

# Developing Appropriate Methods for Cost-Effectiveness Analysis of Cluster Randomized Trials

Manuel Gomes, MSc, Edmond S.-W. Ng, MSc, Richard Grieve, PhD, Richard Nixon, PhD, James Carpenter, PhD, Simon G. Thompson, DSc

---

**Aim.** Cost-effectiveness analyses (CEAs) 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 article compares alternative statistical methods for accommodating clustering in CEAs of CRTs. **Methods.** Our simulation study compared the performance of statistical methods for CEAs of CRTs with 2 treatment arms. The study considered a method that ignored clustering—seemingly unrelated regression (SUR) without a robust standard error (SE)—and 4 methods that recognized clustering—SUR and generalized estimating equations (GEEs), both with robust SE, a “2-stage” nonparametric bootstrap (TSB) with shrinkage correction, and a multilevel model (MLM). 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 (INBs). 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 v. nominal level: 0.95). The MLM and TSB performed well in each scenario (CI coverage, 0.92–0.95). With few clusters, the GEE and SUR (with robust SE) had coverage below 0.90. In the case study, the mean INBs 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 CEAs of CRTs with the characteristics described. SUR and GEE are not recommended for studies with few clusters. **Key words:** randomized trial methodology; statistical methods; cost-effectiveness analysis. (*Med Decis Making* 2012;32:350–361)

---

Received 9 March 2011 from the Department of Health Services Research and Policy, London School of Hygiene and Tropical Medicine, London, UK (MG, ESWN, RG); Modeling and Simulation Group, Novartis Pharma AG, Basel, Switzerland (RN); Department of Medical Statistics, London School of Hygiene and Tropical Medicine, London, UK (JC); MRC Biostatistics Unit, Cambridge, UK (SGT). This work was presented at the Health Economists' Study Group meeting; January 2011; York, UK. Financial support was from a UK Medical Research Council Grant (ESWN, RG) and a PhD scholarship from Fundação para a Ciência e a Tecnologia (MG). Revision accepted for publication 18 June 2011.

Supplementary material for this article is available on the *Medical Decision Making* Web site at <http://mdm.sagepub.com/supplemental>.

Address correspondence to Richard Grieve, PhD, Department of Health Services Research and Policy, London School of Hygiene and Tropical Medicine, 15-17 Tavistock Place, London, WC1H 9SH; telephone: (020) 7927-2255; fax: (020) 7927-2701; e-mail: [richard.grieve@lshtm.ac.uk](mailto:richard.grieve@lshtm.ac.uk).

DOI: 10.1177/0272989X11418372

Cost-effectiveness analyses (CEAs) of group-based interventions often use data from cluster randomized trials (CRTs).<sup>1</sup> A cluster design may be preferred in evaluations of interventions, which operate at a group level (e.g., alternative incentives for health providers), or where there is a high risk of “contamination” among the individuals within clusters (e.g., vaccination programs).<sup>2,3</sup> Agencies such as the National Institute for Health and Clinical Excellence may use these CEAs especially when recommending which public health interventions should be provided.<sup>4</sup> For these studies to provide a sound basis for decision making, appropriate statistical methods need to be developed and used.<sup>5,6</sup> CEAs based on randomized controlled trials (RCTs), where individual patients are randomized, have well-established methods.<sup>7–9</sup> However, statistical

methods for CEAs of CRTs have received limited attention.<sup>10</sup> A review found that less than 10% of published CEAs of CRTs adopted appropriate statistical methods.<sup>1</sup>

A distinct feature of CRTs is that the unit of randomization is the cluster (e.g., 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 multicenter RCTs, where patients within a center 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 intraclass 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,<sup>2,3,11</sup> encourages incorrect inferences,<sup>12–15</sup> and can also lead to bias.<sup>16,17</sup> Appropriate methods for handling clustering in clinical outcomes are well developed and can include multilevel models (MLMs) and generalized estimating equations (GEEs).<sup>18</sup>

CEAs of 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<sup>19–21</sup> and make plausible assumptions about the distribution of costs and outcomes. Based on a conceptual review, we identified 4 main groups of statistical methods that may be appropriate for CEAs of CRTs: seemingly unrelated regression (SUR),<sup>21</sup> GEEs,<sup>22</sup> the nonparametric 2-stage bootstrap (TSB),<sup>23</sup> and MLMs.<sup>20</sup> 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 CEAs of CRTs. The TSB<sup>24</sup> and MLMs<sup>25</sup> have been proposed for CEAs of CRTs, but the only study<sup>26</sup> to compare these methods used data from a single CRT. A simulation study<sup>24</sup> 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 CEAs of CRTs.

The aim of this article was to assess the relative performance of alternative statistical methods for CEAs of 2-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 CEAs of CRTs

We consider 4 methods for CEAs 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 CEAs with 2 alternative treatments, but the models extend to evaluations with more than 2 randomized treatments. Each method takes the common approach of assuming linear additive treatment effects for both costs and outcomes.<sup>9,20,21,27</sup>

### Seemingly Unrelated Regression (SUR)

SUR consists of a system of regression equations that can recognize the correlation between individual costs and outcomes.<sup>9,10,21</sup> The SUR model<sup>1</sup> allows the individual-level error terms ( $\varepsilon$ ) to be correlated through the parameter  $\rho$ :

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

where  $t_j$  is the treatment indicator ( $t_j = 0$  for control and 1 for treatment group). The 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,<sup>28</sup> 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, StataCorp, College Station, TX). Estimates are identical to OLS when the same covariates are included for costs and outcomes.<sup>21,a</sup>

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 CEAs of CRTs. In addition, it is unclear whether the robust SE recognizes the correlation at the cluster level, that is, between cluster-level mean costs and mean outcomes.<sup>29,30</sup> Finally, the asymptotic assumptions underlying the robust variance estimator may not hold in CRTs with few clusters per treatment arm.<sup>31</sup> The problem can be exacerbated by skewed outcomes (or costs) or imbalanced cluster sizes.<sup>18</sup> More details on the robust variance estimator are given in Web Appendix 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.<sup>2,3,32</sup> While multivariate GEEs have been developed to recognize potential correlation between binary end points,<sup>22</sup> they are complex to implement and have not been extended to continuous end points. 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:

$$\begin{aligned} c_{ij} &= \beta_0^c + \beta_1^c t_k + \varepsilon_{ij}^c \\ o_{ij} &= \beta_0^o + \beta_1^o t_k + \varepsilon_{ij}^o \end{aligned} \quad \varepsilon_{ij}^c \perp \varepsilon_{ij}^o; \quad \varepsilon_{ij}^c \perp N(0, \sigma_c^2), \quad \varepsilon_{ij}^o \perp N(0, \sigma_o^2). \quad (2)$$

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, that is, 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 Web Appendix 1 for further details. However, the asymptotic properties required may not hold when there are few clusters.<sup>13,15,33,65-68</sup>

### The Nonparametric 2-Stage Bootstrap (TSB)

Nonparametric bootstrap methods can avoid parametric assumptions and are easy to apply in simple settings (e.g., RCTs).<sup>19</sup> However, the conventional nonparametric bootstrap that resamples individuals has to be extended to recognize the clustering inherent in CRTs. Davison and Hinkley<sup>23</sup> propose a 2-stage routine for CRTs, which resamples clusters as well as individuals, and this approach has been considered for CEAs.<sup>24,26,34</sup> The TSB can recognize the individual-level correlation between costs and outcomes by bivariate resampling, and the resampling can also stratify by treatment group.<sup>24</sup>

*TSB without shrinkage correction.* One proposed TSB algorithm requires resampling clusters and then individuals within each resampled cluster (both with replacement).<sup>23</sup> The resultant data sets are used to calculate the statistics of interest, for example, incremental net benefits (INBs) and confidence intervals (CIs). 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.<sup>23,24,34</sup>

*TSB with shrinkage correction.* Davison and Hinkley<sup>23</sup> 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

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

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 Web Appendix 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 nonnormal.<sup>35,36</sup> Furthermore, the TSB routines described above were only proposed for balanced clusters,<sup>23,24,34</sup> which may make the method inappropriate for CEAs of CRTs with imbalanced clusters.<sup>1</sup> Our implementation therefore extends Davison and Hinkley’s original algorithms to allow for imbalanced clusters (Web Appendix 2).

**Multilevel Models (MLMs)**

MLMs can allow for the correlation between costs and outcomes and recognize clustering.<sup>25</sup> Unlike SUR, MLMs can explicitly recognize clustering by including additional random terms,  $u_j^c, u_j^e$ , which in equation 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.

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

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.<sup>20</sup> A concern with either approach is that the MLM may fail to converge if the CRT has few individuals per cluster.<sup>37,38</sup>

**Monte Carlo Simulations**

*Data-generating process.* The simulation study was designed to test the methods across a wide range of circumstances typically found in CEAs of 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 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.<sup>12,37,39</sup> 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.<sup>21,24,40</sup> We simulated cost ( $c$ ) and outcome data ( $e$ ) from CRTs with  $M$  clusters per arm and  $n_m$  ( $m = 1 \dots M$ ) individuals per cluster. Data were generated firstly at the cluster level and then at the individual level according to equation 4 below.

Cluster level means:

$$\begin{matrix} \phi_j^c \sim \text{dist}(\beta_0^c + \beta_1^c t_j, \tau_c) \\ c_{ij} \sim \text{dist}(\phi_j^c, \sigma_c^c) \end{matrix}$$

Individual-level data:

$$\begin{matrix} \phi_j^e \sim \text{dist}(\beta_0^e + \beta_1^e + \gamma(\phi_j^c - (\beta_0^c + \beta_1^c t_j)), \tau_e) \\ e_{ij} \sim \text{dist}(\phi_j^e + \theta^1(c_{ij} - \phi_j^c), \sigma_e) \end{matrix} \quad (4)$$

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 ( $\beta_0^c + \beta_1^c, \beta_0^e + \beta_1^e + \gamma(\phi_j^c - (\beta_0^c + \beta_1^c t_j))$ ) 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, for example, for costs

**Table 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. of clusters per arm	GEE, SUR, and TSB all rely on asymptotic assumptions	20	3 to 30	15	Base case: 20 clusters per arm suggested for asymptotics to hold. <sup>2,61</sup> SA: takes lower, upper quartiles from literature review. Final case: median number of clusters from literature review.
No. of individuals per cluster	MLM may have convergence issues with few cases per cluster	50	10 to 80	30	Base case: within the range of values from literature review. SA: the lower, upper quartiles from literature review. Final case: median number per cluster from the literature review.
Level of imbalance( $cv_{imb}$ ) of cluster size	GEE, SUR, and TSB have not been assessed with imbalanced clusters	0	0 to 1	0.5	Base case: previous methods articles. <sup>23,24,34</sup> SA: cluster-size imbalance informed by range of values reported across case studies and previous study. <sup>42</sup> 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 articles. <sup>24,34</sup> SA: range of ICCs from case studies and previous study. <sup>62</sup> 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 literature review have ICCs for outcomes $\leq 0.01$ . SA: range from literature review and previous methods studies. <sup>42</sup> Final case: median from literature review.
Coefficient of variation( $cv_{cost}$ ) of cost distribution	SUR, MLM, and GEE 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. <sup>36,63</sup> SA: $\gamma$ distribution; range for $cv_{cost}$ from previous simulation studies. <sup>41,43</sup> Final case: gamma distribution; median $cv_{cost}$ from case studies.
Individual-level correlation of costs and effects	GEE assumes costs and outcomes are independent	0.2	-0.5 to 0.5	-0.2	Base case: plausible level of individual-level correlation. <sup>24,43</sup> SA: based on the range from case studies. Final case: median from the case studies.
Cluster-level correlation of costs and effects	GEE 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. <sup>24</sup> SA: based on the range from case studies. Final case: median from case studies.

Note: SA = sensitivity analyses; ICC = intraclass correlation coefficient; GEE = generalized estimating equation; SUR = seemingly unrelated regression; TSB = nonparametric 2-stage bootstrap; MLM = multilevel model.

$ICC_c = \tau_c^2 / (\sigma_c^2 + \tau_c^2)$ . 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 standard deviation 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 1-way and multiway 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 1). The choices of which parameters to vary in the sensitivity analyses, and which scenarios to combine in the multiway 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,<sup>1</sup> previous methods articles and simulation studies,<sup>24,41–43</sup> and 8 case studies.<sup>44–51</sup> 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 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 INBs (assuming a threshold of £20,000 per QALY) were £500, 0.075, and £1000, 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 Web Appendix 3 for definitions). We used 2000 simulations for each scenario.<sup>b</sup> The performance of each method in estimating incremental costs, incremental QALYs, and INBs was reported.

MLM, GEE, and TSB were implemented in R<sup>52</sup> and SUR in STATA (StataCorp).<sup>53</sup> The SUR was estimated by IFGNLS, without and with a robust SE. The GEE

was estimated with a robust SE, and the TSB was estimated before and then after shrinkage correction. The MLM was estimated by maximum likelihood across all scenarios.<sup>65</sup> 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.<sup>54</sup> The MCMC estimation consisted of 5000 iterations, 3 parallel chains with different starting values, and assuming diffuse priors.<sup>55</sup>

## 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. 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.<sup>56</sup>

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.<sup>46</sup> The CRT included 40 balanced clusters (clinics) randomized to intervention or control. This reanalysis used complete data for 1851 patients. For each patient, the study measured health service costs for 3 months, consisting mainly of the costs of the educational intervention clinic, outpatient visits, and drugs. EQ-5D data were recorded at 3 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 INBs, 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.<sup>57,58</sup> EVPI was calculated assuming that the INB was normally distributed.<sup>56</sup>

b. This was judged to be sufficient to report CI coverage with a margin of MC error of less than 1%; that is, for true coverage of 0.95, 2000 simulations would be 95% certain to give coverage rates of 0.94 to 0.96.

**Table 2** Bias, rMSE, CI Coverage, and Width of the Mean INB for the Base Case (True INB = £1000)

	SUR		GEE With Robust SE	TSB		MLM <sup>a</sup> ML
	Without Robust SE	With Robust SE		Without Shrinkage Correction	With Shrinkage Correction	
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

Note: INB = incremental net benefit; SUR = seemingly unrelated regression; GEE = generalized estimating equation; TSB = nonparametric 2-stage bootstrap; MLM = multilevel model; SE = standard error; ML = maximum likelihood; rMSE = root mean squared error; CI = confidence interval.

a. MLM estimated by Markov chain Monte Carlo (MCMC) in WinBUGS produced similar results.

## RESULTS

### Simulation Study

*Base case.* In the base case, each method reported low bias and similar rMSE for the INB (Table 2). The method that ignored clustering, SUR without the robust SE, performed poorly with CI coverage below 0.9. 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 MLM had coverage close to the nominal level whether estimated by maximum likelihood (Table 2) or MCMC (CI coverage, 0.94). The relative performance across methods was similar for incremental QALYs, incremental costs, and INBs 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 MLM and GEE, and -7.15 (6.29) for the TSB with shrinkage correction (true INB = £1000). 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 3 reports CI coverage for the 1-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 MLM and TSB (following correction), but the SUR and GEE, both with robust SEs, reported poor coverage. With high levels of cluster size imbalance, coverage

levels for these latter 2 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 3). CI coverage also remained close to the nominal level, with high levels of correlation at the individual or cluster level.

*Multiway sensitivity analysis.* The multiway 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 multiway sensitivity analyses. While rMSE increased when fewer clusters were combined with high levels of imbalance, the differences between methods were small.

Figure 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 GEE is worse than for the MLM and TSB if there are 8 or fewer clusters per treatment arm (Figure 1A). With high levels of cluster size imbalance, the coverage levels for the SUR and GEE are poor, with fewer than 10 clusters per arm (Figure 1B). For the MLM 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 GEE reported poor coverage.

**Table 3** Confidence Interval Coverage of the Mean Incremental Net Benefit (Nominal Level Is 0.95) for the 1-Way Sensitivity Analyses

	SUR With Robust SE	GEE With Robust SE	TSB		MLM ML
			Without Shrinkage Correction	With Shrinkage Correction	
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 ( $ICC_c = 0.3$ )	0.936	0.935	0.980	0.944	0.953
High ICC for outcomes ( $ICC_o = 0.3$ )	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

Note: SUR = seemingly unrelated regression; GEE = generalized estimating equation; TSB = nonparametric 2-stage bootstrap; MLM = multilevel model; SE = standard error; ML = maximum likelihood; ICC = intraclass correlation coefficient.

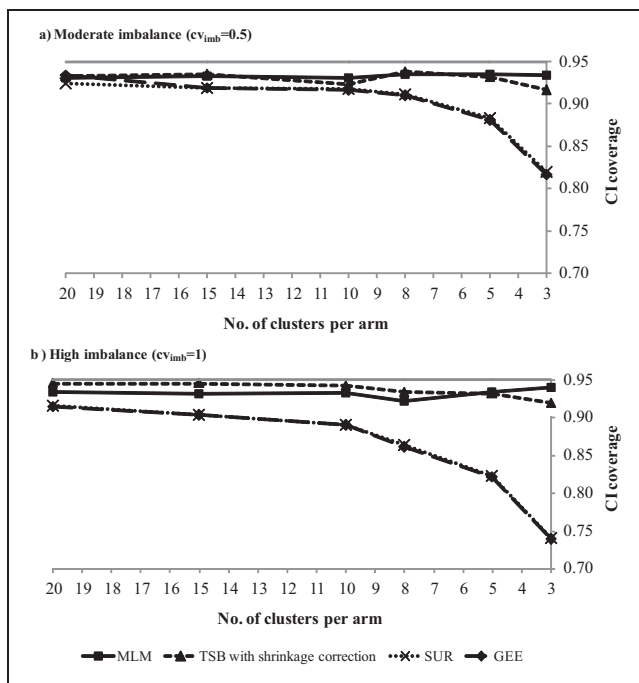


Figure 1 CI coverage (nominal level is 0.95) for multiway sensitivity analyses: high skewness of costs ( $cv_{cost} = 3$ ), decreasing number of clusters combined with (A) moderate and (B) high cluster-size imbalance. The CI coverage is very similar for the GEE and SUR, and hence, their lines show considerable overlap.

*Final “most realistic” scenario.* In the final scenario with all parameters set to their “most realistic” levels (Table 1), bias and rMSE were again similar across methods (Table 4). The SUR without a robust SE and the TSB without correction reported levels of CI coverage that diverged from the nominal, but the

MLM and TSB with correction both had good levels of CI coverage.

**Case Study**

Table 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 INBs. 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.

**DISCUSSION**

This study compares the relative merits of alternative statistical methods for CEAs of CRTs. The simulation study finds that each method reports low bias and similar MSE across the settings considered, with the MLM and TSB (with correction) providing good levels of CI coverage throughout. The simulation study highlights that robust methods (SUR and GEE), 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 (e.g., 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.



**Table 4** Bias, rMSE, CI Coverage, and Width of the Mean INB for the Final Case (True INB = £1000)

	SUR		GEE With Robust SE	TSB		MLM ML
	Without Robust SE	With Robust SE		Without Shrinkage Correction	With Shrinkage Correction	
Mean bias (SE)	6.63 (4.40)	6.63 (4.41)	6.63 (4.40)	7.10 (4.38)	6.85 (4.38)	7.95 (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

Note: INB = incremental net benefit; SUR = seemingly unrelated regression; GEE = generalized estimating equation; TSB = nonparametric 2-stage bootstrap; MLM = multilevel model; SE = standard error; ML = maximum likelihood; rMSE = root mean squared error; CI = confidence interval.

Future studies should not attempt to justify statistical methods that ignore clustering on the basis of low estimated ICCs.

This is the first article to compare a range of statistical methods for CEAs of CRTs. Previous simulation studies<sup>24,34</sup> did not consider MLMs or GEEs, and other studies just compared the methods using a single case study.<sup>26</sup> The design of the simulation study is sufficiently general to consider the methods across common circumstances faced by CEAs of 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 CEAs of CRTs.<sup>1</sup> 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 2-armed CRTs, the findings extend directly to CRTs with 3 or more randomized treatments.

The simulation study finds the TSB performs as well as the MLM 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.<sup>26</sup> 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<sup>24</sup> 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 codes for the TSB, GEE, and MLM are included in Web Appendix 4.

This article 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 CEAs of 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 CEAs where about 40% of such studies have fewer than 15 clusters per treatment arm and 15% less than 8 clusters.<sup>1</sup> The general literature on GEEs has reported similar sample size requirements for asymptotic assumptions to hold,<sup>13,15,18</sup> 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<sup>25</sup> proposed a flexible Bayesian hierarchical model to tackle the main statistical issues faced by CEAs of 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

**Table 5** Case Study Results: Incremental Cost, Incremental QALY, INB (Threshold of R100,000 per QALY),<sup>a</sup> and Individual EVPI

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

Note: QALY = quality-adjusted life year; INB = incremental net benefit; EVPI = expected value of perfect information; SUR = seemingly unrelated regression; GEE = generalized estimating equation; TSB = nonparametric 2-stage bootstrap; MLM = multilevel model; SE = standard error; ML = maximum likelihood.

a. One pound (GBP) corresponded to approximately 6.99 rands (ZAR) in terms of purchasing power parity.<sup>64</sup>

in a different context, this corroborates previous findings that suggest methods assuming normality may be quite robust to skewed cost data.<sup>21,40,41,43</sup> However, it would be worth investigating whether MLMs that 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.<sup>59</sup> Also, CEAs may have more complex variance structures than those considered.<sup>25,60</sup> 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.<sup>18,20,21,43</sup> 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, CEAs of 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 CEAs of 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 CEAs of 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. They 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 work was presented. Finally, they thank the collaborators on the study—Graham Scotland, Patrick Gillespie, Allan Clark, and Mike Gillett—for useful discussions.

## REFERENCES

- Gomes M, Grieve R, Edmunds J, Nixon R. Statistical methods for cost-effectiveness analyses that use data from cluster randomised trials: a systematic review and checklist for critical appraisal. *Med Decis Making*. 2011 May 24. [Epub ahead of print].
- Donner A, Klar N. *Design and Analysis of Cluster Randomization Trials in Health Research*. London: Hodder Arnold Publishers; 2000.
- Hayes R, Moulton L. *Cluster randomised trials*. Boca Raton, Florida, US: CRC Press, Taylor & Francis Group; 2009.
- National Institute for Health and Clinical Excellence (NICE). *Methods for development of NICE public health guidance*. Available from URL: [http://www.nice.org.uk/media/CE1/F7/CPHE\\_Methods\\_manual\\_LR.pdf](http://www.nice.org.uk/media/CE1/F7/CPHE_Methods_manual_LR.pdf)
- Eckermann S, Willan AR. Expected value of information and decision making in HTA. *Health Econ*. 2007;16(2):195–209.
- Griffin SC, Claxton KP, Palmer SJ, Sculpher MJ. Dangerous omissions: the consequences of ignoring decision uncertainty. *Health Econ*. 2011;20(2):212–24.
- Glick HA, Doshi JA, Sonnad SS, Polsky D. *Economic Evaluation in Clinical Trials*. Oxford: Oxford University Press; 2007.
- Gold MR. *Cost-Effectiveness in Health and Medicine*. New York: Oxford University Press; 1996.

9. Willan A, Briggs A. *Statistical Analysis of Cost-Effectiveness Data*. Chichester, UK: John Wiley & Sons Ltd.; 2006.
10. Willan A. Statistical analysis of cost-effectiveness data from randomised clinical trials. *Expert Rev Pharmacoecon Outcomes Res*. 2006;6:337–46.
11. Campbell MJ, Donner A, Klar N. Developments in cluster randomized trials and statistics in medicine. *Stat Med*. 2007;26(1):2–19.
12. Austin PC. A comparison of the statistical power of different methods for the analysis of cluster randomization trials with binary outcomes. *Stat Med*. 2007;26(19):3550–65.
13. Feng ZD, McLerran D, Grizzle J. A comparison of statistical methods for clustered data analysis with Gaussian error. *Stat Med*. 1996;15(16):1793–806.
14. Nixon RM, Thompson SG. Baseline adjustments for binary data in repeated cross-sectional cluster randomized trials. *Stat Med*. 2003;22(17):2673–92.
15. Ukoumunne OC, Thompson SG. Analysis of cluster randomized trials with repeated cross-sectional binary measurements. *Stat Med*. 2001;20(3):417–33.
16. Middleton J. Bias of the regression estimator for experiments using clustered random assignment. *Stat Probab Lett*. 2008;78(16):2654–9.
17. Panageas KS, Schrag D, Russell Localio A, Venkatraman ES, Begg CB. Properties of analysis methods that account for clustering in volume-outcome studies when the primary predictor is cluster size. *Stat Med*. 2007;26(9):2017–35.
18. Omar RZ, Thompson SG. Analysis of a cluster randomized trial with binary outcome data using a multi-level model. *Stat Med*. 2000;19(19):2675–88.
19. Briggs AH, Mooney CZ, Wonderling DE. Constructing confidence intervals for cost-effectiveness ratios: an evaluation of parametric and non-parametric techniques using Monte Carlo simulation. *Stat Med*. 1999;18(23):3245–62.
20. Nixon RM, Thompson SG. Methods for incorporating covariate adjustment, subgroup analysis and between-centre differences into cost-effectiveness evaluations. *Health Econ*. 2005;14(12):1217–29.
21. Willan AR, Briggs AH, Hoch JS. Regression methods for covariate adjustment and subgroup analysis for non-censored cost-effectiveness data. *Health Econ*. 2004;13(5):461–75.
22. Lipsitz S, Fitzmaurice G, Ibrahim J, Sinha D, Parzen M, Lipschultz S. Joint generalized estimating equations for multivariate longitudinal binary outcomes with missing data: an application to acquired immune deficiency syndrome data. *J R Stat Soc Ser A*. 2009;172(Pt 1):3–20.
23. Davison AC, Hinkley DV. *Bootstrap methods and their application*. Cambridge, UK: Cambridge University Press; 1997.
24. Flynn T, Peters T. Conceptual issues in the analysis of cost data within cluster randomized trials. *J Health Serv Res Policy*. 2005;10(2):97–102.
25. Grieve R, Nixon R, Thompson SG. Bayesian hierarchical models for cost-effectiveness analyses that use data from cluster randomized trials. *Med Decis Making*. 2010;30(2):163–75.
26. Bachmann MO, Fairall L, Clark A, Mugford M. Methods for analyzing cost effectiveness data from cluster randomized trials. *Cost Eff Resour Alloc*. 2007;5:12.
27. O'Hagan A, Stevens JW. A framework for cost-effectiveness analysis from clinical trial data. *Health Econ*. 2001;10(4):303–15.
28. Singh B, Ullah A. Estimation of seemingly unrelated regressions with random coefficients. *J Am Stat Assoc*. 1974;69(345):191–5.
29. Zellner A, Ando T. A direct Monte Carlo approach for Bayesian analysis of the seemingly unrelated regression model. *J Econom*. 2010;159(1):33–45.
30. Zellner A, Ando T. Bayesian and non-Bayesian analysis of the seemingly unrelated regression model with Student-t errors, and its application for forecasting. *Int J Forecast*. 2010;26(2):413–54.
31. Smeeth L, Ng ES. 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*. 2002;23(4):409–21.
32. Hardin JW, Hilbe JM. *Generalized Estimating Equations*. Boca Raton, FL: Chapman & Hall/CRC; 2003.
33. Bellamy SL, Gibberd R, Hancock L, Howley P, Kennedy B, Klar N, et al. Analysis of dichotomous outcome data for community intervention studies. *Stat Methods Med Res*. 2000;9(2):135–59.
34. Flynn TN, Peters TJ. Use of the bootstrap in analysing cost data from cluster randomised trials: some simulation results. *BMC Health Serv Res*. 2004;4:33–43.
35. O'Hagan A, Stevens JW. Assessing and comparing costs: how robust are the bootstrap and methods based on asymptotic normality? *Health Econ*. 2003;12(1):33–49.
36. Thompson SG, Nixon RM. How sensitive are cost-effectiveness analyses to choice of parametric distributions? *Med Decis Making*. 2005;25(4):416–23.
37. Austin PC. 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*. 2010;6(1).
38. Rodriguez G, Goldman M. An assessment of estimation procedures for multilevel models with binary responses. *J R Stat Soc Ser A*. 1995;158:73–89.
39. Turner RM, White IR, Croudace T. Analysis of cluster randomized cross-over trial data: a comparison of methods. *Stat Med*. 2007;26(2):274–89.
40. Pinto EM, Willan AR, O'Brien BJ. Cost-effectiveness analysis for multinational clinical trials. *Stat Med*. 2005;24(13):1965–82.
41. Briggs A, Nixon R, Dixon S, Thompson S. Parametric modelling of cost data: some simulation evidence. *Health Econ*. 2005;14(4):421–8.
42. Eldridge SM, Ashby D, Kerry S. Sample size for cluster randomized trials: effect of coefficient of variation of cluster size and analysis method. *Int J Epidemiol*. 2006;35(5):1292–300.
43. Nixon RM, Wonderling D, Grieve RD. Non-parametric methods for cost-effectiveness analysis: the central limit theorem and the bootstrap compared. *Health Econ*. 2010;19(3):316–33.
44. Cheyne H, Hundley V, Dowding D, Bland JM, McNamee P, Greer I, et al. Effects of algorithm for diagnosis of active labour: cluster randomised trial. *BMJ*. 2008;337:a2396.
45. Davies MJ, Heller S, Skinner TC, Campbell MJ, Carey ME, Craddock S, et al. 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*. 2008;336(7642):491–5.
46. Fairall LR, Zwarenstein M, Bateman ED, Bachmann M, Lombard C, Majara BP, et al. Effect of educational outreach to nurses on tuberculosis case detection and primary care of respiratory illness: pragmatic cluster randomised controlled trial. *BMJ*. 2005; 331(7519):750–4.
47. Morgan K, Thompson J, Dixon S, Tomeny M, Mathers N. Predicting longer-term outcomes following psychological treatment for hypnotic-dependent chronic insomnia. *J Psychosom Res*. 2003;54(1):21–9.
48. Morrell CJ, Slade P, Warner R, Paley G, Dixon S, Walters SJ, et al. Clinical effectiveness of health visitor training in psychologically informed approaches for depression in postnatal women: pragmatic cluster randomised trial in primary care. *BMJ*. 2009; 338:a3045.
49. 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. *J Epidemiol Community Health*. 2004;58(12):1004–10.
50. Murphy AW, Cupples ME, Smith SM, Byrne M, Byrne MC, Newell J. Effect of tailored practice and patient care plans on secondary prevention of heart disease in general practice: cluster randomised controlled trial. *BMJ*. 2009;339:b4220.
51. Oluboyede Y, Goodacre S, Wailoo A. Cost effectiveness of chest pain unit care in the NHS. *BMC Health Serv Res*. 2008; 8:174.
52. R. The R project for statistical computing. Available from URL: <http://www.r-project.org/>
53. STATA. Stata Programming Reference Manual, Version 11. College Station, TX: StataCorp; 2009.
54. Spiegelhalter D, Thomas A, Best N, Lunn D. WinBUGS User Manual, Version 1.4. Cambridge, UK: MRC Biostatistics Unit; 2003.
55. Kass R, Wasserman L. The selection of prior distributions by formal rules. *J Am Stat Assoc*. 1996;91(435):1343–70.
56. Claxton K. The irrelevance of inference: a decision-making approach to the stochastic evaluation of health care technologies. *J Health Econ*. 1999;18(3):341–64.
57. Fenwick E, Marshall DA, Blackhouse G, Vidaillet H, Slee A, Shemanski L, et al. 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*. 2008;11(3):365–75.
58. Willan AR. Clinical decision making and the expected value of information. *Clin Trials*. 2007;4(3):279–85.
59. Thompson SG, Nixon RM, Grieve R. Addressing the issues that arise in analysing multicentre cost data, with application to a multinational study. *J Health Econ*. 2006;25(6):1015–28.
60. Turner RM, Omar RZ, Thompson SG. Bayesian methods of analysis for cluster randomized trials with binary outcome data. *Stat Med*. 2001;20(3):453–72.
61. Donner A. Some aspects of the design and analysis of cluster randomization trials. *Appl Stat*. 1998;47:95–113.
62. Campbell MK, Fayers PM, Grimshaw JM. Determinants of the intracluster correlation coefficient in cluster randomized trials: the case of implementation research. *Clin Trials*. 2005;2(2):99–107.
63. Nixon RM, Thompson SG. Parametric modelling of cost data in medical studies. *Stat Med*. 2004;23(8):1311–31.
64. OECD. Country statistical profiles: South Africa. Available from URL: <http://stats.oecd.org/index.aspx?queryid=23123>
65. White H. A heteroskedasticity-consistent covariance-matrix estimator and a direct test for heteroskedasticity. *Econometrica*. 1980;48(4):817–38.
66. Huber P. The behavior of maximum likelihood estimates under nonstandard conditions. In: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1. Le Cam LM, Neyman J, eds. Berkeley: University of California Press; 1967:221–233.
67. Williams RL. A note on robust variance for cluster-correlated data. *Biometrics*. 2000;56(June):645–6.
68. Skrondal A, Rabe-Hesketh S. Generalized Latent Variable Modeling: Multilevel, Longitudinal and Structural Equation Models. Boca Raton, FL: Chapman & Hall/CRC; 2004.
69. Ng ESW. A review of mixed-effects models in S-plus (version 6.2). Available from URL: <http://www.bristol.ac.uk/cmm/learning/mmsoftware/reviewsplus.pdf>