Work in progress – not for citation without authors' permission

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

Manuel Gomes[a], Richard Grieve[a] and John Edmunds[b c]

Corresponding author:

Manuel Gomes

PhD student in Health Economics

Department of Public Health and Policy,

London School of Hygiene and Tropical Medicine,

Room G5, 51 Bedford Square, London. WC1B 3DP

Email: manuel.gomes@lshtm.ac.uk Tel: 02072994734

a HSRU – Department of Public Health and Policy, LSHTM.

b IDEU – Department of Epidemiology and Population Health, LSHTM.

c Health Protection Agency, London.

Paper presented at the Health Economics Study Group, London School of Economics, January 2010.

Abstract

Introduction

Policy makers require economic evaluations of group-level interventions to decide for example, which public health programmes to provide. The best data for economic evaluations of group-level interventions often comes from cluster randomised trials (CRTs), where randomisation is by cluster (e.g. hospitals, GP practices) not individual. While statistical methods for cost-effectiveness analyses (CEA) are well developed, methods for studies that use CRTs have received little attention. General checklists for critically appraising CEA may be insufficient for economic evaluations that use CRTs. This paper develops and applies a new checklist for critical appraisal of economic evaluations that use data from CRTs.

Methods

We reviewed statistical methods for CRTs and CEA. We used the findings from this conceptual review to develop a checklist for critical appraisal of economic evaluations that use CRTs. This checklist included criteria such as whether the analysis allowed for both clustering and the correlation between costs and effects. We applied this checklist alongside the 'Drummond checklist' in a structured literature review to appraise the quality of published economic evaluations that use CRTs.

Results

We identified 47 full economic evaluations that used data from CRTs. On average, these studies met 90% of the criteria in the Drummond checklist. However, 40% of the studies totally ignored clustering in the analysis, and 90% disregarded any correlation between costs and effects. Only one study used appropriate statistical methods that allowed for clustering and correlation in the joint estimation of costs and effects.

Conclusions

The quality of published studies was 'good' when assessed against a general checklist. However, most studies were judged of 'poor' quality when appraised against our more specific criteria. This new checklist can supplement generic checklists and raise awareness of poor research practice. Statistical methods that account for both clustering and correlation between costs and effects are available, and their use should be encouraged. Current research is investigating the relative performance of alternative statistical methods for CEA that use CRTs.

Introduction

Policy-makers worldwide require economic evaluations to decide which health care technologies to provide [1-4]. Economic evaluations are also used to recommend which public health interventions to prioritise [5], and to evaluate different ways of organising health services [6, 7]. 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 different health care technologies [8-11]. However, for the evaluation of group-level interventions, a cluster randomised trial (CRT), where the unit of randomisation is the 'cluster' (e.g. NHS Trust, GP practice, geographical area), is often preferable. 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., between different strategies to reduce hospital acquired infection) [12]. However, the use of CRTs in economic evaluation raises important methodological issues which have received little attention.

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. The costs or outcomes for patients within each cluster will be therefore be correlated [13]. General guidance for the design and analysis of CRTs clearly encourages researchers to recognise clustering [12-17]. However, the structure of cost-effectiveness data raises additional challenges. In particular, statistical methods in economic evaluations need to recognise clustering but also the correlation between costs and effects [18-20]. The methods should also be sufficiently flexible to accommodate cost and effectiveness data with non-normal distributions [11, 21-23]. While appropriate statistical methods for CEA that use data from CRTs have been developed [24, 25], it is unclear whether these methods are used in practice. These methodological concerns are relevant for economic evaluations based on single CRTs, but also studies that use data from CRTs in decision-analytical models. Ignoring the clustering will underestimate the statistical uncertainty surrounding the results and can also lead to biased point estimates [26-28].

Concerns about methodological standards in CEA have provoked policy-makers and academics to develop a plethora of methodological guidelines and critical appraisal criteria [9, 29-34]. These initiatives aim to improve the uptake of appropriate methods and reporting transparency. Generic checklists [9, 29, 32] can help researchers, peer reviewers, and journal editors to gauge studies' methodological quality. However, these general checklists do not include sufficiently detailed criteria for appraising health economic evaluations with

particular designs such as decision-analytic models, and hence additional checklists were required [33, 35]. For economic evaluations from CRTs, the general checklists do not cover fundamental issues raised by the hierarchical nature of the data, and a more specific tool for appraising the quality of these studies is warranted.

This paper aims to develop a new checklist for economic evaluations that use data from CRTs and apply it to critically assess published studies. The next section presents the key concepts that underly the criteria in the new checklist, and describes the methods used to critically appraise published economic evaluations. We then present the results of the review, discuss the findings, the importance of the proposed checklist and the implications for future research.

Methods

To develop criteria for assessing the methodological quality of economic evaluations that use CRTs we considered the statistical issues that arise with this study design. We therefore undertook a conceptual review that covered methodological guidance for analysing CRTs [12, 13, 16, 36-38], and statistical methods for economic evaluation [8-11, 39].

Conceptual review

The conceptual review emphasised that the form of clustering in CRTs is different and potentially stronger than in multicentre RCTs where patients are individually randomised [40, 41]. In CRTs, all individuals within a cluster receive the same randomised treatment, whereas although data in multicentre RCTs may be clustered within centres, individuals are randomised to different treatments within each centre. The specific form of clustering in CRTs needs to be anticipated, and accounted for in the sample size calculation, otherwise the study will be underpowered [17]. The clustering also needs to be recognised in the statistical analysis, or else the study will underestimate the statistical uncertainty surrounding the results. Ignoring clustering can also provide biased estimates, particularly when the CRT has unequal numbers per cluster (imbalance) and there is a relationship between cluster size and the mean outcome in each cluster [27]. While the importance of recognising this form of clustering has been well recognised in the medical statistics literature, it has received little attention in methods for economic evaluation [39].

The particular characteristics of cost and effectiveness data have further implications for the design and analysis of CEA that use CRTs [10, 11]. Firstly, CEA tend to use data from trials (RCTs or CRTs) that have sample sizes calculated to detect differences in clinical endpoints. Conceptual issues from basing sample size calculations on costs, with their larger variances,

are well acknowledged in the literature [42-44]. However, sample size calculations for CEA would ideally consider the variances of both costs and effects, with both inflated to anticipate clustering [45]. Certainly, the study should anticipate *a priori* the impact of clustering in both costs and effects on the width of the uncertainty intervals that will eventually be reported [46].

Secondly, as cost function theory suggests there may be relatively high levels of heterogeneity in resource use, unit costs and efficiency across clusters [47, 48]. This leads to higher intra-cluster correlation coefficients (ICCs)[1] than those for clinical outcomes, which can have greater implications for biasedness and precision of the cost and cost effectiveness estimations if clustering is neglected [27]. Thirdly, as production function theory suggests, costs and effects are often correlated at the individual level. For example, patients with worse health typically incur in greater costs. This is a central issue both for CEAs based only on CRTs and for decision-analytical models. Methods must acknowledge the correlation between costs and effects, when reporting mean incremental cost-effectiveness and measures of statistical uncertainty [18-20, 49]. Fourthly, commentators have emphasised the importance of using statistical methods that make correct assumptions about the distributions of the data [11, 21-23, 39, 50, 51]. Studies that make inappropriate assumptions, for example that costs have normal or lognormal distributions when they have a gamma distribution, can report imprecise or biased estimates [21].

In summary, CEA based on CRTs should use statistical methods that recognise both the clustering inherent with CRTs and the correlation between costs and effects (see Figure 1, area D). Studies may fail to meet these criteria for several reasons: studies may account for correlation but not clustering (area C). For example, studies may use seemingly unrelated regression [20, 52] and allow for correlation between costs and effects, but ignore clustering. Studies may account for clustering in both costs and effects but assume they are uncorrelated (area B), for example by estimating incremental costs and effects with separate multilevel models. Alternatively, the study may fail to account appropriately for clustering and correlation (area A). For example, the study may completely neglect clustering and correlation, or only allow for clustering in one of the univariate measures (e.g. incremental effects). While it is fundamental that studies use methods that lie in area D, they are also required to make appropriate assumptions about the distribution of the data.

<< Figure 1 here >

Appropriate Statistical methods for CEA

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

Methods based on cluster-level summary statistics (e.g. two-sample *t*-test, Wilcoxon rank sum test) or non-parametric tests adjusted for clustering (e.g. adjusted χ^2-test, ratio estimator) may be appropriate for analysing clinical outcome data from CRTs [12, 13, 16], but they lack sufficient flexibility to address all the core statistical issues in CEA from CRTs (e.g. correlation between individual costs and effects). The conceptual review identified flexible statistical methods that could handle clustering and correlation, whilst also accommodating data with non normal distributions, the three groups of methods were: multilevel models [53, 54], generalised estimating equations [55] and the two-stage, non-parametric bootstrap [56].

Multilevel models (MLMs)

MLMs can easily accommodate the hierarchical structure of data from CRT [19, 57]. The Model (1) illustrates the flexibility of MLMs for CEA that use CRTs. The clustering explicitly by including random parameters (u_j^c, u_j^e) to account for the cluster-specific effects. The correlation between the individual costs and effects is recognised by the parameter θ. The model can then report incremental costs (β_1^c) and effects (β_1^e) after allowing for the clustering and correlation.

$$\begin{aligned} c_{ijk} &\sim dist(\mu_{ijk}^c, \sigma_c) & \mu_{ijk}^c &= \beta_0^c + \beta_1^c t_k + u_j^c \\ e_{ijk} &\sim dist(\mu_{ijk}^e, \sigma_e) & \mu_{ijk}^e &= \beta_0^e + \beta_1^e t_k + u_j^e + (\beta_2 + \theta)(c_{ijk} - \mu_{ijk}^c) \end{aligned} \quad (1)$$

Model (1) can be estimated under a Frequentist approach, generally implemented by maximum likelihood and least squares, or Bayesian perspective typically implemented by MCMC. Conceptually, either approach can be appropriate for CEA from CRTs, but the wide range of models that can be fitted in WinBUGS means that implementing the model from a Bayesian perspective currently offers greater flexibility [58]. However, Bayesian estimates may be sensitive to the choice of priors, and models may be more complex to implement [59].

Generalised Estimating Equations (GEEs)

GEEs are a flexible extension of likelihood-based generalised linear models (GLMs) to accommodate clustered data [55, 60]. GEEs take a marginal (population-averaged) rather than a conditional approach, which means they can be relatively more efficient and robust to misspecification of the covariance structure [61], and have been used to analyse CRTs [38, 62-64]. Bivariate GEEs have been developed to recognise correlations between different

outcomes [65]. A concern with GEEs is whether the underlying asymptotic assumptions are plausible for CEA alongside CRTs, particularly when there are few clusters [63, 66].

Two-Stage Bootstrap (TSB)

The TSB is a non-parametric sampling approach which involves resampling clusters (with replacement) and then individuals within each sampled cluster [56, 67]. This two-stage procedure accounts for clustering by recognising sampling variation within and between clusters. The TSB can account for the correlation between costs and effects by sampling them in pairs, and it also avoids assuming the data are drawn from specific parametric distributions [24, 25]. As with GEEs, statistical inference with the TSB relies on asymptotic assumptions [68]. An additional concern is whether the TSB can still provide unbiased, efficient estimates when the numbers per cluster vary widely, i.e. the CRT is unbalanced [25].

Each of these estimation methods can be implemented in bivariate models which estimate incremental costs and effects whilst allowing for clustering in their joint distribution. Alternatively, the correlation between individual costs and effects could be recognised by calculating a univariate measure of net benefit for each individual [18]. Any of the estimation methods described above could then be used to allow for clustering in the net benefits. This approach does recognise correlation and clustering and would be defined as 'type D'. However, this univariate method is more restrictive than keeping costs and outcomes on their original scale [19, 20]. For instance, it is necessary to assess whether correct distributional assumptions have been made about the component measures of cost and effects, but also about the distribution of the net benefit.[2]

Proposed checklist for economic evaluations that use data from CRTs

Based on the conceptual review, we developed criteria for assessing the methodological quality of economic evaluations that use data from CRTs. Provisional versions of the checklist were reviewed by a panel of experts (medical statisticians, health economists, epidemiologists), and the revised checklist is given in Table 1. The checklist includes criteria that cover fundamental statistical issues not included in general checklists. The checklist is intended to identify whether studies have both used and reported appropriate statistical methods as incited by decision makers [1, 9, 33, 69]. To ascertain whether a study has undertaken appropriate methods but failed to report them in the economic evaluation the

[2] Estimating incremental net benefits with OLS regression and applying a 'sandwich estimator' to allow the standard error to recognise clustering does not allow for appropriate adjustment of the point estimates.

checklist can also be completed using information from additional papers and reports (such as the main clinical paper reporting the clinical outcomes from the CRT).

The first question in the checklist recognises that sample size calculations need to incorporate any anticipated clustering in outcomes and costs, otherwise the study will be underpowered [17]. The second question emphasises that univariate analyses of costs and effects should acknowledge clustering [27, 63, 70, 71]. If the study ignores clustering it will underestimate uncertainty, and can provide biased estimates [26-28].

Question 3 considers whether the statistical methods acknowledged any correlation between individual costs and effects [18-20, 49]. Even if a study has allowed for clustering in the univariate endpoints (Question 2), or correlation between costs and effects (Question 3), it must also allow for clustering in the joint estimation of costs and effects or in estimating the incremental net benefit (Question 4) [24, 25, 58]. Question 5 assesses whether the study has made explicit, appropriate assumptions about the distribution of the data, e.g. by adequately assuming that costs and effects are normally distributed, or by using plausible non-normal distributions or methods that make tenable asymptotic assumptions [21-23].

<< Table 1 here >>

A reviewer can use the checklist to score each studygiving a '1' for each criterion which is met (0.5 for each subquestion). The scores are then summed, with each crierion receiving equal weight, to give a total score which ranges from 0 (study did not meet any of the criteria) to 5 (study met all the criteria).

Review of applied literature

We conducted a structured literature review of economic evaluations that used data from CRTs. Although it was not a systematic literature review (for example it was based only on published sources), it did satisfy most requirements of a good quality literature review, according to the PRISMA statement [72]. For example, a consistent and transparent database search strategy was undertaken that could be replicated by another reviewer. The search strategy included a wide range of databases in health economics, public health and medicine used in previous systematic reviews in economic evaluation of healthcare programmes [73-75]. These included Health Economic Evaluations Database (HEED), Cochrane library, NHS Economic Evaluation Database (NEED), EconLit, EMBASE, PubMed, MedLine, Scopus and Web of Knowledge. The search was not restricted to a specific period of time and included all the available evidence up to May 2009. The search strategy combined two sets of free-text terms (related to 'cluster randomised trials' and "economic

evaluations"), slightly adjusted according to the specific database. Title and abstracts were screened to check whether the study met the following inclusion criteria:

1 - Study must be undertaken alongside a cluster randomised trial

2 - Study must compare both cost and effects of alternative interventions

3 - Results must be reported on an incremental basis

4 - The study was a cost-effectiveness, cost-utility or cost-benefit analysis, but not cost minimisation or cost-consequences analysis

5 - Study published in any language but at least with an abstract in English

The selected articles were critically appraised by one reviewer using the Drummond checklist ([9], page 28), and the new checklist.[3] The initial assessments used the information reported in the economic evaluation paper. To examine whether an appropriate method was used but reported elsewhere, the review also considered information from additional sources such as the main clinical paper reporting the CRT.

Results

Study characteristics

Our database search strategy yielded 839 unique articles, 740 of which were excluded after screening the title and abstract and a further 42 after full text review, because they did not satisfy the inclusion criteria: studies either were not based on CRTs or not full economic evaluations (Appendix 1). A total of 47 studies fully satisfied the inclusion criteria, and were included in the review. The selected studies were published between 2001 and 2009, in medical (64%), public health (17%), health economics (17%) and statistical (2%) journals.

The studies reviewed covered evaluations of a wide range of group-level interventions: management of health care services (e.g. vaccination programmes) (36%), disease management (e.g., diabetes) (19%), screening strategies (e.g. breast cancer) (17%), health promotion (e.g., smoking cessation) programmes (17%), and alternative clinical guidelines (11%). Almost three quarters of studies directly used individual patient data. The remaining studies used summary data in decision-analytic models or simply reported aggregated, deterministic measures.

Results from applying the checklists

[3] In the next version of this review all studies will be assessed by a second reviewer

The reviewed economic evaluations answered 'Yes' to, on average 90% of the questions in the Drummond checklist. When the studies were assessed using the proposed checklist, on average studies answered 'Yes' to 20% of the questions, with a median total quality score of 1 (IQR=1) out of 5 (Table 2). The median quality score for studies published in more quantitative journals (specialist health economics and statistical journals) was not statistically different from those published in public health and medical journals (Mann-Whitney U test; p=0.60). The quality scores were similar for studies conducted before or after 2005 (p=0.50). Studies that used patient-level data from the CRT had a higher median quality score than those using summary estimates from the CRT in decision-analytical models or other forms of aggregated analyses (p <0.01).

Table 2 presents disaggregated results according to whether the study met each criterion in the checklist, firstly according to the information reported in the main economic evaluation paper and secondly according to information available in the clinical and other supplementary papers. The results did not differ significantly across these two sources apart from the criteria for the sample size calculation for effects, where 70% of the economic evaluations did not report allowing for clustering whereas using the supplementary papers as well, 40% failed this criterion.

<< Table 2 >>

Over both sources very few studies (8.5%) adjusted the sample size estimates for clustering in costs (Table 2). In the economic evaluation paper, almost 40% of the studies totally ignored clustering in the analysis (30% when supplementary information was also considered). For the univariate analyses, 53% studies ignored clustering in the analysis of effects (economic evaluation paper), and 21% in the analysis of costs (both sources). A total of 28 studies reported ICCs for effects, of which 7% suggested that levels of clustering were 'high' (ICC>0.1), 53% that they were 'moderate' (0.01<ICC<=0.1) and 23% that they were 'low' (ICC<=0.01). Only two studies reported ICCs for costs. More than 90% of economic evaluations neglected the correlation between costs and effects (Table 2). We found only one study [24] that used statistical methods that allowed for both clustering and correlation in the estimation of costs and effects.

The majority of studies (63.8%) did not report appropriate assumptions about the distribution of the cost and effectiveness data (Table 2). Even when looking at both the economic evaluations and clinical papers it was difficult to assess whether the data were truly normal or whether the study had inappropriately assumed normality. Of the nine studies that suggested their effectiveness data may not follow a normal distribution, eight went used non parametric methods. Of the 17 studies that suggested their costs were skewed only three assumed non-

normal distributions (Lognormal, Gamma and Inverse Gaussian), the other 14 rescaled the cost data by taking logarithm transformations or applied non-parametric methods (e.g. Mann-Whitney U tests).

Figure 2 summarises the main findings. The majority of studies (76.6%) were defined as 'type A' 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). Almost 15% of studies reported accounting for clustering in the univariate analyses of both costs and effects but did not allow for correlation (type B), three studies recognised correlation but failed to correctly acknowledge clustering (type C). Only one study correctly recognised both correlation and clustering (type D).

<< Figure 2 >>

Discussion

This study developed and applied a tool for assessing the quality of published economic evaluations that used data from CRTs. Our review found that the methodological quality of these studies was satisfactory judged against the Drummond checklist. Studies used all relevant comparators, valued costs and effects appropriately, discounted when necessary, undertook an incremental analysis, and presented the uncertainty surrounding the results. However, when the studies were assessed against our more stringent criteria they performed relatively badly, with a median score of 1 out of 5. Most studies ignored the correlation between costs and effects and failed to acknowledge clustering in the analyses of both costs and outcomes. Indeed, only one study allowed for clustering and correlation in the estimation of costs and effects. Ignoring clustering in the joint estimation of costs and effects is a major concern as it underestimates uncertainty and can lead to bias [25, 27].

The checklist was based on a careful review of the methods literature and covers fundamental statistical issues in CEA from CRTs. However, it does not attempt to embrace all the statistical issues that arise when designing or analysing economic evaluations. For example, it does not include questions on whether the method appropriately accounted for missing and censored data. This new checklist is intended to be used alongside generic checklists and guidelines for statistical methods in CEA. It is important that such checklists encourage rather than hinder methodological progress, and so these criteria will be updated in response to future methodological developments.

Decision-makers require health economic evaluations that use sound methods and are transparent in their reporting. However, it is conceivable that the studies reviewed *used* appropriate 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 used. This suggests that the studies were failing to *use* appropriate methods, which raises the question: were appropriate methods available?

As part of the conceptual review we identified three groups of potential methods (MLMs, GEEs and the TSB) that can handle both correlations and the specific form of clustering present in CRTs. Methodological guidance on the use of these methods to analyse data from CRTs is established in the medical statistics literature [12, 13, 16, 17, 76]. However, these methods have yet to permeate applied health economic evaluations that use CRTs. 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.

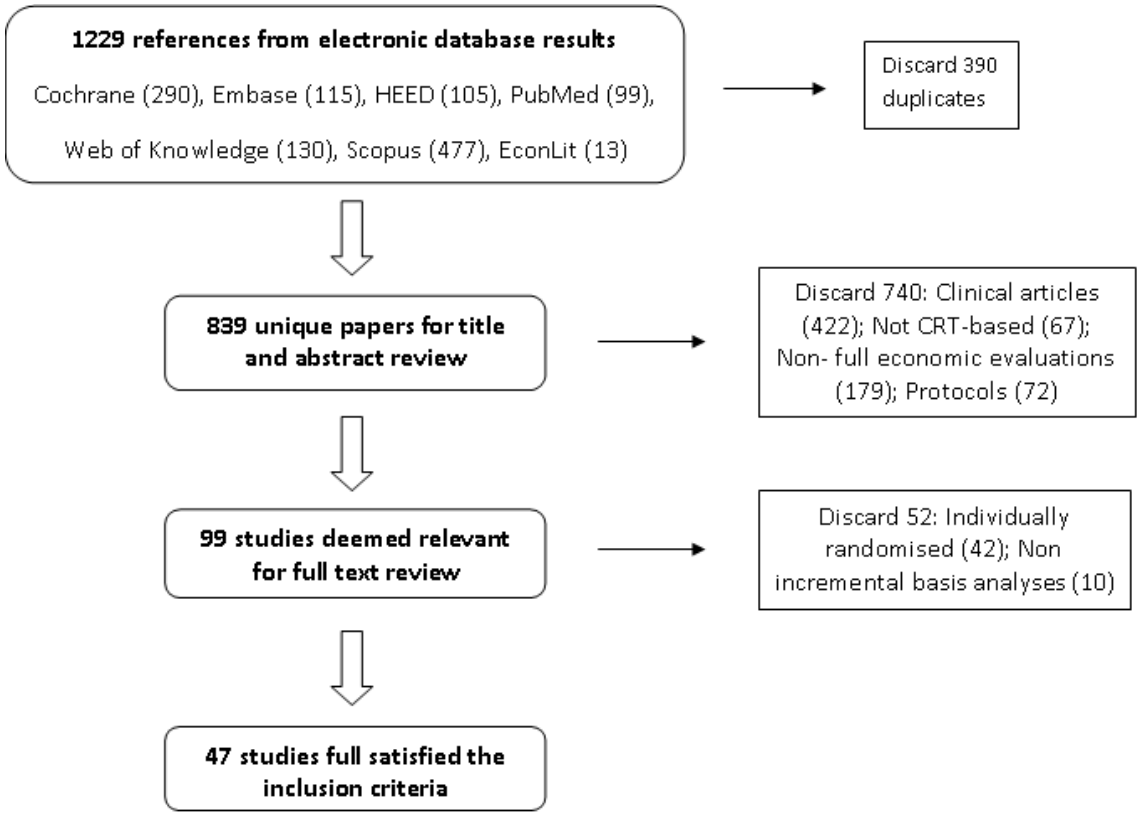
The low uptake of appropriate methods by applied researchers may be explained by the relatively little attention given to this issue in the health economic evaluation methods literature. A recent study proposed MLMs for economic evaluations from CRTs and suggested this approach lead to different cost-effectiveness results compared to methods that ignored clustering [58]. However, the study did not consider alternative approaches (e.g. GEEs or TSB). Likewise, Flynn and Peters [25] only considered the TSB, and limited their investigation to relatively favourable circumstances for the TSB, in particular, when CRT had balanced clusters (equal numbers per cluster). In our survey 85% of studies had imbalanced clusters. Only one study [24] 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 relatively large numbers of balanced clusters balanced clusters (n=50), and so the methodological comparison offered limited generalisability. Research is currently investigating the relative merits of alternative statistical methods for CEA from CRTs. This work will combine simulations with case-studies representative of the studies reviewed, i.e. with small numbers of unbalanced clusters. Economic evaluations with these complex characteristics are prevalent in the literature, and are potentially being used to inform policy-making.

While this paper has developed criteria for assessing economic evaluations that use CRT and has carefully applied them in a comprehensive structural review it has certain limitations. Like previous reviews in health economic evaluations it does not extend to the grey literature and so cannot be termed a 'systematic review' [72]. Furthermore, it is unlikely that the search

strategies captured all published studies that used CRTs. If the economic evaluation did not index the article with terms that enabled us to identify studies that used CRTs, then the study will have been omitted from the review. However, the quality of studies that did not include terms such as 'CRT', 'area-' or 'group-level' in the title or abstract will arguably have been worse than those included. The checklist summarised how well the study addressed the key statistical issues and gives summary scores which can be compared over time and across contexts. The summary score requires that each item is explicitly weighted and it is proposed that each item in the checklist carries equal weight.

In conclusion, economic evaluations that use data from CRTs have overlooked important statistical issues in the analysis. While the quality of the studies appears to be 'good' when evaluated by a generic checklist, it becomes a cause for concern when assessed by more specific criteria that cover fundamental statistical issues for economic evaluations that use CRTs. Methods are available for addressing the specific issues that arise when analysing economic evaluations from CRTs and their use should be actively encouraged if studies are to provide a sound basis for policy-making. Our proposed checklist can help raise awareness of current poor research practice. It can provide a starting point for improving the quality of economic evaluations that use CRT. Now that advisory groups such as NICE are trying to use economic evaluations of group-level interventions, this may provide more impetus to improving methods in this area. Certainly future methodological guidelines for the evaluation of public health interventions could go beyond general criteria and include some of the additional criteria listed.

Appendix 1 - Study inclusion criteria diagram



References

1. NICE, *Methods for Technology Appraisals*. 2008, National Institute for Health and Clinical Excellence. London, UK.
2. CADTH, *Guidelines for the Economic Evaluation of Health Technologies: Canada.* 3rd ed. 2006, Canadian Agency for Drugs and Technologies in Health. Ottawa, Canada.
3. PBAC, *Guidelines for preparing submissions to the Pharmaceutical Benefits Advisory Committee*. 2008, Australian Government - Department of Health and Ageing. Canberra, Australia.
4. IQWIG, *Methods for Assessment of the relation of Benefits to Costs in the German Statutory Health Care System*. 2009, Institute for Quality and Efficiency in Health Care. Cologne, Germany.
5. NICE, *Methods for the development of NICE public health guidance*. 2009, National Institute for Health and Clinical Excellence. London, UK.
6. Grieve, R., et al., *Evaluating health care programs by combining cost with quality of life measures: a case study comparing capitation and fee for service.* Health Serv Res, 2008. **43**(4): p. 1204-22.
7. Hutchings, A., et al., *Evaluation of modernisation of adult critical care services in England: time series and cost effectiveness analysis.* BMJ, 2009. **339**: p. b4353.
8. Briggs, A.H., K. Claxton, and M.J. Sculpher, *Decision modelling for health economic evaluation*. Handbooks in health economic evaluation series. 2006, Oxford: Oxford University Press. x, 237 p.
9. Drummond, M., et al., *Methods for the Economic Evaluation of Health Care Programmes* 3rd ed. 2005, Oxford, UK: Oxford University Press.
10. Glick, H.A., et al., *Economic evaluation in clinical trials*. 2007, Oxford, UK: Oxford University Press.
11. Willan, A. and A. Briggs, *Statistical Analysis of cost-effectiveness data*. 2006, Chichester, UK: John Wiley & Sons, Ltd. .
12. Hayes, R. and L. Moulton, *Cluster randomised trials*. Chapman & Hall/CRC Interdisciplinary Statistics Series, ed. N. Keiding, et al. 2009, Boca Raton - Florida, US: CRC Press, Taylor & Francis Group.
13. Donner, A. and N. Klar, *Design and analysis of cluster randomization trials in health research*. 2000, London, UK: Hodder Arnold Publishers.
14. Donner, A. and N. Klar, *Pitfalls of and controversies in cluster randomization trials.* Am J Public Health, 2004. **94**(3): p. 416-22.
15. Klar, N. and A. Donner, *Current and future challenges in the design and analysis of cluster randomization trials.* Stat Med, 2001. **20**(24): p. 3729-40.
16. Murray, D.M., S.P. Varnell, and J.L. Blitstein, *Design and analysis of group-randomized trials: a review of recent methodological developments.* Am J Public Health, 2004. **94**(3): p. 423-32.
17. Ukoumunne, O.C., et al., *Methods for evaluating area-wide and organisation-based interventions in health and health care: a systematic review.* Health Technol Assess, 1999. **3**(5): p. iii-92.
18. Hoch, J.S., A.H. Briggs, and A.R. Willan, *Something old, something new, something borrowed, something blue: a framework for the marriage of health econometrics and cost-effectiveness analysis.* Health Econ, 2002. **11**(5): p. 415-30.
19. Nixon, R.M. and S.G. Thompson, *Methods for incorporating covariate adjustment, subgroup analysis and between-centre differences into cost-effectiveness evaluations.* Health Econ, 2005. **14**(12): p. 1217-29.
20. Willan, A.R., A.H. Briggs, and J.S. Hoch, *Regression methods for covariate adjustment and subgroup analysis for non-censored cost-effectiveness data.* Health Econ, 2004. **13**(5): p. 461-75.

21. Briggs, A., et al., *Parametric modelling of cost data: some simulation evidence.* Health Econ, 2005. **14**(4): p. 421-8.
22. Mihaylova, B., et al., *Review of statistical methods for analysing healthcare resources and costs, applicable to clinical trial data.* Submitted to Health Economics,, 2008.
23. Thompson, S.G. and R.M. Nixon, *How sensitive are cost-effectiveness analyses to choice of parametric distributions?* Med Decis Making, 2005. **25**(4): p. 416-23.
24. Bachmann, M.O., et al., *Methods for analyzing cost effectiveness data from cluster randomized trials.* Cost Eff Resour Alloc, 2007. **5**: p. 12.
25. Flynn, T.N. and T.J. Peters, *Cluster randomized trials: Another problem for cost-effectiveness ratios.* International Journal of Technology Assessment in Health Care, 2005. **21**(3): p. 403-409.
26. Feng, Z.D., D. McLerran, and J. Grizzle, *A comparison of statistical methods for clustered data analysis with Gaussian error.* Statistics in Medicine, 1996. **15**(16): p. 1793-1806.
27. Flynn, T.N. and T.J. Peters, *Use of the bootstrap in analysing cost data from cluster randomised trials: some simulation results.* Bmc Health Services Research, 2004. **4**: p. 33-43.
28. Nixon, R.M. and S.G. Thompson, *Baseline adjustments for binary data in repeated cross-sectional cluster randomized trials.* Stat Med, 2003. **22**(17): p. 2673-92.
29. Evers, S., et al., *Criteria list for assessment of methodological quality of economic evaluations: Consensus on Health Economic Criteria.* International Journal of Technology Assessment in Health Care, 2005. **21**(2): p. 240-5.
30. Graves, N., et al., *Cost data for individual patients included in clinical studies: no amount of statistical analysis can compensate for inadequate costing methods.* Health Econ, 2002. **11**(8): p. 735-9.
31. Hjelmgren, J., F. Berggren, and F. Andersson, *Health economic guidelines--similarities, differences and some implications.* Value Health, 2001. **4**(3): p. 225-50.
32. Ofman, J.J., et al., *Examining the value and quality of health economic analyses: implications of utilizing the QHES.* J Manag Care Pharm, 2003. **9**(1): p. 53-61.
33. Philips, Z., et al., *Good practice guidelines for decision-analytic modelling in health technology assessment: a review and consolidation of quality assessment.* Pharmacoeconomics, 2006. **24**(4): p. 355-71.
34. Soto, J., *Health economic evaluations using decision analytic modeling. Principles and practices--utilization of a checklist to their development and appraisal.* Int J Technol Assess Health Care, 2002. **18**(1): p. 94-111.
35. Philips, Z., et al., *Review of guidelines for good practice in decision-analytic modelling in health technology assessment.* Health Technol Assess, 2004. **8**(36): p. iii-iv, ix-xi, 1-158.
36. Campbell, M.J., A. Donner, and N. Klar, *Developments in cluster randomized trials and Statistics in Medicine.* Stat Med, 2007. **26**(1): p. 2-19.
37. Donner, A., N. Klar, and G. Zou, *Methods for the statistical analysis of binary data in split-cluster designs.* Biometrics, 2004. **60**(4): p. 919-25.
38. Ma, J., et al., *Comparison of Bayesian and classical methods in the analysis of cluster randomized controlled trials with a binary outcome: The Community Hypertension Assessment Trial (CHAT).* BMC Med Res Methodol, 2009. **9**(1): p. 37.
39. Willan, A., *Statistical Analysis of cost-effectiveness data from randomised clinical trials.* Expert Revision Pharmacoeconomics Outcomes Research, 2006. **6**: p. 337-346.
40. van Breukelen, G.J., M.J. Candel, and M.P. Berger, *Relative efficiency of unequal versus equal cluster sizes in cluster randomized and multicentre trials.* Stat Med, 2007. **26**(13): p. 2589-603.
41. Moerbeek, M., G.J. van Breukelen, and M.P. Berger, *Design issues for experiments in multilevel populations.* Journal of Educational and Behavioral Statistics, 2000. **25**(3): p. 271-284.

42. Al, M.J., et al., *Sample size calculation in economic evaluations.* Health Econ, 1998. **7**(4): p. 327-35.
43. Drummond, M. and B. O'Brien, *Clinical importance, statistical significance and the assessment of economic and quality-of-life outcomes.* Health Econ, 1993. **2**(3): p. 205-12.
44. O'Sullivan, A.K., D. Thompson, and M.F. Drummond, *Collection of health-economic data alongside clinical trials: is there a future for piggyback evaluations?* Value Health, 2005. **8**(1): p. 67-79.
45. Briggs, A., *Economic evaluation and clinical trials: size matters.* BMJ, 2000. **321**(7273): p. 1362-3.
46. Bland, J.M., *The tyranny of power: is there a better way to calculate sample size?* BMJ, 2009. **339**: p. b3985.
47. Campbell, M.K., P.M. Fayers, and J.M. Grimshaw, *Determinants of the intracluster correlation coefficient in cluster randomized trials: the case of implementation research.* Clinical Trials, 2005. **2**(2): p. 99-107.
48. Thompson, S.G., R.M. Nixon, and R. Grieve, *Addressing the issues that arise in analysing multicentre cost data, with application to a multinational study.* J Health Econ, 2006. **25**(6): p. 1015-28.
49. O'Hagan, A. and J.W. Stevens, *A framework for cost-effectiveness analysis from clinical trial data.* Health Econ, 2001. **10**(4): p. 303-15.
50. Manning, W.G., A. Basu, and J. Mullahy, *Generalized modeling approaches to risk adjustment of skewed outcomes data.* J Health Econ, 2005. **24**(3): p. 465-88.
51. Manning, W.G. and J. Mullahy, *Estimating log models: to transform or not to transform?* J Health Econ, 2001. **20**(4): p. 461-94.
52. Willan, A.R., D.Y. Lin, and A. Manca, *Regression methods for cost-effectiveness analysis with censored data.* Stat Med, 2005. **24**(1): p. 131-45.
53. Goldstein, H., *Multilevel Statistical Models*. 3rd ed. 2003, Oxford, UK: Oxford University Press.
54. Leyland, A. and H. Goldstein, *Multilevel Modelling of Health Statistics*. 2001, Chichester, UK: John Wiley & Sons, Ltd.
55. Hardin, J.W. and J.M. Hilbe, *Generalized Estimating Equations*. 2003, Boca Raton, Florida. US: Chapman & Hall/CRC.
56. Davison, A.C. and D.V. Hinkley, *Bootstrap methods and their application*. Cambridge series on statistical and probabilistic mathematics, ed. R. Gill, et al. Vol. 2nd. 1997, Cambridge, UK: Cambridge University Press.
57. Manca, A., et al., *Cost-effectiveness analysis using data from multinational trials: the use of bivariate hierarchical modeling.* Med Decis Making, 2007. **27**(4): p. 471-90.
58. Grieve, R., R.M. Nixon, and S.G. Thompson, *Bayesian hierarchical models for cost-effectiveness analyses that use data from cluster randomised trials.* Medical Decision Making, 2009. **(in press)**.
59. Lambert, P.C., et al., *How vague is vague? A simulation study of the impact of the use of vague prior distributions in MCMC using WinBUGS.* Stat Med, 2005. **24**(15): p. 2401-28.
60. Liang, K.Y. and S. Zeger, *Longitudinal data analysis using generalized linear models.* Biometrika, 1986. **73**(1): p. 13-22.
61. Austin, P.C., *A comparison of the statistical power of different methods for the analysis of cluster randomization trials with binary outcomes.* Stat Med, 2007. **26**(19): p. 3550-65.
62. Turner, R.M., I.R. White, and T. Croudace, *Analysis of cluster randomized cross-over trial data: a comparison of methods.* Stat Med, 2007. **26**(2): p. 274-89.
63. Ukoumunne, O.C. and S.G. Thompson, *Analysis of cluster randomized trials with repeated cross-sectional binary measurements.* Stat Med, 2001. **20**(3): p. 417-33.
64. Ziegler, A., C. Kastner, and M. Blettner, *The Generalised Estimating Equations: An annotated bibliography.* Biom J, 1998. **40**(2): p. 115-139.

65. Lipsitz, S., et al., *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, 2009. **172**(Part 1): p. 3-20.
66. Braun, T.M., *A mixed model-based variance estimator for marginal model analyses of cluster randomized trials.* Biom J, 2007. **49**(3): p. 394-405.
67. Efron, B. and R. Tibshirani, *An introduction to Bootstrap*. 1993, New York, US: Chapman and Hall.
68. O'Hagan, A. and J.W. Stevens, *Assessing and comparing costs: how robust are the bootstrap and methods based on asymptotic normality?* Health Economics, 2003. **12**(1): p. 33-49.
69. Ramsey, S., et al., *Good research practices for cost-effectiveness analysis alongside clinical trials: the ISPOR RCT-CEA Task Force report.* Value Health, 2005. **8**(5): p. 521-33.
70. Spiegelhalter, D.J., *Bayesian methods for cluster randomized trials with continuous responses.* Stat Med, 2001. **20**(3): p. 435-52.
71. Turner, R.M., R.Z. Omar, and S.G. Thompson, *Bayesian methods of analysis for cluster randomized trials with binary outcome data.* Stat Med, 2001. **20**(3): p. 453-72.
72. Moher, D., et al., *Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement.* BMJ, 2009. **339**: p. b2535.
73. Briggs, A.H. and A.M. Gray, *Handling uncertainty when performing economic evaluation of healthcare interventions.* Health Technol Assess, 1999. **3**(2): p. 1-134.
74. Sculpher, M.J., et al., *Generalisability in economic evaluation studies in healthcare: a review and case studies.* Health Technol Assess, 2004. **8**(49): p. iii-iv, 1-192.
75. Turner, D., et al., *Systematic review and economic decision modelling for the prevention and treatment of influenza A and B.* Health Technol Assess, 2003. **7**(35): p. iii-iv, xi-xiii, 1-170.
76. Turner, R.M., R.Z. Omar, and S.G. Thompson, *Modelling multivariate outcomes in hierarchical data, with application to cluster randomised trials.* Biom J, 2006. **48**(3): p. 333-45.

Table 1 – Proposed checklist for CEA that use data from CRTs

Questions	*Score*
1. Did the sample size calculation adjust for any clustering in: [42-45]	
a) effects?	***Yes*** ☐ ***No*** ☐
b) costs?	***Yes*** ☐ ***No*** ☐
2. Was clustering recognised and accounted for in: [27, 63, 70, 71]	
a) the univariate analysis of effects?	***Yes*** ☐ ***No*** ☐
b) the univariate analysis of costs?	***Yes*** ☐ ***No*** ☐
3. Was the correlation between costs and effects recognised and accounted for in the statistical analysis? [18-20, 49]	***Yes*** ☐ ***No*** ☐
4. Were clustering and correlation accounted for in the joint estimation of costs and effects? [18-20, 49]	***Yes*** ☐ ***No*** ☐
5. Did the study explicitly make appropriate assumptions about the distribution of [11, 21-23, 50]	
a) effects?	***Yes*** ☐ ***No*** ☐
b) costs?	***Yes*** ☐ ***No*** ☐

Table 2: Results from applying the CRT checklist to a) the economic evaluation, and b) the economic evaluation and supplementary papers (n=47). The N(%) studies that meet each criterion and total scores.

Question	economic evaluation	economic evaluation and supplementary papers
1. Adjusted for clustering in sample sample size calculations for		
a) effects	14 (29.8%)	29 (61.7%)
b) costs	4(8.5%)	4 (8.5%)
2. Adjusted for clustering in univariate analyses of		
a) effects	25 (53.2%)	29 (61.7%)
b) costs	10 (21.3%)	10 (21.3%)
3. Allowed for correlation between costs and effects	4 (8.5%)	6 (12.8%)[4]
4. Accounted for clustering and correlation	1 (2.1%)	1 (2.1%)
5. Appropriate assumptions about distributions of:		
a) effects	17 (36.2%)	19 (40.5%)
b) costs	17 (36.2%)	17 (36.2%)
Median (IQR) total score	1 (1)	1 (1)
Mean (sd) total score	1.12 (0.89)	1.22 (0.91)

[4] In two studies using decision-analytical models additional information was available which suggested correlations between costs and effects had been recognised.

Figure 1: Summary of key statistical issues in economic evaluations that use data from CRTs

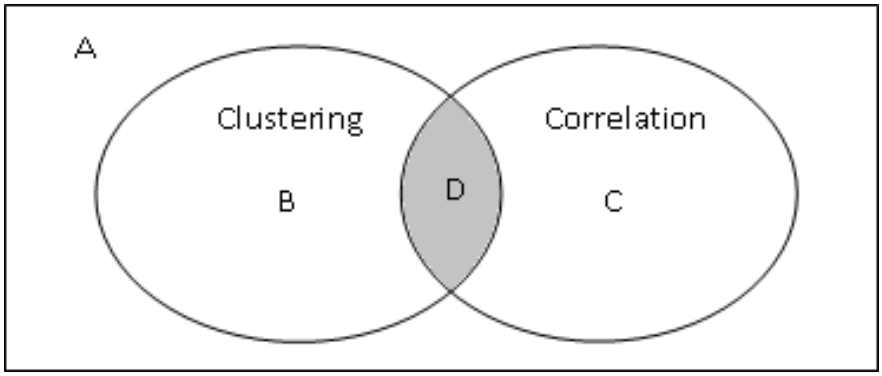


Figure 2: The proportion of the economic evaluations that use CRTs that addressed the key statistical issues (n=47).