Sullivan Thomas (Orcid ID: 0000-0002-6930-5406)

Title: Multiple imputation for handling missing outcome data in randomized trials involving a mixture of independent and paired data

Authors: Thomas R. Sullivan[1, 2], Lisa N. Yelland[1, 2], Margarita Moreno-Betancur[3, 4], Katherine J Lee[4, 3]

[1] SAHMRI Women & Kids, South Australian Health & Medical Research Institute, Adelaide, Australia

[2] School of Public Health, The University of Adelaide, Adelaide, Australia

[3] Department of Paediatrics, The University of Melbourne, Melbourne, Australia

[4] Clinical Epidemiology and Biostatistics Unit, Murdoch Childrens Research Institute, Melbourne, Australia

Corresponding author: Thomas R Sullivan; Email thomas.sullivan@sahmri.com; Phone +61 8 8128 4419; Address SAHMRI Women & Kids, South Australian Health & Medical Research Institute, Women's and Children's Hospital, 72 King William Road, South Australia 5006, Australia

Abstract: Randomized trials involving independent and paired observations occur in many areas of health research, for example in paediatrics, where studies can include infants from both single and twin births. Multiple imputation (MI) is often used to address missing outcome data in randomized trials, yet its performance in trials with independent and paired observations, where design effects can be less than or greater than one, remains to be explored. Using simulated data and through application to a trial dataset, we investigated the performance of different methods of MI for a continuous or binary outcome when followed by analysis using generalized estimating equations to account for clustering due to the pairs. We found that imputing data separately for independent and paired data, with paired data imputed in wide format, was the best performing MI method, producing unbiased point and standard error estimates for the treatment effect throughout. Ignoring clustering in the imputation model performed well in settings where the design effect due to the inclusion of paired data was close to one, but otherwise led to moderately biased variance estimates. Including a random cluster effect in the imputation model led to slightly biased point estimates for binary outcome data and variance estimates that were too small in some settings. Based on these results, we recommend researchers impute independent and paired data separately where feasible to do so. The exception is if the design effect due to the inclusion of paired data is close to one, where ignoring clustering may be appropriate.

This is the author manuscript accepted for publication and has undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1002/sim.9166

Keywords: multiple imputation, clinical trials, clustered data, missing outcome data

1. INTRODUCTION

Randomized trials involving a mixture of independent and paired data occur in many areas of health research. Examples include paediatrics, where studies can include infants from single and multiple births,[1] and ophthalmology, where one or both eyes may require treatment.[2] An important consideration in the design and analysis of such trials is the partial clustering that arises due to observations from the same pair being correlated but other observations being independent. At the design stage, the required sample size assuming all observations are independent can be multiplied by an appropriate design effect (DEFF) to account for the inclusion of paired data. Previous work on two-arm trials has shown that DEFFs in this setting are most sensitive to the degree of correlation between paired observations, the proportion of observations in the dataset belonging to a pair, and the method used to randomize the pairs.[3] Broadly, DEFFs tend to be close to one when members of a pair are randomized individually, greater than one when pairs are randomized as clusters (so that both members of a pair are always allocated to the same treatment group), and less than one when members of a pair are always randomized to opposite groups.[3] At the analysis stage, statistical methods for analysing clustered data, such as generalized estimating equations (GEEs) and mixed effects models, can be applied to account for the pairs when estimating treatment effects.[4]

Another consideration in the analysis of randomized trials involving independent and paired data is how to address missing data, a problem affecting most randomized trials.[5] Although missing data can occur in baseline covariates, the main problem in the trial setting is that of missing outcome data. Among the more principled approaches to handling missing outcome data, multiple imputation (MI) has emerged as a popular approach in contemporary randomized trials due to its considerable flexibility and availability in statistical software packages.[6] Under this approach, missing values are imputed multiple times using a statistical model fitted to the observed data. For "proper" imputation, uncertainty in both parameter values from the fitted statistical model and individual predicted values should be reflected in the imputed values. Following imputation, the multiple completed datasets are then analyzed separately, with results combined across datasets using Rubin's combination rules.[7] In its standard implementation, MI provides unbiased estimates when the imputation model is compatible with the analysis model and data are missing at random (MAR).[8] In a setting with a single incomplete outcome variable, MAR is the assumption that the probability of a missing value is independent of the value itself given observed data.[9] Such an assumption is often sensible for the primary analysis of a randomized trial.[10]

In cluster randomized trials or observational studies with clustered data, in which the analysis accounts for clustering, a compatible imputation model should also account for clustering. Importantly, failure to account for clustering in the imputation model can result in standard error estimates for regression coefficients that are biased downwards.[11,12] Two common strategies to account for clustering are to include either a fixed or a random effect term for cluster in the imputation model. Although easily implemented, the fixed effect approach has been criticized for inflating the standard error estimates of regression coefficients, particularly in settings with smaller cluster sizes.[12-14] As such, "multilevel MI" incorporating a random effect for cluster in the imputation model is often considered the preferred approach for addressing clustering during imputation.[11-17] Whether such findings are applicable to randomized trials involving a mixture of independent and paired data is unclear, however. Unlike cluster randomized trials and observational studies with clustered data, randomized trials with a mixture of independent and paired data can involve DEFFs less than one. Additionally, the inclusion of independent data and the maximum cluster size of two may present problems for the multilevel MI approach and facilitate the use of alternative imputation strategies, such as imputing separately by cluster size and imputing paired data in wide format. The performance of different MI approaches for addressing missing outcome data in trials with a combination of independent and paired data remains to be explored.

Our aim in this paper was to evaluate the performance of different methods of MI when addressing missing outcome data in randomized trials including both independent and paired data. Specifically, we considered the performance of MI for imputing continuous and binary outcomes, where completed datasets are analyzed using linear and logistic GEEs to estimate the effect of the randomized treatment while accounting for the clustering within pairs. The remainder of the paper is structured as follows. In the next section, we describe in more detail the different methods of MI applicable in this setting, drawing attention to potential limitations of each. This is followed by an outline of the simulation methods used to evaluate the MI approaches and a summary of simulation results. The MI approaches are then applied to a real clinical trial dataset. Finally, we conclude the article by discussing key findings and providing suggestions for practice.

2. METHODS

2.1 Setting

Let Y_{ij} denote an outcome of interest for the j^{th} member of the i^{th} cluster of size $n_i = 1$ or 2, and X_{ij} a corresponding baseline variable. Although only a single baseline variable is considered for

simplicity, the methods are easily extended to the case of multiple baseline variables. Suppose an analysis will be performed to estimate the effect of randomization to treatment T_{ij} (0 = control, 1 = intervention) based on the adjusted mean model $g(\mu_{ij}) = \beta_0 + \beta_1 T_{ij} + \beta_2 X_{ij}$, where $\mu_{ij} = E(Y_{ij}|T_{ij}, X_{ij})$ and g is an appropriate link function (identity for continuous outcomes, logit for binary outcomes). Note we focus on adjusted treatment effect estimates in this paper, since adjustment can lead to substantial increases in power for testing the effect of treatment[18,19] and is important when randomization involves stratification or minimization.[20] To account for clustering, suppose estimation will be performed using GEEs, with a robust sandwich estimator of the variance and an independence or exchangeable working correlation structure. In order for the robust sandwich estimator to produce reliable variance estimates, at least 40 clusters are recommended[21]; for studies with fewer clusters, adjustments to this estimator have been proposed.[22] There are several reasons for choosing GEEs for analysis in this setting. In the absence of missing data, GEEs produce consistent parameter estimates provided the mean model is correctly specified, even if the working correlation structure is misspecified.[23] Compared to mixed models, GEEs are less likely to encounter convergence problems (i.e. fail to produce parameter estimates) or produce biased regression parameter estimates when analyzing binary outcomes.[4,24] GEEs may also be preferred for their ability to produce population averaged treatment effects, which are often more relevant for health policy decision making than the cluster-specific treatment effects provided by mixed models.[25,26]

2.2 Multiple imputation approaches

Suppose that missing data occur in the outcome Y but not in the baseline variable X, as is often the case in randomized trials, and that MI will be employed to address missing data under a MAR assumption. Suppose also that an additional variable W associated with the outcome, and possibly also the probability of missingness in the outcome, will be included in the imputation model as an auxiliary variable to improve the efficiency of MI. Below we describe several approaches for implementing MI in this setting. Given the large number of small-sized clusters in trials containing a mixture of independent and paired data, we do not consider the approach of including a fixed effect for cluster in the imputation model.

2.2.1 MI assuming independence

Clearly, one could simply ignore the clustering due to paired observations and assume independence in the imputation model. With missing data confined to the outcome, imputations could be generated

by regressing observed values of Y_{ij} on X_{ij}, T_{ij} and W_{ij} and taking draws from the posterior predictive distribution of the model (i.e. using a univariate imputation model). In the more general setting of multiple incomplete variables (e.g. missing data also in the auxiliary variable or in other outcome variables) the missing values could instead be imputed using fully conditional specification (FCS),[27,28] in which a univariate imputation model is specified for each variable with missing data, or using a joint imputation model such as the multivariate normal.[29] By failing to account for the dependency between observations within pairs, assuming independence in the imputation model could lead to biased standard error estimates for the treatment effect.[11,12]

2.2.2 Multilevel MI

Clustering could also be accounted for in the imputation process by fitting a multilevel imputation model including a random effect for cluster. Assuming missing data are confined to continuous Y_{ij}, imputation could be based on a linear mixed effects model of the form $Y_{ij} = \delta_0 + \delta_1 T_{ij} + \delta_2 X_{ij} + \delta_3 W_{ij} + \alpha_i + e_{ij}$, where $a_i \sim N(0, \sigma_a^2)$ is a random cluster effect and $e_{ij} \sim N(0, \sigma_{ei}^2)$ the error term in the i^{th} cluster. A similarly specified logistic mixed effects model could be used in the case of binary Y_{ij}. Various methods have been proposed for drawing imputed values from a random effects model; for a review of available approaches and technical considerations, see Audigier et al.[17] To avoid over-fitting issues with small-sized clusters, in the current article we focus on an approach that assumes the error variance σ_{ei}^2 is constant across clusters. Using Bayesian mixed models with non-informative priors, the "FCS-GLM" approach can accommodate both continuous and binary variables that may be entirely or partially missing within clusters.[17] A potential drawback of this approach is the suboptimal performance of mixed effects models for analysing binary outcomes in trials with independent and paired data, with mixed models previously associated with biased estimation and problems with convergence, particularly when the proportion of paired observations is low or the ICC is high.[4,24] This could have flow-on effects to the quality of the imputed values.

2.2.3 MI by cluster size

Given the maximum cluster size of two, an alternative strategy is to fit separate imputation models to independent and paired observations. For independent observations, the imputation of missing values could proceed as described in Section 2.2.1. For paired observations, standard imputation procedures such as FCS or a joint imputation model could be applied once data have been rearranged into wide format; that is, with a single row for each pair and separate columns for Y_{i1} and Y_{i2} (and similarly for

X, T and W). This approach is consistent with recommendations for the application of MI in longitudinal studies with a fixed number of repeated measurements, where the wide format allows the imputation model to account for the correlation between repeated measurements.[30-32] Following imputation, the completed datasets for the paired data would be rearranged into long format and appended with completed datasets for the independent observations for subsequent analysis. As well as accounting for clustering, a benefit of imputing pairs in this way is that observed data from the opposing member of a pair is conditioned on during imputation. A potential disadvantage of splitting the data for imputation is a loss of efficiency in the imputation process due to increased uncertainty about parameters in each imputation model.[33] The approach also relies on having enough independent and paired observations to allow for the fitting of separate imputation models.

2.3 Simulation study

The performance of the three different MI approaches for addressing missing outcome data was evaluated in a simulation study, the details of which are described below (for a continuous outcome) or in Appendix S1 (for a binary outcome). The statistical code for implementing the simulation study is included in Appendix S1.

2.3.1 Data generation

For each simulation scenario, 2,000 datasets of 500 observations were generated. A sample size of 500 observations was chosen as this provides approximately 90% power under individual randomization to detect a standardized mean difference of 0.3 for the treatment effect. In a first step, clusters were assigned to be of size two with probability 0.2 or 0.4, corresponding to the expected proportion of mothers with twins in a preterm population and the mean proportion of participants contributing data for both eyes in ophthalmology trials.[1,2] Next, observations were allocated to treatment group T_{ij} using simple randomization, with members of a pair randomized individually (i.e. T_{i1} and T_{i2} assigned independently), cluster randomized to the same treatment group (i.e. $T_{i1} = T_{i2}$) or randomized to opposite groups (i.e. $T_{i2} = 1 - T_{i1}$). Outcome data were then generated according to the model

$$Y_{ij} = 0.3T_{ij} + 0.2X_i + a_i + e_{ij},$$

with $X_i \sim N(0,1)$ a cluster-level baseline covariate, $a_i \sim N(0, \sigma_a^2)$ a random cluster effect and $e_{ij} \sim N(0, \sigma_e^2)$ the error term. The values of (σ_a^2, σ_e^2) were fixed to give a total variance of one (i.e. $\sigma_a^2 + \sigma_e^2 = 1$) and an intra-cluster correlation coefficient (ICC) of 0.4 or 0.8 (with $\text{ICC} = \sigma_a^2/(\sigma_a^2 + \sigma_e^2)$). Although ICCs rarely exceed 0.1 in studies with larger clusters,[34] they are typically much higher for paired outcomes such as those measured in ophthalmology studies and paediatric trials involving twins.[2,35] Assuming GEEs with an independence working correlation structure will be used for analysis (see Section 2.3.2) and letting γ_p denote the proportion of observations belonging to a pair, the theoretical DEFF due to the inclusion of paired data is 1 for individual randomization, $1 + \text{ICC} \times \gamma_p$ for cluster randomization, and $1 - \text{ICC} \times \gamma_p$ for opposite randomization.[3] Our chosen values for the ICC and proportion of clusters of size two therefore resulted in theoretical DEFFs that ranged between 0.54 for opposite randomization and 1.46 for cluster randomization (see Table 1).[3]

<Insert Table 1 here>

Next, a continuous auxiliary variable was generated according to the model $W_{ij} = 0.95Y_{ij} + z_{ij}$, with $z_{ij} \sim N(0,1)$ so that the correlation between W_{ij} and Y_{ij} was approximately 0.70. Although the magnitude of this correlation was not expected to affect the relative performance of the MI approaches, we chose a strong correlation since auxiliary variables for incomplete outcomes tend to be of little benefit for correlations less than 0.5.[36,37]

After the generation of complete datasets, 40% of outcome values (Y_{ij}) were set to missing according to one of the following mechanisms:

1) Missing completely at random (MCAR): $P(R_{ij} = 0) = 0.4$, with R_{ij} denoting whether Y_{ij} is observed (0 = missing, 1 = observed).
2) MAR-individual: $\text{logit}\, P(R_{ij} = 0) = \gamma_0 + \log(3)W_{ij}$, with γ_0 chosen to give 40% missing data.
3) MAR-cluster: $\text{logit}\, P(R_i = 0) = \tau_0 + \log(3)\bar{W}_{i\cdot}$, with $R_i = 0$ indicating missing outcome data for all j in cluster i, $\bar{W}_{i\cdot}$ the average of W_{ij} in cluster i, and τ_0 chosen to give 40% missing data. Unlike the MCAR and MAR-Individual mechanisms which involve setting individual values to be missing, entire clusters are set to missing under the MAR-cluster mechanism.

Combined with the two proportions of pairs in the dataset, three methods of randomization for pairs and two ICCs for the outcome, the three missing data mechanisms led to a total of 36 simulation scenarios for investigation.

2.3.2 Analysis methods and performance measures

For each simulated dataset, missing outcome values were imputed using the MI assuming independence, multilevel MI and MI by cluster size strategies, with 40 imputations used throughout based on the extent of missing data.[38] In the absence of an interaction effect involving treatment group in the data generation model (not considered for brevity), imputation was carried out across all observations rather than performed separately by treatment group.

For MI assuming independence and the imputation of independent data under MI by cluster size, missing values were imputed from the linear regression of Y_{ij} on X_i, T_{ij} and W_{ij} using the mi impute regress command in Stata 16.0. For multilevel MI, missing values were imputed from a linear mixed effects model including a random cluster effect and fixed effects for X_i, T_{ij} and W_{ij}. We used the FCS-GLM approach (method "mice.impute.2l.glm.norm") in the R package micemd[39] to draw imputed values, as this method is recommended over alternative multilevel MI methods for studies involving small sized clusters.[17] For MI by cluster size, paired data were rearranged into wide format and missing values imputed using FCS with linear regression models for Y_{i1} and Y_{i2} (including T_{i1}, T_{i2}, X_{i1}, X_{i2}, W_{i1} and W_{i2}). In scenarios involving cluster or opposite randomization, note that T_{i2} was omitted from the imputation model given its collinearity with T_{i1}. Additionally, W_{i2} was excluded from the conditional imputation model for Y_{i1}, since $Y_{i1} \perp W_{i2}|Y_{i2}$ under our data generation mechanism (and similarly for W_{i1} when imputing Y_{i2}). Imputation was performed using the mi impute chained command in Stata 16.0, with the default of 10 cycles used for generating each imputed dataset. For comparison with the MI approaches, a complete case analysis (CCA) of observations with complete data on the outcome (i.e. where $R_{ij} = 1$) was also performed. With the inclusion of a strongly predictive and fully observed auxiliary variable in the imputation model, we expected CCA to produce treatment effects with wider confidence intervals than the MI approaches for all scenarios and potentially introduce bias when missing data were induced under the two MAR mechanisms.

Following imputation, the treatment effect in each completed dataset was estimated using linear regression of Y_{ij} on T_{ij}, with adjustment for X_i and using GEEs with an independence working correlation structure to account for clustering. An exchangeable structure could have been applied under our data generating mechanism, but we note this structure is not recommended for estimating average treatment effects when the effect of treatment depends on cluster size, which is often a possibility in trials involving independent and paired data.[40]

The performance of the different analysis approaches in estimating the true treatment effect of 0.3 was evaluated in terms of relative bias, the coverage of estimated 95% confidence intervals, and empirical and model-based standard errors. Since the main concern in this setting is producing appropriate variance estimates for the treatment effect, coverage and the ratio of empirical to model-based standard errors were the key measures of interest.

2.3.3 Sensitivity analyses

Additional analyses were undertaken to explore whether findings were sensitive to choices made during model fitting or to the simulation parameters considered. Specifically, we considered additional scenarios where X was specified to be an individual-level covariate with an ICC of 0.5 rather than a cluster level covariate, where completed datasets were analyzed using linear mixed models rather than GEEs, or where multivariate normal imputation was used instead of FCS for the MI by cluster size approach.

2.3.4 Binary outcome data

To investigate whether the relative performance of the MI approaches depends on variable type, we repeated the simulation study with binary W_{ij} and binary Y_{ij} generated under a logistic mixed effects model. Given the similarities with the continuous simulation study described above, we provide details of the binary simulation study in Appendix S1.

3. RESULTS

3.1 Continuous outcome data

There were no issues with non-convergence in the simulation study for continuous outcome data, with all MI approaches producing treatment effect estimates throughout. Figure 1 displays the relative bias of the treatment effect estimates under the different approaches for addressing missing data across the 36 simulation scenarios investigated. The three methods of MI produced treatment effect estimates with minimal bias, with absolute relative bias at most 3% for MI assuming independence and MI by cluster size, and 4% for multilevel MI. Conversely, CCA produced treatment effects that were moderately biased towards the null under the two MAR mechanisms.

<Insert Figure 1 here>

CCA was also the worst performing approach according to the average of the model-based standard errors for the estimated treatment effect (Figure 2). Again, this can be attributed to the presence of a correlated auxiliary variable to inform the prediction of missing values in MI. Conversely, multilevel MI produced the smallest average model-based standard errors across scenarios involving cluster or individual randomization. Compared to MI assuming independence, MI by cluster size produced larger average model-based standard errors when there was cluster randomization and smaller standard errors when pairs were randomized to opposite groups. In line with the theoretical DEFFs presented in Table 1, average model-based standard errors for all analysis approaches were highest under cluster randomization when the ICC was 0.8 and the proportion of pairs 0.4, and lowest for the same values of the ICC and proportion of pairs when pairs were randomized to opposite groups.

<Insert Figure 2 here>

Figure 3 displays the ratio of average model-based to empirical standard errors across the 36 simulation scenarios for the four missing data approaches, our key indicator of the unbiasedness of model-based standard errors for the treatment effect. As shown in the figure, average model-based standard errors for CCA and MI by cluster size remained close to empirical standard errors throughout. In contrast, MI assuming independence produced average model-based standard errors that were too small in scenarios involving cluster randomization and too large when pairs were randomized to opposite groups, with performance deficits most noticeable in scenarios where the proportion of pairs was 0.4 and the ICC 0.8 (i.e. where the DEFF was further away from 1). Like MI assuming independence, multilevel MI produced average model-based standard errors that were around 5% smaller than empirical standard errors in scenarios involving cluster randomization. There was also a tendency for multilevel MI to produce conservative standard errors in scenarios when pairs were randomized to opposite groups, but not to the same degree as MI assuming independence. Coverage results for the MI approaches produced similar findings (data not shown), with MI by cluster size producing confidence intervals with the best coverage performance across the 36 simulation scenarios (range 94.2% to 95.9%), followed by multilevel MI (range 92.9% to 96.0%) and then MI assuming independence (range 92.8% to 97.4%).

<Insert Figure 3 here>

In sensitivity analyses, a similar pattern of results was observed when multivariate normal imputation was used instead of FCS for imputing paired data in the MI by cluster size approach, or when X was specified to be an individual-level covariate with an ICC of 0.5 rather than a cluster-level covariate. Similar results were also observed when completed datasets were analyzed using linear mixed models rather than GEEs, albeit the degree to which average model-based standard errors were too conservative for MI assuming independence and multilevel MI under opposite randomization was more pronounced for analysis with linear mixed models (data not shown).

3.2 Binary outcome data

Results for binary outcome data were mostly consistent with those for continuous outcomes, with the main exception being the poorer performance of multilevel MI relative to the other imputation approaches. Of note, multilevel MI produced moderately biased treatment effect estimates (range in relative bias 1.4% to 13.0%), resulted in average standard errors up to 10% lower than empirical standard errors under cluster randomization, and was less precise than MI by cluster size for individual and opposite randomization. These performance deficits may have been a product of incompatibility between a mixed effects model for imputation and the use of GEEs for analysis, given these models estimate a different treatment effect in the case of logistic regression. Overall, MI by cluster size remained the optimal approach in the binary simulation study, while MI assuming independence again produced standard errors for the treatment effect that were too small under cluster randomization and too large under opposite randomization. Full results for the binary simulation study are presented in Appendix S1.

3.3. Applied example

The N−3 Fatty Acids for Improvement in Respiratory Outcomes (N3RO) trial was a blinded, randomized controlled trial conducted in 13 centers across Australia, New Zealand, and Singapore.[41] 1273 infants born less than 29 weeks gestation were randomized within 3 days of commencing enteral feeds to receive an enteral emulsion providing docosahexaenoic acid (DHA) at a dose of 60mg/kg/day or a control emulsion without DHA until 36 weeks of postmenstrual age (i.e. gestational age plus chronological age). Infants from a single or multiple birth were eligible to participate and infants from a multiple birth were randomized individually. The primary outcome was physiological bronchopulmonary dysplasia (BPD, a type of chronic lung disease) among surviving infants at 36 weeks postmenstrual age or at the time of discharge home, whichever occurred first. To illustrate the application of the MI approaches, here we consider analysis of the composite outcome of clinical

BPD or death at 36 weeks postmenstrual age. Only one infant had missing data on this outcome, a very low count for a randomized trial, offering the opportunity to induce missing data and investigate the ability of the imputation methods to recover the (almost) full-data estimate.

To simplify the N3RO trial data for illustration purposes, we excluded the single infant with missing outcome data along with 10 sets of triplets; we return to the issue of paediatric trials with triplet and higher order births in the discussion. Following exclusions, the example dataset consisted of 614 infants in the DHA group and 628 infants in the control group, including 155 sets of twins and 932 singletons. Overall, 345/614 (56.2%) and 327/628 (52.1%) infants experienced clinical BPD or death in the DHA and control groups, respectively. Using logistic regression with GEEs assuming an independence working correlation structure, and with adjustment for gestational age at birth (a key prognostic variable), the log odds ratio of clinical BPD or death due to DHA treatment (compared to control) was estimated to be 0.147 in the full dataset (standard error 0.130). The ICC for clinical BPD or death in the full dataset was estimated to be 0.40, as obtained from a logistic mixed effects model with a random cluster effect and fixed effects for treatment group and gestational age at birth (calculated on the logistic scale as $\sigma_a^2/[\sigma_a^2 + \pi^2/3]$, with σ_a^2 denoting the variance of the random cluster effect).[42]

To match the simulation study, clinical BPD or death was set to missing with probability 0.4. An MCAR mechanism was used, as the performance of the MI approaches in the simulation study did not appear sensitive to the specific MAR mechanism considered. Missing values were imputed using the MI assuming independence, multilevel MI and MI by cluster size strategies, with 40 imputations used throughout. Days of respiratory support between randomization and the assessment of clinical BPD or death was included as an auxiliary variable in imputation models given its strong association with the outcome in the full dataset (odds ratio per 10-day increase = 2.05; 95% confidence interval 1.91 to 2.21), with the measure available for 1173/1242 infants (94.4%). Imputation models were fit using a similar approach to the binary simulation study (see Appendix S1), with Y_{ij} denoting clinical BPD or death, X_i gestational age at birth, W_{ij} days of respiratory support, and j determined according to birth order (i.e. $j = 2$ for second born of a twin pair). The main departure from the simulation study is that W_{ij} is continuous and partially missing here. Consequently, FCS with 10 cycles was applied in all imputation approaches, with W_{ij} imputed using linear regression in the MI assuming independence and MI by cluster size strategies (mi impute chained command in Stata), and using a linear mixed effects model for multilevel MI ("mice.impute.2l.glm.norm" method in R).

In line with simulation results for individual randomization, we found little difference in average treatment effect estimates and model based standard errors between the MI assuming independence and MI by cluster size approaches (see Table 2). As expected with the inclusion of a correlated auxiliary variable in the imputation model, both MI approaches were more precise than CCA. Multilevel MI produced treatment effect estimates that were moderately biased towards the null (relative bias = -18.2%) with smaller model-based standard errors than the other MI approaches. The bias of multilevel MI was in the opposite direction to what was observed in the binary simulation study and may be related to the use of a continuous and partially missing auxiliary variable here (the direction of association between the outcome and the auxiliary variable was the same as in the simulation study).

<Insert Table 2 here>

4. DISCUSSION

In this article we investigated the performance of different methods of MI for addressing missing outcome data in randomized trials including both independent and paired observations, where completed datasets are analyzed using GEEs. MI by cluster size, where separate imputation models are fitted to independent and paired observations, was the best performing approach, producing unbiased point and standard error estimates for the treatment effect across all the scenarios considered. MI assuming independence performed well in scenarios involving individual randomization but led to moderately biased standard error estimates for the treatment effect under cluster and opposite randomization. Lastly, multilevel MI produced slightly biased treatment effect estimates in the context of binary outcome data and standard errors that were a little too small when randomization was at the cluster level. Based on these results, MI by cluster size can be recommended where numbers of independent and paired observations are large enough to permit such a strategy. Otherwise, MI assuming independence can be adopted in trials where the DEFF due to the inclusion of paired data is close to one, as occurs in trials using individual randomization or where the proportion of pairs or the ICC is low.

Although the simulation study identified shortcomings with the MI assuming independence and multilevel MI strategies, it should be emphasized that performance deficits were relatively small in magnitude. Unlike some other clustered data settings, the maximum DEFF for trials involving a mixture of independent and paired observations is two, which may limit the adverse effects of an inappropriate imputation strategy. Additionally, provided the substantive analysis model accounts for

clustering, the consequences of inadequately addressing clustering during imputation on standard error estimates will be proportional to the amount of outcome data requiring imputation. As such, the choice of imputation strategy may be less influential in settings with lower amounts of missing outcome data than the 40% considered in the simulation study.

Although it performed well, an important drawback of the MI by cluster size approach is that it relies on having enough independent and paired observations to allow for the fitting of separate imputation models in these two groups. In practice, imputation models may involve many more variables than the small number considered in our simulation study and analysis of the N3RO trial. For example, it is not uncommon for researchers to collect data on a long list of secondary outcomes, all of which may be included in a single imputation model. Trial datasets may also be smaller or involve fewer paired or independent observations. All these factors could contribute to over-fitting or convergence problems when applying the MI by cluster size approach. Another practical shortcoming is that imputation is often performed separately by treatment group in randomized trials to allow for potential effect modification.[33] Further stratification by cluster size may not be feasible, particularly in small trials, and the approach would only be feasible when paired observations are cluster randomized to the same treatment group (otherwise members of a pair can end up in separate strata for the imputation process). Finally, MI by cluster size is not easily applied in trials predominantly involving independent and paired data but where the maximum cluster size exceeds two, as with the N3RO trial where a small number of triplets participated. However, outside of paediatric trials, such a pattern of clustering may be uncommon.

CCA was noticeably less efficient than the MI approaches in the majority of simulation scenarios considered and in the example analysis of the N3RO trial. Except for scenarios involving an MCAR mechanism, CCA also led to biased treatment effect estimates in the continuous outcome simulation study. This pattern of results can be attributed to the inclusion of an auxiliary variable in each imputation model that was strongly associated with the outcome.[37] In the absence of auxiliary variables, CCA has been shown to produce similar treatment effect estimates to MI when the probability of missing outcome data depends on fully observed baseline characteristics and those characteristics are adjusted for in the analysis model.[43,44] In settings where useful auxiliary variables for the outcome are lacking and an assumption of covariate-dependent missingness is plausible, CCA can be applied in preference to MI.

Several methods of multilevel MI have been proposed in the literature, but no single approach appears suitable in all situations.[17,31,45] We focused on the FCS-GLM method for multilevel MI in this article

since it directly handles missing data in binary variables and has been recommended for datasets involving small-sized clusters.[17] A limitation of the approach is that it can introduce bias and a lack of variability in imputed values for binary variables where some but not all members of a cluster have missing data.[17] Biased point and standard error estimates with FCS-GLM were evident in our simulation study for binary outcome data and have been seen in other clustered data settings.[17,31] As well as the specific method of multilevel MI applied, the performance deficits in our binary simulation study may have been due to incompatibility between a logistic mixed effects model for imputation and subsequent analysis using logistic GEEs, or known limitations of logistic mixed effects models in settings with independent and paired data.[4,24]

A practical consideration when imputing missing data for trials with independent and paired observations is choosing which member of a pair should be assigned the first position in the cluster. Throughout this article we assumed that observations within a cluster were exchangeable and ordering did not matter, but such an assumption might not always be plausible. In practice, we suggest assigning cluster position in a systematic manner, for example assigning the first born of a twin pair or the left eye to the first position in a cluster. Any association between cluster position and outcome could then be accounted for during imputation through the addition of a fixed effect for cluster position in the case of MI assuming independence or multilevel MI. Such an association is implicitly taken into account when performing MI by cluster size and hence no modifications to this approach are necessary.

Though we anticipate findings will be broadly applicable to many trial settings, a clear limitation of the current article is that conclusions were based on a restricted set of simulation scenarios. For instance, the simulation study only considered missing data in a single outcome variable in settings where cluster size had no relationship with the outcome. Additionally, we only considered simple randomization, whereas randomly permuted blocks or minimization are often used in practice.[46] For individual randomization using randomly permuted blocks, the probability that members of a pair are allocated to opposite groups increases with a decreasing block size. With only paired data in the study and blocks of size two, for example, members of a pair will always be allocated to opposite groups (assuming members of a pair are randomized sequentially). As a result, MI assuming independence and multilevel MI could produce conservative standard error estimates for the treatment effect in trials involving individual randomization and small block sizes. Another limitation is that we did not evaluate alternatives to MI for handling missing outcome data, for example inverse probability weighting. Although inverse probability weighting tends to be less efficient than MI, particularly in the presence of auxiliary variables, it may have merit in settings where MI by cluster size is infeasible and MI assuming independence and multilevel MI are expected to produce biased standard error

estimates. Finally, we focused on the FCS-GLM method of multilevel MI as this has been recommended for datasets involving small-sized clusters,[17] however other methods of multilevel MI might also be evaluated in this setting.

In conclusion, provided it is feasible to fit separate imputation models to the independent and paired data, MI by cluster size is recommended for handling missing outcome data in randomized trials involving a combination of independent and paired data. Often this approach may not be practical, however, for example in small trials, in settings with few pairs or where imputation is implemented separately by randomized group. In such cases, MI assuming independence can be applied provided the DEFF due to the inclusion of paired data is expected to be close to one; generally this will be the case for trials using individual randomization or where the proportion of pairs or the ICC is low. Despite holding much promise, the FCS-GLM method of multilevel MI exhibits some performance deficits in trials involving a combination of independent and paired data, particularly in the case of binary outcome data, and cannot be recommended without further development.

REFERENCES

1. Yelland LN, Sullivan TR, Makrides M. Accounting for multiple births in randomised trials: a systematic review. *Arch Dis Child Fetal Neonatal Ed*. 2015;100(2):F116-20.
2. Karakosta A, Vassilaki M, Plainis S, Elfadl NH, Tsilimbaris M, Moschandreas J. Choice of analytic approach for eye-specific outcomes: one eye or two? *Am J Ophthalmol*. 2012;153(3):571-579.
3. Yelland LN, Sullivan TR, Price DJ, Lee KJ. Sample size calculations for randomised trials including both independent and paired data. *Stat Med*. 2017;36(8):1227-1239.
4. Yelland LN, Salter AB, Ryan P, Makrides M. Analysis of binary outcomes from randomised trials including multiple births: when should clustering be taken into account? *Paediatr Perinat Epidemiol*. 2011;25(3):283-97.
5. Bell ML, Fiero M, Horton NJ, Hsu CH. Handling missing data in RCTs; a review of the top medical journals. *BMC Med Res Methodol*. 2014;14:118.
6. Hayati Rezvan P, Lee KJ, Simpson JA. The rise of multiple imputation: a review of the reporting and implementation of the method in medical research. *BMC Med Res Methodol*. 2015;15:30.
7. Rubin D. *Multiple imputation for nonresponse in surveys*. Wiley & Sons; 1987.
8. Meng X-L. Multiple-imputation inferences with uncongenial sources of input. *Stat Sci*. 1994;9(4):538-558.
9. Rubin D. Inference and missing data. *Biometrika*. 1976;63(3):581-592.
10. Little RJ, D'Agostino R, Cohen ML, et al. The prevention and treatment of missing data in clinical trials. *N Engl J Med*. 2012;367(14):1355-60.
11. Taljaard M, Donner A, Klar N. Imputation strategies for missing continuous outcomes in cluster randomized trials. *Biom J*. 2008;50(3):329-45.
12. Andridge RR. Quantifying the impact of fixed effects modeling of clusters in multiple imputation for cluster randomized trials. *Biom J*. 2011;53(1):57-74.
13. DiazOrdaz K, Kenward MG, Gomes M, Grieve R. Multiple imputation methods for bivariate outcomes in cluster randomised trials. *Stat Med*. 2016;35(20):3482-96.
14. Ludtke O, Robitzsch A, Grund S. Multiple imputation of missing data in multilevel designs: A comparison of different strategies. *Psychol Methods*. 2017;22(1):141-165.
15. Caille A, Leyrat C, Giraudeau B. A comparison of imputation strategies in cluster randomized trials with missing binary outcomes. *Stat Methods Med Res*. 2014;
16. Graham JW. *Missing data: analysis and design*. Springer New York; 2012.
17. Audigier V, White IR, Jolani S, et al. Multiple imputation for multilevel data with continuous and binary variables. *Stat Sci*. 2017;33(2):160-183.
18. Hernandez AV, Steyerberg EW, Habbema JD. Covariate adjustment in randomized controlled trials with dichotomous outcomes increases statistical power and reduces sample size requirements. *J Clin Epidemiol*. 2004;57(5):454-60.

19. Kahan BC, Jairath V, Dore CJ, Morris TP. The risks and rewards of covariate adjustment in randomized trials: an assessment of 12 outcomes from 8 studies. *Trials*. 2014;15:139.
20. Kahan BC, Morris TP. Improper analysis of trials randomised using stratified blocks or minimisation. *Stat Med*. 2012;31(4):328-40.
21. Murray DM, Varnell SP, Blitstein JL. Design and analysis of group-randomized trials: a review of recent methodological developments. *Am J Public Health*. 2004;94(3):423-32.
22. Mancl LA, DeRouen TA. A covariance estimator for GEE with improved small-sample properties. *Biometrics*. 2001;57(1):126-34.
23. Liang K-Y, Zeger SL. Longitudinal data analysis using generalized linear models. *Biometrika*. 1986;73(1):13-22.
24. Sauzet O, Peacock JL. Binomial outcomes in dataset with some clusters of size two: can the dependence of twins be accounted for? A simulation study comparing the reliability of statistical methods based on a dataset of preterm infants. *BMC Med Res Methodol*. 2017;17(1):110.
25. Preisser JS, Young ML, Zaccaro DJ, Wolfson M. An integrated population-averaged approach to the design, analysis and sample size determination of cluster-unit trials. *Stat Med*. 2003;22(8):1235-54.
26. Turner EL, Yao L, Li F, Prague M. Properties and pitfalls of weighting as an alternative to multilevel multiple imputation in cluster randomized trials with missing binary outcomes under covariate-dependent missingness. *Stat Methods Med Res*. 2019;29(5):1338-1353.
27. Raghunathan T, Lepkowski J, Van Hoewyk J, Solenberger P. A multivariate technique for multiply imputing missing values using a sequence of regression models. *Surv Methodol*. 2001;27(1):85-95.
28. van Buuren S. Multiple imputation of discrete and continuous data by fully conditional specification. *Stat Methods Med Res*. 2007;16(3):219-42.
29. Schafer JL. *Analysis of incomplete multivariate data*. Chapman & Hall; 1997.
30. Allison P. *Missing data*. Sage Publications Inc; 2002.
31. Huque MH, Carlin JB, Simpson JA, Lee KJ. A comparison of multiple imputation methods for missing data in longitudinal studies. *BMC Med Res Methodol*. 2018;18(1):168.
32. Wijesuriya R, Moreno-Betancur M, Carlin JB, Lee KJ. Evaluation of approaches for multiple imputation of three-level data. *BMC Med Res Methodol*. 2020;20(1):207.
33. Sullivan TR, White IR, Salter AB, Ryan P, Lee KJ. Should multiple imputation be the method of choice for handling missing data in randomized trials? *Stat Methods Med Res*. 2018;27(9):2610-2626.
34. Murray DM, Blistein JL. Methods to reduce the impact of intraclass correlation in group-randomized trials. *Eval Rev*. 2003;27(1):79-103.
35. Yelland LN, Schuit E, Zamora J, et al. Correlation between neonatal outcomes of twins depends on the outcome: secondary analysis of twelve randomised controlled trials. *BJOG*. 2018;125(11):1406-1413.
36. Graham JW. Missing data analysis: making it work in the real world. Review. *Annu Rev Psychol*. 2009;60:549-76.
37. Collins LM, Schafer JL, Kam CM. A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychol Methods*. 2001;6(4):330-51.

38. White IR, Royston P, Wood AM. Multiple imputation using chained equations: Issues and guidance for practice. Research Support, Non-U.S. Gov't. *Stat Med*. 2011;30(4):377-99.
39. Audigier V, Resche-Rigon M. *micemd: Multiple imputation by chained equations with multilevel data: R package version 1.6.0*. 2019.
40. 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.
41. Collins CT, Makrides M, McPhee AJ, et al. Docosahexaenoic Acid and Bronchopulmonary Dysplasia in Preterm Infants. *N Engl J Med*. 2017;376(13):1245-1255.
42. Austin PC, Merlo J. Intermediate and advanced topics in multilevel logistic regression analysis. *Stat Med*. 2017;36(20):3257-3277.
43. Groenwold RH, Donders AR, Roes KC, Harrell FE, Jr., Moons KG. Dealing with missing outcome data in randomized trials and observational studies. Comparative Study. *Am J Epidemiol*. 2012;175(3):210-7.
44. Hossain A, Diaz-Ordaz K, Bartlett JW. Missing continuous outcomes under covariate dependent missingness in cluster randomised trials. *Stat Methods Med Res*. 2017;26(3):1543-1562.
45. Enders CK, Hayes T, Du H. A Comparison of Multilevel Imputation Schemes for Random Coefficient Models: Fully Conditional Specification and Joint Model Imputation with Random Covariance Matrices. *Multivariate behavioral research*. 2018;53(5):695-713.
46. Kahan BC, Morris TP. Reporting and analysis of trials using stratified randomisation in leading medical journals: review and reanalysis. *BMJ*. 2012;345:e5840.

TABLES

Table 1: Design effects due to the inclusion of paired data.

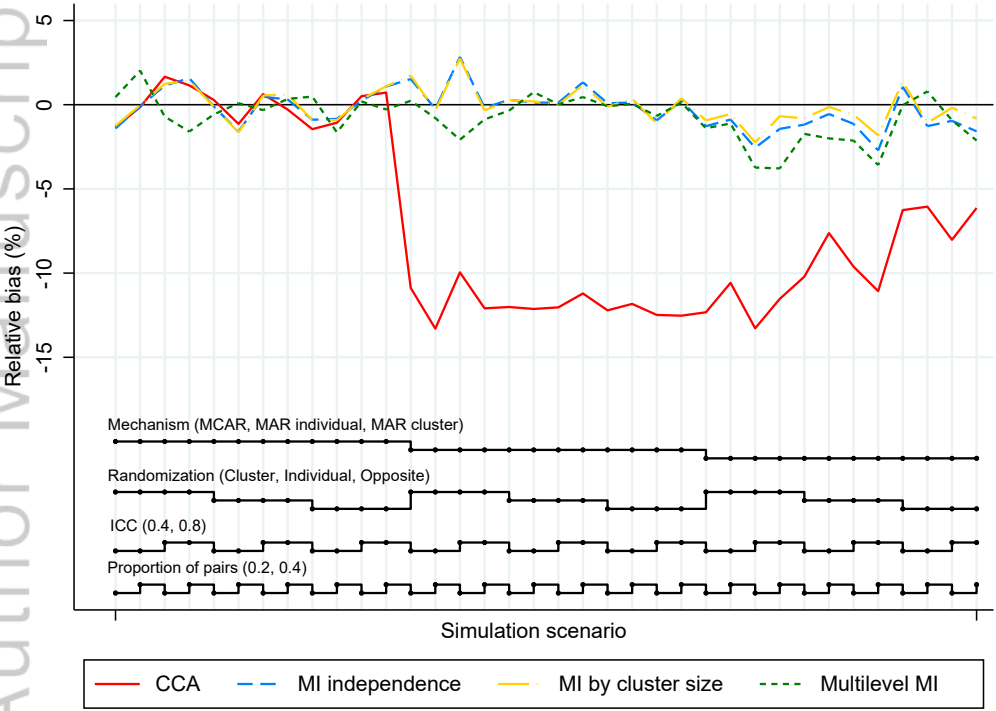
Proportion of clusters of size 2	Proportion of observations belonging to a pair (γ_p)	ICC	DEFF cluster Rx	DEFF individual Rx	DEFF opposite Rx
0.2	0.33	0.4	1.13	1	0.87
0.2	0.33	0.8	1.27	1	0.73
0.4	0.57	0.4	1.23	1	0.77
0.4	0.57	0.8	1.46	1	0.54

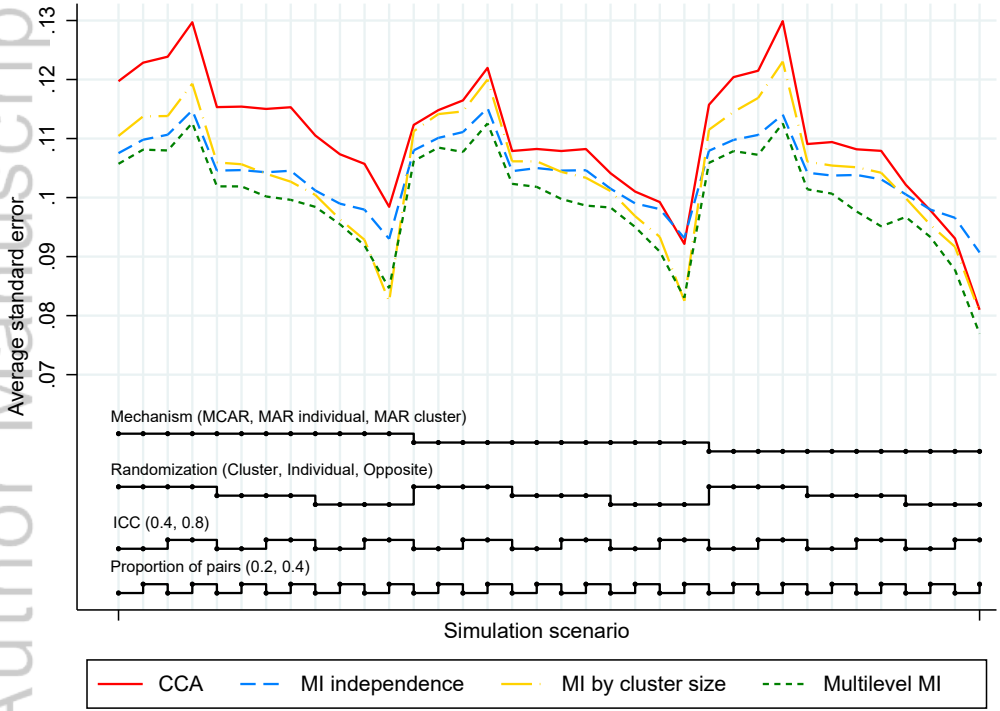
Abbreviations: ICC = intra-cluster correlation coefficient, DEFF = design effect, Rx = randomization.

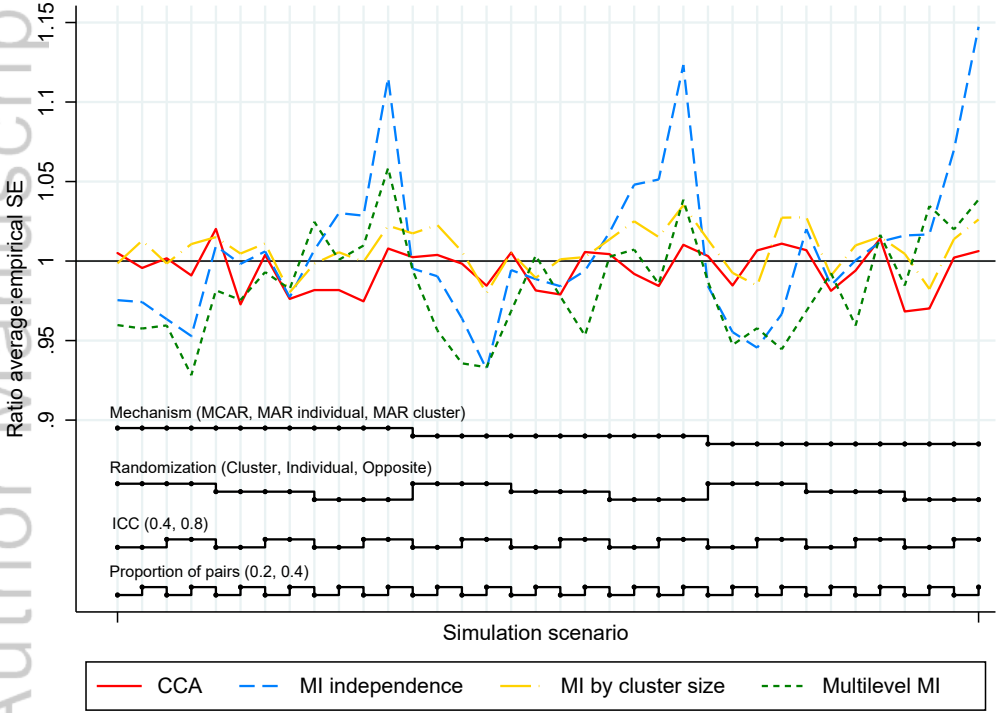
Table 2: Treatment effect estimates for clinical BPD or death from the N3RO trial.

Analysis	Mean estimate	Relative bias (%)	Average standard error
Full dataset	0.147	-	0.130
Complete case analysis	0.146	-1.0	0.167
MI assuming independence	0.143	-2.7	0.154
MI by cluster size	0.143	-2.8	0.154
Multilevel MI	0.121	-18.2	0.147

Abbreviations: BPD = bronchopulmonary dysplasia, MI = multiple imputation.







Author/s:
Sullivan, TR;Yelland, LN;Moreno-Betancur, M;Lee, KJ

Title:
Multiple imputation for handling missing outcome data in randomized trials involving a mixture of independent and paired data

Date:
2021-11-30

Citation:
Sullivan, T. R., Yelland, L. N., Moreno-Betancur, M. & Lee, K. J. (2021). Multiple imputation for handling missing outcome data in randomized trials involving a mixture of independent and paired data. STATISTICS IN MEDICINE, 40 (27), pp.6008-6020. https://doi.org/10.1002/sim.9166.

Persistent Link:
http://hdl.handle.net/11343/298844