
Downloaded from: http://researchonline.lshtm.ac.uk/id/eprint/20729/

DOI: https://doi.org/10.1177/0272989X11418372

Usage Guidelines:

Please refer to usage guidelines at https://researchonline.lshtm.ac.uk/policies.html or alternatively contact researchonline@lshtm.ac.uk.

Available under license: http://creativecommons.org/licenses/by-nc-nd/2.5/
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.

Cost-effectiveness analyses (CEAs) of group-based interventions often use data from cluster randomized trials (CRTs). 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). 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. For these studies to provide a sound basis for decision making, appropriate statistical methods need to be developed and used. CEAs based on randomized controlled trials (RCTs), where individual patients are randomized, have well-established methods. However, statistical
methods for CEAs of CRTs have received limited attention.\textsuperscript{10} A review found that less than 10% of published CEAs of CRTs adopted appropriate statistical methods.\textsuperscript{1}

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 intracluster 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,\textsuperscript{2,5,11} encourages incorrect inferences,\textsuperscript{12–15} and can also lead to bias.\textsuperscript{16,17} Appropriate methods for handling clustering in clinical outcomes are well developed and can include multilevel models (MLMs) and generalized estimating equations (GEEs).\textsuperscript{18}

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\textsuperscript{19–21} 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),\textsuperscript{21} GEEs,\textsuperscript{22} the nonparametric 2-stage bootstrap (TSB),\textsuperscript{23} and MLMs.\textsuperscript{20} 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\textsuperscript{24} and MLMs\textsuperscript{25} have been proposed for CEAs of CRTs, but the only study\textsuperscript{26} to compare these methods used data from a single CRT. A simulation study\textsuperscript{24} 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.\textsuperscript{9,20,21,27}

Seemingly Unrelated Regression (SUR)

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

$$

e_{ij} = \beta_{0} + \beta_{1}t_{i} + c_{ij} \left( e_{ij} \right) - BVN \left( 0 \right), \left( \sigma_{1}^{2} \right), \left( \sigma_{2}^{2} \right)
$$

where $t_i$ is the treatment indicator ($t_i = 0$ for control and 1 for treatment group). The parameters of interest, the incremental costs ($\beta_{1}^{*}$) and outcomes ($\beta_{1}^{*}$), can be estimated by ordinary least squares (OLS). SUR assumes the individual error terms ($e_i$) have a bivariate normal distribution (BVN), with variances $\sigma_{1}^{2}$ and $\sigma_{2}^{2}$. Conceptually, SUR can be extended to accommodate clustering by including random effects,\textsuperscript{28} 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.\(^a\)

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.\(^{29,30}\) Finally, the asymptotic assumptions underlying the robust variance estimator may not hold in CRTs with few clusters per treatment arm.\(^31\) The problem can be exacerbated by skewed outcomes (or costs) or imbalanced cluster sizes.\(^{18}\) 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.\(^2,3,32\) While multivariate GEEs have been developed to recognize potential correlation between binary end points,\(^22\) 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:

\[
\begin{align*}
    e_{ij} & = \beta_0 + \beta_1 t_k + e_{ij}, \\
    e_{ij} & = \beta_0 + \beta_1 t_k + e_{ij}, \\
    e_{ij} & \sim N(0, \sigma^2), \\
    e_{ij} & \sim N(0, \sigma^2).
\end{align*}
\]  

(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.\(^{13,15,33,65-68}\)

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

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

**TSB without shrinkage correction.** One proposed TSB algorithm requires resampling clusters and then individuals within each resampled cluster (both with replacement).\(^23\) 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.\(^23,24,34\)

**TSB with shrinkage correction.** Davison and Hinkley\(^23\) 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 shrunk cluster means and secondly resampling the standardized individual-level residuals across all clusters. Bootstrap data sets are
that the MLM may fail to converge if the CRT has

Both bootstrap routines rely on asymptotic
assumptions, and it is unclear whether they are
satisfied with few clusters, particularly if data
are nonnormal.\textsuperscript{35,36} Furthermore, the TSB routines
described above were only proposed for balanced
clusters,\textsuperscript{23,24,34} which may make the method inap-
propriate for CEAs of CRTs with imbalanced clusters.\textsuperscript{1}
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.\textsuperscript{25} Unlike
SUR, MLMs can explicitly recognize clustering by
including additional random terms, $u^c_i, u^e_i$, which
in equation 3 below represent the differences in
the cluster mean costs and outcomes from the over-
tall means in each treatment group. These random
effects are assumed to follow a bivariate normal
distribution, with variances $\tau^2_c$ and $\tau^2_e$. MLMs
acknowledge individual- and cluster-level correlation
between costs and outcomes through the parameters
$\rho$ and $\psi$. The coefficients $\beta^c$ and $\beta^e$ still represent
incremental costs and outcomes after allowing for
clustering. Like the SUR model, this particular
MLM assumes that the individual error terms ($e$) are normally distributed, but more generally, alter-
native distribution assumptions can be made for
costs, outcomes, or both.

\begin{align*}
  c_{ij} &= \beta^c_0 + \beta^c_1 t_j + u^c_i + \epsilon^c_{ij} \\
  e_{ij} &= \beta^e_0 + \beta^e_1 t_j + u^e_i + \epsilon^e_{ij}
\end{align*}

MLMs can be estimated and interpreted from a fre-
quenstist perspective, generally implemented with
maximum likelihood or with a Bayesian approach
typically using Markov chain Monte Carlo (MCMC)
methods. Current software options for MCMC esti-
mation afford a wide choice of distributional
assumptions.\textsuperscript{20} A concern with either approach is
that the MLM may fail to converge if the CRT has
few individuals per cluster.\textsuperscript{37,38}

Monte Carlo Simulations

Data-generating process. The simulation study was
designed to test the methods across a wide range of cir-
cumstances typically found in CEAs of CRTs. Our con-
ceptual 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
design size imbalance, ICCs, skewness in the cost distribu-
tion, 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 con-
structed to reflect the specific form of clustering found
in CRTs.\textsuperscript{12,37,39} The design allowed for a wide range of
parameters to be varied and could accommodate differ-
ent parametric distributions for costs and outcomes.
As in previous simulation studies in economic evalu-
ation, we assumed a linear additive treatment effect
throughout.\textsuperscript{23,24,40} We simulated cost ($c$) and outcome
data ($e$) from CRTs with $M$ clusters per arm and $n_m$
($m = 1 \ldots M$) individuals per cluster. Data were gener-
ated firstly at the cluster level and then at the individ-
ual level according to equation 4 below.

Cluster level means:

\begin{align*}
  \phi^c_j &\sim \text{dist}(\beta^c_0 + \beta^c_1 t_j, \tau_c) \\
  c_{ij} &\sim \text{dist}(\phi^c_j, \sigma^c)
\end{align*}

Individual-level data:

\begin{align*}
  \phi^e_j &\sim \text{dist}(\beta^e_0 + \beta^e_1 + \gamma(\phi^c_j - (\beta^c_0 + \beta^c_1 t_j)), \tau_e) \\
  e_{ij} &\sim \text{dist}(\phi^e_j + \theta(e^c_{ij} - \phi^c_j), \sigma_e)
\end{align*}

Cluster-level mean costs ($\phi^c_j$) and outcomes ($\phi^e_j$)
were simulated for the $j$th cluster. These were
assumed to follow a certain distribution charac-
terized by the cluster means for the control
($\beta^c_0, \beta^c_1 + \gamma(\phi^c_j - (\beta^c_0 + \beta^c_1 t_j))$) and treatment
($\beta^e_0 + \beta^e_1, \beta^e_1 + \gamma(\phi^c_j - (\beta^c_0 + \beta^c_1 t_j))$) groups and the corres-
ponding cluster-level standard deviations ($\tau_c, \tau_e$). This
mechanism allowed costs and outcomes to be corre-
lated 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 $\theta = \rho(\sigma_c/\sigma_e)$. ICCs
were set to recognize the proportion of the total vari-
ance 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

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Rationale</th>
<th>Base Case</th>
<th>SA</th>
<th>Final Case</th>
<th>Justification for Parameter Levels</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of clusters per arm</td>
<td>GEE, SUR, and TSB all rely on asymptotic assumptions</td>
<td>20</td>
<td>3 to 30</td>
<td>15</td>
<td>Base case: 20 clusters per arm suggested for asymptotics to hold.2,61 SA: takes lower, upper quartiles from literature review. Final case: median number of clusters from literature review.</td>
</tr>
<tr>
<td>No. of individuals per cluster</td>
<td>MLM may have convergence issues with few cases per cluster</td>
<td>50</td>
<td>10 to 80</td>
<td>30</td>
<td>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.</td>
</tr>
<tr>
<td>Level of imbalance($cv_{imb}$) of cluster size</td>
<td>GEE, SUR, and TSB have not been assessed with imbalanced clusters</td>
<td>0</td>
<td>0 to 1</td>
<td>0.5</td>
<td>Base case: previous methods articles.23,24,34 SA: cluster-size imbalance informed by range of values reported across case studies and previous study.42 Final case: median from the case studies.</td>
</tr>
<tr>
<td>ICC for costs</td>
<td>To assess if methods can handle high levels of clustering</td>
<td>0.01</td>
<td>0 to 0.3</td>
<td>0.05</td>
<td>Base case: Start with low ICC as per previous methods articles.24,34 SA: range of ICCs from case studies and previous study.62 Final case: median from case studies.</td>
</tr>
<tr>
<td>ICC for outcomes</td>
<td>As above</td>
<td>0.01</td>
<td>0 to 0.3</td>
<td>0.02</td>
<td>Base case: 30% of studies from literature review have ICCs for outcomes ≤0.01. SA: range from literature review and previous methods studies.42 Final case: median from literature review.</td>
</tr>
<tr>
<td>Coefficient of variation($cv_{cost}$) of cost distribution</td>
<td>SUR, MLM, and GEE assume errors follow a normal distribution</td>
<td>0.2</td>
<td>0.25 to 3</td>
<td>0.5</td>
<td>Base case: start with normal distribution; no skewness as per previous simulation studies.36,63 SA: $\gamma$ distribution; range for $cv_{cost}$ from previous simulation studies.41,43 Final case: gamma distribution; median $cv_{cost}$ from case studies.</td>
</tr>
<tr>
<td>Individual-level correlation of costs and effects</td>
<td>GEE assumes costs and outcomes are independent</td>
<td>0.2</td>
<td>−0.5 to 0.5</td>
<td>−0.2</td>
<td>Base case: plausible level of individual-level correlation.54,43 SA: based on the range from case studies. Final case: median from the case studies.</td>
</tr>
<tr>
<td>Cluster-level correlation of costs and effects</td>
<td>GEE as above; SUR ignores cluster-level correlation</td>
<td>0</td>
<td>−0.5 to 0.5</td>
<td>0.1</td>
<td>Base case: conservative value assuming no correlation at the cluster level.24 SA: based on the range from case studies. Final case: median from case studies.</td>
</tr>
</tbody>
</table>

Note: SA = sensitivity analyses; ICC = intracluster correlation coefficient; GEE = generalized estimating equation; SUR = seemingly unrelated regression; TSB = nonparametric 2-stage bootstrap; MLM = multilevel model.
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. 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. The MCMC estimation consisted of 5000 iterations, 3 parallel chains with different starting values, and assuming diffuse priors.

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 increase in net benefits from taking the optimal decision after resolving current uncertainty.

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. 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, 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. EVPI was calculated assuming that the INB was normally distributed.

\[ ICC_c = \frac{\tau_c^2}{\sigma_c^2 + \tau_c^2} \]

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, previous methods articles and simulation studies, and 8 case studies. In the 62 studies, previous methods articles and simulation studies,24,41–43 and 8 case studies.44–51

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. The performance of each method in estimating incremental costs, incremental QALYs, and INBs was reported.

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

---

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.
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 (cvimb of 0.5 and 1), combined with highly skewed costs (cvcost = 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 2  Bias, rMSE, CI Coverage, and Width of the Mean INB for the Base Case (True INB = £1000)

<table>
<thead>
<tr>
<th>Method</th>
<th>Without Robust SE</th>
<th>With Robust SE</th>
<th>Without Shrinkage Correction</th>
<th>With Shrinkage Correction</th>
<th>MLMa ML</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean bias (SE)</td>
<td>–1.999 (2.45)</td>
<td>–1.999 (2.45)</td>
<td>–1.999 (2.45)</td>
<td>–2.108 (2.45)</td>
<td>–1.999 (2.45)</td>
</tr>
<tr>
<td>rMSE</td>
<td>109.45</td>
<td>109.45</td>
<td>109.45</td>
<td>109.52</td>
<td>109.45</td>
</tr>
<tr>
<td>CI coverage</td>
<td>0.891</td>
<td>0.940</td>
<td>0.933</td>
<td>0.943</td>
<td>0.950</td>
</tr>
<tr>
<td>Mean CI width</td>
<td>353.6</td>
<td>423.7</td>
<td>417.7</td>
<td>539.1</td>
<td>427.5</td>
</tr>
<tr>
<td>Lower tail error rate</td>
<td>0.048</td>
<td>0.030</td>
<td>0.033</td>
<td>0.009</td>
<td>0.028</td>
</tr>
<tr>
<td>Upper tail error rate</td>
<td>0.051</td>
<td>0.029</td>
<td>0.035</td>
<td>0.011</td>
<td>0.030</td>
</tr>
</tbody>
</table>

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.
**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.
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\textsuperscript{24,34} did not consider MLMs or GEEs, and other studies just compared the methods using a single case study.\textsuperscript{26} 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.\textsuperscript{1} 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.\textsuperscript{26} We find that without the shrinkage correction, the TSB overstates the uncertainty, but once the correction is applied, the method gives good CI coverage. This finding contrasts with those of a previous simulation study\textsuperscript{24} 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.\textsuperscript{1} The general literature on GEEs has reported similar sample size requirements for asymptotic assumptions to hold,\textsuperscript{13,15,18} 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\textsuperscript{25} 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 4

<table>
<thead>
<tr>
<th></th>
<th>SUR Without Robust SE</th>
<th>SUR With Robust SE</th>
<th>GEE Without Robust SE</th>
<th>GEE With Robust SE</th>
<th>TSB Without Shrinkage Correction</th>
<th>TSB With Shrinkage Correction</th>
<th>MLM ML</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean bias (SE)</td>
<td>6.63 (4.40)</td>
<td>6.63 (4.41)</td>
<td>6.63 (4.40)</td>
<td>6.63 (4.40)</td>
<td>7.10 (4.38)</td>
<td>6.85 (4.38)</td>
<td>7.95 (4.33)</td>
</tr>
<tr>
<td>rMSE</td>
<td>197</td>
<td>197</td>
<td>197</td>
<td>197</td>
<td>196</td>
<td>196</td>
<td>194</td>
</tr>
<tr>
<td>CI coverage</td>
<td>0.858</td>
<td>0.921</td>
<td>0.920</td>
<td>0.920</td>
<td>0.978</td>
<td>0.944</td>
<td>0.938</td>
</tr>
<tr>
<td>Mean CI width</td>
<td>583</td>
<td>726</td>
<td>724</td>
<td>724</td>
<td>924</td>
<td>778</td>
<td>754</td>
</tr>
<tr>
<td>Lower tail error</td>
<td>0.072</td>
<td>0.041</td>
<td>0.041</td>
<td>0.041</td>
<td>0.014</td>
<td>0.029</td>
<td>0.033</td>
</tr>
<tr>
<td>Upper tail error</td>
<td>0.120</td>
<td>0.038</td>
<td>0.039</td>
<td>0.039</td>
<td>0.010</td>
<td>0.028</td>
<td>0.030</td>
</tr>
</tbody>
</table>

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.
in a different context, this corroborates previous findings that suggest methods assuming normality may be quite robust to skewed cost data. 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. Also, CEAs may have more complex variance structures than those considered. 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. 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**

52. R. The R project for statistical computing. Available from URL: http://www.r-project.org/