelly, P. M., Mcdaid, D., Ludbrook, A. & Powell, J. 2005. Economic appraisal of public health interventions. *NHS Health Development Agency* Briefing Paper.

Lambert, P. C., Billingham, L. J., Cooper, N. J., Sutton, A. J. & Abrams, K. R. 2008. Estimating the cost-effectiveness of an intervention in a clinical trial when partial cost information is available: a Bayesian approach. *Health Econ,* 17, 67-81.

Little, R. J. A. & Rubin, D. B. 1987. *Statistical analysis with missing data,* New York ; Chichester, Wiley.

Liu, J. X. & Gustafson, P. 2008. On Average Predictive Comparisons and Interactions. *International Statistical Review,* 76, 419-432.

Manca, A., Hawkins, N. & Sculpher, M. J. 2005. Estimating mean QALYs in trial-based cost-effectiveness analysis: the importance of controlling for baseline utility. *Health Econ,* 14, 487-96.

Manning, W. 2006. Dealing with skewed data on costs and expenditures. *In:* JONES, A. (ed.) *The Elgar Companion to Health Economics.* Cheltenham, UK: Edward Elgar.

Manning, W. G. & Mullahy, J. 2001. Estimating log models: to transform or not to transform? *Journal of Health Economics,* 20, 461-494.

Mihaylova, B., Briggs, A., O'hagan, A. & Thompson, S. G. 2011. Review of statistical methods for analysing healthcare resources and costs. *Health Econ,* 20, 897-916.

Mullahy, J. & Manning, W. 1994. Statistical issues in cost-effectiveness analysis. *In:* SLOAN, F. (ed.) *Valuing Health Care.* Cambridge, UK: Cambridge University Press.

NICE 2008. Methods for Technology Appraisal. *National Institute for Health and Clinical Excellence,* London, UK.

NICE 2009. Methods for development of NICE public health guidance. *National Institute for Health and Clinical Excellence.,* London, UK.

Nixon, R. M. & Thompson, S. G. 2005. Methods for incorporating covariate adjustment, subgroup analysis and between-centre differences into cost-effectiveness evaluations. *Health Econ,* 14, 1217-29.

Nixon, R. M., Wonderling, D. & Grieve, R. D. 2010. Non-parametric methods for cost-effectiveness analysis: the central limit theorem and the bootstrap compared. *Health Econ,* 19, 316-33.

O'Hagan, A. & Stevens, J. W. 2003. Assessing and comparing costs: how robust are the bootstrap and methods based on asymptotic normality? *Health Economics,* 12, 33-49.

Omar, R. Z. & Thompson, S. G. 2000. Analysis of a cluster randomized trial with binary outcome data using a multi-level model. *Stat Med,* 19, 2675-88.

Pan, W. & Wall, M. M. 2002. Small-sample adjustments in using the sandwich variance estimator in generalized estimating equations. *Statistics in Medicine,* 21, 1429-1441.

Pocock, S. J., Assmann, S. E., Enos, L. E. & Kasten, L. E. 2002. Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: current practice and problems. *Stat Med,* 21, 2917-30.

Rasbash, J., Steele, F., Browne, W. & Goldstein, H. 2004. A user's guide to MLwinN. Version 2.10. *Centre for Multilevel Modelling,* University of Bristol.

Rotnitzky, A. & Robins, J. 1997. Analysis of semi-parametric regression models with non-ignorable non-response. *Statistics in Medicine,* 16, 81-102.

Rubin, D. 1987. *Multiple imputation for nonresponse in surveys,* New York, US, Wiley.

Sculpher, M. 2008. Subgroups and heterogeneity in cost-effectiveness analysis. *Pharmacoeconomics,* 26, 799-806.

Skene, S. S. & Kenward, M. G. 2010. The analysis of very small samples of repeated measurements II: a modified Box correction. *Stat Med,* 29, 2838-56.

Thompson, S. G. & Barber, J. A. 2000. How should cost data in pragmatic randomised trials be analysed? *British Medical Journal,* 320, 1197-1200.

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.

Turner, R. M., Omar, R. Z. & Thompson, S. G. 2001. Bayesian methods of analysis for cluster randomized trials with binary outcome data. *Stat Med,* 20, 453-72.

Vaness, D. & Mullahy, J. 2006. Perspectives on mean-based evaluation of health care. *In:* JONES, A. (ed.) *The Elgar Companion to Health Economics.* Cheltenham, UK: Edward Elgar.

Weatherly, H., Drummond, M., Claxton, K., Cookson, R., Ferguson, B., Godfrey, C., Rice, N., Sculpher, M. & Sowden, A. 2009. Methods for assessing the cost-effectiveness of public health interventions: key challenges and recommendations. *Health Policy,* 93, 85-92.

Willan, A. 2006. Statistical Analysis of cost-effectiveness data from randomised clinical trials. *Expert Revision Pharmacoeconomics Outcomes Research,* 6, 337-346.

Willan, A. & Briggs, A. 2006. *Statistical Analysis of cost-effectiveness data,* Chichester, UK, 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 Econ,* 13, 461-75.