A Comparison of Methods of Estimation for Binary Outcomes in Paired Cluster Randomized Trials

Yuan Tian

A thesis

submitted in partial fulfillment of the

requirements for the degree of

Master of Science

University of Washington

2024

Committee:

James Hughes

Deborah Donnell

Program Authorized to Offer Degree:

Biostatistics

University of Washington

Abstract

A Comparison of Methods of Estimation for Binary Outcomes in Paired Cluster Randomized Trials

Yuan Tian

Chair of the Supervisory Committee:

James Hughes

Department of Biostatistics

The randomized controlled trial is often considered the “gold standard” for evaluating the effectiveness of treatments. However, when individuals are naturally aggregated into groups (e.g., families, clinics, schools) it can be challenging to allocate individuals within the same group to different treatments. To fix this, Cluster Randomized Trials (CRTs) have widely been applied in various fields. Additionally, by matching clusters in CRTs, between-cluster variation can be reduced, which increases the precision of estimation of the treatment effect. The objective of this thesis is to compare different methods in estimating the treatment effect for paired CRTs via a series of simulations, including Restricted Maximum Likelihood (REML), Maximum Likelihood (ML) and two-stage analysis. The simulation results indicate that REML and ML provide unbiased estimation overall. However, REML demonstrates better performance than ML in estimating standard error and in controlling the type I error. Additionally, the two-stage analysis is more robust in maintaining the type I error rate, suggesting that it may be a good choice for controlling the type I error for large cluster size.

1 Introduction

Randomized Controlled Trials (RCT) have been accepted as the “gold standard” for evaluation of the effectiveness of treatments in evidence-based medicine. In a typical RCT, the unit of randomization is the individual and the outcome measurements are collected on the individual level. However, certain interventions are designed to operate at the group level, for instance, educational programs targeting patient groups or initiatives aimed at altering social environments. It is either impossible to allocate individuals in the same group to different treatment arms, or attempting to do so will cause contamination. In other cases when investigators are interested in the population level treatment effect, such as achievement of herd immunity, it is more efficient or more convenient to perform randomization on a group level. Such designs are known as Cluster Randomized Trials (CRTs). [Hayes and Moulton, 2009]

The application of CRTs has significantly expanded across various fields such as public health initiatives, educational interventions, and community-based programs. This increase is driven by the unique advantage of CRTs in addressing group-level interventions and outcomes, where individual-level randomization is impractical or ethically questionable.

Designing and analyzing CRTs present distinct challenges, primarily due to the inherent clustering of data and the smaller number of clusters compared to individual participants. These challenges include achieving balance between treatment arms, managing intra-cluster correlation, and ensuring sufficient power for detecting treatment effects. Strategies such as matching and stratification for cluster selection, advanced statistical models for clustered data, and careful planning of sample sizes are essential to address these issues.

One of the major advantages of randomized controlled trials is that with large number of randomization units, randomization can form treatment groups that are well-balanced with respect to both known and unknown potential confounders. However, there is usually a small number of clusters to be randomized in CRTs. As a result, randomization may not achieve balance between treatment arms. Matching helps reduce differences in baseline characteristics between treatment arms by grouping similar clusters into pairs with respect to some characteristics we choose to match on. If the pairing is effective, i.e., the matching variables are strongly correlated with the outcome of interest and the clusters within each pair are similar on the matching variables, matching reduces between-cluster variation, which in turn increases the precision of estimation of treatment effects and increases the power of paired cluster randomized trials. [Hayes and Moulton, 2009]

If the data are assumed normally distributed and are analyzed with an identity link, we can fit the following linear mixed effects model for a paired analysis using Restricted Maximum Likelihood (REML) procedure, which, unlike Maximum Likelihood (ML) estimation, adjusts for the degrees of freedom lost by estimating the fixed effects. [Millar, 2011]

$$Y_{ijk} = \beta_0 + \beta_i + X_{ij}\beta^* + a_{ij} + e_{ijk}$$

$$a_{ij} \sim N(0, \tau^2)$$

$$e_{ijk} \sim N(0, \sigma^2)$$

where Y_{ijk} is the outcome of the kth individual ($k = 1,2, \dots, K$) in the jth cluster ($j = 0,1$) in the ith pair ($i = 1,2, \dots, I$), X_{ij} is the treatment indicator, β^* is the treatment effect, β_i is the fixed effect for pair i, a_{ij} is the cluster random effect and e_{ijk} is the random error. Compared with linear mixed effects model in a typical RCT, a fixed intercept for pairing effect and a random intercept for the cluster effect are added to the model.

Note that REML is necessary in the analysis of paired CRTs. ML will give biased estimates of the variance component(s) as additional parameters for matched pairs (which increase as more clusters are added) are included in the model. The regression analysis is equivalent to a paired t-test on the aggregated data on a cluster level (i.e., within-cluster means) if the cluster sizes are equal. It can be readily extended to include individual-level and cluster-level covariates.

However, in CRTs the outcome of interest is often binary or counts. Generalized Linear Mixed Models (GLMM) may be used to model correlated binary or count data. In the context of a paired CRT with binary outcomes, for example, a GLMM can be written as

$$\eta_{ij} = \beta_0 + \beta_i + X_{ij}\beta^* + a_{ij}$$

$$g(p_{ij}) = \eta_{ij}$$

$$Y_{ijk} \sim Bernoulli(p_{ij})$$

where Y_{ijk} is the outcome of the kth individual ($k = 1,2, \ldots, K$) in the jth cluster ($j = 0,1$) in the ith pair ($i = 1,2, \ldots, I$), $p_{ij} = P(Y_{ijk} = 1 | a_{ij})$, X_{ij} is the treatment indicator, β^* is the treatment effect, β_i is the fixed effect for pair i, a_{ij} is the cluster random effect and $g(.)$ is the logit link function.

Unfortunately, the extension of REML in GLMM is not straightforward. Unlike REML estimator in the Gaussian case, there is no well-defined REML procedure for GLMM. The computation of REML estimate typically requires analytically intractable integrals. Many methods to avoid the integration have been suggested including various Monte Carlo methods, Gibbs sampling methods and Monte Carlo EM methods, but the performance has not been satisfactory. [Lee et al., 2007]. Recently, however, Lee and Nelder proposed an h-likelihood method as a REML procedure for GLMM [Lee and Nelder, 1996]. It avoids calculating a complex integration which is necessary when using marginal likelihood. This procedure has the potential to extend the mixed model framework to the context of matched pair cluster randomized trials with binary outcomes.

As an alternative, Bennett et al. introduced a two-stage procedure, in which, in the first stage, parameters for pairing and all other covariates are regressed against the outcome of interest, regardless of the treatment arm, to obtain the expected number of events in each cluster; then, in the second stage, a paired t-test is performed on the log ratio of the observed to the expected (based on the first stage model) number of events. [Bennett et al., 2002] [Hayes and Moulton, 2009] Logistic regression and Poisson regression can be utilized in the first stage for binary and count data, respectively.

We will investigate the use of the GLMM framework for paired CRTs with binary endpoints using h-likelihood method described by Lee et al and compare it to Bennett et al's two-stage procedure via simulation. In section 2 we will briefly summarize the h-likelihood theory, the extension of REML estimation in GLMM, and the two-stage procedure. In section 3 we will describe the simulation settings. In section 4 we will show the simulation results. Section 5 will be a discussion of the simulation results and section 6 will provide conclusions.

2 Methods

Suppose that the binary outcome $\mathrm{Y} = [Y_{000}, Y_{001}, \ldots, Y_{I1K}\ ]^T$, where Y_{ijk} is the observation for the *k*th individual ($k = 0,1, \ldots, K$) in *j*th cluster ($j = 0,1$) in the *i*th matched pair ($i = 0,1, \ldots, I$), follows the Bernoulli distribution with a linear predictor, conditional on the random effects ν

$$\eta = \log\{\mathrm{p}/(1-\mathrm{p})\} = \mathrm{X}\beta + \mathrm{Z}\nu$$

$$Y_{ijk} \sim Bernoulli\left(p_{ijk}\right)$$

where $p = P(\mathrm{Y} = 1|\nu)$, X is the $n \times p$ design matrix for the fixed effects β and Z is the $n \times q$ design matrix for the random effects ν. This is a GLMM with binomial responses and a logit link function.

Let $f_\theta(\nu)$ denote the density of ν where θ are the parameters for the random effects. Let $f_{\beta,\theta}(y,\nu)$ and $f_\beta(y|\nu)$ be the joint density of the data and the random effects and the conditional density of *y* given ν, respectively. Then we can write the marginal likelihood of β as

$$L(\beta) = \int f_{\beta,\theta}(y,\nu)\, d\nu$$

$$f_{\beta,\theta}(y,\nu) = f_\beta(y|\nu) f_\theta(\nu)$$

Laplace Approximation[1]

[Shun and McCullagh, 1995] [Kristensen et al., 2015]

Let $g_\theta(\beta,\nu)$ denote the logarithm of $f_{\beta,\theta}(y,\nu)$. The data is considered a known constant and thus is not made explicit in this part for simplicity of notation. Then the marginal likelihood, where the random effects are integrated out, is

$$L(\beta) = \int exp\left(g_\theta(\beta,\nu)\right) d\nu.$$

Denote $\tilde{\nu}(\beta)$ as the maximizer of $g_\theta(\beta,\nu)$ w.r.t. ν,

$$\tilde{\nu}_\theta(\beta) = argmax\ g_\theta(\beta,\nu).$$

Denote $H(\beta)$ as the Hessian of $g_\theta(\beta,\nu)$ w.r.t. ν and evaluated at $\tilde{\nu}(\beta)$,

$$H(\beta) = g_{\theta}{}''_{\nu\nu}(\beta,\nu)|_{\tilde{\nu}(\beta)}.$$

The Laplace approximation of the marginal likelihood is

$$L^*(\beta) = exp\left(g_\theta\left(\beta,\tilde{\nu}(\beta)\right)\right) det(-H(\beta)/2\pi)^{-1/2}.$$

The maximum likelihood estimator $\tilde{\beta}$ maximizes the logarithm of the Laplace approximation

$$\log\ L^*(\beta) = g_\theta\left(\beta,\tilde{\nu}(\beta)\right) - 1/2\ \log\ (\det(-H(\beta)/2\pi)).$$

H-likelihood theory[2]

[Noh & Lee, 2007] [Lee and Nelder, 1996]

Lee and Nelder proposed using an h-likelihood in the form

$$h = logf_\beta(y|\nu) + logf_\theta(\nu),$$

which is the logarithm of the joint density of the data and the random effects $f_{\beta,\theta}(y,\nu)$.

Then the marginal log likelihood is

$$m = \log L(\beta) = log \int exp(h)\, dv.$$

The restricted log-likelihood is defined as

$$r = log\, f_{\theta}(y|\tilde{\beta}),$$

where $\tilde{\beta}$ is the maximum likelihood estimator that maximizes m.

Let l be a likelihood with parameter α. Lee and Nelder considered the function class

$$p_{\alpha}(l) = \left[l - \frac{1}{2} log(det\, D(l, \alpha)/2\pi)\right]|_{\alpha=\tilde{\alpha}},$$

where $D(l, \alpha)$ is the negative Hessian of l w.r.t. α and $\tilde{\alpha}$ maximizes l. Then $p_{\beta}(m)$ is the first order approximation of r, and $p_{v}(h) = \log L^{*}(\beta)$ is the first order Laplace approximation of m.

In principle, we should use h, m and $p_{\beta}(m)$ for inferences concerning v, β and θ respectively. However, we can now apply the $p_{v}(h)$ and $p_{\beta,v}(h)$ as approximates for m and $p_{\beta}(m)$ in the estimation of β and θ, respectively. This approach assumes that $p_{v}(h)$ approximates m and $p_{\beta,v}(h)$ approximates $p_{\beta}(m)$.

Note that in $p_{\beta,v}(h)$ both the fixed effects and the random effects were eliminated simultaneously from h. Thus we obtain the REML estimators, as opposed to the ML estimators based on ordinary Laplace approximation.

Two-stage estimation[3]

[Bennett et al., 2002] [Hayes and Moulton, 2009]

As an alternative to GLMMs, estimation in paired cluster randomized trials can be achieved with a two-stage procedure. For binary outcomes, the first stage is to fit a logistic regression model to the data, including terms for matched pairs and all the covariates of interest except the treatment effect.

$$log\left(p_{ijk}/(1-p_{ijk})\right) = \alpha_0 + \alpha_i + \sum_l \gamma_l z_{ijkl}$$

where p_{ijk} is the probability of event for the *k*th individual in the *j*th cluster ($j = 0, 1$ for the control and treated cluster) in the *i*th matched pair, α_0 is the grand mean, α_i is the coefficient for the *i*th matched pair, z_{ijkl} is the *l*th covariate for each individual, and γ_l is the coefficient for the *l*th covariate.

The expected number of events in each cluster if there was no treatment effect can be computed from the fitted model as

$$e_{ij} = \sum_k \hat{p}_{ijk} = \sum_k \frac{exp(\widehat{\alpha_i} + \sum_l \hat{\gamma}_l z_{ijkl})}{1 + exp(\widehat{\alpha_i} + \sum_l \hat{\gamma}_l z_{ijkl})}$$

The adjusted risk ratio for each pair is

$$\widehat{\theta'}_i = \frac{d_{i1}/e_{i1}}{d_{i0}/e_{i0}}$$

where d_{ij} is the observed number of events in in the *j*th cluster in the *i*th pair.

Note that if no events are observed, then the $\widehat{\theta'}_i$ will be infinity and taking logarithm will be problematic.

Since the adjusted risk ratio is likely skewed, we take logarithms of the ratios $\widehat{\theta'}_i$. We can then compare the observed number of events and the expected from the two clusters in each pair using paired t-test with a null hypothesis that mean of $log\left(\widehat{\theta'}_i\right)$ is 0, or mean difference between $log(d_{i1}/e_{i1})$ and $log(d_{i0}/e_{i0})$ is 0.

The null hypothesis is $\theta_i' = 1$, or equivalently $log\ \theta_i' = 0$.

3 Simulation

We generate binary data both with and without individual level covariates.

For the simulations without individual level covariates, the binary outcomes y_{ijk} are generated using the logistic model

$$logit\ p_{ijk} = \beta_0 + \beta_1 X_{ij} + \gamma_i + \alpha_{ij} \quad (1)$$

$$\alpha_{ij} \sim N(0, \tau_1^2),$$

$$y_{ijk} \sim Bernoulli\left(p_{ijk}\right),$$

where y_{ijk} is the outcome for the *k*th individual ($k = 1, \ldots, K$) in *j*th cluster ($j = 0{,}1$) in the *i*th matched pair ($j = 1, \ldots, J$), X_{ij} is the treatment indicator, β_1 is the treatment effect, γ_i is the fixed effect for pair *i* (we use $\gamma_i \sim N(0, \tau_2^2)$ for data simulation) and α_{ij} is the random effect for cluster *j* in pair *i*. τ_1^2 is the cluster dispersion parameter.

For simulations with individual level covariates, the binary outcomes y_{ijk} are generated using the logistic model with two covariates:

$$logit\ p_{ijk} = \beta_0 + \beta_1 X_{ij} + \beta_2 X_{ijk1} + \beta_3 X_{ijk2} + \gamma_i + \alpha_{ij} \qquad (2)$$

$$\alpha_{ij} \sim N(0, \tau_1^2),$$

$$\mathrm{X_{ijk1}} \sim \mathrm{N}(\mu, \sigma^2),\ X_{ijk2} \sim Bernoulli(p)$$

$$y_{ijk} \sim Bernoulli\left(p_{ijk}\right),$$

where β_2 and β_3 are the fixed effects for the individual-level covariates X_{ijk1} and X_{ijk2}. X_{ijk1} is generated as a normal covariate (e.g. age) with mean μ and variance σ^2 and X_{ijk2} is binary (e.g. sex) with parameter p, for all individuals in all the clusters.

We fit the GLMMs using `glmmTMB` for both the analyses with and without individual level covariates and compare the REML estimators (REML=TRUE) and ML estimators (REML=FALSE). We also perform the two-stage analysis for individual-level covariate adjustment (this approach is even available when there are no individual-level covariates).

Table 1 and Table 2 present the simulation scenarios without and with individual level covariates for Model (1) and Model (2) respectively. Each scenario depicts different treatment effect, baseline prevalence or intraclass correlation. We will conduct ML, REML and two-stage analysis for each scenario.

Table 1: Simulation scenarios for analyses without covariates

No.	β_0	β_1	τ_1^2	τ_2^2
1	-2	0	0.01	0.01
2	-1.5	0	0.01	0.01
3	-1	0	0.01	0.01
4	-2	0.1	0.01	0.01
5	-1.5	0.1	0.01	0.01
6	-1	0.1	0.01	0.01
7	-2	0	0.05	0.01
8	-1.5	0	0.05	0.01
9	-1	0	0.05	0.01
10	-2	0.1	0.05	0.01
11	-1.5	0.1	0.05	0.01
12	-1	0.1	0.05	0.01

Table 2: Simulation scenarios for analyses with covariates

No.	β_0	β_1	β_2	β_3	τ_1^2	τ_2^2	μ	σ^2	p
13	-2	0	0.05	0.05	0.05	0.01	30	25	0.5
14	-1.5	0	0.05	0.05	0.05	0.01	30	25	0.5
15	-1	0	0.05	0.05	0.05	0.01	30	25	0.5

4 Results

In Table 3 through 17 we conduct the simulations under 15 scenarios defined in Table 1 and Table 2 respectively. Each numbered table has sub-tables corresponding to different estimation methods, ML, REML or two-stage methods. In each table we simulate datasets of 6, 10, 20 or 40 pairs (2 clusters in each pair assigned to treatment or control group) and 20, 50 or 200 individuals in each cluster. Each row of the tables summarizes the results of 1000 runs.

Table 3 through 14 show the results of the simulations without covariates. The results of REML estimator and ML estimator are given in the two sub-tables in each table together with the hypothesis testing results. We also present the average model-based standard error of $\hat{\beta}_1$ as well as the observed standard error of the estimator se($\hat{\beta}_1$) across simulations. In Table 3 through 8, we assume the dispersion parameter of the random effect $\tau_1^2 = 0.01$. In Table 9 through 14, we assume $\tau_1^2 = 0.05$. Table 3 through 5 and 9 through 11 are results under the null hypothesis $\beta_1 = 0$. Simulation results when the alternative hypothesis is true ($\beta_1 = 0.1$) are presented in Table 6 through 8 and 12 through 14. We also perform two-stage analysis under the above scenarios.

In general, both REML and ML give unbiased estimates of the treatment effect β_1 except for some slight bias under small number of pairs ($J = 6$) and small cluster size ($K = 20$). We evaluate the bias of se($\hat{\beta}_1$) by comparing the average model-based se($\hat{\beta}_1$) with the observed se($\hat{\beta}_1$). The bias of REML for se($\hat{\beta}_1$) is regularly positive, implying that the REML slightly overestimates se($\hat{\beta}_1$) in most cases. The exception is that, for large cluster size (K=200), the REML underestimates the se($\hat{\beta}_1$). In contrast, ML slightly underestimates se($\hat{\beta}_1$), indicating the bias of ML is negative, except for a few cases with a small number of pair (J=10) and cluster size

(K=20). The estimation of se($\hat{\beta}_1$) by using ML does not improve across different scenarios as the number of clusters and/or cluster size increases. The bias of se($\hat{\beta}_1$) approaches 0 under REML estimation as the size of cluster increases. The observed se($\hat{\beta}_1$) gets smaller with larger number of pairs and larger cluster size, as expected. The magnitudes of over-estimation and under-estimation are similar in REML and ML. Performances of both REML and ML improve as the baseline prevalence increases and are similar across different levels of dispersion.

REML tends to overestimate the dispersion parameter τ_1^2 across most scenarios, with the degree of overestimation being more pronounced in simulations with a smaller number of pairs (J=6) and smaller cluster sizes (K=20). Conversely, ML estimation generally results in an underestimation of τ_1^2, with this bias being relatively constant regardless of the number of pairs or cluster size. Notably, the bias in $\hat{\tau}_1^2$ estimation for REML diminishes as the number of pairs and the size of clusters increase, indicating a trend towards unbiased estimation in larger study designs.

Surprisingly, biases of ML estimates of τ_1^2 are negative and quite small in most cases, in this sense, it is wise to use the relative biases of τ_1^2 (bias relative to the size of τ_1^2) as the measurement of bias. The sign of relative biases of τ_1^2 are negative, the largest relative bias is -1. Additionally, biases of ML estimates of τ_1^2 does not improve with changing number of clusters or cluster size.

Under the null hypothesis, the REML estimation method demonstrate relatively low type I error rates in scenarios with smaller cluster sizes. However, as cluster size increases to 200 individuals, the type I error rate exceeds the nominal 5% level, indicating an inflation that persists irrespective of the number of clusters. ML produces inflated type I error in most cases and greater inflation than REML. In contrast, the two-stage analysis generally preserves Type I error rates well. However, we encountered a significant number of failed estimations, particularly when dealing with smaller clusters. These failures can be attributed to the zero observed event dataset, which impedes the initial computational stage from yielding valid outcomes. In this study, there is a large proportion of zero events in the datasets randomly generated by simulating low prevalence and small cluster scenarios. These zero-event datasets were excluded from subsequently analysis, indicating that two-stage analysis may be problematic with small clusters and low prevalence. Under the alternative hypothesis, ML has greater power than REML and power increases with increasing number of clusters and cluster size. This is because, as we have previously discussed, the type I error rate inflation is larger for ML. Power also rises with larger baseline prevalence. To enhance the accuracy of power computations, we recommend adjusting the critical value based on the alpha-correction method. This approach entails determining the critical z or p value that aligns with a 5% rejection rate under the null hypothesis, thereby ensuring that our power estimates more accurately reflect controlled type I error rates. While power in the two-stage analysis also increases with larger number of cluster and larger cluster sizes, it is generally lower than REML and ML.

Table 15 through 17 present the simulation results of analyses with individual-level covariate adjustment. We compare the results of GLMMs and the two-stage procedure in the presence of covariates. In GLMM, REML and ML estimation follow similar pattern as in the unadjusted analysis.

Under the null hypothesis, the two-stage procedure gives generally unbiased risk ratio. Both REML and ML return generally unbiased odds ratios as in the unadjusted analysis.

Since the estimands are different in GLMM and the two-stage procedure, we focus our comparison between the two methods on type I error rate. REML inflates type I error rate in the scenarios with larger cluster size. In comparison, ML has the larger inflation of type I error compared to REML. The test performances are similar under different baseline prevalence.

5 Discussion

Cluster randomized trials are important designs for assessing the effects of interventions, especially when it is impossible to randomly assign individuals. In this design, matching is used to enhance comparability across groups and increase statistical power of the study by reducing heterogeneity across groups.

Matching is attractive in cluster randomized trials, but there is a trade-off to consider in the design stage. On the one hand, matching can increase power by reducing study population heterogeneity and can guarantee balance on selected prognostic factors. However, matching decreases the degrees of freedom and may lead to a loss of statistical power when matching is not effective (i.e., when the outcomes and the matching variables are poorly correlated) [Hayes and Moulton, 2009]. In the design phase, the researcher should determine the correlation between matching variables and outcome variables through pre-analysis (when possible) to evaluate the potential benefits of matching strategies. If associations are expected to be low, it may be more appropriate to consider other matching strategies, such as stratified randomization, which is a varied version of matching, or no matching.

In addition to the decision whether or not to match, matching also poses statistical difficulty in the analysis stage. While REML estimation and other analytical techniques are available for continuous outcomes, valid and efficient estimation method for matched pair binary or count data is less studied and less clear. For matched pair binary outcome data, REML estimation provides an effective way to deal with the increasing degrees of freedom lost due to matching, but its performance and applicability still need to be further explored.

In this simulation study, we compare one proposed REML estimation method with ML and the two-stage procedure in the matched pair cluster randomized trial setting when the outcome measurement is binary. We simulate data in different number of clusters and different cluster sizes. Since the Intra-class Correlation Coefficient (ICC) for binary outcome data depends on between-cluster variance and the overall proportion [Hayes and Moulton, 2009], we also consider different variances of the cluster random effect τ_1^2, or dispersion parameter and different baseline prevalence.

We demonstrate that in different simulation scenarios without individual level covariates both REML and ML give unbiased estimation of the treatment effect while REML has better performance in estimation of the standard error of the treatment effect as well as preservation of type I error. The precision of estimation increases when there are more clusters and the cluster sizes are larger. However, REML generally inflates type I error when the cluster sizes are large (n=200), but the inflation is smaller than ML. Our findings suggest that the reason why the REML method tends to overstate type I errors at large cluster sizes may be related to the way it estimates standard errors. In large clusters, inter-individual correlation may have a significant impact on the variability of outcomes, and REML estimates may not adequately adjust for this within-cluster correlation, leading to overly optimistic estimates of standard errors and thus increasing the probability of rejecting the true null hypothesis (that is, producing a type I error) in statistical tests.

The two-stage procedure is better at preservation of type I error than REML and ML, and also returns a generally unbiased estimation of the treatment effect. Compared with ML or REML methods, the two-stage procedure accounted for the effects of within-cluster correlation and individual level covariates more effectively, thus providing more robust standard error estimates when tested statistically. Hence, the two-stage procedure seems to be best at controlling type I error in the face of large clusters and complex data structures.

The precision of estimation is not comparable between the two-stage procedure and REML/ML since REML and ML estimate the odds ratio while the two-stage procedure estimates the risk ratio. Although both measures can reflect the size of the treatment effect, they differ in their statistical nature, potentially leading to differences in the estimates of standard errors.

The performance of preserving Type I error for two stage, REML and ML is robust under different baseline prevalence and dispersion parameter. All methods are able to accommodate different levels of inter-cluster variation and baseline prevalence, but for specific methods such as REML and ML, finer adjustments may be required to optimize performance under specific conditions.

Other matching models can be taken into account, such as stratified randomization. Researchers may consider stratification as a way to reduce heterogeneity across trial populations. It allows adjustment of multiple prognostic factors at the analysis stage and thus potentially improving statistical power without sacrificing as many degrees of freedom. . Our study has several limitations. We do not take into consideration of cluster size for two-stage procedure (that is, all clusters are given equal weight in the second stage t-test). Imbens points out that it is particularly helpful to stratify or pair based on cluster size if cluster size varies substantially [Imbens, 2011]. Matching may be more attractive and widely applied in circumstances of unequal cluster sizes to address several significant challenges. Firstly, it helps to neutralize the impact of varying baseline prevalence rates that are associated with differing cluster size. For instance, larger clusters might exhibit different initial levels of the outcome variable, which could skew the estimated treatment effect. Through matching, researchers can ensure that these baseline variations do not unfairly influence the analysis, leading to more accurate estimates of the treatment effect. Secondly, matching is crucial for addressing the variability in treatment effectiveness across different cluster sizes. The impact of an intervention might not be uniformly distributed across all cluster sizes; hence, matching allows researchers to detect and comprehend these differences in treatment effect. This understanding is crucial for interpreting the results and extrapolating them to other populations or settings. Given the varying sample size, future research should explore strategies for matching or stratification based on group size. In addition, including interaction terms provides a deeper understanding of heterogeneity in intervention effects, yet we do not discuss extensively. Lastly, we do not investigate the performance of the methods in more practical and logistical situations such as individual or cluster lost to follow-up. In cluster randomized trials with matched pairs, one cluster lost to follow-up may affect the validity of the entire pair. Multiple missing data handling methods may be applied to assess their impact on study results.

6 Conclusion

We compared different estimation methods (REML, ML and two-stage estimation) for estimating treatment effect for a binary response in paired random trials by using simulations. Through the comparison, this thesis reveals that the Restricted Maximum Likelihood (REML) outperformed Maximum Likelihood (ML) in terms of the estimation of standard error and type I error control for binary outcomes, particularly when dealing with larger cluster sizes. In this sense, REML seems a valid alternative to ML when estimating treatment effects for Cluster Randomized Trials (CRTs) with binary endpoints.

Additionally, we compared the use of the two-stage analysis method, REML, and ML (Maximum Likelihood) when the outcome measurement is binary. The comparison demonstrates the robustness of the two-stage analysis method in maintaining the desired type I error rate. We found that the type I error rate of the two-stage analysis does not inflate with a varying number of clusters. Additionally, the two-stage analysis method exhibits better preservation of the type I error rate when the cluster size is small, in comparison with REML and ML.

Further research could investigate the performance of these estimation methods under different conditions, such as the performance of two-stage analysis given the varying cluster sizes, and the performance of REML, ML and two-stage analysis deal with stratification or matching based on the group size.

In conclusion, we compare different methods for estimating GLMMs in CRTs in binary outcomes through simulations. Overall, given the binary outcome variable, the REML and ML give unbiased estimation of treatment effect, yet the REML outperforms the ML in estimating the standard error of treatment effect and in preserving the type I error rate. Furthermore, the two-stage analysis provides unbiased estimation and reveals better performance of preservation the type I error rate compared with REML and ML.

References

Bennett, J. B. , & Lehman, W. E. K. . (2002). Supervisor tolerance-responsiveness to substance abuse and workplace prevention training: use of a cognitive mapping tool. Health Educ Res(1), 27-42.

Bright Wood, S. N. . (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. Journal of the Royal Statistical Society, 73(1), 3-36.

Hayes, R. J. , & Moulton, L. H. . (2009). Cluster Randomised Trials.

Imbens GW. Experimental design for unit and cluster randomized trials. Prepared for the International Initiative for Impact Evaluation. 2011

Kristensen, K. , Nielsen, A. , Berg, C. W. , Skaug, H. , & Bell, B. . (2015). TMB: Automatic Differentiation and Laplace Approximation. 10.18637/jss.v070.i05.

Lee, Y. , & Nelder, J. A. . (1996). Hierarchical generalized linear models. Journal of the Royal Statistical Society: Series B (Methodological), 58(4).

Lee, S. Y. , & Song, X. Y. . (2007). A unified maximum likelihood approach for analyzing structural equation models with missing nonstandard data. Sociological Methods & Research, 35(3), 352-381.

Leyrat, Clemence, Morgan, Katy, E., & Leurent, et al. (2018). Cluster randomized trials with a small number of clusters: which analyses should be used? (vol 47, pg 321, 2018). International Journal of Epidemiology Official Journal of the International Epidemiological Association.

Millar, & Russell, B. . (2011). Maximum likelihood estimation and inference (with examples in R, SAS and ADMB) || tools of the trade. 286-298.

Noh, M. , & Lee, Y. . (2007). REML estimation for binary data in GLMMs. Journal of Multivariate Analysis, 98(5), 896-915.

Offorha Bright C., Walters Stephen J. & Jacques Richard M.. (2023). Analysing cluster randomised controlled trials using GLMM, GEE1, GEE2, and QIF: results from four case studies. BMC Medical Research Methodology (1), 293-293.

Shun, Z. , & Mccullagh, P. . (1995). Laplace approximation of high-dimensional integrals. Journal of the Royal Statistical Society. Series B: Methodological, 57(4), 749-760.

Tables

Table 3: results of 1000 simulations per row; $\beta_1 = 0$; baseline prevalence = 0.119 ($\beta_0 = -2$); τ_1^2 = 0.01; τ_2^2 = 0.01

Table 3.A REML analysis for scenario No. 1

Number of pairs	Size of cluster	bias($\hat{\beta}_1$)	Model-based se($\hat{\beta}_1$)	se($\hat{\beta}_1$)	bias(se($\hat{\beta}_1$))	bias($\hat{\tau}_1^2$)	Type I error
6	20	0.009	0.501	0.436	0.065	0.296	0.037[a]
6	50	0.002	0.294	0.273	0.021	0.058	0.045
6	200	0.003	0.146	0.142	0.004	0.008	0.059
10	20	0.001	0.366	0.346	0.020	0.164	0.045[b]
10	50	-0.006	0.219	0.203	0.016	0.038	0.036
10	200	0.001	0.111	0.106	0.005	0.005	0.047
20	20	-0.011	0.244	0.229	0.015	0.09	0.037[c]
20	50	0.006	0.151	0.146	0.005	0.027	0.046
20	200	-0.002	0.078	0.075	0.003	0.003	0.058
40	20	-0.014	0.166	0.159	0.007	0.052	0.031[d]
40	50	-0.005	0.105	0.101	0.004	0.017	0.03
40	200	-0.001	0.054	0.057	-0.003	0.002	0.064

Number of runs that did not converge: *a* = 32; *b* = 65; *c* = 137; *d* = 230

Runs that did not converge because one of the variance components was estimated zero have been excluded.

Table 3.B ML analysis for scenario No. 1

Number of pairs	Size of cluster	bias($\hat{\beta}_1$)	Model-based se($\hat{\beta}_1$)	se($\hat{\beta}_1$)	bias(se($\hat{\beta}_1$))	bias($\hat{\tau}_1^2$)	Type I error
6	20	0.018	0.414	0.42	-0.006	-0.006	0.043[a]
6	50	0.009	0.257	0.266	-0.009	-0.008	0.054
6	200	-0.001	0.127	0.137	-0.010	-0.009	0.076
10	20	-0.01	0.316	0.322	-0.006	-0.009	0.046[b]
10	50	0.003	0.197	0.207	-0.010	-0.01	0.057
10	200	0.005	0.098	0.108	-0.010	-0.009	0.07
20	20	0.016	0.016	0.225	-0.209	-0.01	0.055[c]
20	50	0.005	0.139	0.143	-0.004	-0.01	0.055
20	200	-0.002	0.069	0.072	-0.003	-0.01	0.061
40	20	0.009	0.157	0.172	-0.015	-0.01	0.076[d]
40	50	0.005	0.098	0.103	-0.005	-0.01	0.066
40	200	-0.001	0.049	0.053	-0.004	-0.01	0.071

Number of runs that did not converge: *a* = 9; *b* = 8; *c* = 12; *d* = 49

Table 3.C Two-stage analysis for scenario No. 1

Number of pairs	Size of cluster	Risk ratio	Type I error
6	20	1.045	0.034[a]
6	50	1.032	0.044[b]
6	200	1.008	0.042
10	20	1.021	0.055[c]
10	50	1.018	0.041[d]
10	200	1.006	0.057
20	20	1.029	0[e]
20	50	1.006	0.058[f]
20	200	1.003	0.058
40	50	0.999	0.047[g]
40	200	1.002	0.049
40	20	NA	NA[h]

Number of failed estimation due to zero observed event: *a* = 647; *b* = 29; *c* = 837; *d* = 48; *e* = 965; *f* = 83; *g* = 178; *h* = 1000

Cases failed due to zero observed events have been excluded.

Table 4: results of 1000 simulations per row; $\beta_1 = 0$; baseline prevalence = 0.182 ($\beta_0 = -1.5$); τ_1^2 = 0.01; τ_2^2 = 0.01

Table 4.A REML analysis for scenario No. 2

Number of pairs	Size of cluster	bias($\hat{\beta}_1$)	Model-based se($\hat{\beta}_1$)	se($\hat{\beta}_1$)	bias(se($\hat{\beta}_1$))	bias($\hat{\tau}_1^2$)	Type I error
6	20	-0.026	0.403	0.366	0.037	0.137	0.039[a]
6	50	-0.003	0.246	0.229	0.017	0.041	0.049
6	200	0.002	0.127	0.122	0.005	0.007	0.058
10	20	0	0.297	0.27	0.027	0.086	0.031[b]
10	50	0.003	0.183	0.167	0.016	0.024	0.033
10	200	0.002	0.096	0.094	0.002	0.004	0.058
20	20	-0.007	0.203	0.199	0.004	0.054	0.047[c]
20	50	0.001	0.127	0.128	-0.001	0.015	0.048
20	200	-0.001	0.067	0.066	0.001	0.002	0.05
40	20	0.001	0.14	0.132	0.008	0.035	0.039[d]
40	50	-0.001	0.088	0.084	0.004	0.011	0.042
40	200	0	0.047	0.049	-0.002	0.001	0.056

Number of runs that did not converge: *a* = 3; *b* = 4; *c* = 5; *d* = 17

Table 4.B ML analysis for scenario No. 2

Number of pairs	Size of cluster	bias($\hat{\beta}_1$)	Model-based se($\hat{\beta}_1$)	se($\hat{\beta}_1$)	bias(se($\hat{\beta}_1$))	bias($\hat{\tau}_1^2$)	Type I error
6	20	0.01	0.345	0.366	-0.021	-0.006	0.062
6	50	0.005	0.215	0.227	-0.012	-0.008	0.06
6	200	-0.004	0.107	0.116	-0.009	-0.009	0.077
10	20	0.006	0.265	0.274	-0.009	-0.009	0.05
10	50	-0.011	0.165	0.176	-0.011	-0.01	0.062
10	200	0	0.083	0.096	-0.013	-0.01	0.09
20	20	0.009	0.186	0.196	-0.010	-0.01	0.052[a]
20	50	0.003	0.117	0.122	-0.005	-0.01	0.056
20	200	0.002	0.058	0.067	-0.009	-0.01	0.092
40	20	-0.004	0.131	0.132	-0.001	-0.01	0.047[b]
40	50	-0.001	0.082	0.088	-0.006	-0.01	0.061
40	200	-0.001	0.041	0.047	-0.006	-0.01	0.096

Number of runs that did not converge: *a* = 1; *b* = 7

Table 4.C Two-stage analysis for scenario No. 2

Number of pairs	Size of cluster	Risk ratio	Type I error
6	20	1.055	0.057[a]
6	50	1.013	0.039
6	200	1.01	0.049
10	20	1.026	0.056[b]
10	50	1.006	0.05[c]
10	200	1.005	0.047
20	20	1.018	0.028[d]
20	50	1.009	0.055[e]
20	200	1.003	0.053
40	20	1	0.032[f]
40	50	1.004	0.049[g]
40	200	1	0.043

Number of failed estimation due to zero observed event: *a* = 217; *b* = 306; *c* = 3; *d* = 534; *e* = 4; *f* = 812; *g* = 9

Table 5: results of 1000 simulations per row; $\beta_1 = 0$; baseline prevalence = 0.269 ($\beta_0 = -1$); τ_1^2 = 0.01; τ_2^2 = 0.01

Table 5.A REML analysis for scenario No. 3

Number of pairs	Size of cluster	bias($\hat{\beta}_1$)	Model-based se($\hat{\beta}_1$)	se($\hat{\beta}_1$)	bias(se($\hat{\beta}_1$))	bias($\hat{\tau}_1^2$)	Type I error
6	20	-0.025	0.343	0.32	0.023	0.087	0.041
6	50	-0.003	0.214	0.197	0.017	0.028	0.048
6	200	0	0.112	0.113	-0.001	0.004	0.068
10	20	0.012	0.255	0.245	0.010	0.059	0.044
10	50	-0.002	0.161	0.147	0.014	0.019	0.036
10	200	-0.004	0.085	0.086	-0.001	0.002	0.06
20	20	0.001	0.174	0.167	0.007	0.034	0.048
20	50	0.001	0.111	0.101	0.010	0.012	0.036
20	200	-0.002	0.06	0.06	0.000	0.001	0.053
40	20	0.002	0.122	0.115	0.007	0.026	0.037
40	50	-0.005	0.077	0.073	0.004	0.006	0.042
40	200	0	0.042	0.042	0.000	0	0.059

Table 5.B ML analysis for scenario No. 3

Number of pairs	Size of cluster	bias($\hat{\beta}_1$)	Model-based se($\hat{\beta}_1$)	se($\hat{\beta}_1$)	bias(se($\hat{\beta}_1$))	bias($\hat{\tau}_1^2$)	Type I error
6	20	0.004	0.299	0.308	-0.009	-0.007	0.059
6	50	0.001	0.187	0.196	-0.009	-0.009	0.069
6	200	-0.01	0.094	0.108	-0.014	-0.009	0.087
10	20	0.002	0.23	0.228	0.002	-0.009	0.042
10	50	0.001	0.144	0.151	-0.007	-0.01	0.064
10	200	-0.006	0.072	0.083	-0.011	-0.009	0.078
20	20	0.002	0.162	0.164	-0.002	-0.01	0.044
20	50	0.002	0.102	0.106	-0.004	-0.01	0.049
20	200	-0.001	0.051	0.06	-0.009	-0.01	0.099
40	20	-0.003	0.114	0.117	-0.003	-0.01	0.053
40	50	-0.002	0.072	0.076	-0.004	-0.01	0.062
40	200	-0.001	0.036	0.044	-0.008	-0.01	0.118[a]

Number of runs that did not converge: *a* = 1

Table 5.C Two-stage analysis for scenario No. 3

Number of pairs	Size of cluster	Risk ratio	Type I error
6	20	1.032	0.048[a]
6	50	1.006	0.055
6	200	1.002	0.054
10	20	1.021	0.051[b]
10	50	1.01	0.043
10	200	1.002	0.056
20	20	1.011	0.056[c]
20	50	1.006	0.048
20	200	1.002	0.053
40	20	1.008	0.039[d]
40	50	1.002	0.037
40	200	1	0.047

Number of failed estimation due to zero observed event: *a* = 26; *b* = 39; *c* = 73; *d* = 171

Table 6: results of 1000 simulations per row; $\beta_1 = 0.1$; baseline prevalence = 0.130 ($\beta_0 = -2$); $\tau_1^2 = 0.01$; $\tau_2^2 = 0.01$

Table 6.A REML analysis for scenario No. 4

Number of pairs	Size of cluster	bias($\hat{\beta}_1$)	Model-based se($\hat{\beta}_1$)	se($\hat{\beta}_1$)	bias(se($\hat{\beta}_1$))	bias($\hat{\tau}_1^2$)	Power
6	20	0.02	0.497	0.436	0.061	0.269	0.029[a]
6	50	-0.004	0.285	0.246	0.039	0.053	0.043
6	200	0.001	0.147	0.136	0.011	0.011	0.111
10	20	0.016	0.361	0.321	0.040	0.164	0.053[b]
10	50	0.004	0.217	0.204	0.013	0.042	0.076
10	200	-0.003	0.111	0.107	0.004	0.007	0.142
20	20	0.008	0.24	0.228	0.012	0.09	0.062[c]
20	50	0	0.148	0.14	0.008	0.022	0.084
20	200	0.004	0.076	0.077	-0.001	0.003	0.283
40	20	-0.001	0.164	0.164	0.000	0.051	0.1[d]
40	50	0	0.103	0.095	0.008	0.016	0.14[e]
40	200	-0.002	0.053	0.053	0.000	0.001	0.461

Number of runs that did not converge: *a* = 27; *b* = 55; *c* = 109; *d* = 201; *e* = 1

Table 6.B ML analysis for scenario No. 4

Number of pairs	Size of cluster	bias($\hat{\beta}_1$)	Model-based se($\hat{\beta}_1$)	se($\hat{\beta}_1$)	bias(se($\hat{\beta}_1$))	bias($\hat{\tau}_1^2$)	Power
6	20	-0.021	0.407	0.414	-0.007	-0.006	0.046[a]
6	50	0.013	0.252	0.266	-0.014	-0.008	0.081
6	200	-0.001	0.125	0.135	-0.010	-0.009	0.154
10	20	0.004	0.31	0.326	-0.016	-0.008	0.062[b]
10	50	0.006	0.193	0.204	-0.011	-0.009	0.089
10	200	0	0.096	0.111	-0.015	-0.01	0.2
20	20	-0.007	0.218	0.226	-0.008	-0.01	0.078[c]
20	50	0.003	0.136	0.14	-0.004	-0.01	0.128
20	200	0.003	0.068	0.075	-0.007	-0.01	0.353
40	20	-0.004	0.153	0.155	-0.002	-0.01	0.097[d]
40	50	0.006	0.096	0.1	-0.004	-0.01	0.205
40	200	0.001	0.048	0.054	-0.006	-0.01	0.565

Number of runs that did not converge: *a* = 2; *b* = 6; *c* = 7; *d* = 50

Table 6.C Two-stage analysis for scenario No. 4

Number of pairs	Size of cluster	Risk ratio	Power
6	20	1.127	0.048[a]
6	50	1.134	0.053[b]
6	200	1.102	0.075
10	20	1.128	0.083[c]
10	50	1.112	0.072[d]
10	200	1.101	0.126
20	20	1.126	0.075[e]
20	50	1.11	0.104[f]
20	200	1.093	0.246
40	20	1.034	0[g]
40	50	1.107	0.148[h]
40	200	1.093	0.449

Number of failed estimation due to zero observed event: *a* = 586; *b* = 23; *c* = 794; *d* = 45; *e* = 960; *f* = 58; *g* = 999; *h* = 147

Table 7: results of 1000 simulations per row; $\beta_1 = 0.1$; baseline prevalence = 0.198 ($\beta_0 = -1.5$);

$\tau_1^2 = 0.01$; $\tau_2^2 = 0.01$

Table 7.A REML analysis for scenario No. 5

Number of pairs	Size of cluster	bias($\hat{\beta}_1$)	Model-based se($\hat{\beta}_1$)	se($\hat{\beta}_1$)	bias(se($\hat{\beta}_1$))	bias($\hat{\tau}_1^2$)	Power
6	20	-0.01	0.396	0.346	0.050	0.138	0.037[a]
6	50	-0.001	0.237	0.218	0.019	0.03	0.06
6	200	-0.002	0.127	0.117	0.010	0.008	0.114
10	20	-0.013	0.293	0.277	0.016	0.086	0.052[b]
10	50	0.005	0.182	0.172	0.010	0.026	0.089
10	200	0.004	0.096	0.095	0.001	0.004	0.205
20	20	0.007	0.2	0.19	0.010	0.054	0.084[c]
20	50	-0.001	0.124	0.117	0.007	0.014	0.112
20	200	0.006	0.066	0.066	0.000	0.002	0.356
40	20	0.005	0.138	0.13	0.008	0.037	0.109[d]
40	50	0.005	0.087	0.084	0.003	0.01	0.216
40	200	0.001	0.046	0.045	0.001	0.001	0.571

Number of runs that did not converge: *a* = 2; *b* = 3; *c* = 8; *d* = 13

Table 7.B ML analysis for scenario No. 5

Number of pairs	Size of cluster	bias($\hat{\beta}_1$)	Model-based se($\hat{\beta}_1$)	se($\hat{\beta}_1$)	bias(se($\hat{\beta}_1$))	bias($\hat{\tau}_1^2$)	Power
6	20	0.005	0.34	0.352	-0.012	-0.006	0.053
6	50	0.001	0.212	0.212	0.000	-0.009	0.065
6	200	0.003	0.106	0.119	-0.013	-0.009	0.166
10	20	0.002	0.26	0.285	-0.025	-0.009	0.094
10	50	-0.002	0.163	0.166	-0.003	-0.009	0.064
10	200	0.001	0.081	0.094	-0.013	-0.009	0.179
20	20	0.013	0.183	0.189	-0.006	-0.01	0.1
20	50	-0.001	0.115	0.125	-0.010	-0.01	0.148
20	200	0	0.057	0.064	-0.007	-0.01	0.317
40	20	0.004	0.129	0.137	-0.008	-0.01	0.142[a]
40	50	0.003	0.081	0.086	-0.005	-0.01	0.234
40	200	0.001	0.04	0.045	-0.005	-0.01	0.593

Number of runs that did not converge: *a* = 2

Table 7.C Two-stage analysis for scenario No. 5

Number of pairs	Size of cluster	Risk ratio	Power
6	20	1.121	0.053[a]
6	50	1.115	0.06[b]
6	200	1.091	0.102
10	20	1.1	0.047[c]
10	50	1.103	0.075
10	200	1.088	0.177
20	20	1.1	0.083[d]
20	50	1.094	0.114[e]
20	200	1.082	0.29
40	20	1.094	0.117[f]
40	50	1.092	0.188[g]
40	200	1.085	0.527

Number of failed estimation due to zero observed event: *a* = 195; *b* = 1; *c* = 281; *d* = 472; *e* = 3; *f* = 744; *g* = 2

Table 8: results of 1000 simulations per row; $\beta_1 = 0.1$; baseline prevalence = 0.289 ($\beta_0 = -1$); $\tau_1^2 = 0.01$; $\tau_2^2 = 0.01$

Table 8.A REML analysis for scenario No. 6

Number of pairs	Size of cluster	bias($\hat{\beta}_1$)	Model-based se($\hat{\beta}_1$)	se($\hat{\beta}_1$)	bias(se($\hat{\beta}_1$))	bias($\hat{\tau}_1^2$)	Power
6	20	-0.002	0.336	0.31	0.026	0.079	0.049
6	50	0.01	0.211	0.196	0.015	0.027	0.082
6	200	0.001	0.112	0.108	0.004	0.004	0.173
10	20	-0.003	0.252	0.235	0.017	0.055	0.059
10	50	-0.002	0.158	0.147	0.011	0.016	0.087
10	200	0.002	0.085	0.083	0.002	0.002	0.235
20	20	0	0.173	0.169	0.004	0.036	0.079
20	50	0.003	0.11	0.103	0.007	0.011	0.147
20	200	0.003	0.059	0.06	-0.001	0.001	0.406
40	20	0.009	0.12	0.119	0.001	0.024	0.151
40	50	0.001	0.077	0.077	0.000	0.007	0.257
40	200	0.002	0.042	0.044	-0.002	0.001	0.665

Table 8.B ML analysis for scenario No. 6

Number of pairs	Size of cluster	bias($\hat{\beta}_1$)	Model-based se($\hat{\beta}_1$)	se($\hat{\beta}_1$)	bias(se($\hat{\beta}_1$))	bias($\hat{\tau}_1^2$)	Power
6	20	0.018	0.294	0.299	-0.005	-0.008	0.07
6	50	0.002	0.184	0.199	-0.015	-0.009	0.106
6	200	0	0.093	0.108	-0.015	-0.009	0.242
10	20	0.006	0.227	0.225	0.002	-0.009	0.061
10	50	0.003	0.143	0.147	-0.004	-0.009	0.126
10	200	-0.001	0.072	0.084	-0.012	-0.009	0.29
20	20	0	0.16	0.166	-0.006	-0.01	0.105
20	50	0	0.1	0.104	-0.004	-0.01	0.166
20	200	0	0.05	0.061	-0.011	-0.01	0.524
40	20	-0.001	0.113	0.114	-0.001	-0.01	0.144
40	50	-0.004	0.071	0.074	-0.003	-0.01	0.292
40	200	-0.001	0.035	0.041	-0.006	-0.01	0.782[a]

Number of runs that did not converge: *a* = 2

Table 8.C Two-stage analysis for scenario No. 6

Number of pairs	Size of cluster	Risk ratio	Power
6	20	1.134	0.052[a]
6	50	1.088	0.078
6	200	1.081	0.114
10	20	1.085	0.053[b]
10	50	1.077	0.088
10	200	1.076	0.18
20	20	1.097	0.091[c]
20	50	1.078	0.164
20	200	1.074	0.331
40	20	1.091	0.14[d]
40	50	1.078	0.244
40	200	1.076	0.644

Number of failed estimation due to zero observed event: *a* = 22; *b* = 37; *c* = 80; *d* = 148

Table 9: results of 1000 simulations per row; $\beta_1 = 0$; baseline prevalence = 0.119 ($\beta_0 = -2$); τ_1^2 = 0.05; τ_2^2 = 0.01

Table 9.A REML analysis for scenario No. 7

Number of pairs	Size of cluster	bias($\hat{\beta}_1$)	Model-based se($\hat{\beta}_1$)	se($\hat{\beta}_1$)	bias(se($\hat{\beta}_1$))	bias($\hat{\tau}_1^2$)	Type I error
6	20	-0.006	0.517	0.467	0.050	0.288	0.04[a]
6	50	0	0.309	0.295	0.014	0.052	0.059
6	200	0.013	0.182	0.182	0.000	0.009	0.084
10	20	-0.016	0.369	0.357	0.012	0.146	0.056[b]
10	50	0.004	0.234	0.234	0.000	0.037	0.053
10	200	0.004	0.14	0.143	-0.003	0.004	0.074
20	20	-0.003	0.252	0.241	0.011	0.091	0.036[c]
20	50	-0.001	0.16	0.155	0.005	0.016	0.042
20	200	-0.007	0.1	0.101	-0.001	0.004	0.057
40	20	-0.004	0.172	0.168	0.004	0.047	0.051[d]
40	50	-0.001	0.112	0.111	0.001	0.01	0.049
40	200	-0.001	0.07	0.068	0.002	0.001	0.048

Number of runs that did not converge: *a* = 31; *b* = 77; *c* = 146; *d* = 254

Table 9.B ML analysis for scenario No. 7

Number of pairs	Size of cluster	bias($\hat{\beta}_1$)	Model-based se($\hat{\beta}_1$)	se($\hat{\beta}_1$)	bias(se($\hat{\beta}_1$))	bias($\hat{\tau}_1^2$)	Type I error
6	20	0.011	0.414	0.464	-0.050	-0.045	0.068[a]
6	50	0.009	0.258	0.291	-0.033	-0.045	0.082
6	200	0.002	0.134	0.187	-0.053	-0.043	0.159
10	20	-0.011	0.317	0.346	-0.029	-0.048	0.067[b]
10	50	0.005	0.197	0.225	-0.028	-0.048	0.074
10	200	0.002	0.103	0.14	-0.037	-0.044	0.152
20	20	-0.005	0.222	0.249	-0.027	-0.049	0.078[c]
20	50	0.003	0.139	0.156	-0.017	-0.049	0.082
20	200	-0.006	0.073	0.098	-0.025	-0.044	0.14
40	20	0.004	0.156	0.165	-0.009	-0.05	0.062[d]
40	50	0.002	0.098	0.109	-0.011	-0.05	0.075
40	200	0	0.051	0.072	-0.021	-0.046	0.174

Number of runs that did not converge: *a* = 1; *b* = 7; *c* = 7; *d* = 61

Table 9.C Two-stage analysis for scenario No. 7

Number of pairs	Size of cluster	Risk ratio	Type I error
6	20	1.06	0.046[a]
6	50	1.033	0.045[b]
6	200	1.021	0.049
10	20	1.015	0.041[c]
10	50	1.028	0.044[d]
10	200	1.01	0.035
20	20	1.028	0.057[e]
20	50	1.012	0.051[f]
20	200	1.004	0.054
40	20	1.114	0[g]
40	50	1.006	0.054[h]
40	200	1.003	0.053

Number of failed estimation due to zero observed event: *a* = 677; *b* = 50; *c* = 853; *d* = 73; *e* = 965; *f* = 143; *g* = 999; *h* = 261

Table 10: results of 1000 simulations per row; $\beta_1 = 0$; baseline prevalence = 0.182 ($\beta_0 = -1.5$); τ_1^2 = 0.05; τ_2^2 = 0.01

Table 10.A REML analysis for scenario No. 8

Number of pairs	Size of cluster	bias($\hat{\beta}_1$)	Model-based se($\hat{\beta}_1$)	se($\hat{\beta}_1$)	bias(se($\hat{\beta}_1$))	bias($\hat{\tau}_1^2$)	Type I error
6	20	-0.016	0.425	0.393	0.032	0.161	0.049[a]
6	50	-0.008	0.267	0.256	0.011	0.039	0.058
6	200	-0.01	0.163	0.17	-0.007	0.002	0.09
10	20	0.004	0.308	0.302	0.006	0.084	0.049[b]
10	50	0.005	0.201	0.196	0.005	0.021	0.064
10	200	-0.002	0.128	0.128	0.000	0.003	0.076
20	20	-0.006	0.211	0.202	0.009	0.05	0.038[c]
20	50	0.001	0.139	0.139	0.000	0.01	0.066
20	200	-0.004	0.091	0.089	0.002	0	0.059
40	20	0	0.146	0.147	-0.001	0.03	0.052[d]
40	50	0.002	0.097	0.097	0.000	0.005	0.051
40	200	-0.002	0.065	0.064	0.001	0.001	0.056

Number of runs that did not converge: *a* = 2; *b* = 5; *c* = 17; *d* = 18

Table 10.B ML analysis for scenario No. 8

Number of pairs	Size of cluster	bias($\hat{\beta}_1$)	Model-based se($\hat{\beta}_1$)	se($\hat{\beta}_1$)	bias(se($\hat{\beta}_1$))	bias($\hat{\tau}_1^2$)	Type I error
6	20	0.006	0.345	0.379	-0.034	-0.047	0.07
6	50	0.006	0.216	0.26	-0.044	-0.045	0.107
6	200	-0.003	0.118	0.168	-0.050	-0.041	0.178
10	20	0.006	0.265	0.296	-0.031	-0.047	0.075[a]
10	50	-0.008	0.167	0.191	-0.024	-0.047	0.084
10	200	-0.006	0.091	0.125	-0.034	-0.041	0.161
20	20	0.009	0.186	0.204	-0.018	-0.05	0.066
20	50	0.003	0.117	0.135	-0.018	-0.049	0.085
20	200	0.001	0.064	0.096	-0.032	-0.042	0.194
40	20	-0.002	0.131	0.142	-0.011	-0.05	0.077[b]
40	50	0	0.082	0.095	-0.013	-0.05	0.095
40	200	0	0.046	0.065	-0.019	-0.042	0.176

Number of runs that did not converge: *a* = 1; *b* = 2

Table 10.C Two-stage analysis for scenario No. 8

Number of pairs	Size of cluster	Risk ratio	Type I error
6	20	1.039	0.045[a]
6	50	1.022	0.049[b]
6	200	1.011	0.056
10	20	1.031	0.049[c]
10	50	1.013	0.057[d]
10	200	1.008	0.057
20	20	1.003	0.035[e]
20	50	1.011	0.054[f]
20	200	1	0.051
40	20	1.02	0.055[g]
40	50	1.008	0.049[h]
40	200	1.001	0.047

Number of failed estimation due to zero observed event: *a* = 236; *b* = 3; *c* = 369; *d* = 3; *e* = 605; *f* = 11; *g* = 854; *h* = 19

Table 11: results of 1000 simulations per row; $\beta_1 = 0$; baseline prevalence = 0.269 ($\beta_0 = -1$); $\tau_1^2 = 0.05$; $\tau_2^2 = 0.01$

Table 11.A REML analysis for scenario No. 9

Number of pairs	Size of cluster	bias($\hat{\beta}_1$)	Model-based se($\hat{\beta}_1$)	se($\hat{\beta}_1$)	bias(se($\hat{\beta}_1$))	bias($\hat{\tau}_1^2$)	Type I error
6	20	0.004	0.359	0.342	0.017	0.088	0.053
6	50	-0.009	0.237	0.223	0.014	0.024	0.06
6	200	0	0.156	0.159	-0.003	0.003	0.098
10	20	0.004	0.269	0.253	0.016	0.057	0.045[a]
10	50	-0.001	0.18	0.181	-0.001	0.014	0.066
10	200	-0.005	0.121	0.124	-0.003	0.002	0.069
20	20	0	0.186	0.19	-0.004	0.035	0.063
20	50	-0.005	0.125	0.123	0.002	0.006	0.052
20	200	0.003	0.086	0.089	-0.003	0	0.071
40	20	0.001	0.13	0.13	0.000	0.028	0.058[b]
40	50	0.001	0.088	0.089	-0.001	0.004	0.06
40	200	0	0.062	0.062	0.000	0.002	0.058

Number of runs that did not converge: *a* = 1; *b* = 2

Table 11.B ML analysis for scenario No. 9

Number of pairs	Size of cluster	bias($\hat{\beta}_1$)	Model-based se($\hat{\beta}_1$)	se($\hat{\beta}_1$)	bias(se($\hat{\beta}_1$))	bias($\hat{\tau}_1^2$)	Type I error
6	20	0	0.3	0.325	-0.025	-0.044	0.074
6	50	-0.025	0.19	0.222	-0.032	-0.044	0.095
6	200	-0.002	0.108	0.161	-0.053	-0.04	0.208
10	20	-0.012	0.231	0.258	-0.027	-0.047	0.08
10	50	0.009	0.146	0.176	-0.030	-0.046	0.118
10	200	-0.001	0.084	0.123	-0.039	-0.039	0.195
20	20	0.001	0.162	0.178	-0.016	-0.049	0.072
20	50	0.002	0.102	0.126	-0.024	-0.048	0.117
20	200	0	0.06	0.088	-0.028	-0.039	0.192
40	20	-0.001	0.114	0.127	-0.013	-0.05	0.088[a]
40	50	0.001	0.072	0.092	-0.020	-0.049	0.127
40	200	0.001	0.043	0.063	-0.020	-0.038	0.182

Number of runs that did not converge: *a* = 2

Table 11.C Two-stage analysis for scenario No. 9

Number of pairs	Size of cluster	Risk ratio	Type I error
6	20	1.021	0.053[a]
6	50	1.01	0.041
6	200	1.005	0.046
10	20	1.03	0.044[b]
10	50	1.012	0.04
10	200	1.006	0.043
20	20	1.008	0.043[c]
20	50	1.01	0.037
20	200	1.003	0.05
40	20	1.004	0.035[d]
40	50	1.002	0.05
40	200	1	0.046

Number of failed estimation due to zero observed event: *a* = 42; *b* = 61; *c* = 142; *d* = 257

Table 12: results of 1000 simulations per row; $\beta_1 = 0.1$; baseline prevalence = 0.130 ($\beta_0 = -2$); $\tau_1^2 = 0.05$; $\tau_2^2 = 0.01$

Table 12.A REML analysis for scenario No. 10

Number of pairs	Size of cluster	$\text{bias}(\hat{\beta}_1)$	Model-based $\text{se}(\hat{\beta}_1)$	$\text{se}(\hat{\beta}_1)$	$\text{bias}(\text{se}(\hat{\beta}_1))$	$\text{bias}(\hat{\tau}_1^2)$	Power
6	20	0.011	0.507	0.445	0.062	0.284	0.041[a]
6	50	0.001	0.308	0.293	0.015	0.06	0.06[b]
6	200	-0.005	0.179	0.179	0.000	0.007	0.115
10	20	0.008	0.363	0.336	0.027	0.137	0.055[c]
10	50	0.015	0.231	0.22	0.011	0.035	0.081[d]
10	200	0.002	0.139	0.143	-0.004	0.005	0.141
20	20	-0.016	0.244	0.235	0.009	0.075	0.057[e]
20	50	0.001	0.159	0.155	0.004	0.019	0.098
20	200	0.007	0.098	0.096	0.002	0.001	0.216
40	20	-0.003	0.169	0.158	0.011	0.046	0.078[f]
40	50	0.002	0.111	0.11	0.001	0.009	0.164
40	200	0.001	0.07	0.069	0.001	0.002	0.315

Number of runs that did not converge: *a* = 20; *b* = 1; *c* = 50; *d* = 1; *e* = 102; *f* = 210

Table 12.B ML analysis for scenario No. 10

Number of pairs	Size of cluster	bias($\hat{\beta}_1$)	Model-based se($\hat{\beta}_1$)	se($\hat{\beta}_1$)	bias(se($\hat{\beta}_1$))	bias($\hat{\tau}_1^2$)	Power
6	20	-0.006	0.406	0.458	-0.052	-0.045	0.075[a]
6	50	0.005	0.253	0.286	-0.033	-0.046	0.11
6	200	-0.001	0.134	0.18	-0.046	-0.042	0.21
10	20	0.009	0.31	0.327	-0.017	-0.048	0.071[b]
10	50	-0.002	0.194	0.21	-0.016	-0.047	0.095
10	200	-0.003	0.101	0.136	-0.035	-0.044	0.241
20	20	0.011	0.217	0.23	-0.013	-0.05	0.088[c]
20	50	-0.003	0.136	0.154	-0.018	-0.049	0.149
20	200	0	0.071	0.099	-0.028	-0.045	0.346
40	20	0.009	0.153	0.166	-0.013	-0.05	0.118[d]
40	50	0.001	0.096	0.11	-0.014	-0.05	0.215
40	200	-0.004	0.05	0.069	-0.019	-0.045	0.492

Number of runs that did not converge: *a* = 5; *b* = 8; *c* = 6; *d* = 52

Table 12.C Two-stage analysis for scenario No. 10

Number of pairs	Size of cluster	Risk ratio	Power
6	20	1.148	0.067[a]
6	50	1.144	0.06[b]
6	200	1.102	0.074
10	20	1.103	0.036[c]
10	50	1.131	0.062[d]
10	200	1.097	0.088
20	20	1.09	0.111[e]
20	50	1.114	0.094[f]
20	200	1.097	0.173
40	20	1.101	0[g]
40	50	1.106	0.155[h]
40	200	1.095	0.278

Number of failed estimation due to zero observed event: *a* = 644; *b* = 27; *c* = 805; *d* = 56; *e* = 955; *f* = 134; *g* = 998; *h* = 219

Table 13: results of 1000 simulations per row; $\beta_1 = 0.1$; baseline prevalence = 0.198 ($\beta_0 = -1.5$); $\tau_1^2 = 0.05$; $\tau_2^2 = 0.01$

Table 13.A REML analysis for scenario No. 11

Number of pairs	Size of cluster	bias($\hat{\beta}_1$)	Model-based se($\hat{\beta}_1$)	se($\hat{\beta}_1$)	bias(se($\hat{\beta}_1$))	bias($\hat{\tau}_1^2$)	Power
6	20	0.018	0.415	0.382	0.033	0.144	0.049[a]
6	50	0.001	0.26	0.249	0.011	0.029	0.07
6	200	0.005	0.164	0.166	-0.002	0.005	0.129
10	20	-0.005	0.306	0.295	0.011	0.086	0.062[b]
10	50	-0.005	0.198	0.193	0.005	0.019	0.08
10	200	0.004	0.125	0.127	-0.002	-0.001	0.161
20	20	0.01	0.208	0.201	0.007	0.047	0.082[c]
20	50	0.003	0.137	0.136	0.001	0.008	0.124
20	200	-0.001	0.09	0.091	-0.001	0.001	0.205
40	20	0.003	0.144	0.142	0.002	0.029	0.117[d]
40	50	0.001	0.096	0.093	0.003	0.005	0.179
40	200	-0.001	0.064	0.065	-0.001	0.001	0.332

Number of runs that did not converge: *a* = 5; *b* = 4; *c* = 4; *d* = 14

Table 13.B ML analysis for scenario No. 11

Number of pairs	Size of cluster	bias($\hat{\beta}_1$)	Model-based se($\hat{\beta}_1$)	se($\hat{\beta}_1$)	bias(se($\hat{\beta}_1$))	bias($\hat{\tau}_1^2$)	Power
6	20	0.003	0.34	0.374	-0.034	-0.044	0.064
6	50	0.006	0.213	0.247	-0.034	-0.046	0.113
6	200	0	0.117	0.173	-0.056	-0.041	0.272
10	20	0.001	0.261	0.286	-0.025	-0.048	0.087[a]
10	50	-0.007	0.164	0.192	-0.028	-0.047	0.129
10	200	-0.002	0.091	0.13	-0.039	-0.041	0.281
20	20	0.003	0.183	0.206	-0.023	-0.05	0.112
20	50	-0.007	0.115	0.136	-0.021	-0.049	0.169
20	200	0.001	0.064	0.089	-0.025	-0.041	0.406
40	20	0.006	0.129	0.138	-0.009	-0.05	0.148
40	50	-0.003	0.081	0.1	-0.019	-0.05	0.269
40	200	0.004	0.045	0.063	-0.018	-0.041	0.587

Number of runs that did not converge: *a* = 2

Table 13.C Two-stage analysis for scenario No. 11

Number of pairs	Size of cluster	Risk ratio	Power
6	20	1.127	0.046[a]
6	50	1.124	0.061[b]
6	200	1.101	0.072
10	20	1.113	0.054[c]
10	50	1.101	0.066[d]
10	200	1.086	0.098
20	20	1.117	0.087[e]
20	50	1.09	0.079[f]
20	200	1.09	0.2
40	20	1.095	0.107[g]
40	50	1.092	0.171[h]
40	200	1.089	0.337

Number of failed estimation due to zero observed event: *a* = 233; *b* = 1; *c* = 330; *d* = 3; *e* = 552; *f* = 7; *g* = 776; *h* = 10

Table 14: results of 1000 simulations per row; $\beta_1 = 0.1$; baseline prevalence = 0.289 ($\beta_0 = -1$); $\tau_1^2 = 0.05$; $\tau_2^2 = 0.01$

Table 14.A REML analysis for scenario No. 12

Number of pairs	Size of cluster	bias($\hat{\beta}_1$)	Model-based se($\hat{\beta}_1$)	se($\hat{\beta}_1$)	bias(se($\hat{\beta}_1$))	bias($\hat{\tau}_1^2$)	Power
6	20	0.01	0.35	0.336	0.014	0.074	0.058
6	50	0.016	0.233	0.229	0.004	0.02	0.107
6	200	-0.009	0.156	0.159	-0.003	0.005	0.126
10	20	0.008	0.266	0.26	0.006	0.053	0.079
10	50	-0.001	0.179	0.171	0.008	0.015	0.083
10	200	-0.005	0.12	0.124	-0.004	0	0.162
20	20	-0.003	0.182	0.173	0.009	0.028	0.073
20	50	-0.004	0.125	0.127	-0.002	0.007	0.133
20	200	0.004	0.086	0.088	-0.002	0	0.222
40	20	-0.001	0.128	0.127	0.001	0.02	0.13[a]
40	50	0.003	0.088	0.088	0.000	0.005	0.212
40	200	0.005	0.061	0.063	-0.002	0.001	0.404

Number of runs that did not converge: $a = 2$

Table 14.B ML analysis for scenario No. 12

Number of pairs	Size of cluster	bias($\hat{\beta}_1$)	Model-based se($\hat{\beta}_1$)	se($\hat{\beta}_1$)	bias(se($\hat{\beta}_1$))	bias($\hat{\tau}_1^2$)	Power
6	20	0.005	0.298	0.326	-0.028	-0.044	0.088
6	50	-0.007	0.188	0.224	-0.036	-0.045	0.131
6	200	-0.006	0.107	0.155	-0.048	-0.039	0.259
10	20	-0.003	0.228	0.258	-0.030	-0.047	0.105
10	50	0.003	0.144	0.172	-0.028	-0.047	0.159
10	200	0.006	0.084	0.118	-0.034	-0.039	0.329
20	20	-0.001	0.16	0.181	-0.021	-0.049	0.13
20	50	0	0.101	0.124	-0.023	-0.048	0.225
20	200	0.002	0.059	0.085	-0.026	-0.039	0.44
40	20	0.008	0.113	0.123	-0.010	-0.05	0.182[a]
40	50	0	0.071	0.088	-0.017	-0.049	0.35
40	200	-0.002	0.043	0.062	-0.019	-0.038	0.581

Number of runs that did not converge: *a* = 1

Table 14.C Two-stage analysis for scenario No. 12

Number of pairs	Size of cluster	Risk ratio	Power
6	20	1.116	0.06[a]
6	50	1.098	0.054
6	200	1.082	0.08
10	20	1.115	0.071[b]
10	50	1.089	0.07
10	200	1.08	0.115
20	20	1.092	0.091[c]
20	50	1.085	0.13
20	200	1.077	0.186
40	20	1.093	0.131[d]
40	50	1.083	0.213
40	200	1.075	0.348

Number of failed estimation due to zero observed event: *a* = 40; *b* = 58; *c* = 125; *d* = 200

Table 15: results of 1000 simulations per row; $\beta_1 = 0$; $\beta_2 = 0.05$; $\beta_3 = 0.05$; baseline prevalence = 0.119 ($\beta_0 = -2$); τ_1^2 = 0.05; τ_2^2 = 0.01; include 2 covariates

Table 15.A REML analysis for scenario No. 13

Number of pairs	Size of cluster	bias($\hat{\beta}_1$)	Model-based se($\hat{\beta}_1$)	se($\hat{\beta}_1$)	Bias (se($\hat{\beta}_1$))	bias($\hat{\tau}_1^2$)	Type I error
6	20	-0.01	0.329	0.307	0.022	0.062	0.046
6	50	0	0.22	0.209	0.011	0.015	0.051
6	200	0.002	0.15	0.154	-0.004	0.001	0.088
10	20	-0.015	0.25	0.242	0.008	0.047	0.052
10	50	-0.003	0.17	0.168	0.002	0.012	0.062
10	200	-0.002	0.118	0.118	0.000	0.002	0.085
20	20	0.009	0.173	0.163	0.010	0.031	0.046
20	50	-0.003	0.118	0.121	-0.003	0.005	0.057
20	200	-0.001	0.084	0.086	-0.002	0.001	0.066
40	20	-0.002	0.121	0.121	0.000	0.02	0.054
40	50	0	0.083	0.079	0.004	0.002	0.053
40	200	0.001	0.06	0.061	-0.001	0.001	0.066

Table 15.B ML analysis for scenario No. 13

Number of pairs	Size of cluster	bias($\hat{\beta}_1$)	Model-based se($\hat{\beta}_1$)	se($\hat{\beta}_1$)	bias(se($\hat{\beta}_1$))	bias($\hat{\tau}_1^2$)	Type I error
6	20	-0.016	0.279	0.322	-0.043	-0.044	0.094
6	50	-0.007	0.177	0.209	-0.032	-0.043	0.099
6	200	0.004	0.103	0.151	-0.048	-0.039	0.185
10	20	-0.001	0.213	0.236	-0.023	-0.048	0.07
10	50	-0.008	0.136	0.169	-0.033	-0.046	0.115
10	200	0.003	0.081	0.119	-0.038	-0.038	0.205
20	20	0	0.149	0.164	-0.015	-0.049	0.077
20	50	-0.001	0.095	0.121	-0.026	-0.048	0.13
20	200	0.003	0.058	0.085	-0.027	-0.038	0.197
40	20	0.003	0.105	0.119	-0.014	-0.05	0.087
40	50	0.004	0.066	0.086	-0.020	-0.049	0.13
40	200	0.002	0.041	0.059	-0.018	-0.037	0.178

Table 15.C Two-stage analysis for scenario No. 13

Number of pairs	Size of cluster	Risk ratio	Type I error
6	20	1.022	0.048[a]
6	50	1.014	0.05
6	200	1.009	0.07
10	20	1.019	0.052[b]
10	50	1	0.052
10	200	1.005	0.044
20	20	1	0.052[c]
20	50	1.004	0.053
20	200	1.002	0.056
40	20	1.004	0.047[d]
40	50	1	0.044
40	200	1.002	0.059

Number of failed estimation due to zero observed event: *a* = 6; *b* = 2; *c* = 11; *d* = 16

Table 16: results of 1000 simulations per row; $\beta_1 = 0$; $\beta_2 = 0.05$; $\beta_3 = 0.05$; baseline prevalence = 0.182 (β_0 =-1.5); τ_1^2 = 0.05; τ_2^2 = 0.01; include 2 covariates

Table 16.A REML analysis for scenario No. 14

Number of pairs	Size of cluster	bias($\hat{\beta}_1$)	Model-based se($\hat{\beta}_1$)	se($\hat{\beta}_1$)	bias(se($\hat{\beta}_1$))	bias($\hat{\tau}_1^2$)	Type I error
6	20	0	0.319	0.319	0.000	0.054	0.055
6	50	-0.005	0.217	0.22	-0.003	0.016	0.072
6	200	-0.001	0.148	0.153	-0.005	0.002	0.092
10	20	0.003	0.244	0.24	0.004	0.044	0.055
10	50	-0.002	0.164	0.17	-0.006	0.006	0.074
10	200	0	0.117	0.119	-0.002	0.002	0.075
20	20	-0.009	0.169	0.163	0.006	0.029	0.056
20	50	-0.005	0.115	0.12	-0.005	0.003	0.069
20	200	-0.001	0.084	0.084	0.000	0.002	0.059
40	20	0.001	0.117	0.118	-0.001	0.014	0.059
40	50	0	0.081	0.085	-0.004	0.001	0.066
40	200	-0.001	0.059	0.059	0.000	0.001	0.057

Table 16.B ML analysis for scenario No. 14

Number of pairs	Size of cluster	bias($\hat{\beta}_1$)	Model-based se($\hat{\beta}_1$)	se($\hat{\beta}_1$)	bias(se($\hat{\beta}_1$))	bias($\hat{\tau}_1^2$)	Type I error
6	20	0.002	0.272	0.301	-0.029	-0.043	0.073
6	50	-0.024	0.173	0.208	-0.035	-0.043	0.107
6	200	-0.001	0.102	0.156	-0.054	-0.038	0.228
10	20	0.001	0.208	0.226	-0.018	-0.047	0.061
10	50	-0.007	0.133	0.162	-0.029	-0.045	0.104
10	200	0.001	0.08	0.117	-0.037	-0.038	0.189
20	20	-0.002	0.146	0.17	-0.024	-0.049	0.087
20	50	0.001	0.092	0.116	-0.024	-0.048	0.116
20	200	0.003	0.057	0.087	-0.030	-0.037	0.209
40	20	0.003	0.103	0.113	-0.010	-0.05	0.078
40	50	0	0.065	0.082	-0.017	-0.049	0.111
40	200	0.001	0.041	0.059	-0.018	-0.036	0.192

Table 16.C Two-stage analysis for scenario No. 14

Number of pairs	Size of cluster	Risk ratio	Type I error
6	20	1.008	0.05
6	50	1.003	0.043
6	200	1.004	0.061
10	20	1.016	0.041
10	50	1.004	0.039
10	200	1.003	0.047
20	20	1.002	0.04[a]
20	50	1.003	0.047
20	200	1	0.045
40	20	1.002	0.05[b]
40	50	1.001	0.045
40	200	1	0.057

Number of failed estimation due to zero observed event: *a* = 2; *b* = 2

Table 17: results of 1000 simulations per row; $\beta_1 = 0$; $\beta_2 = 0.05$; $\beta_3 = 0.05$; baseline prevalence = 0.269 ($\beta_0 = -1$); τ_1^2 = 0.05; τ_2^2 = 0.01; include 2 covariates

Table 17.A REML analysis for scenario No. 15

Number of pairs	Size of cluster	bias($\hat{\beta}_1$)	Model-based se($\hat{\beta}_1$)	se($\hat{\beta}_1$)	bias(se($\hat{\beta}_1$))	bias($\hat{\tau}_1^2$)	Type I error
6	20	0.004	0.334	0.321	0.013	0.071	0.051
6	50	0.006	0.223	0.211	0.012	0.019	0.06
6	200	0.003	0.149	0.157	-0.008	0.001	0.103
10	20	-0.003	0.251	0.233	0.018	0.047	0.046
10	50	0.001	0.169	0.168	0.001	0.009	0.071
10	200	0	0.117	0.118	-0.001	0	0.07
20	20	-0.002	0.171	0.176	-0.005	0.021	0.066
20	50	0.003	0.118	0.115	0.003	0.003	0.055
20	200	0.006	0.084	0.083	0.001	0	0.06
40	20	-0.005	0.121	0.119	0.002	0.022	0.039
40	50	-0.001	0.084	0.085	-0.001	0.003	0.062
40	200	0.002	0.06	0.06	0.000	0.001	0.058

Table 17.B ML analysis for scenario No. 15

Number of pairs	Size of cluster	bias($\hat{\beta}_1$)	Model-based se($\hat{\beta}_1$)	se($\hat{\beta}_1$)	bias(se($\hat{\beta}_1$))	bias($\hat{\tau}_1^2$)	Type I error
6	20	-0.005	0.28	0.331	-0.051	-0.044	0.092
6	50	-0.009	0.178	0.218	-0.040	-0.044	0.112
6	200	0	0.103	0.152	-0.049	-0.039	0.205
10	20	0.001	0.215	0.235	-0.020	-0.047	0.067
10	50	0.002	0.137	0.166	-0.029	-0.045	0.11
10	200	-0.001	0.082	0.12	-0.038	-0.037	0.192
20	20	-0.009	0.15	0.169	-0.019	-0.049	0.08
20	50	-0.007	0.095	0.122	-0.027	-0.048	0.121
20	200	-0.003	0.058	0.089	-0.031	-0.037	0.202
40	20	0.01	0.106	0.118	-0.012	-0.05	0.084
40	50	-0.001	0.067	0.081	-0.014	-0.049	0.103
40	200	0.006	0.042	0.06	-0.018	-0.037	0.184

Table 17.C Two-stage analysis for scenario No. 15

Number of pairs	Size of cluster	Risk ratio	Type I error
6	20	1.003	0.053
6	50	1	0.043
6	200	1.004	0.049
10	20	1.004	0.039
10	50	0.998	0.05
10	200	1.002	0.054
20	20	0.998	0.051
20	50	1	0.053
20	200	1.001	0.039
40	20	1	0.053
40	50	1.001	0.049
40	200	1	0.061