Analysing cluster randomised controlled trials using MLE, GEE, GEE2 and QIF: results from four case studies

Bright C. Offorha (bcofforha1@sheffield.ac.uk)
The University of Sheffield
Stephen J. Walters
The University of Sheffield
Richard M. Jacques
The University of Sheffield

Research Article

Keywords: Cluster Randomised controlled trial, Statistical models, SAS, Intracluster correlation coefficient, Statistical methods

Posted Date: December 28th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-2015033/v1

Additional Declarations: No competing interests reported.

Analysing cluster randomised controlled trials using MLE, GEE, GEE2 and QIF: results from four case studies

Bright C. Offorha[1*], Stephen J. Walters[1] and Richard M. Jacques[1]

[1]School of Health and Related Research, The University of Sheffield, Sheffield.
[*]Corresponding author

ABSTRACT

Background: Using four case studies, we aim to provide practical guidance and recommendations for the analysis of cluster randomised controlled trials.

Methods: Four modelling approaches (Generalized Linear Mixed Models with parameters/coefficients estimated by Maximum likelihood; Generalized Linear Models with parameters/coefficients estimated by Generalized Estimating Equations (1st order or second order) or Quadratic Inference Function) for the analysis of correlated individual participant level outcomes in cluster randomised controlled trials were identified after we reviewed the literature. These four methods are applied to four case studies of cluster randomised controlled trials with the number of clusters ranging from 10 to 100 and individual participants ranging from 748 to 9,207. Results are obtained for both continuous and binary outcomes using the statistical packages, R and SAS.

Results: The intracluster correlation coefficient (ICC) for each of the case studies was small (<0.05) indicating little dependence of the outcomes related to cluster allocation. In most cases the four methods produced similar results. However, in a few analyses quadratic inference function produced different results compared to the other three methods.

Conclusion: This paper demonstrates the analysis of cluster randomised controlled trials with four modelling approaches. The results obtained were similar in most cases, a plausible reason could be the negligible correlation (small ICCs) observed among responses in the four case studies.

Due to the small ICC values obtained the generalisability of our results is limited. It is important to conduct simulation studies to comprehensively investigate the performance of the four modelling approaches.

Keywords: Cluster Randomised controlled trial, Statistical models, SAS, Intracluster correlation coefficient, Statistical methods

BACKGROUND

Randomisation is used in clinical trials to achieve balance between the treatment arms in chance variations caused by both known and unknown prognostic factors. If done properly, it may minimise the effect of the prognostic factors so that researchers can controllably study the effect of the intervention(s) of interest. Instead of randomising individuals to the treatment arms as often done in individually randomised controlled trials (IRCTs), groups/clusters of individuals are randomised in cluster randomised controlled trials (CRCTs).

In CRCTs there are two levels; the distinctive cluster level and the individual level (with correlated outcomes) that are nested within the clusters. An appropriate statistical method for analysing CRCTs will be any method that considers this hierarchical nature of the CRCT design. Ignoring the correlated outcomes within a cluster and using standard statistical methods that treat the outcomes as being independent, might lead to underestimating the standard errors of the true parameters and consequently obtaining narrower confidence intervals, false small p-values and incorrectly over stating the effect of the intervention.

Some of the common issues in CRCT design and analysis are (a) Ignoring clustering(1), (b) inadequate handling of missing data(2), (c) and poor reporting of results (1,3). Newer analytical methods for handling clustering have been proposed in the literature of other study designs with clustered data, such as longitudinal study design. Notable ones are targeted maximum likelihood estimation (TMLE), quadratic inference function (QIF) and alternating logistic regression (ALR). It is worth noting that these recent alternatives have not been comprehensively compared to the existing methods used in CRCTs and might account for their slow uptake. This study aims to contribute to the literature (in the context of CRCTs) on the performance of one of the newer methods compared to the existing methods with the aim of promoting its use in CRCTs (if necessary).

This paper reviews and describes the selected statistical methods for analysing both continuous and binary outcomes in CRCTs. We focus on statistical methods for analysing individual participant level outcomes which are correlated within a cluster. The paper explores the performance of all the models given the settings of our case studies. The objectives of this study are to demonstrate the practical application of these selected modelling approaches for analysing CRCTs, to contrast and discuss their methodological differences and to make general comments based on our findings.

METHODS

Literature Search

We conducted a review of the literature using a systematic searching approach from 1st January 2003 to 19th December 2020. This was a year prior to the publication of the CONSORT statement 2004 extension for cluster randomised controlled trials(4). A standardised pre-piloted data collection tool was used to extract information on the study and methodological characteristics from the included articles. We used search terms developed with the inputs of an information specialist (see, **Additional file 1**). We searched the online bibliography databases of MEDLINE, EMBASE, PsycINFO (via OVID); and CINAHL (via EBSCO) and SCOPUS. In addition to searching published literature databases, OpenGrey, web-of-science and Scopus databases for conference proceedings were also searched to identify difficult-to-locate (grey) literature. One reviewer, BO, carried out the search and extraction of the relevant information; two other independent reviewers, SW and RJ, supervised and validated the process. We discussed extensively to reach a consensus on issues presenting during the review process.

Literature search results

The literature search identified 1573 articles and after duplicates were removed 1073 articles were remaining. After screening the title and abstract of each of the identified articles 104 were shortlisted and 64 (these include 12 articles from citation tracking) were finally included in the list of relevant articles to review while 64 others were excluded for other reasons (**Figure 1**). These articles are methodological and application papers and are referenced throughout.

The study selection process is presented in **Figure 1**. Among the included 64 papers; 49/64 (77%) compared already existing methods (of which 31% (13/49) compared different models with their parameters estimated using a single method), 14% of the papers proposed new statistical estimators and 9% refined already existing ones. There is no clear pattern in the development, advancement, or comparison of statistical methods for analysing CRCTs in the last two decades approximately (see, **Additional file 2**).

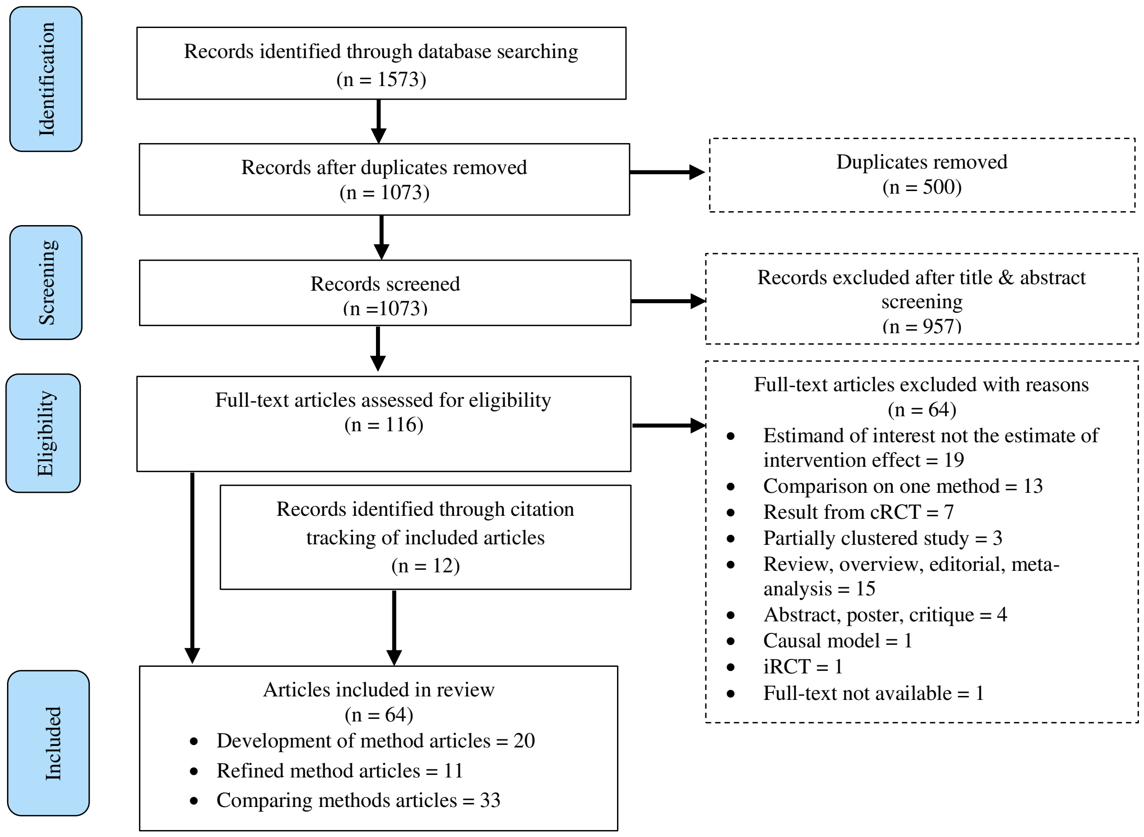
Identification
Records identified through database searching
(n = 1573)
Records after duplicates removed
(n = 1073)
Duplicates removed
(n = 500)
Screening
Records screened
(n =1073)
Records excluded after title & abstract screening
(n = 957)
Eligibility
Full-text articles assessed for eligibility
(n = 116)
Full-text articles excluded with reasons
(n = 64)
• Estimand of interest not the estimate of intervention effect = 19
• Comparison on one method = 13
• Result from cRCT = 7
• Partially clustered study = 3
• Review, overview, editorial, meta-analysis = 15
• Abstract, poster, critique = 4
• Causal model = 1
• iRCT = 1
• Full-text not available = 1
Records identified through citation tracking of included articles
(n = 12)
Included
Articles included in review
(n = 64)
• Development of method articles = 20
• Refined method articles = 11
• Comparing methods articles = 33

Figure 1 Flow chart of the search and selection process of the included articles in the review

The number of the times each method was studied in the 64 articles and their references are summarised in **Table S1** (See, **Additional file 3**). This review identified 34 unique statistical methods for analysing CRCTs and the methods were studied 134 times in total. Regression models with parameters estimated by GEE1 was the most studied (30/134, 22%) followed by MLE (19%). Among the newer methods QIF was the most studied (6%).

Four statistical regression models for the analysis of correlated individual participant level outcomes in cluster randomised controlled trials were selected based on the findings of the literature review. They are:

1. Generalized Linear Mixed Models (GLMM) with parameters/coefficients estimated by Maximum likelihood (MLE).

2. Marginal Generalized Linear Models (mGLM) with parameters/coefficients estimated by 1st order Generalized Estimating Equations (mGLM-GEE1).

3. Marginal Generalized Linear Models (mGLM) with parameters/coefficients estimated by 2nd order Generalized Estimating Equations (mGLM-GEE2).

4. Marginal Generalized Linear Models (mGLM) with parameters/coefficients estimated by Quadratic Inference Function (mGLM-QIF).

Specifically, GLMM-MLE and mGLM-GEE1 were selected based on the results of the literature search, which indicated that they are the two most studied regression methods while mGLM-GEE2 and mGLM-QIF were selected based on findings that suggested that they are the two most promising improvements on the mGLM-GEE1(5–8). The methodological properties of the four selected regression models are compared in **Table 1**.

While GEE2 and QIF estimation methods are not commonly used for analysing CRCTs, QIF has been extensively studied and applied in the context of longitudinal studies where outcomes measured repeatedly over time from a particular individual are likely to be correlated/clustered. It is worth investigating the capabilities of the QIF in relation to other estimation methods like MLE, GEE1 and GEE2 in the context of CRCTs. This has recently attracted the attention of researchers, but these studies majorly focused on comparing GEE1 and QIF (9–14). Our study aims to contribute to this debate in the literature using evidence from real world example data from four CRCTs.

The choice of a statistical modelling approach for analysing a CRCT is often motivated by the scientific question the study is investigating, the type of data of the primary outcome and the assumptions being made. The generalized linear mixed model (GLMM) and the marginal generalized linear model (mGLM) are the two most common regression models when the research interest is about the effect of the intervention on the individual participants and/or across sub populations. The maximum likelihood estimator (MLE) is often used to estimate the parameters of GLMM while the generalized estimating equation (GEE) is used for that of mGLM. An alternative to the MLE is restricted maximum likelihood (REML)(15).

The first order GEE denoted as GEE1 considers the correlation among the outcomes in a cluster as a nuisance and only adjusts for its effect but does not estimate it. A “working” covariance matrix which is solved separately from the mean model is used to achieve this. The GEE1 estimator is consistent even when the working correlation structure is wrongly specified but may suffer some loss in efficiency.

Statistical efficiency is a desirable property of a good estimator after unbiasedness has been established. Among all unbiased competing estimators an efficient estimator is the one that produces the smallest standard error estimate, which is indicative of a lesser variability and a higher degree of precision. The MLE as a parametric method does not suffer this problem. However, the MLE makes a strong Normality assumption about the unobserved cluster-level random effect. If this assumption does not hold in practice, MLE is likely to produce invalid results. This makes the other two semi-parametric methods (GEE2 and QIF) natural alternatives to the MLE and GEE1, most especially when the population averaged intervention effect is of interest.

Table 1 Similarities and differences in the methodological properties of the four selected statistical models for analysing CRCTs

S/NO	Feature	GLMM-MLE	mGLM-GEE1	mGLM-GEE2	mGLM-QIF
1	**Ad-hoc adjustment for clustering**	Adjustment for clustering is done within the estimating algorithm of the regression parameters hence it affects the parameter estimates	Same as GLMM-MLE	Same as GLMM-MLE	Same as GLMM-MLE
2	**Covariate adjustment**	Allows for both cluster and individual level covariates to be adjusted for via a link function	Same as GLMM-MLE	Same as GLMM-MLE	Same as GLMM-MLE
3	**Adjustment for clustering**	Clustering is accounted for via a random effects variable(s), which is incorporated into a univariate mean model equation.	Achieved using a working covariance matrix (characterised by the correlation parameter) and it is specified separately from the mean model.	Same as mGLM-GEE1	Avoids the direct use of the correlation parameter in its algorithm instead uses linear combination of the product of basis matrices and some constants.
4	**Assumption on the distribution of the cluster-level random effects**	As a full likelihood method, it assumes that the cluster-level random effects variable follows a parametric Normal distribution.	As a semi-parametric method, it makes no assumption about the distribution of the cluster-level random effects variable.	Same as mGLM-GEE1	Same as mGLM-GEE1
5	**Multiple forms of clustering**	Allows multiple forms of correlation to be investigated by incorporating them as random effects in the mean model.	Allows multiple forms of correlation but through a complex procedure of including higher forms of clustering as fixed effects in the mean model.	Same as mGLM-GEE1	Same as mGLM-GEE1
6	**Assumption of missing data mechanism required to obtain consistent parameter estimates**	Missing completely at random and missing at random.	Missing completely at random	Same as mGLM-GEE1	Same as mGLM-GEE1
7	**Heterogenous correlation**	Can fit models that assume different correlation structures across the treatment arms/clusters.	Cannot fit models that assume different correlation structures across the treatment arms/clusters.	Same as GLMM-MLE	Same as GLMM-GEE1
8	**Improvement on efficiency**	Gain in efficiency compared to GLM by including random effects component in the mean model to account for correlation among outcomes from a cluster.	Gain in efficiency compared to GLM by using a "working covariance matrix" which accounts for the effect of correlation among outcomes from a cluster and treat it as a nuisance/noise.	Gain in efficiency compared to mGLM-GEE1 by adjusting and estimating the nuisance effect of the correlation among outcomes from a cluster.	Compared to mGLM-GEE1, this method weights the information contributed by each cluster using an empirical weighting matrix, clusters with large variation/variance are given less weights and vice versa.
9	**Moment specification**	Not applicable	First and second order moments need to be specified.	First four order moments[1], but the third and fourth can be specified as a function of the first two moments since a working correlation is being used.	Same as mGLM-GEE1
10	**Approximation technique**	Laplace/Adaptive Gauss-Hermite Quadrature[2]	Modified Fisher scoring algorithm	Alternate between Modified Fisher scoring algorithm and method of moment.	Newton-Raphson algorithm
11	**Goodness of fit**	All the model selection criteria that are based on maximum likelihood theory are applicable, such as the LRT, AIC and the BIC.	Uses a modification to the AIC based on quasi-likelihood theory known as QIC (and QICu[3]) for model and working correlation selections.	Same as mGLM-GEE1	Provides an objective function (which is analogue to likelihood ratio test) that follows a chi-square distribution.
12	**Availability in selected statistical software, function(package)**	R = glmer(lme4) and SAS =glimmix(proc)	R = glmgee(geepack) and SAS = genmod(proc).	R = geese(geepack) only	R = qif(qif) and SAS = qif(macro)

GLMM: Generalized linear mixed model; mGLM: marginal generalized linear model; GEE: Generalized estimating equations; QIF: Quadratic inference function; LRT = likelihood ratio test; AIC = Akaike information criteria; BIC = Bayesian information criteria; QIC = Quasi-likelihood independence criterion.

1. The first four order moments of the outcome of interest are the mean, variance, skewness and kurtosis.

2. Adaptive Gauss-Hermite Quadrature equals the Laplace approximation when the quadrature point/node is 1. Other techniques do exist.

3. QICu is a variant of QIC (16).

Notation

A boldface letter denotes either a vector or a matrix or as otherwise specified. The general notation is established as; let y_{ij} denote an outcome variable for the j^{th} subject in the i^{th} cluster ($i = 1, \dots, N;\ j = 1, \dots, n_i$); N is the number of the independent clusters in the study and n_i denotes the different number of subjects in each cluster (i.e., the i^{th} cluster size), y_{ij} has a corresponding set of p-dimensional vector covariates $\boldsymbol{X}_{pij}^{T} = (1, x_{1ij}, \cdots, x_{(p-1)ij})$ where x_{1ij} denotes an indicator variable for the treatment arms which a cluster belongs to (x_{1ij} = 0 indicates the control group and x_{1ij} = 1 the intervention group) and $\boldsymbol{Y}_i = (y_{i1}, \cdots, y_{in_i})^T\ is$ a $n_i \times 1$ vector of the collection of the individual level outcomes for the i^{th} cluster. Also, $\boldsymbol{\beta}_p = (\beta_0, \beta_1, \cdots, \beta_{p-1})$ is an unknown p-dimensional vector of regression parameters and $\boldsymbol{\mu}_i = (\mu_{i1}, \cdots, \mu_{in_i})^T$ is an $n_i \times 1$ vector of true means with $\mu_{ij} = E(y_{ij}\,|\boldsymbol{X}_{pij}^{T})$ being the conditional expectation for the j^{th} subject in the i^{th} cluster with covariates $\boldsymbol{X}_{pij}^{T}$.

Cluster-Level Approach (CLA)

This approach is also known as cluster-level analysis, it is akin to solving a problem by avoiding it. In this approach clustering is handled by reverting the originally correlated data in each group/cluster to independent data (where outcomes from subjects in the group are no longer correlated) by simply collapsing the outcomes in a cluster to a univariate data point. The first step is to obtain a summary measure, such as the mean outcome for all the subjects or the proportion of the event of interest for each cluster. The second step is to apply standard statistical methods that assume independence among outcomes in a cluster, such as the *t-test or linear regression* with the summary measure for each cluster as the response value. If y_i is a continuous summary measure for the i^{th} cluster, the cluster-level mean model is specified as

$$y_i = \beta_0 + \beta_k x_{kij} + \varepsilon_i, \quad i = 1, \dots, N; k = 1, \dots, p-1;\ \ \varepsilon_i \sim N(0, \sigma_\varepsilon^2) \tag{1}$$

β_0 is the mean outcome for the control group; β_1 is the difference in mean outcomes between the intervention and the control group known as the intervention effect, x_{1ij} is the treatment arm indicator variable, x_{kij} is a vector of other fixed cluster-level covariates with coefficients β_k and ε_i is the independent Normally distributed cluster-level residuals.

This approach is not popularly recommended for analysing CRCTs for some obvious reasons; it leads to loss of information contained in the original data and consequently a reduced sample size. This approach mainly allows only cluster level covariates to be adjusted for. In medical research such as clinical trials it is ideal/conventional to account for the effect of known prognostic factors, like age and smoking status of the individual participants. Another major reason for not recommending this approach is that it does not reflect the true study design of a CRCT. Since this current paper is aimed at illustrating the application of statistical methods that appropriately account for clustering, we do not carry this method further in this study. For examples of implementing this modelling approach see Campbell & Walters(17) and Walters et al., (18).

Individual-Level Approach (ILA)

Here, outcomes from all the participating subjects in the trial are used as response values. The problems posed by using an aggregate value for each cluster as done in CLA are circumvented. This approach is further categorised according to how the regression model adjusts for clustering of the response values of subjects in a cluster. The different regression models and statistical methods used for estimating the regression coefficients in the models are explained in the subsequent subsections.

Conditional/subject-specific model

The models classed under this category adjust for clustering by using the outcome for each individual subjects and conditionally relate it to a fixed effects component and a random effects component of the model. The parameter estimates of the fixed effects and the random effects components of the model are obtained simultaneously, and inferences are made regarding the individual subjects in the trial. The generalized linear mixed model (GLMM) is a common example under this approach.

GLMM with coefficients estimated by MLE (GLMM-MLE)

The generalized linear mixed model is also called a random (or mixed) effects model and is the most used conditional/subject-specific model for analysing CRCTs (1,2). In a GLMM, a univariate equation is specified to assess the fixed effects of some covariates of interest and the random effects of the randomly selected clusters on the outcome(s) of interest in the study. The parameters of the two components of the GLMM are estimated simultaneously using MLE. An alternative likelihood-based estimation method like the restricted maximum likelihood estimation (REML) can be utilised. A major drawback of the GLMM-MLE method is that a strong Normality assumption is made about the unobserved random effects component. It is likely that this assumption may not hold in practice which might result in invalid inferences. Let y_{ij} denote a continuous outcome from a jth individual in an ith cluster, a GLMM model is thus specified as

$$y_{ij} = \beta_0 + \beta_k x_{kij} + \tau_i + \varepsilon_{ij}, \quad i = 1, \dots, N; \; j = 1, \dots, n_i; \; k = 1, \dots, p-1 \tag{2}$$

$$\tau_i \sim N(0, \sigma_a^2); \; \varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2)$$

where β_0, β_1 and x_{1ij} is as defined in equation

(1); x_{kij} is a $p \times 1$ vector of individual-level and/or cluster-level fixed effects covariates with coefficients β_k, n_i is the ith cluster size, N is the total number of the independent clusters, τ_i is the random effects term for the ith cluster which causes each cluster mean to vary and ε_{ij} is the random error or residual for each individual. The model specified in equation

(2) is a linear mixed model for a Normally distributed outcome, when y_{ij} is a non-Gaussian outcome such as a binary or count outcome, equation

(2) can be generalized. This explains the "generalized" in GLMM and is given as follows

$$\eta\left(E(y_{ij})\right) = \eta(\mu_{ij}) = \beta_0 + \beta_k x_{kij} + \tau_i \tag{3}$$

where y_{ij} is a non-Gaussian outcome, $\eta(.)$ is a link function that linearly relates the response values to the fixed effects component and the random effects component of the model. β_0, β_1, x_{1ij}, β_k, x_{kij} and τ_i are the same as defined in equationyij= $\beta_0 + \beta_k x_{kij} + \tau_i + \varepsilon_{ij}, \quad i = 1, \dots, N; \; j = 1, \dots, n_i; \; k = 1, \dots, p-1$

(2). For example, if $y_{ij} \sim Bi(1, \mu_{ij})$ then equation

(2) is specified using a logit link function as

$$logit\ (\mu_{ij}) = \beta_0 + \beta_k x_{kij} + \tau_i \tag{4}$$

where μ_{ij} is the probability of $y_{ij} = 1$ and $logit\ (\mu_{ij}) = \frac{\mu_{ij}}{(1-\mu_{ij})}$, all other parameters of equation μ▯ij▯▯ $= \beta_0 + \beta_k x_{kij} + \tau_i$

(

4

)

(3)). The general full likelihood (19) of equation

(3) and μ▯ij▯▯ $= \beta_0 + \beta_k x_{kij} + \tau_i$

(4) is given as

$$l(\theta, \tau_i\,; y_{ij}) = \prod_{i=1}^{N} \int \prod_{j=1}^{n_i} \psi(\tau_i, \theta) g(\tau_i; \sigma_a^2) \partial\tau_i \tag{5}$$

Where $l(.)$ is the full likelihood function for y_{ij}, $\psi(.)$ is the probability function for y_{ij}, $g(.)$ is the Normal probability density function for τ_i and $\theta = (\beta_0, \beta_1, \beta_k)$. The maximum likelihood parameter estimates are obtained by taking the first derivatives of the log of $l(.)$ with respect to each parameter, while the second derivatives produce the standard errors. It is difficult to obtain a closed form solution analytically for equation (5) due to the high dimension of the integral involved, a numerical likelihood approximation method is often used to circumvent this problem. We used the Adaptive Gauss-Hermite Quadrature (AGHQ) to perform the numerical approximation (20). The GLMM models were implemented using the SAS 9.4 procedure; *PROC GLIMMIX.*

Population-Averaged/Marginal Model

The regression models under this class are appropriate for assessing the population averaged intervention effect. Inferences are made across the sub populations of the treatment arms rather than on the individual subjects. They are formulated based on the marginal likelihoods of the correlated response values for the i^{th} cluster, $\boldsymbol{Y}_i$, hence are considered as semi-parametric models. The marginal distribution of $\boldsymbol{Y}_i$ is modelled using a generalized linear model like equation $\eta E(y_{ij})$▯ $= \eta(\mu_{ij}) = \beta_0 + \beta_k x_{kij} + \tau_i$

(3), but without the random effects component τ_i, hence, the correlation among any pair of outcomes within a cluster are accounted for using a separate working covariance matrix. In general, a marginal model is given as

$$\eta\big(E(\boldsymbol{Y}_i)\big) = \eta(\boldsymbol{\mu}_i) = \boldsymbol{X}_{pij}^{T}\,\boldsymbol{\beta}_p \tag{6}$$

The marginal variance of a univariate response value y_{ij} is often specified as $\phi v(\mu_{ij})$, where $v(.)$ is a known variance function and ϕ is a scale parameter which equals 1 for a binary outcome and σ^2 (needs to be estimated) for a continuous outcome. Equation
(6) is similar to
(1) but different in that $corr(\varepsilon_{ij}, \varepsilon_{ij'}) \neq 0$ but rather $corr(\varepsilon_{ij}, \varepsilon_{ij'}) = \rho(x_{ij}, x_{ij'}; \boldsymbol{P})\ \forall\, j \neq j'$, $\boldsymbol{P}$ is the true correlation matrix to be approximated by a "working" correlation matrix, $\boldsymbol{R}$, which is characterised by the intracluster correlation coefficient (ICC), α. The formula for ICC is given in equation
(7) below, obtained from Campbell and Walters (Chapter 5) (17).

mGLM with coefficients estimated by GEE1 (mGLM-GEE1)

The first order generalized estimating equations (GEE1) is the most common multilevel statistical method used for obtaining the parameter estimates of a marginal generalized linear model (mGLM)(i.e., a population-averaged model) specified in equation
(6), and we denote it as the mGLM-GEE1 onward. Here, the clusters are assumed to be a random sample from the population of clusters of interest, hence the variations within a cluster, quantified by the ICC is not of much interest and it is treated as a nuisance/noise. However, the nuisance effect is accounted for by using a "working" covariance matrix, the ICC characterises the working covariance matrix.

The ICC quantifies the correlation between the response values of any pair of subjects within a cluster, when the ICC is zero it indicates that any randomly paired outcome values from any randomly paired subjects in a cluster are independent giving rise to the "independence" working covariance matrix. However, in most CRCTs the ICC is assumed to be the same and nonzero across all the clusters which gives rise to the "exchangeable" working covariance matrix. The independence and the exchangeable working covariance matrices are the two most assumed in CRCTs. Since we are interested in accounting for the correlations among outcomes in a particular cluster, we discussed and considered only the exchangeable working covariance matrix in this study (18,21).

The ICC is given as

$$\hat{\alpha} = \frac{\hat{\sigma}_a^2}{\hat{\sigma}_a^2+\hat{\sigma}_\varepsilon^2} \text{ for continuous outcomes, or } \hat{\alpha} = \frac{\hat{\sigma}_a^2}{\hat{\sigma}_a^2+\frac{\pi^2}{3}} \text{ for binary outcomes} \tag{7}$$

where $\hat{\sigma}_a^2$ is the intracluster variation, $\hat{\sigma}_\varepsilon^2$ is individual subject variation and $\pi = 3.141593$ (22). Let the univariate response value y_{ij} be as defined in equation

(2) and $\eta E\left(y_{ij}\right)$⍰ $= \eta\left(\mu_{ij}\right) = \beta_0 + \beta_k x_{kij} + \tau_i$

(3), if it's marginal probability density function (or probability mass function for discrete distribution) can be expressed as belonging to the linear exponential family distribution, then the first and second moments of y_{ij} can be solved by taking the partial derivative of the log of the moment generating function (MGF) parameterized in the mean. It is worth noting that the nuisance parameter is also contained in the MGF but without itself being estimated. The mGLM-GEE1 draws its strength from the linear exponential family distribution (23). Liang and Zeger (21) proposed a class of estimating equations that uses a working covariance matrix to obtain the parameter estimates of equation

(6) given as

$$U(\boldsymbol{\beta}) = \sum_{i=1}^{N} \left(\frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}}\right)^T \boldsymbol{V}_i^{-1}\left(\boldsymbol{Y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta})\right) = 0 \tag{8}$$

where $\boldsymbol{V}_i$ is the $n_i \times n_i$ covariance matrix for $\boldsymbol{Y}_i$ (i.e., $\boldsymbol{V}_i = Cov(Y_i)$) specified by the working correlation matrix $\boldsymbol{R}(\alpha)$ and defined as

$$\boldsymbol{V}_i = \phi \boldsymbol{G}_i^{\frac{1}{2}} \boldsymbol{R}_i(\alpha) \boldsymbol{G}_i^{\frac{1}{2}} \tag{9}$$

where $\boldsymbol{G}_i = diag\{v(\mu_{i1}), \cdots, v(\mu_{in_i})\}$ is a diagonal matrix with the diagonal elements $v(\mu_{ij})$ the variance function for each response y_{ij}, and $\boldsymbol{R}_i(\alpha)$ is an $n_i \times n_i$ working correlation matrix specified by the ICC, α. The mGLM-GEE1 estimator computes asymptotically consistent estimates $\widehat{\boldsymbol{\beta}}$, regardless of the choice of $\boldsymbol{R}_i(\alpha)$ but provided that the mean structure is correct. However, it may suffer some loss in efficiency if the choice of $\boldsymbol{R}_i(\alpha)$ is wrong (24). The parameter estimates $\widehat{\boldsymbol{\beta}}$ are iteratively obtained by alternating between a modified Fisher scoring algorithm for $\boldsymbol{\beta}$ and the moment estimation of α and ϕ, and its residual $N^{\frac{1}{2}}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ is multivariate Normally distributed with mean zero and a robust sandwich variance-covariance matrix $\boldsymbol{\xi}_i$. The mGLM-GEE1 models were fitted using the SAS 9.4 procedure, *PROC GENMOD.*

mGLM with coefficients estimated by GEE2 (mGLM-GEE2)

The class of regression models under mGLM-GEE2 attempt to leverage on the major drawback of the mGLM-GEE1 which may improve the efficiency of the parameter estimates, by producing parameter estimates with smaller standard errors (i.e., less variable, and more precise). The mGLM-GEE2 model adjust and estimate the correlation parameter (i.e., the nuisance parameter in mGLM-GEE1) and the mean parameter simultaneously in its model specification (6–8,25,26). Also, if accurate assessment of how the association among subjects influences the outcome (s) of interest could be of interest in a study. For example, in a family study to assess the impact of the genetic relatedness of the family members on their alcohol dependence, then mGLM-GEE2 is highly recommended (8).

If the marginal density of $\boldsymbol{Y}_i$ conditioned on the mean vector $\boldsymbol{\mu}_i$ and the covariance matrix $\boldsymbol{V}_i$, can be expressed as belonging to the quadratic exponential family distribution, then this allows for the mean and the covariance of $\boldsymbol{Y}_i$ to be obtained simultaneously. The class of models under the mGLM-GEE2 draws its strength from the quadratic exponential family distribution (23). Several mGLM-GEE2 estimators have been proposed for estimating the mean and correlation parameters simultaneously(6,7,25,26), however that of Yan and Fine (8) used separate link functions to model the mean, the scale, and the correlation structures and generated their corresponding set of estimating equations to obtain the parameter estimates simultaneously and was shown to improve inferences. This is also known as the three-estimating equations (3EE), and it is applied in this paper.

To establish the model specification, let $\boldsymbol{X}_{1i}, \boldsymbol{X}_{2i}\ and\ \boldsymbol{X}_{3i}$ be the $n_i \times p, n_i \times r$ and $\frac{n(n+1)}{2} \times q$ design matrices for the mean, the scale, and the correlation parameters of the vector of outcomes $\boldsymbol{Y}_i$, respectively. The specific link function for the mean, the scale and correlation parameters to $\boldsymbol{X}_{1i}, \boldsymbol{X}_{2i}\ and\ \boldsymbol{X}_{3i}$ is given as

$$\begin{gathered} \eta_1(\boldsymbol{\mu}_i) = \boldsymbol{X}_{1i}\boldsymbol{\beta} \\ \eta_2(\boldsymbol{\phi}_i) = \boldsymbol{X}_{2i}\boldsymbol{\pi} \\ \eta_3(\boldsymbol{\rho}_i) = \boldsymbol{X}_{3i}\boldsymbol{\alpha} \end{gathered} \tag{10}$$

where $\boldsymbol{\mu}_i$ is a $n_i \times 1$ mean vector specified by $\boldsymbol{\beta}$, $\boldsymbol{\phi}_i$ is a $n_i \times 1$ scale vector specified by $\boldsymbol{\pi}$ and $\boldsymbol{\rho}_i$ is a $\frac{n_i(n_i+1)}{2} \times 1$ pairwise correlation vector specified by $\boldsymbol{\alpha}$. The unified corresponding set of estimating equations for equation

$$\eta 3\boldsymbol{\rho} i= \boldsymbol{X}_{3i}\boldsymbol{\alpha}$$

(10) to be solved simultaneously are

$$U(\boldsymbol{\beta},\boldsymbol{\pi},\boldsymbol{\alpha}) = \begin{array}{l} \sum_{i=1}^{N}\left(\frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}}\right)^T \boldsymbol{V}_i^{-1}\left(\boldsymbol{Y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta})\right) = 0 \\ \sum_{i=1}^{N}\left(\frac{\partial \boldsymbol{\phi}_i}{\partial \boldsymbol{\pi}}\right)^T \boldsymbol{V}_{2i}^{-1}(\boldsymbol{Z}_i - \boldsymbol{\phi}_i(\boldsymbol{\pi})) = 0 \\ \sum_{i=1}^{N}\left(\frac{\partial \boldsymbol{\rho}_i}{\partial \boldsymbol{\alpha}}\right)^T \boldsymbol{V}_{3i}^{-1}\left(\boldsymbol{S}_i - \boldsymbol{\rho}_i(\boldsymbol{\alpha})\right) = 0 \end{array} \quad (11)$$

where $\boldsymbol{Y}_i$ and $\boldsymbol{V}_{1i}$ is as defined in the mean model of equations (8) and (9), $\boldsymbol{Z}_i$ is the $n_i \times 1$ vector of the scales, $\boldsymbol{S}_i$ is the $\frac{n_i(n_i+1)}{2} \times 1$ vector of the pairwise correlations, $\boldsymbol{V}_{1i}$ and $\boldsymbol{V}_{2i}$ are the working covariance matrices of $\boldsymbol{Z}_i$ and $\boldsymbol{S}_i$ respectively.

The mGLM-GEE2 estimator of equation (11) requires the specification of the first four central moments of the outcome vector (mean response, variance, skewness, kurtosis). Yan and Fine (8) suggested a way around it to avoid the problem of convergence and it is implemented using the *geese* (27) function in R package *geepack* (28). In general, the third and fourth moments can be specified as functions of the first and second moments, thereby avoiding the direct estimation of higher order moments (7). The mGLM-GEE2 estimator consistently estimate the mean parameters $\boldsymbol{\beta}$ regardless of whether the scale and correlation structures are wrong, the estimates for scales $\boldsymbol{\pi}$ are consistent regardless of whether the working correlation is mis-specified, but provided that the mean and scale structures are correct.

The major merit of the 3EE variant of the mGLM-GEE2 estimator is that it allows for separate covariates in the mean, the scale and the correlation structures to be adjusted for, this is important when investigating heterogeneity across the clusters or sub populations (e.g., the treatment arms). Where each cluster or treatment arm presents a different degree of correlation α_i among subjects, possibly due to cluster sizes and covariates imbalance. Taking this into account may improve efficiency, instead of assuming a constant correlation value across the clusters or treatment arms (5). The solutions of equation (11) are obtained interactively by alternating between a modified Fisher scoring algorithm and moment estimation method. The mGLM-GEE2 models were fitted using the R's *geese* function in the *geepack* package.

mGLM with coefficients estimated by QIF (mGLM-QIF)

The quadratic inference function (QIF) was proposed to circumvent the major pitfalls of the mGLM-GEE1 just like the mGLM-GEE2 estimator. mGLM-QIF avoids the direct estimation of the correlation parameter (the ICC) that characterises the correlation matrix. Instead, it uses a linear combination of basis matrices and some constants to replace the inverse of the working correlation matrix. This strategy enables the mGLM-QIF estimator to obtain consistent and efficient parameter estimates compared to the mGLM-GEE1 estimator, even when the working covariance structure is not correct, and equal mGLM-GEE1 if it is correct (9,29).

Let $\boldsymbol{Y}_i$, $\boldsymbol{X}_i$, $\boldsymbol{\mu}_i$, and $\boldsymbol{V}_i$ be the same as defined in equations (8) and (9). In mGLM-QIF equation the inverse of $\boldsymbol{R}$ specified in equation (9) of the covariance matrix $\boldsymbol{V}_i$ is approximated using a linear combination of a set of several basis matrices $\boldsymbol{R}_h^{-1} \approx k_h\boldsymbol{I}_h + \cdots + k_m\boldsymbol{M}_m;$ $(h = 0, \dots, m)$; $\boldsymbol{I}_h$ is the identity matrix, $\boldsymbol{M}_m$ are known basis matrices and $\boldsymbol{k}_m$ are unknown constants that need to be estimated. For the exchangeable and autoregressive working covariance matrix, $h = 1\ and\ 2$ would suffice, respectively (13,29). Using this new information, we can rewrite the estimating equations (8) of the mGLM-GEE1 as an extended score vector given as

$$\overline{\boldsymbol{g}}_N(\boldsymbol{\beta}) = \frac{1}{N}\sum_{i=1}^{N} g_i(\boldsymbol{\beta}) \approx \frac{1}{N}\begin{pmatrix} \sum_{i=1}^{N}\left(\frac{\partial\boldsymbol{\mu}_i}{\partial\boldsymbol{\beta}}\right)^T \boldsymbol{G}_i^{-1}(\boldsymbol{Y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta})) \\ \sum_{i=1}^{N}\left(\frac{\partial\boldsymbol{\mu}_i}{\partial\boldsymbol{\beta}}\right)^T \boldsymbol{G}_i^{-1/2}\boldsymbol{M}_1\boldsymbol{G}_i^{-1/2}(\boldsymbol{Y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta})) \\ \vdots \\ \sum_{i=1}^{N}\left(\frac{\partial\boldsymbol{\mu}_i}{\partial\boldsymbol{\beta}}\right)^T \boldsymbol{G}_i^{-1/2}\boldsymbol{M}_m\boldsymbol{G}_i^{-1/2}(\boldsymbol{Y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta})) \end{pmatrix} \tag{12}$$

In equation (12), the constants $\boldsymbol{k}_m$ are considered as nuisance and are not included. The mGLM-QIF uses the generalised method of moments (GMM) (30) to optimally combine the multiple estimating equations in (12) which are more than the unknown parameters. The estimate $\widehat{\boldsymbol{\beta}}$ is obtained by minimising the weighted length of $\overline{\boldsymbol{g}}_N$ using GMM and given as

$$\widehat{\boldsymbol{\beta}} = arg\overline{\boldsymbol{g}}_N^T\boldsymbol{\Sigma}_N^{-1}\,\overline{\boldsymbol{g}}_N \tag{13}$$

where the true covariance matrix $\boldsymbol{\Sigma}_N$ is replaced by the empirically estimated covariance matrix $\boldsymbol{C}_N$ in equation

(13), and its inverse $\boldsymbol{C}_N^{-1}$ serves as a weighting function. $\boldsymbol{C}_N^{-1}$ is the main reason behind QIF's efficiency advantage, because it weights the information each i^{th} cluster contributes to the estimating equation, clusters with large variation are given less weight than the ones with small variation. The QIF estimator is thus defined as

$$Q_N(\boldsymbol{\beta}) = \bar{\boldsymbol{g}}_N^T \mathbf{C}_N^{-1} \, \bar{\boldsymbol{g}}_N \tag{14}$$

where $\boldsymbol{C}_N = (1/N^2) \sum_i^N g_i(\boldsymbol{\beta}) g_i^T(\boldsymbol{\beta})$. The estimate $\widehat{\boldsymbol{\beta}}$ is obtained iteratively using the Newton – Raphson algorithm (29). The mGLM-QIF models were fitted using the SAS 9.4 macro: *qif.*

Main analysis

The sample size characteristics of our case studies are summarised using frequency and percentage. All the models were fitted using only complete cases. Among the case studies, the range of the missing data was from 0% to 7% which is negligible, hence no sensitivity analysis was conducted. In clinical trials, it is a common strategy to fit unadjusted and adjusted regression models containing different numbers of covariates to efficiently assess the effect of the intervention administered compared to the control. The unadjusted/univariate model is a model containing only the indicator variable x_{1ij} for the randomised treatment arms as a covariate. While the adjusted/multivariate models include other known prognostic factors $\boldsymbol{X}_{pij}^T$ (with the treatment arm indicator inclusive), such as baseline outcome values, age and sex. For each of the combined statistical methods (e.g., GLMM-MLE) we fitted both unadjusted and adjusted models. In each analysis we consider a P-value < 0.05 to mean that the result is statistically significant.

Software

Two statistical software packages were used for demonstrating how to analyse CRCTs using the four selected methods. They are SAS (version 9.4) and R (version 1.4.1717). The GLMM-MLE and mGLM-QIF models were fitted using SAS while mGLM-GEE1 and mGLM-GEE2 models were fitted using R. The SAS syntax and R codes for fitting all the statistical models applied to one case study (the PoNDER trial) is provided (see, **Additional file 4**).

The initial plan was to fit all the models using a free and open software such as R, but we observed that the *qif* command in the R's *qif package (CRAN - Package qif (r-project.org))* could not fit the mGLM-QIF model to trials with clusters of size of 1, the PoNDER and Age gap trials had clusters of size 1, the error message suggests that it is a problem of the incompatibility of the matrices in the matrix multiplication procedure. So, we switched to using SAS which was able to overcome the problem. Also, *lmer* command for fitting linear mixed effects model to continuous outcomes in *lme4* package in R does not have AGHQ as an option but *glmer* for generalized linear mixed modelling does. SAS procedure, *GLIMMIX,* has AGHQ as an option for mixed effects models for both continuous and binary outcomes.

The GLMM-MLE models were fitted using the *GLIMMIX* procedure in SAS and we set the quadrature points (nodes) to 10 for the AGHQ algorithm. Higher nodes increase the complexity of the AGHQ procedure but produces more reliable results than lower nodes (20). The *GLIMMIX* procedure does not produce a value for ICC, so we calculated it using the estimates of the between cluster variation and individual variation from *PROC GLIMMIX* output. The mGLM-GEE1 models were fitted using the *geeglm* function of R's *geepack* package with an exchangeable correlation structure, and so was mGLM-GEE2 using the *geese* function. The mGLM-QIF models were fitted using the *qif* macro in SAS. In the mGLM-GEE2 models no covariate was adjusted for in the working correlation and scale structures. The link function for the mean structure was either identity for continuous outcome or logit for binary outcome, for the scale structure it was the identity, and for the correlation structure it was the modified Fisher's z transformation.

Description of the four CRCT datasets

The PoNDER trial (31)

The PoNDER CRCT aimed to assess the effect of two psychological informed interventions by health visitors on postnatal depression in postnatal women who have recently given birth. One-hundred and one general practices (clusters) in the Trent region of England were included in the trial. The general practices were randomised in a 2:1 ratio to the Intervention group (n=63 clusters) or the control group (n=38 clusters). Health visitors in the intervention clusters were trained to identify depressive symptoms at six to eight weeks postnatally using the Edinburgh postnatal depression scale (EPDS) and were also trained in providing psychologically informed sessions based on cognitive behavioural or person-centred principles for an hour a week for eight weeks. Health visitors in the control group provided usual care.

The primary outcome was the score on the EPDS at six months follow-up. The EPDS consists of 10 questions and generates a score on a 0 to 30 scale with higher scores indicating a great risk of depression. For the PoNDER trial this outcome was dichotomised into a binary outcome of EPDS score < 12 vs >=12 with women with a score of 12 or more classified as "at risk" of postnatal depression. One hundred (n=63 intervention, n= 37 control) clusters, and n=2659 new mothers (1745 Intervention: 913 Control) provided valid primary outcome data at 6 months. Also, one of the secondary outcomes in the PoNDER trial "the mean EPDS score at six months" was used as a continuous outcome in this study. In the original study both outcomes were analysed using mGLM-GEE1 and an exchangeable correlation structure with robust standard errors. The descriptive statistics of the trial size is presented in **Table 2** below.

The Bridging the Age Gap trial (32)

Bridging the age gap CRCT investigated the effects of two decision support interventions (DESI) to support treatment choices in older women (aged >=70 years) with operable breast cancer (32). Forty-six breast cancer units (clusters) in England and Wales were included in the trial. The breast cancer units were randomised to have access to the DESI (Intervention group n=21 clusters) or to continue with usual care (Control group n=25 clusters). The DESI comprised an online algorithm, booklet, and brief decision aid to inform choices between surgery plus adjuvant endocrine therapy versus primary endocrine therapy, and adjuvant chemotherapy versus no chemotherapy.

The primary outcome was the global health status/quality of life (QoL) score (questions 29 and 30) on the cancer specific patient reported outcome the European Organisation for the Research and Treatment of Cancer (EORTC) QoL questionnaire (QLQ)-C30 at 6 months post baseline. The EORTC QLC-C30 global health status/QoL scale is scored on a 0 to 100 scale with a higher score representing a better QoL. Forty-three clusters (n=19 intervention, n= 24 control), and n=748 patients (359 Intervention: 389 Control) provided valid primary outcome data at 6 months.

The primary endpoint was a continuous outcome "Global health status quality of life score" measured 6 months after diagnosis and was analysed using mGLM-GEE1 with sandwich (robust) standard errors and an exchangeable working correlation matrix. The total number of participants included in the trial is 748 distributed across 43 clusters and the cluster size ranged from size 1 to 73. The complete description of the trial size is provided in **Table 2**.

The Informed Choice trial (33)

The Informed Choice (IC) study was aimed at investigating the impact of a set of 10 pairs of evidence-based leaflets – The Midwives' Information and Resource Service (MIDIRS) and NHS Centre for Reviews and Dissemination informed choice leaflets through a survey. The study was designed to cover 8 of the 10 MIDIRS decision points in everyday maternity care. Conducted in 12 large maternity units in Wales, the maternity units were grouped into 10 clusters. Pairs of clusters were randomly assigned to the intervention arm and control arm based on their annual numbers of deliveries, to achieve balance.

The primary objective was to improve the management of women during pregnancy and childbirth, by assessing the effect of an intervention that promotes informed choice. The primary binary outcome was the change in the proportion of women who reported exercising informed choice (yes or no). For illustrations, one of the secondary outcomes "the average of the women's levels of knowledge" on the 10 topics covered in the survey was used as a continuous outcome in this current study. Knowledge of the topics was assessed on a 1 (poor) to 10 (good) scale.

Two samples of different women were surveyed, the antenatal and postnatal samples. The antenatal sample is made up of all women who reached 28 weeks' gestation period during a six-week period and were receiving antenatal care in any setting. The questionnaire used for the cohort covered three decision points that the women may have encountered. The postnatal sample was made up of all women who delivered live babies during a six-week period. A questionnaire that covered the remaining five decision points was used to survey the women postnatally. The postnatal sample had a total of 3,288 women, who were cross sectionally surveyed before (n = 1,741) and after the intervention was administered (n = 1,547). However, to demonstrate the fitting of the statistical methods in this study only the follow-up (i.e., after the intervention) postnatal sample was used and reported. Only women who delivered in all settings and above the age of 16 years were included. Random effects models (i.e., GLMM) were used to analyse the outcomes in the original study. **Table 2** presents the descriptive summary of the trial size.

The Nourishing Start for Health (NOSH) trial (34)

The NOSH CRCT assessed the effect of an area-level financial incentive (shopping vouchers) on breastfeeding in new mothers and babies in areas with low breastfeeding prevalence (34). Ninety-two electoral ward areas (clusters) in England were included in the trial with baseline breastfeeding prevalence at 6 to 8 weeks postnatally of less than 40%. The areas were randomised to the financial incentive plus usual care (n = 46 clusters) or to usual care alone (n = 46 clusters). All 92 clusters provided breastfeeding outcome data on 9,207 mother-infant pairs (4,973 in the NOSH group, 4324 in the control group) (**Table 2**).

The primary outcome was the electoral ward area-level 6 to 8 weeks breastfeeding prevalence, as assessed by clinicians at the routine 6 to 8 weeks postnatal check. This was derived from the number of new mothers who were breastfeeding or not at 6 weeks in each local authority area/cluster**.** A cluster level approach was used to analyse the primary outcome after obtaining a summary measure for each cluster. Specifically, a weighted multiple linear regression model was used in the original study.

Table 2 Summary of the trial size of the four CRCT data sets used in the study

Trial	No. of clusters	No. of participants	Average cluster size	(Min, Max) cluster size	Median cluster size	Missing n (%)
PoNDER	101	2659	27	(1, 101)	21	35 (1)
Informed Choice	10	1547	155	(74, 308)	145	108 (7)
Age Gap	43	748	18	(1, 73)	16	36 (5)
NOSH	92	9207	100	(12, 333)	75	0 (0)

RESULTS

Description of results for each case study

The PoNDER trial

The mean age of all the women in the control and intervention groups were the same (32±5yrs, respectively), the maximum age across all the women was 46yrs. The proportion of women with EPDS score ≥ 12 at 6 months was 16% (150/914) in the control arm and 12% (205/1745) in the intervention arm. For the other outcome “the mean EPDS score at six months”, it was 6.4(SD = 5.0) vs 5.5(SD = 4.9) for the control vs the intervention arms respectively. It is worth noting that for both outcomes, smaller is better.

The results for the unadjusted intervention effect from the analysis of the continuous primary outcome are slightly different among the models except for mGLM-GEE1 and mGLM-GEE2 which were the same. After adjustments were made for the baseline EPDS 6 weeks score, living alone, previous history of major life events and previous history of postnatal depression, the intervention effect became approximately the same across the models (mean difference, -0.78) except for mGLM-QIF (-0.84). The standard errors of the intervention effect estimates are approximately the same across the models, ranging from 0.25 to 0.28 for the unadjusted models and 0.20 to 0.21 for the adjusted models. The intervention effect estimates across all the models were significant as evident by the small P-values (<0.05) and the confidence intervals which excluded zero. Similar results were obtained from the binary primary outcome analysis, the odds ratio was 0.67 in all the unadjusted models and adjusted models, except for mGLM-QIF (0.66 and 0.62 respectively), and all were significant as well, suggested by the small P-values and confidence intervals excluding one (**Table 3**).

These results are graphically compared using forest plots and shown in **Figure a** (& **Figure b**) and **Figure** Error! Reference source not found.**a** (& **Figure** Error! Reference source not found.**b**), in the plots all the point estimates for the intervention effect and the associated 95% confidence intervals (CIs) are to the left-hand side of zero favouring the intervention arm. The distance between the left and right whiskers that indicate the 95% CIs are approximately the same for all the models.

The Age – Gap trial

The mean global health status/quality of life (QoL) score at 6-months follow-up was 68.9 (SD 19.6) for the control arm against 69.0 (SD 19.5) for the intervention arm. The results from the analysis of the primary continuous outcome in the Age – Gap trial is summarised in **Table 4** and graphically shown in **Figure e** & **Figure f**. The Age – Gap trial had a moderate number of clusters (43 clusters) with 748 patients in total. The unadjusted models appear to produce differing intervention effect values ranging from mean difference of -0.28 to 0.12 but became stable after the baseline QoL values (*ql scale*) was adjusted for; the mean difference became 1.71 for all the models except for mGLM-QIF (1.46). However, all the estimates of the intervention effect across the models were not significant (i.e., P > 0.05). The SEs are approximately the same for the adjusted models (1.40) except for mGLM-QIF (1.20).

Informed Choice trial

The Informed choice trial had a small number of clusters (10) with large number of subjects (1547). In the intervention arm 59% (477/816) of the women reported to have exercised informed choice while using the maternity service compared to 57% (346/612) in the control arm. And the mean knowledge of the 10 topics covered in the survey for those in the intervention arm was 3.6 (SD = 1.62) compared to the 3.3 (SD = 1.60) of the control arms. The covariates in the adjusted models were the age of the mother, age the mother left education, parity and the delivering style.

The results of the unadjusted and adjusted models from the analysis of the primary continuous and binary outcomes are presented in **Table 5** and visualised in **Figure c** (&**Figure d**) and **Figure c**(& **Figure d**), respectively. For the continuous outcome the unadjusted intervention effect was the same for the three models (mean difference = 0.20, SE = 0.11) but different for mGLM-QIF (0.03, SE = 0.05). Similarly, the adjusted intervention effects are the same 0.22 (SE = 0.1) for all the models except mGLM-QIF 0.05 (SE = 0.02). The parameter estimates of the intervention effect for the mGLM-QIF models are far more inconsistent with the observed data (difference in mean score = 0.3). The unadjusted models were not significant (i.e., P > 0.05) for all the models. The adjusted models were somewhat significant (i.e., P< 0.05) for all the models except for GLMM-MLE.

Similarly, for the primary binary outcome, the unadjusted odds ratio of women who reported exercising informed choice in the intervention arm compared to the control arm was the same for all the models (odds ratio = 1.12, SE = 0.10 to 0.11) except for mGLM-QIF (1.17, SE = 0.04). The adjusted odds ratio for all the models were the same (odds ratio = 1.1, SE = 0.10 to 0.11). The odds ratios for the unadjusted and adjusted intervention effect were not significant for all the models except mGLM-QIF which was highly significant (P < 0.0001) (see, **Table 5**).

The NOSH trial

Overall, 36% (1869/4973) mothers in the 46 clusters of the NOSH group were breastfeeding at 6 weeks compared to 30% (1299/4324) in the 46 clusters of the control group. The statistical analysis adjusted for cluster level baseline prevalence and local government area as covariates.

For the NOSH case study, only a binary primary outcome was measured. The results from the unadjusted and adjusted models are presented in **Table 6** and are graphically presented in **Figure e** and **Figure f.** The odds ratios that the mothers were breastfeeding at the end of trial were approximately the same for all the unadjusted (1.40) and adjusted (1.30) models and were statistically significant. This was similar for the SEs = 0.08 for all the univariate/unadjusted models and 0.07 for all the multivariate/adjusted models except for mGLM-GEE2 (0.05). The ICCs from the unadjusted and adjusted mGLM-GEE2 models were quite different from the other models. The ICC ranged from 0.02 to 0.04 for the unadjusted models and 0.004 to 0.02 for the adjusted models (see, **Table 6**).

Table 3 Summary of the results obtained from fitting the different statistical models to the PoNDER trial data (N = 2659)

		Continuous outcome[1]				Binary outcome[2]			
Parameter	**Type of modelling**	**GLMM-MLE**	**mGLM-GEE1**	**mGLM-GEE2**	**mGLM-QIF**	**GLMM-MLE**	**mGLM-GEE1**	**mGLM-GEE2**	**mGLM-QIF**
Intervention effect[3]	Unadjusted	-0.97	-0.98	-0.98	-0.94	0.67	0.67	0.67	0.66
	Adjusted*	-0.78	-0.78	-0.78	-0.84	0.67	0.67	0.67	0.62
SE	Unadjusted	0.25	0.28	0.28	0.28	0.13	0.14	0.14	0.14
	Adjusted*	0.20	0.21	0.21	0.20	0.13	0.13	0.13	0.13
P-value	Unadjusted	0.0002	0.0005	0.0005	0.0009	0.0025	0.0032	0.0032	0.0019
	Adjusted*	0.0001	0.0001	0.0001	<0.0001	0.0019	0.0019	0.0019	0.0001
95% CI	Unadjusted	-1.47 to -0.47	-1.53 to -0.43	-1.53 to -0.43	-1.50 to -0.39	0.51 to 0.86	0.51 to 0.87	0.51 to 0.87	0.51 to 0.86
	Adjusted*	-1.17 to -0.39	-1.18 to -0.38	-1.18 to -0.38	-1.24 to -0.44	0.52 to 0.86	0.52 to 0.86	0.52 to 0.86	0.48 to 0.79
ICC	Unadjusted	0.0167	0.0191	0.0382	0.0191	0.0167	0.0063	0.0126	0.0063
	Adjusted*	0.0077	0.0081	0.0162	0.0081	<0.0001	-0.0018	-0.0036	-0.0018
Number of subjects	Unadjusted	2659	2659	2659	2659	2659	2659	2659	2659
	Adjusted*	2624	2624	2624	2624	2624	2624	2624	2624
Number of clusters	Unadjusted	100	100	100	100	100	100	100	100
	Adjusted*	100	100	100	100	100	100	100	100

* Model adjusted for EPDS score at 6 weeks, living alone (no or yes), previous history of major life events (no or yes) and any previous history of postnatal depression (no or yes). SE = Standard error; CI: Confidence interval; ICC: Intracluster correlation coefficient. GLMM: Generalized linear mixed model; mGLM: marginal generalized linear model; GEE: Generalized estimating equations; QIF: Quadratic inference function

1. EPDS score at 6 months postnatally. The EPDS is scored on a 0 to 30 scale with higher scores indicating a greater risk of PND.
2. Dichotomised EPDS score at 6 months postnatally of < 12 or >=12.
3. The intervention effect for the continuous outcome is the difference in the mean 6-month EPDS scores between the intervention and control groups; with a negative mean difference favouring lower scores (better outcomes) in the intervention group. The intervention effect for the binary outcome is the odds ratio for an EPDS score of 12 or more in the intervention group compared to the control group with an odd ratio <1 favouring better outcomes (lower odds of PND) in the intervention group.

Table 4 Summary of the results from fitting the different models to the Age Gap trial data with a continuous primary outcome[1] (N = 748)

	Unadjusted model				**Adjusted model****			
Parameters	**GLMM-MLE**	**mGLM-GEE1**	**mGLM-GEE2**	**mGLM-QIF**	**GLMM-MLE**	**mGLM-GEE1**	**mGLM-GEE2**	**mGLM-QIF**
Intervention effect[2]	0.12	-0.19	-0.19	-0.28	1.71	1.71	1.71	1.46
SE	1.43	1.26	1.26	1.23	1.40	1.37	1.37	1.20
P-value	0.9343	0.8818	0.8810	0.8175	0.2294	0.2127	0.2127	0.2230
95% CI	-2.77 to 3.00	-2.65 to 2.28	-2.65 to 2.28	-2.69 to 2.12	-1.12 to 4.53	-0.98 to 4.39	-0.98 to 4.39	-0.89 to 3.80
ICC	<0.0001	-0.0068	-0.0135	-0.0068	0.0042	0.0028	0.0056	0.0028
Number of subjects	748	748	748	748	712	712	712	712
Number of clusters	43	43	43	43	43	43	43	43

** Model adjusted of global QoL baseline values. SE = Standard error; CI: Confidence interval; ICC: Intracluster correlation coefficient; GLMM: Generalized linear mixed model; mGLM: marginal generalized linear model; GEE: Generalized estimating equations; QIF: Quadratic inference function

1. Global QoL score on the EORTC-C30 at 6 months post-baseline. The EORTC-C30 Global scale is scored on a 0 (poor) to 100 (good health) scale.
2. The intervention effect for the continuous outcome is the difference in the mean 6-month Global QoL scores between the intervention groups; with a positive mean difference favouring higher scores (better outcomes) in the intervention group.

Table 5 Summary of the results obtained from fitting the different statistical models to the Informed Choice postnatal data (N = 1547)

		Continuous outcome[1]				Binary outcome[2]			
Parameters	Type of modelling	GLMM-MLE	mGLM-GEE1	mGLM-GEE2	mGLM-QIF	GLMM-MLE	mGLM-GEE1	mGLM-GEE2	mGLM-QIF
Intervention effect[3]	Unadjusted	0.20	0.20	0.20	0.03	1.12	1.12	1.12	1.17
	Adjusted***	0.22	0.22	0.22	0.05	1.08	1.06	1.06	1.12
SE	Unadjusted	0.11	0.11	0.11	0.05	0.11	0.06	0.06	0.04
	Adjusted***	0.10	0.10	0.10	0.02	0.11	0.05	0.05	0.07
P-value	Unadjusted	0.1030	0.0730	0.0731	0.5306	0.3178	0.0647	0.0647	<0.0001
	Adjusted***	0.0676	0.0324	0.0324	0.0158	0.5206	0.2175	0.2175	<0.0001
95% CI	Unadjusted	-0.05 to 0.46	-0.02 to 0.41	-0.02 to 0.41	-0.07 to 0.13	0.88 to 1.43	0.99 to 1.27	0.99 to 1.27	1.10 to 1.26
	Adjusted***	-0.02 to 0.46	0.02 to 0.42	0.02 to 0.42	0.01 to 0.09	0.84 to 1.38	0.97 to 1.16	0.97 to 1.16	1.08 to 1.15
ICC	Unadjusted	0.0042	0.0027	0.0055	0.0027	0.0000	-0.0029	-0.0058	-0.0029
	Adjusted***	0.0029	0.0018	0.0036	0.0018	0.0000	-0.0036	-0.0072	-0.0032
Number of subjects	Unadjusted	1534	1534	1534	1534	1485	1485	1485	1485
	Adjusted***	1474	1474	1474	1474	1439	1439	1439	1439
Number of clusters	Unadjusted	10	10	10	10	10	10	10	10
	Adjusted***	10	10	10	10	10	10	10	10

***Model adjusted for mother's age, age mother left education, parity and delivering style. SE = Standard error; CI: Confidence interval; ICC: Intracluster correlation coefficient; GLMM: Generalized linear mixed model; mGLM: marginal generalized linear model; GEE: Generalized estimating equations; QIF: Quadratic inference function

1. Knowledge of informed choice leaflets score at 8 weeks postnatally. Knowledge is scored on a 0 to 10 scale with higher scores indicating a greater knowledge of the leaflets.
2. Proportion of women who answered "yes" to the question "Have you had enough information and discussion with midwives or doctors to make a choice together about all the things that happened during maternity care?" with the options "yes," "partly," "no," "there was no choice," and "did not apply."
3. The intervention effect for the continuous outcome is the difference in the mean 6 week knowledge scores between the intervention and control groups; with a positive mean difference favouring (better outcomes) in the intervention group. The intervention effect for the binary outcome informed choice (yes or no) is the odds ratio for yes to overall informed choice in the intervention group compared to the control group with an odd ratio >1 favouring better outcomes (higher odds of an informed choice) in the intervention group.

Table 6 Summary of the results obtained from fitting the different statistical models to the NOSH data with binary outcome[1]

	Unadjusted model				Adjusted model†			
Parameters	**GLMM-MLE**	**mGLM-GEE1**	**mGLM-GEE2**	**mGLM-QIF**	**GLMM-MLE**	**mGLM-GEE1**	**mGLM-GEE2**	**mGLM-QIF**
Intervention effect[2]	1.37	1.36	1.36	1.36	1.31	1.31	1.27	1.28
SE	0.08	0.08	0.08	0.08	0.07	0.07	0.05	0.07
P-value	0.0002	<0.0001	<0.0001	0.0009	0.0002	<0.0001	<0.0001	0.0002
95% CI	1.16 to 1.60	1.17 to 1.58	1.17 to 1.58	1.17 to 1.59	1.14 to 1.51	1.14 to 1.49	1.15 to 1.41	1.12 to 1.46
ICC	0.0262	0.0192	0.0383	0.0192	0.0162	0.0098	0.0042	0.0098
Number of subjects	9207	9207	9207	9207	9207	9207	9207	9207
Number of clusters	92	92	92	92	92	92	92	92

†The statistical models were adjusted for the cluster level baseline breast-feeding rate and local government area. SE = Standard error; CI: Confidence interval; ICC: Intracluster correlation coefficient; GLMM: Generalized linear mixed model; mGLM: marginal generalized linear model; GEE: Generalized estimating equations; QIF: Quadratic inference function

1. The binary outcome was if the mother was breastfeeding her baby at 6 weeks postnatally (response value = 1) or not (response value = 0).
2. The intervention effect for the binary outcome is the odds for breastfeeding at 6 weeks postnatally in the NOSH intervention group compared to the odds of breastfeeding in the control group with an odds ratio >1 favouring better outcomes (higher odds of breastfeeding) in the intervention group.

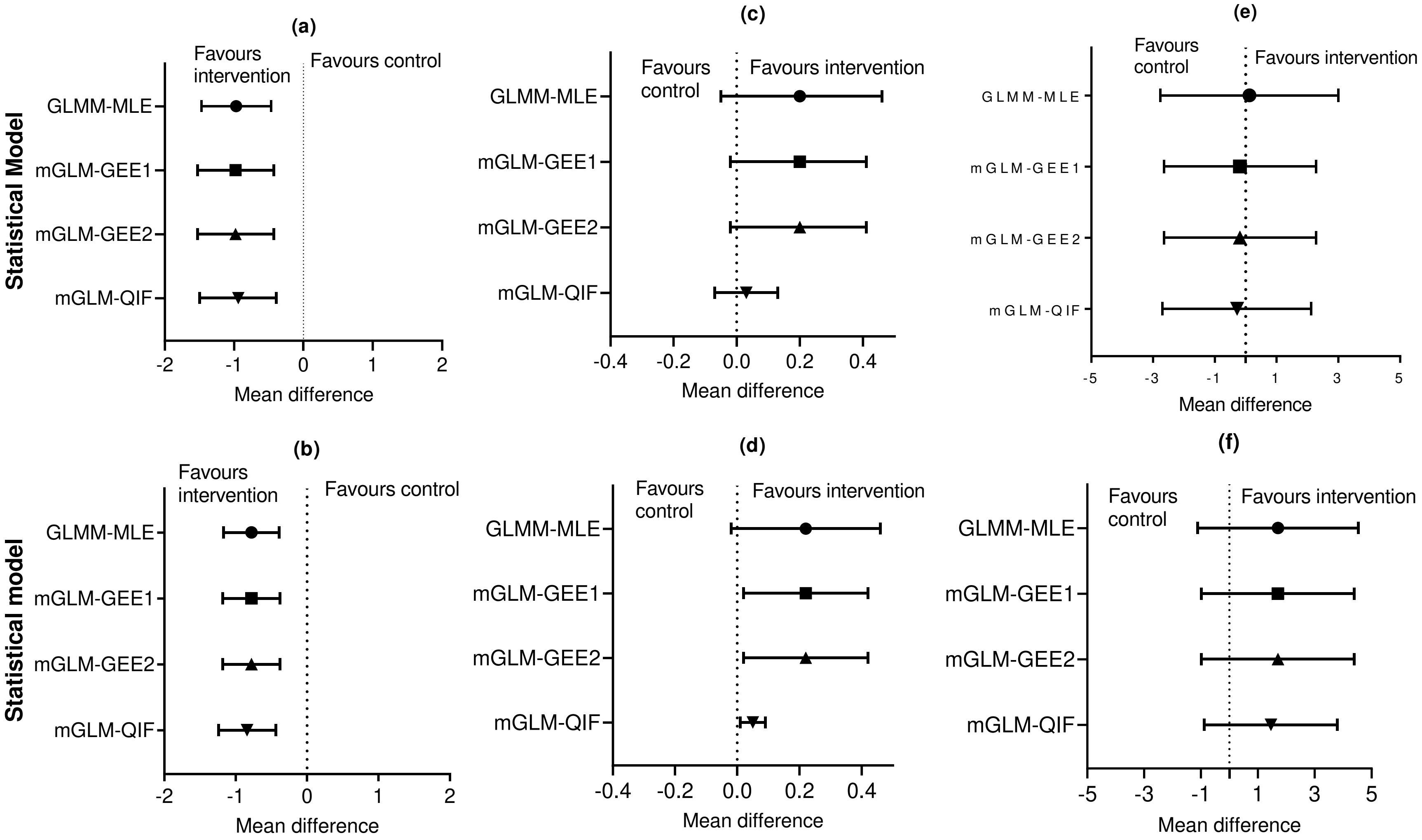

Figure 2 Forest plots showing the intervention effect estimate and its associated 95% CI for the four statistical methods fitted using the continuous primary outcomes of three trial datasets where plot (a) & (b) are the unadjusted and the adjusted models fitted on the PoNDER trial respectively, (c) and (d) is that of the Informed choice and (e) & (f) is of the Age-Gap trial.

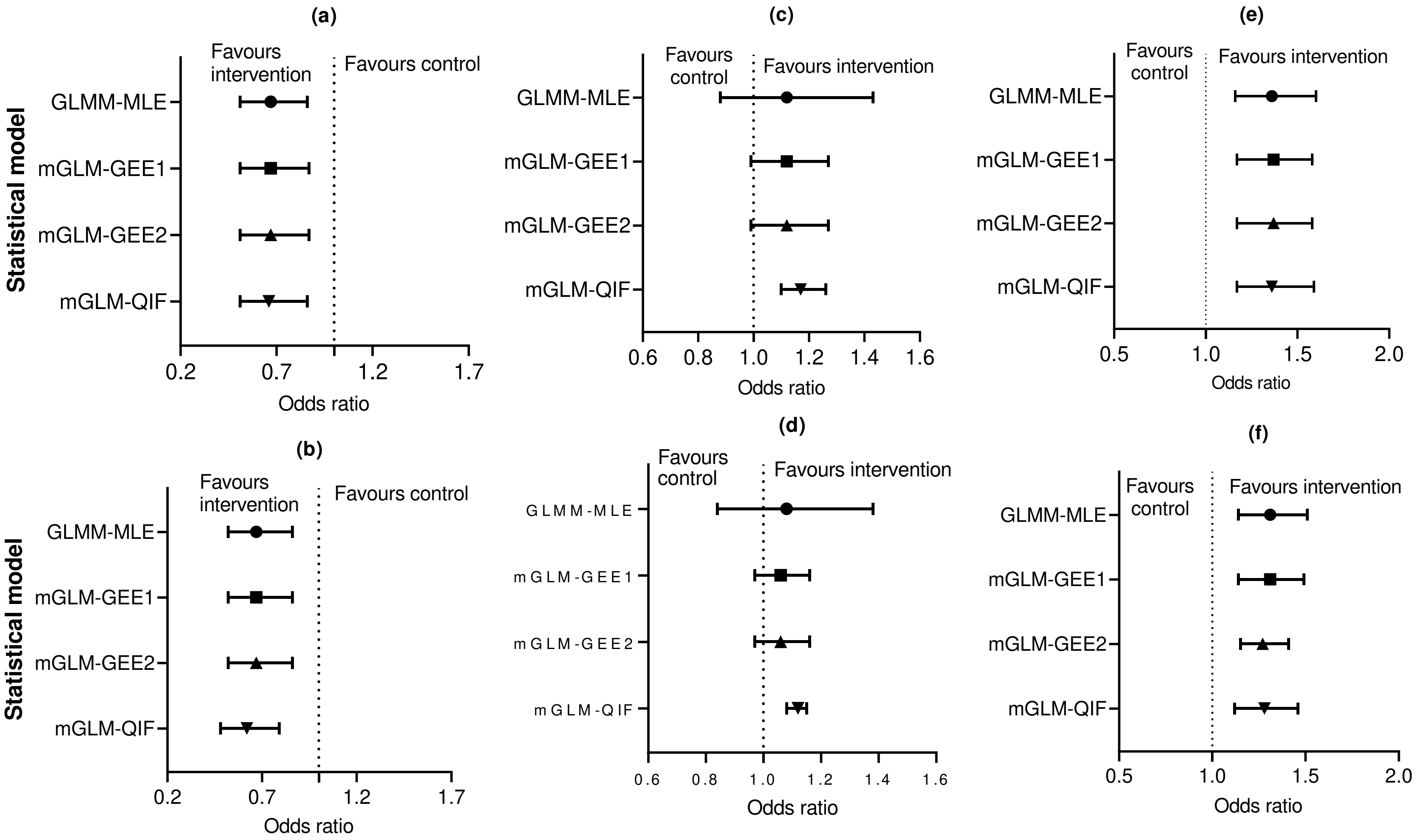

Figure 3 Forest plots showing the intervention effect estimate and its associated 95% CI for each of the statistical model fitted on the binary primary outcomes of three cluster trials datasets where plots (a) & (b) are the unadjusted and the adjusted models fitted on the PoNDER trial respectively, (c) and (d) is that of the Informed Choice trial and (e) & (f) is of the NOSH trial.

DISCUSSION

In this paper, four different approaches for analysing CRCTs with clustering in the treatment arms have been described. All selected four approaches have been applied to the four case studies with different settings to demonstrate their implementation and evaluate their use in practice. The case studies considered have small estimates for the ICC. All had an ICC less than 0.05 and three studies had an ICC less than 0.02. This indicates there was little clustering of outcomes. It is possible that conventional statistical methods that ignore clustering would also have performed well, considering the observed small ICCs and the performance of the four multilevel methods.

Three studies had negative parameter estimates for the ICC, theoretically the ICC is assumed to be bounded between 0 and 1. But in practice negative ICCs can be observed from a real-world trial data. The theoretical set-up of the GLMM-MLE estimator prevents a negative ICC but that is not the same for the marginal models (35), which was the case in this study. Only the marginal models produced negative ICCs. The GLMM-MLE is known to truncate the ICC to zero rather than produce negative ICCs, effectively fitting a generalized linear model (GLM) (36). Sampling error due to limited number of subjects to sample from (sample cluster size) compared to the population cluster size which is unlimited could be the cause for a negative ICC (35). This would cause the sampled clusters not to be good representative of the population clusters. When a negative ICC is obtained, it is recommended to use conventional statistical methods (which ignore clustering), as they are more likely to produce reliable results (35). Another reason this could happen is when there are large discrepancies in the allotment of trial resources within the clusters (17), this would cause large variations in the observed outcomes.

Our results showed that parameter estimates for the intervention effect, SEs, P-values and 95% CI were the same for mGLM-GEE1 and mGLM-GEE2 models in almost all cases, they only differ in their estimates for the ICC. This could possibly be the effect of the observed small ICCs in the case studies. Indicating that both methods are fitting the same models regardless of whether the correlation parameter is estimated or not within the method algorithm. If the observed ICCs happens to be large, it is recommended that models with heterogenous correlation that also allows for covariates adjustments for the corelation structure should be considered, it is likely to improve inferences (5). This happens to be the major merit of Yan & Fine 3EE GEE2 model (8). Obtaining accurate estimates for the correlation parameters was not a major interest in this study. It would be worth investigating to know which of the two methods is producing the more accurate estimate of the corelation parameter using simulation studies.

Our initial intention was to use a free and open-source software package to analyse the datasets such as R, but we resorted to using R to fit mGLM-GEE1 and mGLM-GEE2 only. We used the SAS macro "QIF" because its R sister version could not fit the mGLM-QIF models to datasets of trials with cluster size of 1(i.e., only one outcome was observed in the cluster). The PoNDER and Age Gap trials had clusters with one observed outcome only. We did reach out to one of the authors of both software packages, Peter X.K. Song, through email correspondence and Song promised to investigate this problem with the QIF algorithm in R.

The four case studies have different features which cover most of the different settings of cluster randomised controlled trials in practice. The impact of these features on the estimates of the four statistical models are evident in the results obtained. The PoNDER trial had a large sample size (both in the number of clusters and cluster sizes, 100 clusters with an average cluster size of 26), had small ICCs, both cluster level and individual level covariates were adjusted for in the multivariate models. Hence, the results showed that the intervention effect estimates were apparently equivalent for the four different methods for both the continuous and binary primary outcomes analysed. This was the same for the associated standard errors and 95% CIs. On the aspect of hypothesis testing, the inferences were the same using any of the four statistical models and it was consistent with that of the original analysis by Morrell et al., (31); a significant benefit of training health visitors to adequately manage women with postnatal depressive symptoms (i.e., favouring the intervention arm).

The Age - Gap trial had a moderate sample size (43 clusters with an average cluster size of 18). Among the four statistical methods, estimates for the intervention effect from the univariate/unadjusted models were unstable ranging from -0.28 to 0.12 but became stable and the same (mean difference = 1.78) after adjustment for the individual level baseline values of the primary outcome (baseline Global QoL), except for mGLM-QIF (1.46) which had the smallest SE estimates. This elucidates the importance of accounting for known prognostic factors in clinical trials. This was similar for the SEs and the 95% CIs. All the four statistical models resulted in the same inference and is consistent with that of the original analysis which was "no significant difference in the Global QoL between the control and the intervention arms" (32).

For the Informed Choice trial, continuous and binary outcomes were measured, and the study had small number of clusters (10 clusters) but with large cluster sizes (median cluster size = 145). The original study was a cross-sectional repeated measurement, so the estimate for the intervention effect was the interaction effect term between the treatment group (*group*) and time of measurement (*time*). But for the purpose of demonstration, we used only the "after intervention" postnatal sample. Both cluster and individual level covariates were adjusted for in the multivariate models. All the three methods produced approximately the same parameter estimates which differs from the estimates produced by mGLM-QIF, for both the continuous and binary outcomes.

The impact of the interplay between small number of clusters, covariate and cluster size imbalance on the mGLM-QIF and mGLM-GEE1 has been studied. It was found that the mGLM-QIF is severely affected compared to the mGLM-GEE1(12). A correction was proposed to improve the empirical estimated covariance matrix that causes the mGLM-QIF to be poorly behaved (13). In this study it is more likely that the differing performance of the mGLM-QIF estimator is due to the small number of clusters.

It is of interest to us to carry out simulation studies to comprehensively learn the finite small sample size performance of the mGLM-QIF in relation to the three other statistical models in this study in the context of CRCTs in the future (since the true parameter values would be known). Lastly, for the NOSH trial with only binary primary outcome measured and large sample sizes (92 clusters with an average cluster size of 100). The parameter estimates from the four statistical approaches were approximately the same, hence, their performance was equivalent.

Limitations

This study employed a formal search of relevant literature to capture most of the related work conducted. However, this was not an exhaustive review of all work in this area.

We have used four case studies that have arisen from our work as applied medical statisticians in clinical trials research. The results and inferences made are applicable to data with similar properties to these studies. For example, our results focus only on binary and continuous endpoints and as discussed relate to trials with small ICCs and relatively small clusters. Our analysis of these case studies was on complete cases only, we have ignored any data collected on patients for whom the outcome of interest was not recorded. It is likely that this data limitation (i.e., missing data) might not result to adverse consequences since the proportion of missing data is small. Although, the other data limitations (i.e., small sample size and small ICC) might.

While small number of clusters (and cluster sizes), small ICCs and incomplete data are issues in many real-world data sets, to increase the generalisability of these results to trials with different characteristics to the case studies we hope to conduct a simulation study. This study will explore how our findings might change for varying cluster sizes, varying ICCs, varying number of clusters and differential variance in the control and intervention arms.

CONCLUSION

In summary, we used four cluster randomised controlled trials (CRCTs) as case studies to demonstrate the applications of four statistical methods for analysing CRCTs. The characteristics of the four case studies covered a range of settings in CRCTs in practice; however, the generalizability of our findings should be limited to studies with similar characteristics as our case studies. In most cases the modelling approaches produced similar results which are consistent with the original primary analyses. A plausible reason for this could be the negligible correlation (small ICCs) among responses observed in each of the cases studied.

However, the mGLM-QIF produced differing parameter estimates compared to the other three statistical models in some cases, but they most times reached the same conclusion. These differences are noticeable for studies with small to moderate number of clusters (i.e., less than 46). Although the four statistical methods were compared among each other, we cannot determine a superior method using only this example data analysis. We recommend further research based on simulation studies to comprehensively evaluate the performance of the methods.

Abbreviations

IRCT: Individual randomised controlled trial; CRCT: Cluster randomised controlled trial; TMLE: Targeted maximum likelihood estimator; ALR: Alternating logistic regression; GEE: Generalized estimating equations; QIF: Quadratic inference function; MLE: Maximum likelihood estimator; GLMM: Generalized linear mixed model; mGLM: Marginal generalized linear model; GLM: Generalized linear model; REML: Restricted maximum likelihood; CLA: Cluster level approach; ILA: Individual level approach; AGHQ: Adaptive Gauss-Hermit quadrature; ICC: Intraclusters correlation coefficient; MGF: Moment generating function; 3EE: Three estimating equations; GMM: Generalized method of moments

Acknowledgements

We wish to acknowledge all trial staff, participants and members of the trial review boards—the Data Monitoring and Ethics Committee and Trial Steering Committee. We wish to thank the Trial Chief Investigator for the Bridging the age gap (AgeGap) in breast cancer trial Professor Lynda Wyld, Department of Oncology and Metabolism, University of Sheffield Medical School, Sheffield, UK, for the use of the AgeGap trial data.

Contributors

All authors contributed to the study concept and study design. Literature reviews underpinning the work were conducted by BCO. BCO conducted the data analysis and drafted the manuscript. The first draft of the manuscript was written by BCO with contributions from SJW, and RMJ. All authors critically revised the manuscript and approved the final manuscript.

Funding

BCO's PhD is financially sponsored by the Nigerian Tertiary Education Trust Fund (TETFund) (Grant No. TETF/ES/UNIV/UTURU/TSA/2019). SJW and RMJ received funding across various projects by NIHR. SJW was a National Institute for Health Research (NIHR) Senior Investigator (NF-SI-0617-10012) supported by the NIHR for this research project. The views expressed in this publication are those of the author(s) and not necessarily those of the TETFund, NIHR, NHS or the UK Department of Health and Social Care. These organisations had no role in the study design; in the collection, analysis, and interpretation of the data; in the writing of the report; or in the decision to submit the paper for publication.

Competing interests

The authors declare that they have no competing interests.

Patient and public involvement

Patients and/or the public were not involved in the design, conduct, reporting or dissemination plans of this research.

Consent for publication

Not applicable.

Ethics approval

The need of informed consent from each participant was waived by the University of Sheffield Research Ethics Committee (Reference number 038285). The data analysed in this paper is based on published trials where ethics approvals were obtained by the original trial teams. This paper does not involve recruiting new participants and the original trial participants cannot be identified from this analysis. Additionally, all methods were done in accordance with the relevant guidelines and regulations.

Provenance and peer review

Not commissioned; externally peer reviewed.

Data availability statement

Data are available upon reasonable request from BCO at bcofforha1@sheffield.ac.uk

REFERENCE

1. Offorha BC, Walters SJ, Jacques RM. Statistical analysis of publicly funded cluster randomised controlled trials: a review of the National Institute for Health Research Journals Library. Trials. 2022 Dec;23(1):115.

2. Fiero MH, Huang S, Oren E, Bell ML. Statistical analysis and handling of missing data in cluster randomized trials: A systematic review. Trials. 2016;17(1):1–10.

3. Ivers NM, Taljaard M, Dixon S, Bennett C, McRae A, Taleban J, et al. Impact of CONSORT extension for cluster randomised trials on quality of reporting and study methodology: review of random sample of 300 trials, 2000-8. BMJ. 2011 Sep 26;343(sep26 1):d5886–d5886.

4. Campbell MK, Elbourne DR, Altman DG. CONSORT statement: extension to cluster randomised trials. BMJ. 2004 Mar;328(7441):702 LP – 708.

5. Crespi CM, Wong WK, Mishra SI. Using second-order generalized estimating equations to model heterogeneous intraclass correlation in cluster-randomized trials. Stat Med. 2009;28(5):814–27.

6. Prentice RL. Correlated Binary Regression with Covariates Specific to Each Binary Observation. Biometrics. 1988 Dec;44(4):1033.

7. Prentice RL, Zhao LP. Estimating Equations for Parameters in Means and Covariances of Multivariate Discrete and Continuous Responses. Biometrics. 1991 Sep;47(3):825.

8. Yan J, Fine J. Estimating equations for association structures: ESTIMATING EQUATIONS FOR ASSOCIATION STRUCTURES. Stat Med. 2004 Mar 30;23(6):859–74.

9. Song PXK, Jiang Z, Park E, Qu A. Quadratic inference functions in marginal models for longitudinal data. Stat Med. 2009 Dec;28(29):3683–96.

10. Westgate PM. A bias-corrected covariance estimate for improved inference with quadratic inference functions. Stat Med. 2012;31(29):4003–22.

11. Westgate PM. A readily available improvement over method of moments for intra-cluster correlation estimation in the context of cluster randomized trials and fitting a GEE–type marginal model for binary outcomes. Clin Trials. 2019;16(1):41–51.

12. Westgate PM, Braun TM. The effect of cluster size imbalance and covariates on the estimation performance of quadratic inference functions. Stat Med. 2012;31(20):2209–22.

13. Westgate PM, Braun TM. An improved quadratic inference function for parameter estimation in the analysis of correlated data. Stat Med. 2013 Aug 30;32(19):3260–73.

14. Yu H, Li F, Turner EL. An evaluation of quadratic inference functions for estimating intervention effects in cluster randomized trials. Contemp Clin Trials Commun. 2020;19:100605–100605.

15. Zhang X. A Tutorial on Restricted Maximum Likelihood Estimation in Linear Regression and Linear Mixed-Effects Model. 2015 Jan;11.

16. Pan W. Akaike's Information Criterion in Generalized Estimating Equations. Biometrics. 2001 Mar;57(1):120–5.

17. Campbell MJ, Walters SJ. How to Design, Analyse and Report Cluster Randomised Trials in Medicine and Health Related Research [Internet]. New York, UNITED KINGDOM: John Wiley & Sons, Incorporated; 2014. Available from: http://ebookcentral.proquest.com/lib/sheffield/detail.action?docID=1662762

18. Walters SJ, Morrell CJ, Slade P. Analysing data from a cluster randomized trial (cRCT) in primary care: A case study. J Appl Stat. 2011;38(10):2253–69.

19. McCulloch CE. Maximum Likelihood Algorithms for Generalized Linear Mixed Models. J Am Stat Assoc. 1997 Mar;92(437):162–70.

20. Handayani D, Notodiputro KA, Sadik K, Kurnia A. A comparative study of approximation methods for maximum likelihood estimation in generalized linear mixed models (GLMM). In Jawa Barat, Indonesia; 2017 [cited 2022 Apr 16]. p. 020033. Available from: http://aip.scitation.org/doi/abs/10.1063/1.4979449

21. Liang BYK yee, Zeger SL. Longitudinal data analysis using generalized linear models. Biometrika. 1986;73(1):13–22.

22. Rodríguez G, Elo I. Intra-class correlation in random-effects models for binary data. Stata J. 2003;3(1):32–46.

23. Ziegler A. Generalized estimating equations. New York: Springer; 2011. 144 p. (Lecture notes in statistics).

24. Qu A, Bruce, G. Lindsa. Improving generalised estimating equations using quadratic inference functions. Biometrika. 2000 Dec 1;87(4):823–36.

25. Hall DB, Severini TA. Extended Generalized Estimating Equations for Clustered Data. J Am Stat Assoc. 1998 Dec;93(444):1365–75.

26. Ziegler A, Kastner C, Brunner D, Blettner M. Familial associations of lipid profiles: a generalized estimating equations approach. Stat Med. 2000 Dec 30;19(24):3345–57.

27. Yan J. geepack: Yet Another Package for Generalized Estimating Equations. R-News. 2002 Jan 1;2:12–4.

28. Højsgaard S, Halekoh U, Yan J. The R Package geepack for Generalized Estimating Equations. J Stat Softw. 2005 Dec 22;15(2):1–11.

29. Qu A, Lindsay BG, Bing LI. Improving generalised estimating equations using quadratic inference functions. Biometrika. 2000;87(4):823–36.

30. Hansen LP. Generalized method of moments estimation. In: Durlauf SN, Blume LE, editors. Macroeconometrics and Time Series Analysis [Internet]. London: Palgrave

Macmillan UK; 2010 [cited 2022 Apr 24]. p. 105–18. Available from: http://link.springer.com/10.1057/9780230280830_13

31. Morrell CJ, Warner R, Slade P, Dixon S, Walters S, Paley G, et al. Psychological interventions for postnatal depression: Cluster randomised trial and economic evaluation. The PoNDER trial. Health Technol Assess. 2009;13(30).

32. Wyld L, Reed MWR, Collins K, Burton M, Lifford K, Edwards A, et al. Bridging the age gap in breast cancer: cluster randomized trial of two decision support interventions for older women with operable breast cancer on quality of life, survival, decision quality, and treatment choices. Br J Surg. 2021 May 27;108(5):499–510.

33. O'Cathain A. Use of evidence based leaflets to promote informed choice in maternity care: randomised controlled trial in everyday practice. BMJ. 2002 Mar 16;324(7338):643–643.

34. Relton C, Strong M, Thomas KJ, Whelan B, Walters SJ, Burrows J, et al. Effect of Financial Incentives on Breastfeeding A Cluster Randomized Clinical Trial. JAMA - J Am Med Assoc. 2018;172(2):1–7.

35. Eldridge SM, Ukoumunne OC, Carlin JB. The Intra-Cluster Correlation Coefficient in Cluster Randomized Trials: A Review of Definitions. Int Stat Rev. 2009 Dec;77(3):378–94.

36. Nelder JA, Wedderburn RWM. Generalized Linear Models. J R Stat Soc Ser Gen. 1972;135(3):370.

37. Odueyungbo A, Browne D, Akhtar-danesh N, Thabane L. Comparison of generalized estimating equations and quadratic inference functions using data from the National Longitudinal Survey of Children and Youth (NLSCY) database. BMC Med Res Methodol. 2008;8(28):1–10.

38. Yang W, Liao S. A study of quadratic inference functions with alternative weighting matrices. Commun Stat---Simul Comput-275pt. 2017;46(2):994–1007.

39. Asar O, Ilk O zlem. Flexible multivariate marginal models for analyzing multivariate longitudinal data, with applications in R. Comput Methods Programs Biomed. 2014;115(3):135–46.

40. Austin PC. A comparison of the statistical power of different methods for the analysis of cluster randomization trials with binary outcomes. Stat Med. 2007;26(Jan):3550-3565-3550–65.

41. Austin PC. A comparison of the statistical power of different methods for the analysis of repeated cross-sectional cluster randomization trials with binary outcomes. Int J Biostat [Internet]. 2010;6(1). Available from: https://www.scopus.com/inward/record.uri?eid=2-s2.0-77950524668&doi=10.2202%2F1557-4679.1179&partnerID=40&md5=ca4acf77300c58dbe435f7e86fc641f1

42. Barker D., D'Este C., Campbell M.J., McElduff P. Minimum number of clusters and comparison of analysis methods for cross sectional stepped wedge cluster randomised trials with binary outcomes: A simulation study. Trials. 2017;18(1):119.

43. Borhan S, Mallick R, Pillay M, Kathard H, Thabane L. Sensitivity of methods for analyzing continuous outcome from stratified cluster randomized trials – an empirical comparison study. Contemp Clin Trials Commun. 2019;15:100405–100405.

44. Forbes AB, Akram M, Pilcher D, Cooper J, Bellomo R. Cluster randomised crossover trials with binary data and unbalanced cluster sizes: Application to studies of near-universal interventions in intensive care. Clin Trials. 2015;12(1):34–44.

45. Heo M, Leon AC. Comparison of statistical methods for analysis of clustered binary observations. Stat Med. 2005;24(6):911–23.

46. Hossain A, Bartlett JW. Missing binary outcomes under covariate-dependent missingness in. Stat Methods Med Res. 2017;36(19):3092–109.

47. Kim HY, Preisser JS, Rozier RG, Valiyaparambil JV. Multilevel analysis of group-randomized trials with binary outcomes. Community Dent Oral Epidemiol. 2006;34(4):241–51.

48. Leyrat C., Morgan K.E., Leurent B., Kahan B.C. Cluster randomized trials with a small number of clusters: Which analyses should be used? Int J Epidemiol. 2018;47(1):321–31.

49. Ma J, Thabane L, Kaczorowski J, Chambers L, Dolovich L, Karwalajtys T, et al. Comparison of Bayesian and classical methods in the analysis of cluster randomized controlled trials with a binary outcome: The Community Hypertension Assessment Trial (CHAT). BMC Med Res Methodol. 2009 Dec;9(1):37.

50. McNeish D, Stapleton LM. Modeling Clustered Data with Very Few Clusters. Multivar Behav Res. 2016 Jul 3;51(4):495–518.

51. Morgan KE, Forbes AB, Keogh RH, Jairath V, Kahan BC. Choosing appropriate analysis methods for cluster randomised cross-over trials with a binary outcome. Stat Med. 2016;36(2):318–33.

52. Pacheco GD, Hattendorf J, Colford Jr. JM, Mäusezahl D, Smith T. Performance of analytical methods for overdispersed counts in cluster randomized trials: Sample size, degree of clustering and imbalance. Stat Med. 2009;28(24):2989–3011.

53. Peek N, Goud R, De Keizer N. Handling intra-cluster correlation when analyzing the effects of decision support on health care process measures. In Dept. of Medical Informatics, University of Amsterdam, PO Box 22700, 1100 DD Amsterdam, Netherlands; 2013. p. 22–7. Available from: https://www.scopus.com/inward/record.uri?eid=2-s2.0-84886769645&doi=10.3233%2F978-1-61499-240-0-22&partnerID=40&md5=e7bf1d8a748ede3ab90d55aa59e00689

54. Flórez AJ, Molenberghs G, Verbeke G, Abad AA. A closed-form estimator for meta-analysis and surrogate markers evaluation. J Biopharm Stat. 2019;29(2):318–32.

55. Charvat H, Remontet L, Bossard N, Roche L, Dejardin O, Rachet B, et al. A multilevel excess hazard model to estimate net survival on hierarchical data allowing for non-linear and non-proportional effects of covariates. Stat Med. 2016;35(18):3066–84.

56. Chebon S, Faes C, De Smedt A, Geys H. Marginalized models for right-truncated and interval-censored time-to-event data. J Biopharm Stat. 2019;29(6):1043–67.

57. Ghebretinsae AH, Faes C, Molenberghs G, Geys H, Van Der Leede BJ. Joint modeling of hierarchically clustered and overdispersed non-gaussian continuous outcomes for comet assay data. Pharm Stat. 2012;11(6):449–55.

58. Kang W, Lee MS, Lee Y. HGLM versus conditional estimators for the analysis of clustered binary data. Stat Med. 2005;24(5):741–52.

59. Lam KF, Ip D. REML and ML estimation for clustered grouped survival data. Stat Med. 2003;22(12):2025–34.

60. Lee K, Kang S, Liu X, Seo D. Likelihood-based approach for analysis of longitudinal nominal data using marginalized random effects models. J Appl Stat. 2011;38(8):1577–90.

61. Olsen MK, DeLong ER, Oddone EZ, Bosworth HB. Strategies for analyzing multilevel cluster-randomized studies with binary outcomes collected at varying intervals of time. Stat Med. 2008;27(29):6055–71.

62. Pedroza C, Truong VTT. Estimating relative risks in multicenter studies with a small number of centers - which methods to use? A simulation study. Trials [Internet]. 2017;18(1). Available from: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85032719571&doi=10.1186%2Fs13063-017-2248-1&partnerID=40&md5=71cc5945ce939d8f56555335b5973f41

63. Sauzet O, Wright KC, Marston L, Brocklehurst P, Peacock JL. Modelling the hierarchical structure in datasets with very small clusters: A simulation study to explore the effect of the proportion of clusters when the outcome is continuous. Stat Med. 2013;32(8):1429–38.

64. Tawiah R, Yau KKW, McLachlan GJ, Chambers SK, Ng SK. Multilevel model with random effects for clustered survival data with multiple failure outcomes. Stat Med. 2019;38(6):1036–55.

65. Yelland LN, Sullivan TR, Pavlou M, Seaman SR. Analysis of Randomised Trials Including Multiple Births When Birth Size Is Informative. Paediatr Perinat Epidemiol. 2015;29(6):567–75.

66. Young ML, Preisser JS, Qaqish BF, Wolfson M. Comparison of subject-specific and population averaged models for count data from cluster-unit intervention trials. Stat Methods Med Res. 2007;16(2):167–84.

67. Du R, Lee JH. A weighted Jackknife method for clustered data. Commun Stat - Theory Methods. 2019;48(8):1963–80.

68. Ho MW, Tu W, Ghosh P, Tiwari RC. A Nested Dirichlet Process Analysis of Cluster Randomized Trial Data With Application in Geriatric Care Assessment. J Am Stat Assoc. 2013 Mar;108(501):48–68.

69. Chen BE, Wang J. Joint modeling of binary response and survival for clustered data in clinical trials. Stat Med. 2019;(August):1–14.

70. Brown RL. Modeling impure clusters in a cluster randomized controlled trial. Res Nurs Health. 2013;36(2):216–23.

71. Clark AB, Bachmann MO. Bayesian methods of analysis for cluster randomized trials with count outcome data. Stat Med. 2010;29(2):199–209.

72. Li Z, Xu X, Shen J. Semiparametric Bayesian analysis of accelerated failure time models with cluster structures. Stat Med. 2017;36(25):3976–89.

73. Ma J, Thabane L, Kaczorowski J, Chambers L, Dolovich L, Karwalajtys T, et al. Comparison of Bayesian and classical methods in the analysis of cluster randomized controlled trials with a binary outcome: the Community Hypertension Assessment Trial (CHAT). BMC Med Res Methodol. 2009;9(1):37–37.

74. Müller P, Quintana FA, Rosner GL. Semiparametric Bayesian inference for multilevel repeated measurement data. Biometrics. 2007;63(1):280–9.

75. Pan C, Cai B, Wang L. Multiple frailty model for clustered interval-censored data with frailty selection. Stat Methods Med Res. 2017;26(3):1308–22.

76. Peters TJ, Richards SH, Bankhead CR, Ades AE, Sterne JAC. Comparison of methods for analysing cluster randomized trials: An example involving a factorial design. Int J Epidemiol. 2003;32(5):840–6.

77. Thompson SG, Warn DE, Turner RM. Bayesian methods for analysis of binary outcome data in cluster randomized trials on the absolute risk scale: BAYESIAN METHODS FOR ANALYSIS OF BINARY OUTCOME DATA. Stat Med. 2004 Feb 15;23(3):389–410.

78. Ukoumunne OC, Carlin JB, Gulliford MC. A simulation study of odds ratio estimation for binary outcomes from cluster randomized trials. Stat Med. 2007;26(18):3415–28.

79. Ukoumunne OC, Forbes AB, Carlin JB, Gulliford MC. Comparison of the risk difference, risk ratio and odds ratio scales for quantifying the unadjusted intervention effect in cluster randomized trials. Stat Med. 2008 Nov;27(25):5143–55.

80. Hedeker D. A mixed-effects multinomial logistic regression model. Stat Med. 2003;22(9):1433–46.

81. Wang Z, Louis TA. Marginalized binary mixed-effects models with covariate-dependent random effects and likelihood inference. Biometrics. 2004;60(4):884–91.

82. Prague M, Wang R, Stephens A, Tchetgen Tchetgen E, DeGruttola V, Tchetgen ET, et al. Accounting for interactions and complex inter-subject dependency in estimating treatment effect in cluster-randomized trials with missing outcomes. Biometrics. 2016;72(4):1066–77.

83. Perin J, Preisser JS. Alternating logistic regressions with improved finite sample properties. Biometrics. 2016;73(2):696–705.

84. Johnson JL, Kreidler SM, Catellier DJ, Murray DM, Muller KE, Glueck DH, et al. Recommendations for choosing an analysis method that controls Type I error for

unbalanced cluster sample designs with Gaussian outcomes. Stat Med. 2015;34(27):3531–45.

85. Molas M, Lesaffre E. Hurdle models for multilevel zero-inflated data via h-likelihood. Stat Med. 2010;29(30):3294–310.

86. Lu SE, Wang MC. Marginal analysis for clustered failure time data. Lifetime Data Anal. 2005;11(1):61–79.

87. Bossoli D, Bottai M. Marginal quantile regression for dependent data with a working odds-ratio matrix. Biostatistics. 2018;19(4):529–45.

88. Balzer LB, Petersen ML, van der Laan MJ, Collaboration the S. Targeted estimation and inference for the sample average treatment effect in trials with and without pair-matching. Stat Med. 2016;35(21):3717–32.

89. Murray DM. Statistical models appropriate for designs often used in group-randomized trials. Stat Med. 2001;20(9–10):1373–85.

90. Christian NJ, Ha ID, Jeong JH. Hierarchical likelihood inference on clustered competing risks data. Stat Med. 2016;35(2):251–67.

91. Chen C.-M., Yu C.-Y. A two-stage estimation in the Clayton-Oakes model with marginal linear transformation models for multivariate failure time data. Lifetime Data Anal. 2012;18(1):94–115.

92. Cai J, Kim J. Nonparametric quantile estimation with correlated failure time data. Lifetime Data Anal. 2003;9(4):357–71.

93. Murray DM, Hannan PJ, Pals SP, McCowen RG, Baker WL, Blitstein JL. A comparison of permutation and mixed-model regression methods for the analysis of simulated data in the context of a group-randomized trial. Stat Med. 2006;25(3):375–88.

94. Wang R, De Gruttola V. The use of permutation tests for the analysis of parallel and stepped-wedge cluster-randomized trials. Stat Med. 2017;36(18):2831–43.

95. Flórez AJ, Molenberghs G, Verbeke G, Kenward MG, Mamouris P, Vaes B. Fast two-stage estimator for clustered count data with overdispersion. J Stat Comput Simul. 2019;89(14):2678–93.

96. Borhan S, Kennedy C, Ioannidis G, Papaioannou A, Adachi J, Thabane L. An empirical comparison of methods for analyzing over-dispersed zero-inflated count data from stratified cluster randomized trials. Contemp Clin Trials Commun. 2020;17:100539–100539.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- ADDITIONALFILE1.docx
- ADDITIONALFILE2.docx
- ADDITIONALFILE3.docx
- ADDITIONALFILE4.docx