Multilevel models for cost-effectiveness analyses that use cluster randomised trials: an approach to model choice

Running title: Model choice for CEA of cluster trials

Edmond S.-W. Ng CStat[1], Karla DiazOrdaz PhD[1], Richard Grieve PhD[1], Richard M. Nixon PhD[2], Simon G. Thompson DSc[3], James R. Carpenter DPhil[4]

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

[2]Modeling and Simulation Group, Novartis Pharma AG, Basel, Switzerland.

[3]Department of Public Health and Primary Care, University of Cambridge.

[4]Department of Medical Statistics, London School of Hygiene and Tropical Medicine, London, UK.

Corresponding author

Richard Grieve, Department of Health Services Research and Policy, London School of Hygiene and Tropical Medicine, London, WC1E 7HT, United Kingdom. Tel: +44 (0) 7927 2255, Fax: +44 (0)20 7927 2701, Email: richard.grieve@lshtm.ac.uk

Abstract

Multilevel models (MLMs) provide a flexible modelling framework for cost-effectiveness analysis (CEA) that use cluster randomised trials. However, there is a lack of guidance on how to choose the most appropriate MLMs. This paper illustrates an approach for deciding what level of model complexity is warranted; in particular how best to accommodate complex variance-covariance structures, right-skewed costs and partially-observed data. Our proposed models differ according to whether or not they allow individual-level variances and correlations to differ across treatment arms or clusters, and by the assumed cost distribution (Normal, Gamma, Inverse Gaussian). The models are fitted by Markov chain Monte Carlo methods. Our approach to model choice is based on four main criteria: the characteristics of the data, model pre-specification informed by the previous literature, diagnostic plots and assessment of model appropriateness. This is illustrated by re-analysing a previous CEA that uses data from a cluster randomised trial. We find that the most useful criterion for model choice was the deviance information criterion, which distinguished amongst models with alternative variance-covariance structures, as well as between those with different cost distributions. This strategy for model choice can help CEA provide reliable inferences for policy-making when using cluster trials, including those with partially-observed data.

Key words: Cost-effectiveness analysis, cluster randomised trial, multilevel models

1. Introduction

Policy-makers require cost-effectiveness analyses (CEAs) to help decide which health care programmes to prioritise.[1-5] For interventions that operate at a group-level, CEAs commonly use data from cluster randomised trials (CRTs). A fundamental issue raised by the cluster design is that individuals within a cluster are likely to be more similar in their characteristics and the care they receive than those in different clusters. Methods that accommodate this clustering are well-established for analysing clinical outcomes; however a review of 62 published CEAs that used CRTs found that 37(60%) adopted methods that disregarded clustering, which can underestimate statistical uncertainty, but can also provide misleading point estimates. [6]

CEA provide evidence on the relative costs and health outcomes of alternative health care interventions. This requires studies to report the effects of alternative treatments on the joint distribution of costs and health outcomes, i.e. to estimate the between-treatment differences in mean costs and health outcomes, together with their respective variances and covariances [7]. To meet this requirement, statistical methods for CEA that use data from cluster trials must address specific challenges. They must accommodate clustering in individuals' costs and health outcomes, but also recognise correlation between these variables at both individual and cluster levels. [7-10] Statistical methods are also required that make appropriate assumptions about the distribution of the outcome variables, recognising that costs tend to be heavily right-skewed. [11-13] The variance-covariance structure may be complex; that is individual-level costs may have variances, and correlations with health outcomes, that differ across clusters, because for example of clinical practice variations. [14] Typically, the requisite data are only partially-observed [15], which needs to be handled in the analyses and acknowledged in the definition and calculation of measures of model fit.[16]

Previous research has proposed alternative multi-level models (MLMs) for CEA, and applied them to studies that use data from multicentre and cluster trials ([10, 14, 17, 18]). However, within this framework many plausible MLMs can be specified and there is no guidance available on model choice, particularly in settings where data are partially-observed. This paper focuses on approaches for CEA that use partially-observed data from cluster trials, and assumes these data are Missing At Random (MAR).

The aim of this paper is to propose and illustrate a set of criteria for choosing MLM in CEA that use cluster trial data. The criteria proposed are: visual inspection of the data, pre-specifying models drawing on the previous literature [13, 14, 19-21], diagnostic plots [22], and assessment of model appropriateness using the Deviance Information Criterion [23]. We exemplify these criteria using a case study typical of CEA that use CRT data [6].

The next section gives a brief overview of the case study followed by a proposed strategy for choosing amongst MLMs. In Section 3 we introduce models drawing on the previous literature. Section 4 applies these alternative models to the case study; we present diagnostic plots, assessments of model appropriateness, and CEA results. Section 5 discusses the proposed strategy to model choice and suggests some areas for further research.

2. Features of case study and implications for choice of MLMs

2.1 Case study overview

The aim of the Secondary Prevention of Heart disEase in geneRal practicE trial (SPHERE) study was to assess the effectiveness and cost-effectiveness of a secondary prevention strategy for patients with coronary heart disease (CHD). [24, 25] In SPHERE, 48 General Practices (with 903 patients) were randomised to intervention (practices and patients had access to tailored care plans) or control (patients received usual care). The main endpoints were health

service costs and health-related quality of life (HRQoL) assessed by administering the SF-12 questionnaire and recorded 18 months post randomisation. HRQoL was converted into a utility measure using the SF-6D algorithm [26], and combined with mortality data to report quality-adjusted life years (QALYs) over 18 months. The CEA reported incremental QALYs, costs and the incremental net monetary benefit, known as the INB. The INB reports the relative value for money of alternative health care programmes.[27] The INB is calculated by estimating the difference between the treatment alternatives in the mean health outcomes, in this case QALYs, valuing this difference by the threshold willingness to pay for a unit of health gain λ (€20 000 per QALY in our illustrative analysis), and subtracting from this the incremental cost, so: INB= $\lambda\Delta e - \Delta c$; where $\Delta e = \bar{e}_1 - \bar{e}_0$, and $\Delta c = \bar{c}_1 - \bar{c}_0$ are the incremental health outcomes and costs for treatment (subscript 1) versus control (subscript 0).

In Table 1, to help motivate the subsequent MLMs we report incremental costs, QALYs and the INB with two contrasting approaches. Firstly, we report 'individual level' incremental effects by simply contrasting the means for all individuals within each randomised arm. This approach disregards clustering and assumes that each individual's endpoint is independent, and has equal weight; i.e. the endpoints for a cluster with many patients are given higher weight than a cluster with few patients. Secondly, we calculate the INB with the summary measures of mean costs and QALYs from each cluster. Under this approach, the mean endpoints in each cluster have the same weight irrespective of the numbers per cluster. For both approaches we estimate the variance of the INB with a standard approach which assumes that the central limit theorem applies, and that variances and correlations between costs and outcomes are constant [28].

The alternative ways of weighting the data has little impact on the point estimates of the INB which are around €600 for each approach (Table 1). By contrast, the SEs of the INB are

larger with the summary approach which disregards information at the individual-level versus the individual-level approach which assumes each observation is independent, Rather than weighting the cluster means equally or according to the numbers per cluster, the subsequent MLMs weight the data in each cluster according to the amount of information within versus between clusters. The subsequent MLMs differ according to assumptions they make about whether or not the individual-level variances and covariances are assumed constant within clusters, and according to the assumed distribution of individual-level costs.

TABLE 1

SPHERE illustrates several potential challenges for CEA that use CRT data, beyond the potential clustering of costs and health outcomes. In pragmatic CRTs like SPHERE, the treatment protocols tend to accommodate clinical practice variations across clusters. Hence resource use and costs may have individual-level variances that differ across clusters. Secondly, individual costs and health outcomes tend to be correlated (for example patients with lower health status may incur higher costs), and practice variations can result in individual-level correlations that differ across clusters (for example, some clusters might monitor patients with lower health status more intensively than others). Thirdly, costs tend to be heavily right-skewed; in SPHERE a small proportion of patients have lengthy hospital stays at high cost. Complexities, such as cost skewness and complex variation are present in CEA of CRT more generally.[6]

We now examine the features of the SPHERE data most relevant for the choice of MLM (Figure 1, Table 1). The cost histograms show typically long right tails in both treatment arms (Figure 1). We overlay three alternative distributions (Normal, Gamma and Inverse Gaussian)

previously proposed for cost analysis [13, 19-22]; here the Inverse Gaussian appears to fit the cost data somewhat better. For QALYs assuming Normality appears reasonable (Figure 1).

The strength of clustering is commonly reported by the intracluster correlation coefficient (ICC) assuming constant within-cluster variation. When individual-level variances differ across clusters, the ICC can be uninformative, so we also report the I^2 statistic which was originally developed for measuring the degree of inconsistency in studies' results in a meta-analysis.[29] It is calculated as $I^2=100\%\times(Q-df)/Q$, where Q is Cochran's heterogeneity statistic and df the degrees of freedom and they are generally reported in meta analysis by, for example, the *metaan* command in Stata.[30] For each endpoint, the I^2 describes the percentage of the total variability which is due to heterogeneity, in this case differences in the means for each endpoint across clusters, rather than chance.[31] In SPHERE the I^2 is higher for costs (85%) than for QALYs (I^2=52%). The standard deviations (SDs) for the individual-level QALYs were similar across treatment groups and clusters, but differed for the individual-level costs (Table 1, Figure 2), and tended to increase with mean costs per cluster (Levene's test of equal variances [32] across all clusters showed evidence of heterogeneity, $p<0.0001$). The correlations between individual costs and QALYs appear to differ across clusters (range –0.7 to 0.8) but a forest plot of the Fisher z-transformed correlations [33] suggests these differences are compatible with chance, and the heterogeneity across clusters is low (I^2=0% for control and 27% for treatment clusters). These correlations were calculated using cases without missing endpoint data.

At the cluster-level, the control group mean costs in SPHERE has a higher SD and lower correlation with mean QALYs than the treatment group (Table 1). However, SPHERE is typical of most CEA that use CRT data in that there is limited information at the cluster level

to assess whether parameters such as the cluster-level SD or correlation between mean costs and mean QALYs, differ between treatment groups.

3. MLMs for the CEA

SPHERE illustrates complexities that pervade CEA that use CRT more generally, and these have implications for the choice of MLM. The mean costs but also the individual-level variances can be heterogeneous across clusters. [34] Costs and outcomes tend to be correlated at the individual level, and may differ by cluster. Cluster-level variances but also correlations between costs and health outcomes may differ by treatment group. Bivariate MLMs for CEA of CRT have been previously proposed. [14, 35, 36] In the next stage of the proposed strategy we pre-specify a series of MLMs that extend those previously proposed [14, 35] by allowing individual-level correlations to vary across clusters, and cluster-level SDs and correlations to differ between treatment groups. We present a flexible modelling framework that recognises this complex variance-covariance structure [14], accommodates highly skewed costs, and can be applied to CEA with partially observed endpoint data.

MLMs for handling skewed costs

Cost data tend to be right skewed, and models that assume Normality can provide inefficient estimates of the mean cost. Any approach for addressing cost skewness has to recognise that the prime interest is in the treatment's effect on the arithmetic mean costs (and health outcomes). So if the costs are transformed (for example, by log-transformation), they have to then be back-transformed appropriately, which is not straightforward when there is heteroscedasticity [22, 37, 38]. While there is no consensus on which distributions should be considered for modelling health care costs, it is acknowledged that model specification is inherently complex, and the correct parametric specification may well differ according to the specific dataset [30]. Our strategy for model choice is informed by recommendations from

the literature which encourage the use of Generalised Linear Models (GLMs) as they allow for non-Normal distributions but directly report the effect of treatment on mean costs.[13, 19-22]. Specifically we consider costs to have Gamma and Inverse Gaussian distributions which have variances proportional to increasing power of their means, i.e. μ^2 for Gamma and μ^3 for Inverse Gaussian, and consider their relative fit and appropriateness for the dataset in question [39] There is evidence in SPHERE, and in CEA more generally, that the cost variance is not independent of the mean (Figure 2a). By contrast for QALYs, the variance does not appear to differ according to the mean in each cluster (Figure 2b).

3.1 Range of bivariate MLMs

Each of the following MLMs jointly models costs and health outcomes (e.g. QALYs), includes a linear additive treatment effect, and recognises potential heterogeneity in the mean costs and outcomes with cluster-specific random effects. The range of models considered differs firstly, by choice of cost distribution (Normal versus Gamma versus Inverse Gaussian) and secondly, according to the assumed variance-covariance structure. Following the notation of Nixon and Thompson [10] c_{ijk} and e_{ijk} represent costs and health outcomes for the *i*-th individual in the *j*-th cluster randomised to the *k*-th treatment arm. Here *j* defines *k* as all individuals within a cluster receive the same randomised treatment, and for simplicity *k* takes the value 0 or 1 according to whether the cluster is randomised to the control or treatment group.

Now, we introduce, and assume throughout, bivariate Normal cluster random effects u_j^c and u_j^e for cost and health outcomes respectively,

$$\begin{pmatrix} u_j^c \\ u_j^e \end{pmatrix} \sim BVN\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \tau_{ck}^2 & \phi_k \tau_{ck} \tau_{ek} \\ & \tau_{ek}^2 \end{pmatrix}\right)$$

where variances $\tau_{ck}^2$and τ_{ek}^2 and correlation ϕ_k may be specific to each treatment arm.

We build the bivariate model on the expectations η_{ijk}^c and $\eta_{ijk}^e$of the two outcomes, c_{ijk} and e_{ijk} defined conditionally on the two cluster effects, following the bivariate models introduced by Nixon and Thompson [10, 13]

$$\boldsymbol{\eta_{ijk}^c = \beta_0^c + \beta_1^c t_k + u_j^c} \qquad [1]$$

and health outcomes conditional on costs and cluster effects

$$\boldsymbol{\eta_{ijk}^e = \beta_0^e + \beta_1^e t_k + u_j^e + \gamma_j(c_{ijk} - \eta_{ijk}^c)} \qquad [2]$$

The mean costs and health outcomes for the control group are represented by β_0^c and β_0^e, and the incremental costs and health outcomes of the randomised treatment by β_1^c and β_1^e, and the individual-level correlation is represented by the regression coefficient, γ_j.

3.1.1 Alternative cost distributions

Throughout, we assume that $e_{ijk}\,|c_{ijk} \sim \text{Normal}(\eta_{ijk}^e, \sigma_{ej}^2)$, while for c_{ijk} we consider three possible distributions from the exponential family: Normal, Gamma and Inverse Gaussian.

We begin by assuming that costs follow a Normal distribution, $c_{ijk} \sim \text{Normal}(\eta_{ijk}^c, \sigma_{cj}^2)$ and so, that individual-level residuals, ε_{ijk}^c and ε_{ijk}^e, are drawn from a bivariate Normal distribution (BVN) with variances σ_{cj}^2 and σ_{ej}^2 and correlation ρ_j , all of which may differ across clusters. For these models, $\gamma_j = \rho_j \frac{\sigma_{ej}}{\sigma_{cj}}$.

The Gamma distribution is parameterised here in terms of a rate (r_{ijk}) and a dispersion parameter, shape (s_j), as follows:

$$f(c_{ijk}|s_j, \eta_{ijk}^c) = \frac{1}{\Gamma(s_j)}\left(\frac{s_j}{\eta_{ijk}^c}\right)^{s_j} {c_{ijk}}^{(s_j-1)} \exp\left(-\frac{s_j c_{ijk}}{\eta_{ijk}^c}\right),$$

where $c_{ijk} > 0; s_j > 0, \eta^c_{ijk} > 0$. The mean is $\eta^c_{ijk} = s_j / r_{ijk}$, and the variance = s_j / r^2_{ijk}. So we write $c_{ijk} \sim \text{Gamma}(r_{ijk}, s_j)$.

Finally, the Inverse Gaussian distribution has been proposed for modelling highly skewed costs.[22, 40] When, $c_{ijk} \sim \text{IG}(\eta^c_{ijk}, s_j)$, parameterised by its mean, η, and shape *s*, parameters, where η >0 and *s*>0 (with consequent variance η^3/s).

As before, these parameters (s_j, γ_j) may differ across clusters in the more complex models we consider.

3.1.2 Alternative Variance-covariance matrices

The MLMs considered here allow for different forms of complexity in the variance-covariance matrix. The differences across the MLMs are according to whether we allow for differences between randomised treatment groups and clusters in: the individual-level variances (BVN models) or the corresponding shape parameter s_j (Gamma and Inverse Gaussian models), and in the regression parameter between cost and health outcome, γ_j. The MLMs also differ in whether or not the cluster-level variances τ^2_{ck} and τ^2_{ek} and correlation ϕ_k are allowed to differ by treatment group.

TABLE 2

The simplest bivariate Normal model (1), assumes that all variances and correlations are constant across clusters, denoted by omitting the *j* and *k* subscripts from the variance-covariance matrices (Table 2). Models 2a-c allow the individual-level variances to differ first by treatment (2a), then by cluster using either different fixed (2b) (FEs) or random (2c) (REs) effects. The FE specification assumes that the individual-level variances are different and independent from one another, whereas under REs the individual-level variances are assumed

exchangeable, i.e. drawn from some common distribution. In many CRTs there are some clusters with few patients and here cluster-specific variance estimates can be imprecise. By using REs to model individual-level variances (and later correlations) we "borrow strength" from the larger clusters to estimate the variances (and correlations) of the smaller clusters.[41]

Models 3a-c allow individual-level correlations to differ first by treatment (3a), then by cluster using FEs (3b) or REs (3c). In Model 3c the individual-level correlation, ρ_j, is transformed into z_j using Fisher's z transformation [33]; z_j is modelled by REs with $z_j \sim N(\mu_z,\sigma_z^2)$. Models 4a-c extend models 3a-c in allowing cluster-level variances and correlations to differ by treatment group.

The analogous bivariate MLMs that assume costs follow a Gamma or Inverse Gaussian distribution allow the shape to differ either by treatment (*k*) or cluster (*j*) [13, 14]. For these models, we specified constant variances for health outcomes σ_{ej}^2, but allowed the regression parameter, γ_j, to vary by treatment (*k*) or cluster (*j*).

3.2 Model implementation

We applied each of the BVN models described in Table 2, and the analogous Gamma-Normal and Inverse Gaussian-Normal models to estimate incremental QALYs, costs and INBs in the SPHERE case study. When fitting the models we used rescaled costs (raw costs were divided by a value 4500). The mean (SD) of the scaled costs is 1.045 (1.209). The rescaling improved the stability of the MCMC estimation. Each model was fitted in WinBUGS by Markov chain Monte Carlo (MCMC) methods[42], with three chains, each with a burn-in of 5 000 iterations followed by 10 000 iterations. Convergence was assessed by visual inspection of the mixing of the chains and the Gelman-Rubin statistics.[43] The Inverse Gaussian is not a standard

distribution in WinBUGS, so we used a Bernoulli distribution with success rate $\pi_i = e^{l_i}$, where l_i is the IG log-likelihood . [44]

We assumed vague priors throughout. Wide Normal priors, N(0,10^6), were assumed for the mean QALYs, vague Gamma (0.01,0.01) for mean costs, wide Uniform priors for the SDs of the logarithms of individual QALYs, and for SDs of the individual costs (BVN models), i.e. σ_{ej}~ U(-10,5), and when σ_{cj}~ U(-10,10) for costs. We also assumed wide uniform priors for the cluster-level SDs, i.e. τ_{ek} ~U(0,10) for QALYs, τ_{ck}~ U(0,100) for costs [45], and U(-1,1) for the individual-level correlations ρ_j (BVN models), and throughout for the cluster-level correlations, ϕ_k.

In the models that used REs for modelling the individual-level SDs (i.e. 2c, 3c and 4c), we assumed $\log(\sigma_{cj})$ ~ N($\mu_\sigma^c, \sigma_{c\sigma}^2$) with priors μ_σ^c ~ N(0,10) and $\sigma_{c\sigma}$~ U(0,10). For those that used REs for individual-level Fisher's z transformed correlations (i.e. 3c and 4c), z_j ~ N(μ_z,σ_z^2), we used priors μ_z ~ N(0,10^6) and σ_z ~ U(0,1). In Model 2c, the logarithms of the individual-level SDs were assumed to be Normally distributed, where $\log(\sigma_{cj})$ ~ N($\mu_\sigma^c, \sigma_{c\sigma}^2$) and $\log(\sigma_{ej})$ ~ N($\mu_\sigma^e, \sigma_{e\sigma}^2$); the log scale is used to avoid negative variances.

For the Gamma-Normal models, we assumed wide Uniform priors, U(0,10), for the shape parameters, s_j, and wide Normal N(0,10^6) for the regression parameters, γ_j, when modelled by FEs. When modelled by REs we assumed $\log(s_j)$ ~ N(μ_s, σ_s^2) with priors μ_s ~ N(0,10^4), σ_s ~ U(0,10); and γ_j~ N($\mu_\gamma, \sigma_\gamma^2$) with priors μ_γ ~ N(0,10^6) and σ_γ~ U(0,100). For the Inverse Gaussian-Normal models the same priors were used for the regression and shape parameters but the shape was constrained to be >1.

4. Model comparison

4.1 Diagnostic plots

To inform the choice of distributions for the endpoints [33, 39] we considered normal plots of deviance residuals as a tool for assessing model fit. We used the posterior mean of the *ijk*-th deviance residual as an estimate of the deviance residual for the corresponding individual.

The deviance residual Δ_t was defined as the signed square root of the individual contribution, d_t, to the total deviance D. $\Delta_t = sign(y_t - \mu_t)\sqrt{d_t}$, where y_t is the observed, and μ_t the expected value for individual t, so that $\sum_t \Delta_t^2 = D$.

With η_{ijk}^c defined as in equation [1], the deviance residuals for the three distributions considered for costs, c_{ijk}, were then given as:

$$\text{Normal: } \Delta_{ijk} = c_{ijk} - \eta_{\text{ijk}}^{\text{c}}$$

$$\text{Gamma: } \Delta_{ijk} = sign(c_{ijk} - \eta_{\text{ijk}}^{\text{c}})\sqrt{-\log\left(\frac{c_{ijk}}{\eta_{\text{ijk}}^{\text{c}}}\right) + (c_{ijk} - \eta_{\text{ijk}}^{\text{c}})/\eta_{\text{ijk}}^{\text{c}}}$$

$$\text{Inverse Gaussian: } \Delta_{ijk} = \left(c_{ijk} - \eta_{\text{ijk}}^{\text{c}}\right)/(\eta_{\text{ijk}}^{\text{c}}\sqrt{c_{ijk}})$$

For models that fit the data well, the deviance residuals should approximate a Normal distribution and lie along the line of identity in the normal plots. [22] For each chosen cost distribution, the plots of the residual deviances do not reveal noticeable differences across the models with increasingly complex variance structure as shown by the three illustrative plots for each cost distribution in Figure 3. However, when costs are assumed to follow a Gamma rather than a Normal distribution, the residual plots suggest some improvement in model fit; with further, albeit marginal gains when using an Inverse Gaussian distribution.

4.2 Model appropriateness

Overall model fit and appropriateness can be summarised by measures such as the mean deviance and the deviance information criterion (DIC). The advantage of the DIC is that it reflects predictive accuracy by penalising models which have a greater effective number of parameters [23], and it is useful for comparing the fit of non-nested models. The DIC is calculated as DIC = $D(\bar{\theta}) + 2p_D$, where $D(\bar{\theta})$ is the deviance evaluated at the posterior means of the parameters being estimated, and p_D is the effective number of parameters. The model with the lowest DIC may then be judged the 'most appropriate', although models with DIC within 5 units also warrant consideration [46]. Other related measures include the Akaike Information Criterion (AIC) [47] and Bayesian Information Criterion (BIC) [48, 49], but both require the number of model parameters to be known which is problematic for random effects models.

The DIC is constructed using the likelihood of the parameters given the data, and has been extended to the missing data setting [16, 50, 51]. In the SPHERE example, the data are partially-observed , so we use here the appropriate extension of the DIC, based on the observed data likelihood $L(\boldsymbol{\theta}|\boldsymbol{y_{obs}})$, under ignorability of the missing data [51], but conditional on the cluster. This means that the inferential focus is the cluster, and this is the DIC typically implemented for hierarchical models in WinBUGS. In the context of the missing data setting (DIC_7 in [16]), we have:

$$\begin{aligned} DIC_C = {} & D(\bar{\boldsymbol{\theta}}, \boldsymbol{u}|\boldsymbol{y_{obs}}) + 2p_D \\ & 2\overline{D(\boldsymbol{\theta}, \boldsymbol{u}|\boldsymbol{y_{obs}})} - D(\overline{\boldsymbol{\theta}, \boldsymbol{u}}|\boldsymbol{y_{obs}}) \\ & -4E\{l(\boldsymbol{\theta}, \boldsymbol{u}|\boldsymbol{y_{obs}})\} + 2l(\overline{\boldsymbol{\theta}, \boldsymbol{u}}|\boldsymbol{y_{obs}}) \end{aligned}$$

Where the cluster random effect $\boldsymbol{u}$ is considered as an extra parameter to be estimated,

$\overline{D(\boldsymbol{\theta}, \boldsymbol{u}|\boldsymbol{y_{obs}})} = -2\mathrm{E}_{\theta}[\log f(\boldsymbol{y_{obs}}|u, \theta) \mid \boldsymbol{y_{obs}}]$ and p_D denotes the number of effective parameters, defined as $p_D = \overline{D(\boldsymbol{\theta}, \boldsymbol{u}|\boldsymbol{y_{obs}})} - (\overline{\boldsymbol{\theta}, \boldsymbol{u}}|\boldsymbol{y_{obs}})$. This requires computation of the observed data likelihood for our bivariate models with random effects, which cannot be calculated automatically by WinBUGS. In particular, for those models assuming Gamma or IG distributions, the marginal distribution of the health outcomes does not have a known closed form, and we use Monte Carlo integration to calculate the corresponding likelihood and the DIC_C. The intuitive reason for the need to calculate the DIC for a specific model using the *observed* data alone, rather than using the observed data and current Bayesian draw of the missing data, is that model fit should only be assessed with respect to the observed data.

We report the DIC_C, the deviance (at the posterior mean $\bar{\boldsymbol{\theta}}$, called the plug-in deviance), and the effective number of parameters across the MLMs under comparison.

As Table 3 shows, the MLMs that assumed an Inverse Gaussian distribution for individual costs gave substantially lower deviances and DICs, and were judged more appropriate than any of the bivariate Normal or Gamma-Normal models (Table 3). Within each assumed cost distribution, the simplest MLM that assumed constant variances and correlations gave the highest DIC; moving to MLMs with more complex variance-covariance structures generally improved model fit and appropriateness; these gains were relatively large for the bivariate Normal models (Table 3). The models with the lowest DIC were the Inverse Gaussian-Normal models that allowed individual-level shape and regression parameters to differ by cluster using REs (Models 3c and 4c). Under such circumstances, the results from both models should be reported.

Analogous models (3b and 4b) that allowed the shape and regression parameters to differ by fixed effects gave lower residual deviances, but these models required many more parameters to be fitted and hence gave higher DICs.

TABLE 3

4.3. Cost-effectiveness results

The cost-effectiveness results across all the models considered reported that the intervention had positive point estimates of the INB, with 95% credible intervals that included zero (Figure 4). For the Inverse Gaussian-Normal models, the mean (95% credible intervals) were similar across the MLMs with different variance-covariance matrices (Figure 4). Within the alternative bivariate Normal models, there were more differences in both the point estimates and width of the credible intervals, versus the more appropriate Gamma-Normal and Inverse Gaussian-Normal models. In particular, the bivariate Normal models 2b-c and 3b-c gave much narrower credible intervals than their better fitting Inverse Gaussian and Gamma counterparts (Figure 4). The two Inverse Gaussian-Normal models with the lowest DIC (models 3c and 4c) reported similar INB of around €600 (95% credible intervals from around -500 to 1600). Some of the Inverse Gaussian-Normal models, which considered a simpler variance-covariance structure gave somewhat different INB, for example the simplest Inverse Gaussian-Normal model (1) reported an INB of approximately €400 (95% credible Interval from -1200 to 1800).

5. Discussion

This paper extends previous research on statistical methods for CEA (Willan et al, 2004; Nixon and Thompson 2005; Bachmann et al, 2007; Grieve et al, 2010; Gomes et al, 2012), by providing an approach for choosing amongst MLMs for CEA that use cluster trials. Such studies tend to have complex data structures, and current CEA methods guidance does not offer advice for choosing amongst a range of models .[14] This paper encourages future studies to take a systematic approach to model choice in considering a range of alternative model specifications to address issues such as: missing data, cost skewness, and between-

setting differences in the variances of individual costs. Our paper therefore addresses an important gap in the literature by providing a strategy for model choice with four complementary strands: a) data description, b) pre-specification of MLMs drawing on the literature, c) diagnostic plots, and d) assessment of model fit and appropriateness for partially-observed data. To help future CEA that use cluster trial data chose between MLMs, we provide exemplar WinBUGS code for calculating the DIC in typical settings, i.e. when cost or health outcome data are partially-observed, and the cost data are non-Normally distributed [http://www.lshtm.ac.uk/php/hsrp/ceathatuseclustertrials/index.html: code to add, pre re-submission].

This approach to model choice was illustrated by re-analysis of a representative case study.[6] In particular, the cost data were right-skewed, and clinical practice variations appeared to lead to individual cost variances that differed across clusters. In these circumstances many alternative MLMs warrant consideration and can be fitted by MCMC in WinBUGS or OpenBUGs, which has now implemented the inverse Gaussian distribution. The data description encouraged the specification of MLMs that allowed individual costs to have non-Normal distributions with variances that differed by cluster, but assumed that health outcomes were Normally distributed. SPHERE, like many CRTs, [6] has moderate numbers of clusters and so little information was available on cluster-level parameters such as the correlation between mean costs and health outcomes. Hence, a careful description of the salient features of the data appears a necessary but insufficient criterion for model choice.

We followed recommendations for handling right-skewed cost data, and considered the GLM family of models, such as those that assume a Gamma or Inverse Gaussian distribution [22],These can accommodate common mean-variance relationships, allow the choice of a range of link and variance functions. While in this case study, it was plausible to assume that

treatment had an additive effect on mean costs and health outcomes, if *a priori* reasoning and the data description suggest treatment has a multiplicative effect then a GLM with a log-link could be chosen instead.[18] In principle the proposed approach, of using residual plots and a measure of model appropriateness such as the DIC, can also inform the choice of link function.

In our case study we found that plots of individual-level residuals can help in choosing amongst MLMs that make different distributional assumptions: the MLMs that assumed costs were from Inverse Gaussian distributions appeared to fit the data relatively well. However, these residual plots were less useful for differentiating between models with different variance-covariance matrices. Alternative diagnostic tools such as Bayesian *p*-values could also be considered, but they require simulations from the posterior distribution which is not straightforward for distributions beyond those available by default in the modelling software, WinBUGS.

Of the criteria considered, we found that the DIC was the most useful overall for choosing amongst the plausible MLMs; it differentiated between cost distributions but also amongst models with alternative variance-covariance structures. Other ways of comparing models include Bayes factors [52] and cross-validation.[53, 54] In line with standard practice in WinBUGS, the DIC we have reported here is calculated conditional on the cluster level random effects. Marginalising over these (away from the tractable BVN models) involves extremely time-consuming double numerical integration.

Our approach to model choice is not necessarily intended to lead the analyst to a single MLM. Indeed in this case study at least two models warranted careful consideration (Inverse Gaussian-Normal models 3c and 4c). While in this example plausible models yielded similar cost-effectiveness estimates, more generally in CEA, models with similar fit can yield

radically different inferences.[13] A formal and objective way to synthesise the results from alternative models, while accounting for the uncertainty around model choice, is to employ Bayesian model averaging.[55]

In CEA previous methodological studies have encouraged analysts to fit models by MCMC estimation, in WinBUGS, because this offers a wide choice of distributions.[10, 14, 56] This raises the potential concern that results can be sensitive to the choice of prior distributions particularly for cluster-level parameters such as the random effects.[57] For all model parameters we aimed to choose vague priors such as wide Normal, Uniform or Gamma distributions. However, there is no universally accepted standard for vague priors, and analysts are encouraged to explore the sensitivity of results to alternative choices of prior distributions.[58] In the SPHERE example, we found that the estimates of the INB were very similar when we chose alternative prior distributions.

This paper has some limitations. Firstly the residual plots suggested there was scope for improvement in the fit of the models to the cost data. Secondly, the approach presented was illustrated for a single dataset. Other CEA of CRT may present further challenges; for example they may require covariate adjustment, health outcomes may be binary rather than continuous, and may take the form of repeated measures over time. Thirdly, in common with previous studies the approach did not consider whether it was plausible to assume that cluster-level residuals were Normally distributed.[14, 35] In CEA, the major interest is in the incremental cost-effectiveness, and such fixed parameter estimates have been shown to be generally robust to misspecification of the distribution of random effects.[59] Nonetheless, our models could be expanded to consider non-Normal distributions for the random effects.

The approach presented to model choice for CEA that use CRT data opens up areas for further research. Firstly, the approach to model choice could consider a broader range of

alternative handling skewed costs. These include flexible parametric approaches such as beta-type size distributions that can model cost skewness and heterogeneity as a function of covariates (see Jones et al 2012 for a review), and common alternatives to the GLMs considered, such as the lognormal distribution. Secondly, this paper focuses on CEA that use data from cluster trials, but approaches to model choice are also required for other study designs such as CEA that use multicentre or multinational RCTs.[17, 18] Thirdly, the proposed approach to model choice could be expanded to the context where data are assumed Missing Not at Random. Here, the use and interpretation of measures of model appropriateness such as DIC raise challenges which are currently unresolved [50].

ACKNOWLEDGEMENTS

The authors would like to thank William Browne, Paul Clarke, Paddy Gillespie, Manuel Gomes, Mike Kenward, David Lunn and Nicky Best for useful discussions. We would also like to thank the members of the SPHERE study team including: A Murphy, M Cupples, S Smith, M Byrne, ME Byrne, E O'Shea, P Gillespie, C Leathem, A Houlihan, M O'Malley, V Spillane, H Grealish, P Ryan, M Corrigan, M D'Eath, J Wilson, A Kelly, J Newell, and M Donnelly. The authors would also like to thank the patients and practitioners who participated in the SPHERE study.

FUNDING

This work was supported by the Medical Research Council [grant number G0802321/1].

REFERENCES

1. NICE: **Methods for Technology Appraisal.** *National Institute for Health and Clinical Excellence* 2008, **London, UK.**
2. CADTH: **Guidelines for the Economic Evaluation of Health Tecnologies: Canada. 3rd Ed.** *Canadian Agency for Drugs and Technologies in Health* 2006, **Ottawa, Canada.**
3. PBCA: **Guidelines for preparing submissions to the Pharmaceutical Benefits Advisory Committee.** *Australian Government - Department of Health and Ageing* 2008, **Camberra, Australia.**
4. IQWIG: **Methods for assessment of the relation of Benefits to Costs in the German Statutory Health Care System.** *Institute for Quality and Efficiency in Health Care* 2009, **Cologne, Germany.**
5. Baio G, Dawid AP: **Probabilistic sensitivity analysis in health economics**. *Stat Methods Med Res* 2011.
6. Gomes M, Grieve R, Nixon R, Edmunds WJ: **Statistical methods for cost-effectiveness analyses that use data from cluster randomized trials: a systematic review and checklist for critical appraisal.** *Medical Decision Making; DOI: 101177/0272989x11407341* 2011.
7. Willan AR, Briggs AH, Hoch JS: **Regression methods for covariate adjustment and subgroup analysis for non-censored cost-effectiveness data**. *Health Economics* 2004, **13**(5):461-475.
8. Briggs AH, O'Brien BJ: **The death of cost-minimization analysis?** *Health Economics* 2001, **10**(2):179-184.
9. Hoch JS, Briggs AH, Willan AR: **Something old, something new, something borrowed, something blue: a framework for the marriage of health econometrics and cost-effectiveness analysis**. *Health Economics* 2002, **11**(5):415-430.
10. Nixon RM, Thompson SG: **Methods for incorporating covariate adjustment, subgroup analysis and between-centre differences into cost-effectiveness evaluations**. *Health Economics* 2005, **14**(12):1217-1229.
11. Mihaylova B, Briggs A, O'Hagan A, Thompson SG: **Review of statistical methods for analysing healthcare resources and costs**. *Health Economics* 2010, **DOI: 10.1002/hec.1653**.
12. Nixon RM, Thompson SG: **Parametric modelling of cost data in medical studies**. *Stat Med* 2004, **23**(8):1311-1331.
13. Thompson SG, Nixon RM: **How sensitive are cost-effectiveness analyses to choice of parametric distributions?** *Medical Decision Making* 2005, **25**(4):416-423.
14. Grieve R, Nixon R, Thompson SG: **Bayesian hierarchical models for cost-effectiveness analyses that use data from cluster randomized trials**. *Medical Decision Making* 2010, **30**(2):163-175.
15. Willan AR, Briggs A: **Statistical Analysis of Cost-Effectiveness Data.** Chichester: Wiley-Blackwell; 2006.
16. Celeux G, Forbes F, Robert CP, Titterington DM: **Deviance Information Criteria for Missing Data Models**. *Bayesian Analysis* 2006, **1**(4):651-673.
17. Grieve R, Nixon R, Thompson SG, Normand C: **Using multilevel models for assessing the variability of multinational resource use and cost data**. *Health Economics* 2005, **14**(2):185-196.
18. Thompson SG, Nixon RM, Grieve R: **Addressing the issues that arise in analysing multicentre cost data, with application to a multinational study**. *Journal of Health Economics* 2006, **25**(6):1015-1028.
19. Briggs A, Nixon R, Dixon S, Thompson S: **Parametric modelling of cost data: some simulation evidence**. *Health Economics* 2005, **14**(4):421-428.
20. 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.
21. Thompson SG, Barber JA: **How should cost data in pragmatic randomised trials be analysed?** *BMJ* 2000, **320**(7243):1197-1200.

22. Barber J, Thompson S: **Multiple regression of cost data: use of generalised linear models**. *Journal of health services research & policy* 2004, **9**(4):197-204.
23. Spiegelhalter DJ, Best NG, Carlin BP, van der Linde A: **Bayesian measures of model complexity and fit**. *Journal of the Royal Statistical Society Series B-Statistical Methodology* 2002, **64**:583-616.
24. De Backer G, Ambrosioni E, Borch-Johnsen K, Brotons C, Cifkova R, Dallongeville J, Ebrahim S, Faergeman O, Graham I, Mancia G *et al*: **European guidelines on cardiovascular disease prevention in clinical practice - Third Joint Task Force of European and other Societies on Cardiovascular Disease Prevention in Clinical Practice**. *European Heart Journal* 2003, **24**(17):1601-1610.
25. Gillespie P, O'Shea E, Murphy AW, Byrne MC, Byrne M, Smith SM, Cupples ME: **The cost-effectiveness of the SPHERE intervention for the secondary prevention of coronary heart disease**. *International Journal of Technology Assessment in Health Care* 2010, **26**(3):263-271.
26. Brazier J, Roberts J, Deverill M: **The estimation of a preference-based measure of health from the SF-36**. *Journal of Health Economics* 2002, **21**(2):271-292.
27. Stinnett AA, Mullahy J: **Net health benefits: A new framework for the analysis of uncertainty in cost-effectiveness analysis**. *Medical Decision Making* 1998, **18**(2):S68-S80.
28. Nixon RM, Wonderling D, Grieve RD: **Non-parametric methods for cost-effectiveness analysis: the central limit theorem and the bootstrap compared**. *Health Economics* 2010, **19**(3):316-333.
29. Higgins JP, Thompson SG, Deeks JJ, Altman DG: **Measuring inconsistency in meta-analyses**. *BMJ* 2003, **327**(7414):557-560.
30. Basu A, Manning WG: **Issues for the Next Generation of Health Care Cost Analyses**. *Med Care* 2009, **47**(7_Supplement_1):S109-S114 110.1097/MLR.1090b1013e31819c31894a31811.
31. Higgins JPT, Thompson SG: **Quantifying heterogeneity in a meta-analysis**. *Statistics in Medicine* 2002, **21**(11):1539-1558.
32. Levene H (ed.): **Robust Tests for Equality of Variances**. Palo Alto, CA: Stanford University Press; 1960.
33. Bland JM: **An introduction to medical statistics.**, 3rd edition edn. Oxford: Oxford University Press; 2000.
34. Morrell CJ, Slade P, Warner R, Paley G, Dixon S, Walters SJ, Brugha T, Barkham M, Parry GJ, Nicholl J: **Clinical effectiveness of health visitor training in psychologically informed approaches for depression in postnatal women: pragmatic cluster randomised trial in primary care**. *Br Med J* 2009, **338**.
35. Gomes M, Ng ESW, Grieve R, Nixon R, Carpenter J, Thompson SG: **Developing appropriate methods for cost-effectiveness analysis of cluster randomized trials**. *Medical Decision Making, DOI: 101177/0272989X11418372* 2011.
36. Bachmann MO, Fairall L, Clark A, Mugford M: **Methods for analyzing cost effectiveness data from cluster randomized trials**. *Cost effectiveness and resource allocation* 2007, **5**:12.
37. Manning WG, Mullahy J: **Estimating log models: to transform or not to transform?** *Journal of Health Economics* 2001, **20**(4):461-494.
38. Manning WG, Basu A, Mullahy J: **Generalized modeling approaches to risk adjustment of skewed outcomes data**. *J Health Econ* 2005, **24**(3):465-488.
39. McCullagh P, Nelder JA: **Generalized Linear Models** Second edn. Cambridge Chapman and Hall/CRC; 1989.
40. Moran JL, Solomon PJ, Peisach AR, Martin J: **New models for old questions: generalized linear models for cost prediction**. *Journal of Evaluation in Clinical Practice* 2007, **13**(3):381-389.
41. Goldstein H: **Multilevel Statistical Models** 4th edn. London: Wiley-Blackwell; 2010.
42. Lunn DJ, Thomas A, Best N, Spiegelhalter D: **WinBUGS - A Bayesian modelling framework: Concepts, structure, and extensibility**. *Statistics and Computing* 2000, **10**(4):325-337.

43. Brooks SP, Gelman A: **General methods for monitoring convergence of iterative simulations**. *Journal of Computational and Graphical Statistics* 1998, **7**(4):434-455.
44. Ntzoufras I: **Bayesian Modeling Using WinBUGS** Wiley-Blackwell; 2009.
45. Gelman A: **Prior distributions for variance parameters in hierarchical models(Comment on an Article by Browne and Draper)**. *Bayesian Analysis* 2006, **1**(3):515-533.
46. Spiegelhalter DJ, Best NG, Carlin BP, Van Der Linde A: **Bayesian measures of model complexity and fit**. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 2002, **64**(4):583-639.
47. Akaike H: **Information theory and an extension of the maximum likelihood principle**. In: *Second International Symposium on Information Theory: 1973; Budapest*: Akademiai Kiado; 1973: 267-281.
48. Ward EJ: **A review and comparison of four commonly used Bayesian and maximum likelihood model selection tools**. *Ecol Model* 2008, **211**(1-2):1-10.
49. Schwarz G: **Estimating dimension of a model**. *Ann Stat* 1978, **6**(2):461-464.
50. Mason A, Richardson S, Best N: **Two-pronged Strategy for Using DIC to Compare Selection Models with Non-Ignorable Missing Responses**. *Bayesian Analysis* 2012, **7**(1):109-146.
51. Daniels MJ, Hogan JW: **Missing Data In Longitudinal Studies Strategies for Bayesian Modeling and Sensitivity Analysis**: Chapman & Hall; 2008.
52. Kass RE, Raftery AE: **Bayes factors**. *J Am Stat Assoc* 1995, **90**(430):773-795.
53. Geisser S, Eddy WF: **Predictive approach to model selection**. *J Am Stat Assoc* 1979, **74**(365):153-160.
54. Stone M: **Cross-validatory choice and assessment of statistical predictions**. *Journal of the Royal Statistical Society Series B-Statistical Methodology* 1974, **36**(2):111-147.
55. Jackson CH, Sharples LD, Thompson SG: **Structural and parameter uncertainty in Bayesian cost-effectiveness models**. *J R Stat Soc Ser C-Appl Stat* 2010, **59**:233-253.
56. Manca A, Sculpher MJ, Goeree R: **The Analysis of Multinational Cost-Effectiveness Data for Reimbursement Decisions A Critical Appraisal of Recent Methodological Developments**. *Pharmacoeconomics* 2010, **28**(12):1079-1096.
57. Lambert PC, Sutton AJ, Burton PR, Abrams KR, Jones DR: **How vague is vague? A simulation study of the impact of the use of vague prior distributions in MCMC using WinBUGS**. *Statistics in Medicine* 2005, **24**(15):2401-2428.
58. Turner RM, Omar RZ, Thompson SG: **Modelling multivariate outcomes in hierarchical data, with application to cluster randomised trials**. *Biometrical Journal* 2006, **48**(3):333-345.
59. Neuhaus JM, Hauck WW, Kalbfleisch JD: **The effects of mixture distribution misspecification when fitting mixed-effects logistic-models**. *Biometrika* 1992, **79**(4):755-762.
60. Deb P, Burgess JFJ: **A Quasi-experimental Comparison of Econometric Models for Health Care Expenditures**. In.: Hunter College: Department of Economics; 2003.

Table 1. Descriptive statistics for the SPHERE case study. Crude means, SDs, correlations (at individual and cluster level), incremental costs, QALYs and net benefits. Incremental effects are reported by taking means at the individual-level ignoring clustering and from cluster-level summary statistics.

	Control	Treatment	Overall
Number Individuals (n)	455	437	892
Clusters	24	24	48
Mean (SD) cost [€]	5 066 (5 966)	4 324 (4 810)	4 704 (5 442)
Mean (SD) QALY	1.020 (0.121)	1.014 (0.141)	1.017 (0.131)
SD of mean costs across clusters	1 365	1 317	1 363
SD of mean QALYs across clusters	0.032	0.043	0.037
ICC costs (SD)	<0.001 (0.021)	0.014 (0.021)	0.007 (0.013)
ICC QALYs (SD)	<0.001 (0.024)	0.005 (0.026)	<0.001 (0.017)
Correlation (individual costs and QALYs)[I]	-0.04	-0.04	-0.04
Correlations (cluster mean costs and QALYs)	-0.34	-0.08	-0.18

Cost-effectiveness results

	Individual-level	Cluster-level
Incremental costs	-742	-619
Incremental QALYs	-0.0058	-0.0031
INB (SE)[II]	626	556
	(409) [II]	(580)[II]

Notes: I: Correlations estimated from cases without missing endpoints. II: INB = λ × incremental QALY – incremental cost where λ is the willingness to pay threshold of €20 000 per QALY. II. Standard error of INB is as defined in Nixon et al (2010) and uses all available data. [28]

Table 2. Specification of the bivariate Normal MLMs according to assumptions made for costs and health outcomes about the individual- and cluster-level variances and correlations

Model		σ^2_{cj}	σ^2_{ej}	ρ	τ^2_{ck}	τ^2_{ek}	ϕ_k
Group	**Number**						
Basic	1	σ^2_c	σ^2_e	ρ	τ^2_c	τ^2_e	ϕ
+ Individual-level variances differ	2a	σ^2_{ck}	σ^2_{ek}	•	•	•	•
	2b	fixed σ^2_{cj}	fixed σ^2_{ej}	•	•	•	•
	2c	random σ^2_{cj}	random σ^2_{ej}	•	•	•	•
+ Individual-level correlations differ	3a	σ^2_{ck}	σ^2_{ek}	ρ_k	•	•	•
	3b	fixed σ^2_{cj}	fixed σ^2_{ej}	fixed ρ_j	•	•	•
	3c	random σ^2_{cj}	random σ^2_{ej}	random ρ_j	•	•	•
+ Cluster-level variances and correlations differ	4a	σ^2_{ck}	σ^2_{ek}	ρ_k	τ^2_{ck}	τ^2_{ek}	ϕ_k
	4b	fixed σ^2_{cj}	fixed σ^2_{ej}	fixed ρ_j	•	•	•
	4c	random σ^2_{cj}	random σ^2_{ej}	random ρ_j	•	•	•

Note: • = same as the cell above*; k* refers to randomised treatment; *j* refers to cluster

Table 3. Model fit and appropriateness (deviances at posterior parameter means, effective numbers of parameters and DICs)

Model		bivariate Normal			Gamma-Normal			Inverse Gaussian-Normal		
Group[II]	**Number**	**$D(\bar{\theta})$**	**p_D**	**DIC**	**$D(\bar{\theta})$**	**p_D**	**DIC**	**$D(\bar{\theta})$**	**p_D**	**DIC**
Basic	**1**	2015.09	20.68	2033.82	811.15	39.61	890.37	369.95	33.77	437.49
Varying individual-level variances	**2a**	1989.93	21.59	2033.12	791.89	41.23	874.35	414.71	81.34	577.40
	2b	1153.95	116.73	1387.40	590.30	84.43	759.16	245.46	70.63	386.71
	2c	1167.41	103.86	1375.12	567.88	87.33	742.54	231.13	89.92	410.96
Varying individual-level correlations	**3a**	1987.40	23.39	2034.18	770.45	42.90	856.26	324.01	51.21	426.44
	3b	1148.01	128.57	1405.15	407.65	132.69	673.03	96.38	136.94	370.25
	3c	1156.92	108.44	1373.81	456.49	113.84	684.18	159.26	93.76	**346.78**
Varying cluster-level variances and correlations	**4a**	1980.05	28.46	2036.97	769.92	45.66	861.23	260.86	116.16	493.18
	4b	1138.35	132.91	1404.17	423.59	122.14	667.87	94.91	131.14	357.19
	4c	1147.63	112.81	1373.25	464.50	105.29	675.07	159.38	95.18	**349.74**

Notes: I. $D(\bar{\theta})$ = deviance evaluated at posterior mean of parameters; p_D = effective number of parameters; DIC = deviance information criterion
II. Variance and correlation are substituted by shape, *s*, and regression, γ, parameters for the non-Normal models.
The DICs in bold correspond are the smallest, and therefore represent best fit for the data.

Figure 1. SPHERE case study: individual QALYs and costs, with fitted densities of Normal, Gamma and Inverse Gaussian distributions, by treatment arm (values in plots are log-likelihoods)

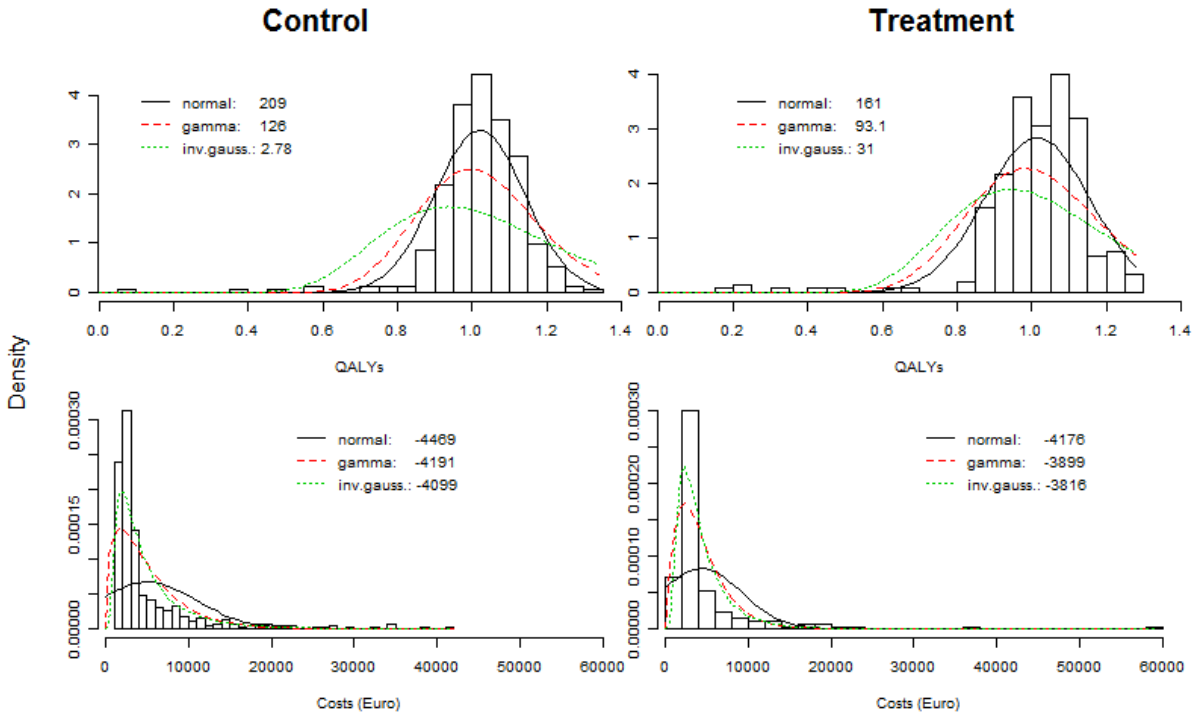
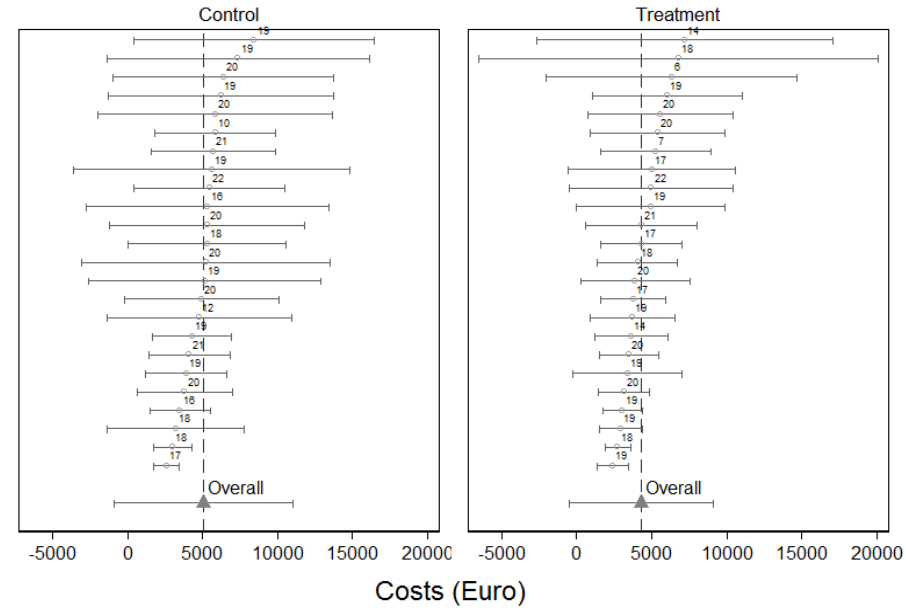


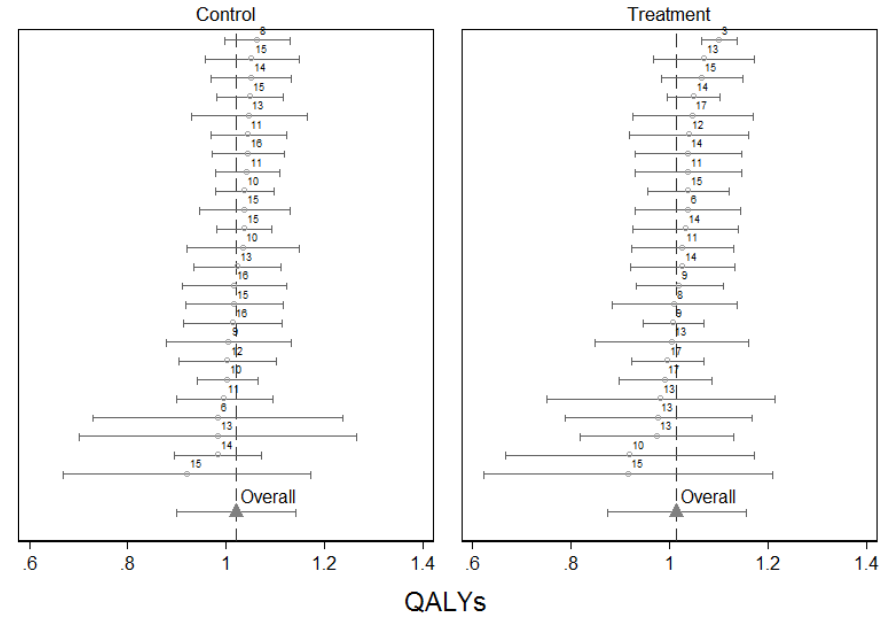
Figure 2. SPHERE case study endpoints (all patients, n=892)

a) Mean costs ± 1 SD by cluster and treatment arm



error bar = ± 1SD; numbers in plots are cluster sizes

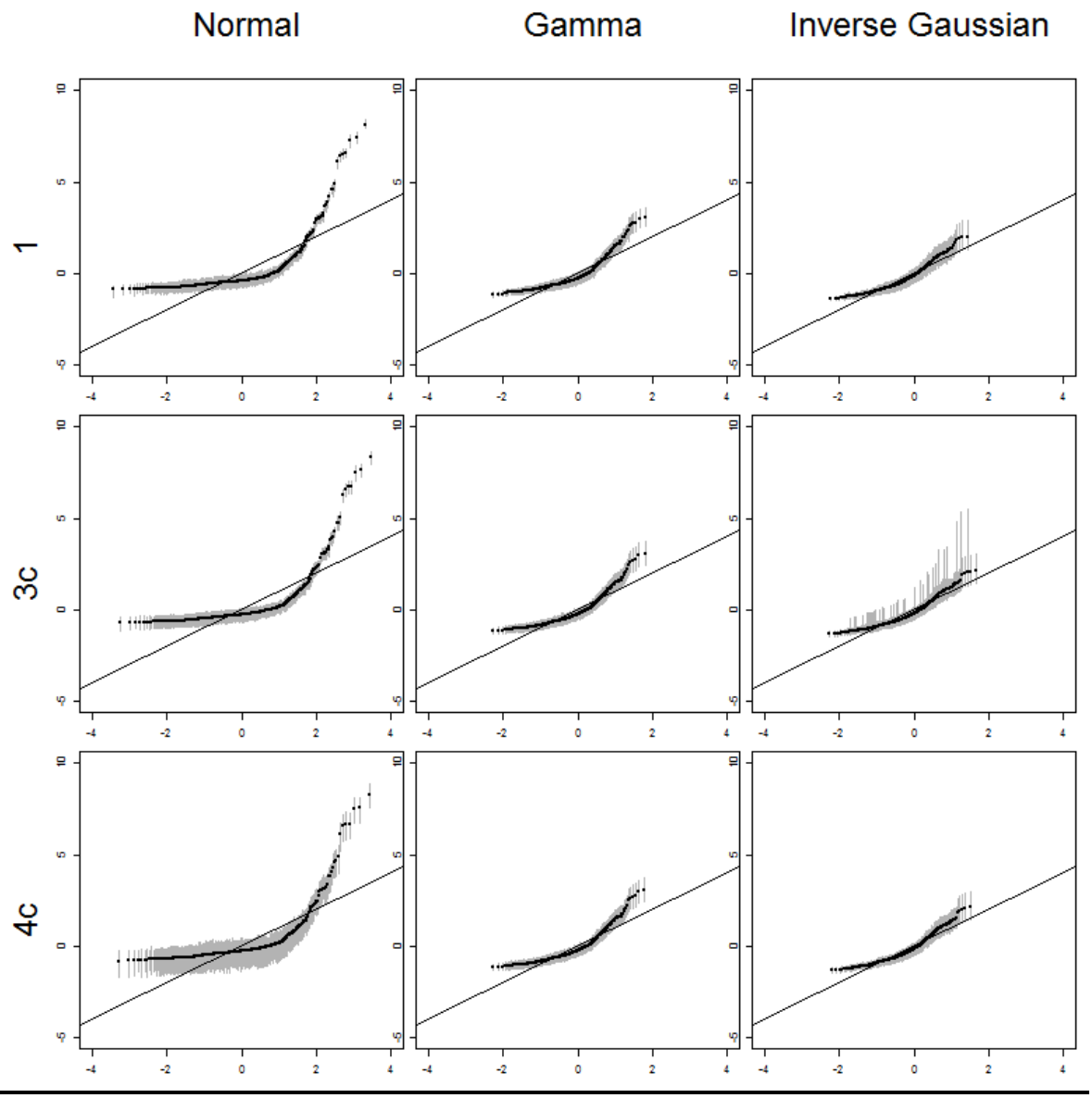
b) <u>Mean QALYs ± 1 SD by cluster and treatment arm</u>



error bar = ± 1SD; numbers in plots are numbers of observations

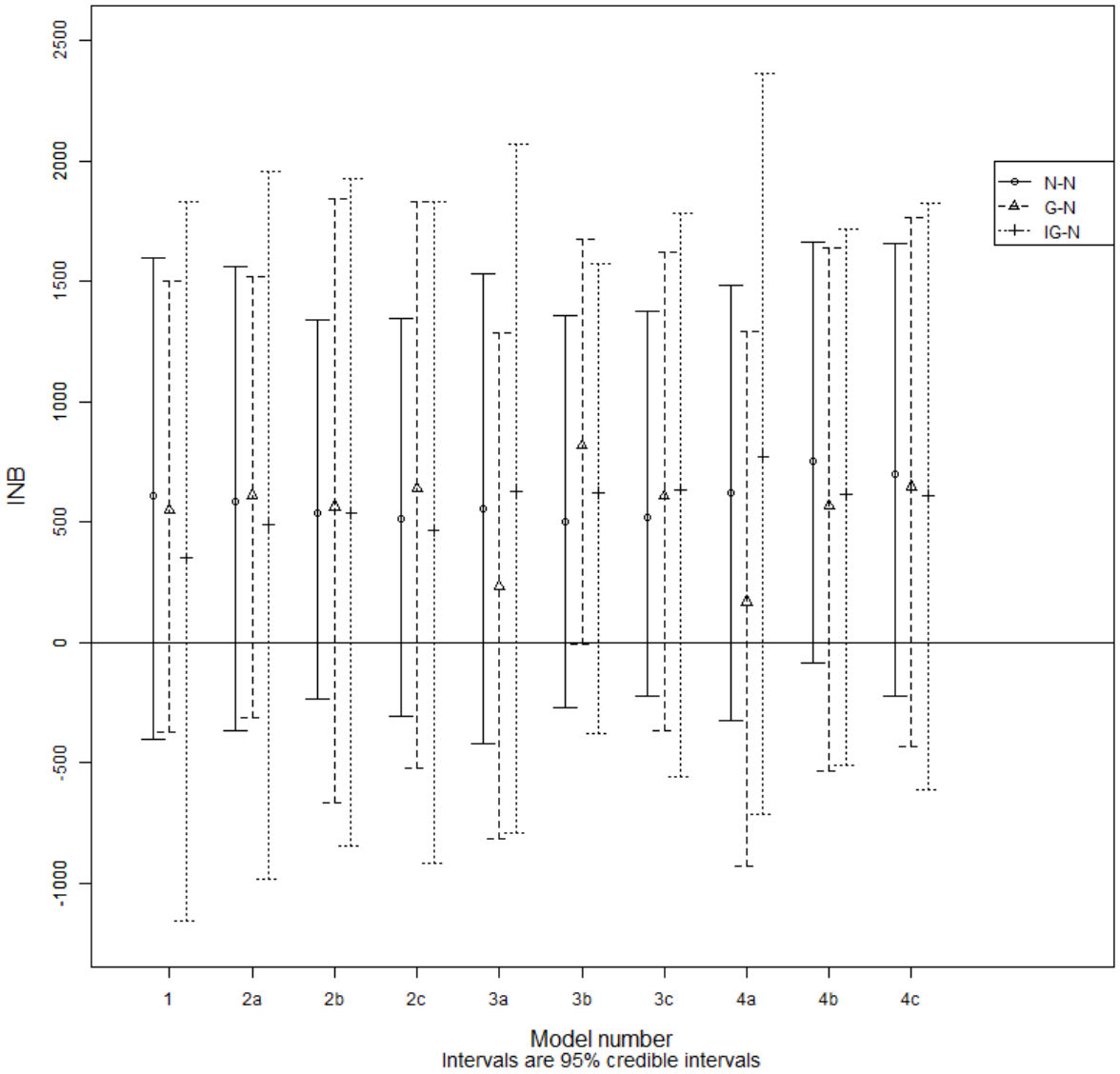
Note: Clusters are ordered by their means.

Figure 3. Normal plots of deviance residuals for costs from a subset of bivariate MLMs assuming that costs have Normal, Gamma and Inverse Gaussian distributions, and QALYs have Normal distributions (95% credible intervals in grey shade).



Note: Model 1 assumes constant variances and correlations, Model 3c allows individual variances (shapes) and correlations (regression parameters) that differ by cluster with random effects, and Model 4c extends 3c by allowing cluster level variances (shapes) and correlations (regression parameters) that differ by treatment group

Figure 4. SPHERE cost-effectiveness results according to MLMs that assume different cost distributions and levels of complexity in their variance-covariance matrices. Results are reported as mean INB (95% credible intervals).



Notes:
I. N-N: bivariate Normal; G-N: Gamma-Normal; IG-N: Inverse Gaussian-Normal
II. INBs estimated at the willingness to pay threshold value of €20 000 per QALY