Appendix 1 **Chondros Patty (Orcid ID: 0000-0003-0393-8734)**

Appendix 2

Appendix 3

When should matching be used in the design of cluster randomised trials?

Authors

Patty Chondros[1], Obioha C Ukoumunne[2], Jane M Gunn[1], John B Carlin[3]

Affiliations

1. Department of General Practice, Melbourne Medical School, The University of Melbourne, AUSTRALIA. E-mail: p.chondros@unimelb.edu.au
2. NIHR Applied Research Collaboration South West Peninsula (PenARC), University of Exeter, UK
3. Clinical Epidemiology and Biostatistics Unit, Murdoch Childrens Research Institute, Melbourne and Department of Paediatrics, The University of Melbourne, AUSTRALIA.

Abstract

For cluster randomised trials (CRT) with a small number of clusters, the matched-pair (MP) design, where clusters are paired before randomising one to each trial arm, is often recommended to minimise imbalance on known prognostic factors, add face-validity to the study and increase efficiency, provided the analysis recognises the matching. Little evidence exists to guide decisions on when to use matching. We used simulation to compare the efficiency of the MP design with the stratified and simple designs, based on the mean confidence interval width of the estimated intervention effect. Matched and unmatched analyses were used for the MP design; a stratified analysis was used for the stratified design; and analyses without and with post-stratification adjustment for factors that would otherwise have been used for restricted allocation were used for the simple design. Results showed the MP design was generally the most efficient for CRTs with 10 or more pairs when the correlation between cluster-level outcomes within pairs (matching correlation) was moderate to strong (0.3 to 0.5). There was little gain in efficiency for the matched-pair or stratified designs compared to simple randomisation when the matching correlation was weak (0.05 to 0.1). For trials with 4 pairs of clusters, the simple and stratified designs were more efficient than the MP design because greater degrees of freedom were available for the analysis, although an unmatched analysis of the MP design recovered precision for weak matching correlations. Practical guidance on choosing between the matched-pair, stratified and simple designs is provided.

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.9152

Key words: Cluster randomised trial; randomisation; matched-pair; stratified; completely randomised; simple randomisation

1. INTRODUCTION

Cluster randomised trials are studies in which clusters of individuals, such as general practices and communities, rather than individuals themselves, are allocated to the trial arms and outcomes are measured on the individuals within those clusters.[1,2] These studies are characterised by the correlation between outcomes of individuals that belong to the same cluster, quantified by the intra-cluster correlation coefficient (ρ_y).[1]

Three established methods for allocating clusters are simple randomisation, stratified randomisation and matched-pair randomisation.[3-11] Simple randomisation (also referred to as complete randomisation) involves allocating clusters at random to each trial arm, without any form of restriction to balance baseline characteristics. Post-stratification adjustment for factors that would otherwise have been used to balance the allocation can be used when estimating the intervention effect under this design. The stratified and matched-pair designs are restricted allocation approaches. Under stratified randomisation, clusters are stratified on baseline characteristics and then randomly allocated to each arm within strata. Stratification can be based on cluster characteristics, such as summaries (i.e., means or proportions) of the outcome measure at baseline, socio-economic status, cluster size or geographic area.[12] Less commonly, the matched-pair design is used to allocate clusters.[13-16] This involves pairing the clusters on baseline characteristics and then randomly allocating one cluster from each pair to each trial arm; thus, the design can be considered a special case of a stratified design, with only two clusters in each stratum.

The matched-pair design is intuitively attractive, especially when the number of clusters is small, to avoid imbalance on factors that are known to affect the outcome.[17,18] Matching clusters on factors strongly associated with the outcome may also increase the efficiency and power of the study, provided the analysis recognises the pairing of clusters.[2,3,17] Furthermore, balancing on factors that are not necessarily prognostic of the outcome but are considered important for logistical or political reasons may also help avoid criticism of the study and maintain "face validity".[1,17,19,20] However, the matched-pair design has unique design and analytical challenges (see Supplement 1). Unless clusters are paired on factors that are strongly correlated with the outcome, the efficiency of a matched-pair study that recognises the matching of clusters in the analysis can be considerably reduced when the number of pairs is less than 10.[17] This is because the degrees of freedom used to calculate the confidence interval

and p-value for the intervention effect is based on the number of pairs of clusters rather than the total number of clusters.[1,21] The efficiency of the matched-pair design can be increased by ignoring the matching of clusters in the analysis.[17,22]

Guidance on the use of the matched-pair design for cluster randomised trials remains unclear, particularly for trials with few clusters.[17,20,22-31] The US National Institutes of Health Collaboratory review noted the lack of specific guidance to choose between the matched-pair, stratified and simple randomisation methods for cluster randomised trials and recommended a comprehensive simulation study to assess their relative efficiency.[30]

This paper uses simulation to identify design scenarios where matched allocation is more efficient than the use of simple and stratified randomisation in cluster randomised trials with a normally distributed outcome, based on using cluster-level analysis methods to estimate the intervention effect. Section 2 describes the simulation model used to define the framework from which we generated datasets from studies that use simple, stratified or matched-pair allocation, in a manner that reflects the decision context in real cluster randomised trials. The results are summarised in Section 3 and a discussion of the findings, with guidance on when to use matching, is presented in Section 4.

2. SIMULATION STUDY

Design

Simulation was used to quantify the relative gain in precision of the matched-pair design compared to the stratified and simple randomisation designs. The simulation model was used to define scenarios in which potential matched pairs of clusters exist (matched-pair design), with the stratified and simple randomisation designs embedded in the data structure, making it possible to assess the ramifications of choosing one design over the others. A normally distributed outcome variable Y was generated for cluster randomised controlled trials with two arms ($i = 1,2$). For simplicity, we specified p strata, q matched pairs of clusters within each stratum, an equal number of clusters in each trial arm (pq) and an equal number of participants sampled from each cluster (cluster size n).

Individual-level datasets were generated using a random effects model (2.1)

$$Y_{ijkl} = \mu + \beta + R_j + T_{jk} + C_{ijk} + e_{ijkl} \qquad (2.1)$$

where $j = 1, \ldots, p$ indexes the strata; $k = 1, \ldots, q$ indexes the cluster pairs within strata, and $l = 1,2, \ldots, n$ indexes the individual participants within clusters. μ is the overall mean response in the control group and β is the difference in mean response (or intervention effect) between the trial arms. R_j is the random effect shared by all clusters within the j^{th} stratum, with variance σ_r^2. T_{jk} is the additional random effect shared by clusters within the k^{th} matched pair in the j^{th} stratum, with variance σ_t^2. C_{ijk} and e_{ijkl} represent random effects for the cluster and individual participants, with variance σ_c^2 and σ_e^2, respectively. The random effects were drawn from independent normal distributions, where

$$R_j = \text{stratum effect} \sim N(0, \sigma_r^2)$$
$$T_{jk} = \text{matched-pair effect} \sim N(0, \sigma_t^2)$$
$$C_{ijk} = \text{cluster effect} \sim N(0, \sigma_c^2)$$
$$e_{ijkl} = \text{residual errors} \sim N(0, \sigma_e^2)$$

$(R_j + T_{jk})$ is the random effect shared by clusters in the k^{th} matched pair of the j^{th} stratum, with variance $\sigma_s^2 = \sigma_r^2 + \sigma_t^2$. To simplify the model, μ and β were set to zero, and the total variance of the outcome, $\sigma_T^2 = \sigma_r^2 + \sigma_t^2 + \sigma_c^2 + \sigma_e^2$ was set to 1 without any loss of generality. The mean of the outcome for each cluster (average over the n individuals) is:

$$\overline{Y}_{ijk.} = \mu + R_j + T_{jk} + C_{ijk} + \overline{e}_{ijk.} \text{ where } \overline{e}_{ijk.} \sim N\left(0, \frac{\sigma_e^2}{n}\right).$$

For each generated dataset, categorical variables were specified to identify the matched-pair and the stratum that the clusters belong to. These variables will be referred to as the matching variable and the stratification variable, respectively. After each dataset was generated for the matched-pair design, the corresponding dataset for the stratified randomisation design was created by breaking the matches and then randomly allocating clusters within strata. The corresponding dataset for the simple randomisation design was created from the matched-pair design dataset by breaking both the matches and strata and then randomly allocating clusters. With this approach, outcome data were generated with the correct correlation structure for the matched-pair, stratified and simple randomisation designs, respectively.

To generate the data, the chosen parameter values for the correlation between cluster mean outcomes within pairs (matching correlation, ρ_M), the correlation between cluster mean outcomes for clusters that are in different pairs within the same stratum (intra-stratum

correlation for the unmatched clusters $\rho_U = f\rho_M$, where f is the proportion of the variation within strata that is *not* attributable to the pairing of clusters) and the intra-cluster correlation coefficient (ρ_y) were used to determine the variance components $(\sigma_s^2, \sigma_c^2, \sigma_e^2)$ specified in the simulation (Model 2.1), using the following equations (derivations provided in Supplement 2):

$$\boldsymbol{\sigma_e^2 = 1 - \rho_y}$$

$$\boldsymbol{\sigma_c^2 = \rho_y - \rho_M \left[\frac{1 + (n-1)\rho_y}{n}\right]}$$

$$\boldsymbol{\sigma_s^2 = \rho_M \left[\frac{1+(n-1)\rho_y}{n}\right]} \qquad (2.2)$$

Given that $\sigma_s^2 = \sigma_r^2 + \sigma_t^2$ is the variance shared by clusters in the same pair, the variance shared by clusters within a stratum can expressed as $\sigma_r^2 = f\sigma_s^2$ and the variance component specifically attributable to matching as $\sigma_t^2 = (1-f)\sigma_s^2$.

Parameter values

Parameter values for the total number of clusters and cluster sizes were chosen to reflect typical scenarios found in primary care and community health settings,[10,21] and in community trials where even fewer clusters are recruited.[25,32] Motivated by knowledge of these settings, we simulated data for cluster randomised trials with a small number (8 or 20) of large clusters (n = 20, 50, 100) and for trials with a large number (40) of small clusters (n = 5, 20, 50).

Table 1 presents six study design scenarios defined in combination by the total number of clusters and the number of strata, where the clusters could be allocated using the simple, stratified or matched-pair randomisation methods. For studies with 8 clusters, the simple randomisation design (4 clusters in each arm) and the stratified design with 2 strata were compared with the matched-pair design with 4 paired clusters (Scenario 1). For studies with 20 clusters, the simple randomisation design (10 clusters per arm) and the stratified designs with 2 strata (Scenario 2) and 5 strata (Scenario 3) were compared with the matched-pair design with 10 paired clusters. For studies with 40 clusters the simple randomisation design (20 clusters per arm) and the stratified designs with 2, 5 and 10 strata (Scenarios 4, 5, and 6, respectively) were compared with the matched-pair design with 20 paired clusters.

For each of the six study design scenarios, we specified values of the intra-cluster correlation coefficient of the outcome (ρ_y) that are typically reported for patient outcomes ($\rho_y = 0.001, 0.01, 0.05$) and process measures ($\rho_y = 0.1, 0.3$) in primary care and community health settings.[33-38] There is little empirical evidence for the likely range of matching correlations.[1,3,16,39] We specified values for the matching correlation considered to be indicative of weak ($\rho_M = 0, 0.05, 0.1$) or strong ($\rho_M = 0.3, 0.5$) matching.[1,17,22] For each value of the matching correlation, we set the correlation between the unmatched cluster means within strata as $\rho_U = f\rho_M$, by specifying the values $f = {}^1/_3, {}^1/_2, {}^2/_3$ or 1. As f increases, clusters from different pairs within the same stratum become more alike and matched pair membership provides a decreasing amount of additional information beyond that already provided by stratum membership.

Under Model (2.1), the values that ρ_y and ρ_M can take are constrained because the variance components ($\sigma_s^2, \sigma_c^2, \sigma_e^2$) cannot be negative. For the variance components to be non-negative, ρ_y can only take values between 0 and 1[40] and ρ_M can only take values between 0 and $\frac{n\rho_y}{1+(n-1)\rho_y}$, the upper bound constrained by both the cluster size (n) and the intra-cluster correlation coefficient (ρ_y). Table 2 shows the upper bound for ρ_M for specific parameter values of ρ_y and n. When the cluster sizes are imbalanced, n in the upper bound is replaced with the smallest cluster size $n_{\min}$.

Altogether, 1530 combinations of the simulation design parameters were originally specified. These consisted of 90 combinations (6 design scenarios by 3 cluster sizes by 5 intra-cluster correlations) when the matching correlation was zero and 1,440 combinations when the matching correlations were greater than zero (90 combinations as above by 4 non-zero matching correlations (ρ_M) by 4 correlations between unmatched clusters within strata (ρ_U)). Four hundred and thirty-two design parameter combinations were excluded because some values of the matching correlation could not be achieved for certain configurations of the intra-cluster correlation coefficient and cluster size, leaving 1098 combinations. Two thousand replications for each combination of design parameter values were used, sufficient to estimate coverage of the nominal 95% confidence interval for the intervention effect with a 95% confidence interval of width less than ±1 percent.

Analysis methods

For each simulated dataset we used cluster-level analyses, applying the *t*-test or linear regression, using the means of the outcomes in the clusters as the observations, to estimate the intervention effect. The method is easy to implement, provides unbiased estimates and reasonable efficiency even with small sample sizes,[1,41] and has been shown to be robust to deviations from model assumptions.[1] The 95 percent confidence interval for the estimated intervention effect was calculated using the *t*-distribution, based on the number of clusters.

The matched-pair design with matched analysis (**Pair/M – paired design, matched analysis**) used the cluster-level paired *t*-test with degrees of freedom equal to the number of pairs minus one ($\mathrm{df} = pq - 1$) and the simple randomised design (**Simple**) was analysed using the cluster-level two-sample *t*-test with degrees of freedom equal to the number of clusters minus two ($\mathrm{df} = 2pq - 2$). The stratified design (**Stratified**) was analysed using the extension of the cluster-level two-sample *t*-test, where a separate estimate of the intervention effect is calculated for each stratum and the overall intervention effect is obtained by taking the average of the stratum-specific estimates.[41] For the stratified design, the degrees of freedom was the total number of clusters ($2pq$) minus the number of study arm and stratum combinations ($2p$).[1,41]

The cluster-level two sample *t*-test was also applied to the matched-pair design (**Pair/U – paired design, unmatched analysis**) ($\mathrm{df} = 2pq - 2$) because it has been shown to provide nominal coverage and narrower confidence intervals than the cluster-level paired *t*-test when the number of clusters is small and the true matching correlation is weak (0.1 or less).[22,25]

Finally, for the simple randomised trial design we also used post-stratification adjustment for stratum membership (**Simple/PS**). This method was investigated because it can potentially improve the efficiency of the simple randomisation design.[1,19] The intervention effect was estimated using linear regression of the cluster-level means on trial arm status, with the stratification variable fitted as a categorical covariate. The number of degrees of freedom for calculating the 95% confidence interval was $2pq - p - 1$.

Unless otherwise stated, the analytical methods described as **Pair/M**, **Stratified** and **Simple** are implied when referring to the matched-pair, stratified and simple randomised designs,

respectively. Supplement 3 provides a summary of the study designs with the respective analytical methods and degrees of freedom.

The bias and standardised bias (the bias expressed as a percentage of the empirical standard error) of the intervention effect estimate and the coverage of the 95% confidence interval were estimated to validate the simulation model and confirm that the analytical methods were unbiased. The randomisation methods were evaluated by comparing the precision of the estimated intervention effect, based on the mean width of the 95 percent confidence intervals. Relative efficiency, quantified as the ratio of the mean confidence interval width for each study design to the mean confidence interval width of the matched-pair design (using a matched-pair analysis), was reported. A ratio of one indicates that the study design provides the same mean confidence interval width (i.e., is as efficient) as the matched pair design, a ratio less than one indicates the study design is more efficient than the matched-pair design, and a ratio greater than one indicates the study design is less efficient than the matched pair design.

4. RESULTS

Bias and coverage

Table 3 reports the standardised bias of the effect estimators and shows that the bias was never greater than 10 percent of the empirical standard error, and not large enough to impact on the coverage of the analytical methods.[42] Consistent with this, the coverage was close to the nominal 95 percent level for all study designs and analysis methods, with the exception of the matched-pair design with the unmatched analysis (**Pair/U**) for which the coverage ranged between 0.955 and 0.996 across the different design parameter combinations when the matching correlation was 0.3 or 0.5 (Supplementary Table 1).

Mean width of the confidence intervals and relative efficiency of the study designs

Overall, the mean confidence interval width increased for all study designs as the intra-cluster correlation coefficient increased, and decreased with increasing number of clusters, sample cluster size and correlation of clusters within the stratum. The mean confidence interval width for the matched-pair design was consistent across the range of ratios of the correlation between unmatched clusters within strata to the correlation between the matched clusters. Although the actual width of the confidence intervals depended on the parameter values for the intra-cluster correlation and sample cluster size (Supplementary Table 2), the general patterns of the relative

efficiency between the designs were similar. Thus, for brevity, only the results for an intra-cluster correlation coefficient of 0.3 and sample cluster size of 50 (available for all combinations of values for the other design parameters) are presented.

Figure 1 shows the mean width of the confidence intervals for each study design with the respective analytical methods and Figure 2 shows the ratio of the mean confidence interval width of each study design to that of the matched-pair design with the matched analysis. For studies with 8 clusters (Scenario 1), the figures show that when the matching correlation was weak ($\rho_M = 0.05, 0.1$), the matched-pair design with the matched analysis (**Pair/M**) had the widest confidence intervals, followed by the stratified design (**Stratified**); the simple randomisation design (**Simple**) and the matched-pair design with the unmatched analysis (**Pair/U**) had similar mean confidence interval widths and were the most efficient study designs. The **stratified** design was generally more efficient than the matched-pair design when the matching correlation was 0.3, but the matched-pair design (**Pair/M**) was more efficient than the **simple** and **stratified** designs when the matching correlation was strong ($\rho_M = 0.5$) and the correlation between unmatched clusters from the same stratum was weak relative to the correlation between the matched clusters ($f = 1/3$).

The pattern for the relative efficiency between the designs was similar across studies with 20 (Scenarios 2 and 3) or 40 (Scenarios 4, 5 and 6) clusters. There was little or no difference in efficiency between the matched-pair (**Pair/M**), **stratified** and **simple** designs when the matching correlation was weak ($\rho_M = 0.05, 0.1$). However, for stronger matching correlations ($\rho_M = 0.3, 0.5$), the difference in the mean confidence interval width across the study designs was marked. Generally, the matched-pair design (**Pair/M**) was more efficient than the other designs, except when the correlation between unmatched clusters in the same stratum was equal to the matching correlation ($\rho_U = \rho_M$). In this instance, the **stratified** design had slightly narrower confidence intervals than the matched-pair design (**Pair/M**), but the difference diminished as either the number of clusters increased, or number of strata increased for a fixed number of clusters.

For all studies with 8, 20 or 40 clusters, the simple design with adjustment for the stratification variable in the analysis (**Simple/PS**) had similar study efficiency to the **stratified** design, when there were 2 strata (Scenario 1, 2 and 4), but the mean width of the confidence intervals

increased for this method compared to the stratified design as the number of strata increased relative to the number of clusters ($k = 5, 10$).

5. DISCUSSION

Our simulation design and approach to generating datasets from studies with matched-pair, stratified and simple randomisation designs reflects the decision context in real trials. The three allocation methods, used with appropriate analytical methods, were compared in the same simulation study for different combinations of design parameter values that were informed by published trials. To our knowledge no other study has addressed this issue as comprehensively.

Table 4 summarises practical guidance on choosing between the allocation methods, based on the precision of the estimators in our simulation study and the existing literature. For cluster randomised trials with 10 or more pairs, there is little or no measurable gain in efficiency when a matched-pair or stratified design is adopted compared to simple randomisation when the correlation between cluster-level outcomes within pairs (the matching correlation) is 0.05 or 0.1. The matched-pair design is most efficient for cluster randomised trials with at least 10 clusters per arm when the matching correlation is stronger (0.3 and 0.5), except when the correlation between unmatched clusters within strata is as great as the correlation between matched clusters, in which case the stratified design has similar efficiency.

For studies with fewer than 10 clusters per arm, simple randomisation is the most efficient design when the stratification factors are weakly correlated with the outcome. Our simulation study showed that the simple and stratified randomisation designs were more efficient than the matched-pair design when the number of clusters was less than 10 per arm, due to there being greater degrees of freedom available for estimation using the *t*-distribution,[3] except when the matching correlation was 0.5. Similarly, because greater degrees of freedom were available for estimation, the simple design was more efficient than the stratified design when there were fewer than 10 clusters per arm, particularly when the matching correlation of the outcomes was weak, as has been found by others.[3,22] When imbalance on weak risk factors for studies with fewer than 10 clusters in each arm is a concern, a stratified design may be a good compromise between a matched-pair design, which may lack efficiency, and a simple design.[19]

Our findings showed that an unmatched analysis of a matched-pair design with fewer than 10 pairs of clusters when matching was weak (0.05 or 0.1) recovered precision of the estimates, supporting previous findings by Diehr *et al.*[22] When the matched-pair design is used to maintain "face validity" or for social or logistical reasons, and the matching factors are not strongly correlated with the outcome, an unmatched analysis may provide a more efficient estimate of the intervention effect when there are few clusters. The decision to break the matches in the analysis needs to be stated *a priori* at the design stage of the study.[22]

Post-stratification adjustment, where potential risk factors are adjusted for in the analysis for the simple design, is an alternative approach to the stratified design,[1,19] when there are at least 5 clusters per stratum and study arm combination. Post-stratification adjustment for too many cluster-level risk factors in the regression model relative to the number of clusters can, however, lead to significant loss in study efficiency. The reason for this is that when the number of clusters is fixed, the parameters in the model are estimated with greater imprecision as the number of cluster-level risk factors included in the model increases because there is less information in the data for estimation.

Trials with fewer than four clusters per study arm should not be considered.[22,43] Challenges of studies with few clusters include greater chance of imbalance on risk factors, lack of study power to detect realistic clinically important effect sizes[19,22,26] and limited generalisability of results.[22,43,44] Furthermore, analytical methods that rely on large sample asymptotic theory are not reliable when the number of clusters is small.[43,45]

Caution is needed when generalising the results beyond the values of the design parameters set in the simulation. The findings may potentially be extrapolated to trials with more than two study arms, variable cluster sizes,[46] and non-continuous outcomes, such as binary or count data, but need to be applied carefully when the cluster sizes are small, or the outcomes are highly sparse or skewed. Furthermore, the simulations specified an equal number of clusters across the stratum and study arm combinations, for both the matched-pair and stratified designs. In practice, this is not always the case, even for the matched-pair design.

The simulation study examined two restricted allocation methods. Other forms of restricted randomisation such as minimisation, a dynamic (adaptive) random allocation method,[12,47] and covariate-constrained randomisation[31] may produce tighter balance on multiple risk factors

between study arms, but may also be prone to similar limitations as those for the stratified and matched-pair designs, particularly when there are few clusters.

We have shown algebraically that the maximum matching correlation that can be achieved depends on the intra-cluster correlation coefficient of the outcome and sample cluster size (Table 2). For instance, when planning a study with an anticipated intra-cluster correlation coefficient of 0.001, effective matching can potentially capture up to 50 percent of the outcome variability between clusters when the sample cluster size is 1000, compared to only 9 percent when the sample cluster size is 100. To achieve effective matches, the between-cluster variance of summary measures needs to be large.[21,26] Effective matches are difficult to achieve as the intra-cluster correlation coefficient approaches zero because most of the outcome variation is within clusters and there is relatively little variation between cluster means. But as the sample cluster size increases, closer matches can be achieved on cluster-level summary outcomes because the observed variability on these observations diminishes.[41] In trials where the number of clusters available is limited, but entire communities are sampled, such as large-scale public policy evaluation[48] and community trials,[1,32] effective matches may be achieved.

Matching is intuitively appealing, but we have confirmed in this study that unless it is possible to create very strong matches it is less useful in practice than might be expected, especially for studies with small numbers of clusters. Furthermore, there is often little information on the potential strength of matching, and researchers may be overly optimistic about the effectiveness of their matching. The true strength of the correlation between the matching variables and the outcome of interest is seldom known.[3,39] For matched or stratified designs to be effective, good prior knowledge of the strength of the correlation between potential stratification factors and the outcome variable is needed. There is also the need to consider the complex manner in which the between-cluster variation and the sample cluster size together impact on the variance in the observed cluster means. The rationale for using the matched-pair design and justification for selecting the factors used to balance the randomisation should be reported in trial protocols. Similar considerations apply for trials that use other restricted randomisation schemes. The matching correlation and the correlation between the stratification factors and the outcome should be reported with confidence intervals, both as measures of effectiveness of the randomisation strategy and to inform the design of future trials.

Acknowledgments

Patty Chondros was supported by the Australian National Health and Medical Research Council Public Health Postgraduate Scholarship. Obi Ukoumunne was supported by the National Institute for Health Research Applied Research Collaboration South West Peninsula. The views expressed in this publication are those of the authors and not necessarily those of the National Institute for Health Research or the Department of Health and Social Care.

References

1. Donner A, Klar N. *Design and analysis of cluster randomization trials in health research.* London: Arnold; 2000.
2. Murray DM. *Design and analysis of group-randomized trials.* New York: Oxford University Press; 1998.
3. Klar N, Donner A. The merits of matching in community intervention trials: A cautionary tale. *Stat Med.* 1997;16(15):1753-1764.
4. Donner A, Brown KS, Brasher P. A methodological review of non-therapeutic intervention trials employing cluster randomization, 1979-1989. *Int J Epidemiol.* 1990;19(4):795-800.
5. Simpson JM, Klar N, Donner A. Accounting for cluster randomization: a review of primary prevention trials, 1990 through 1993. *Am J Public Health.* 1995;85(10):1378-1383.
6. Hayes RJ, Alexander ND, Bennett S, Cousens SN. Design and analysis issues in cluster-randomized trials of interventions against infectious diseases. *Stat Methods Med Res.* 2000;9(2):95-116.
7. Puffer S, Torgerson D, Watson J. Evidence for risk of bias in cluster randomised trials: review of recent trials published in three general medical journals. *Br Med J.* 2003;327(7418):785-787.
8. Isaakidis P, Ioannidis JPA. Evaluation of cluster randomized controlled trials in sub-Saharan Africa. *Am J Epidemiol.* 2003;158(9):921-926.
9. Varnell SP, Murray DM, Janega JB, Blitstein JL. Design and analysis of group-randomized trials: A review of recent practices. *Am J Public Health.* 2004;94(3):393-399.
10. Eldridge SM, Ashby D, Feder GS, Rudnicka AR, Ukoumunne OC. Lessons for cluster randomized trials in the twenty-first century: a systematic review of trials in primary care. *Clin Trials.* 2004;1(1):80-90.
11. Murray DM, Pals SL, Blitstein JL, Alfano CM, Lehman J. Design and analysis of group-randomized trials in cancer: A review of current practices. *J Natl Cancer Inst.* 2008;100(7):483-491.
12. Eldridge S, Kerry S. *A Practical Guide to Cluster Randomised Trials in Health Services Research.* 2nd ed: Hoboken: Wiley; 2012.
13. Fiero MH, Huang S, Oren E, Bell ML. Statistical analysis and handling of missing data in cluster randomized trials: a systematic review. *Trials.* 2016;17(1):72.
14. Hodgkinson A, Abbott J, Hurley MA, Lowe N, Qualter P. An educational intervention to prevent overweight in pre-school years: a cluster randomised trial with a focus on disadvantaged families. *BMC Public Health.* 2019;19(1):1430.
15. Croker H, Lucas R, Wardle J. Cluster-randomised trial to evaluate the 'Change for Life' mass media/ social marketing campaign in the UK. *BMC Public Health.* 2012;12(1):404.
16. Semrau KEA, Hirschhorn LR, Kodkany B, et al. Effectiveness of the WHO Safe Childbirth Checklist program in reducing severe maternal, fetal, and newborn harm in Uttar Pradesh, India: study protocol for a matched-pair, cluster-randomized controlled trial. *Trials.* 2016;17(1):576.
17. Martin DC, Diehr P, Perrin EB, Koepsell TD. The Effect of Matching on the Power of Randomized Community Intervention Studies. *Stat Med.* 1993;12(3-4):329-338.
18. Campbell MJ, Donner A, Klar N. Developments in cluster randomized trials and Statistics in Medicine. *Stat Med.* 2007;26(1):2-19.

19. Bloom HS. *Learning More from Social Experiments: Evolving Analytic Approaches.* New York: Russel Sage Foundation; 2005.
20. Imai K, King G, Nall C. The Essential Role of Pair Matching in Cluster-Randomized Experiments, with Application to the Mexican Universal Health Insurance Evaluation. *Stat Sci.* 2009;24(1):29-72.
21. Campbell MJ. Cluster randomized trials in general (family) practice research. *Stat Methods Med Res.* 2000;9(2):81-94.
22. Diehr P, Martin DC, Koepsell T, Cheadle A. Breaking the matches in a paired t-test for community interventions when the number of pairs is small. *Stat Med.* 1995;14(13):1491-1504.
23. Klar N, Donner A. Current and future challenges in the design and analysis of cluster randomization trials. *Stat Med.* 2001;20(24):3729-3740.
24. Thompson SG. The merits of matching in community intervention trials: A cautionary tale (Letter). *Stat Med.* 1998;17(18):2149-2151.
25. Feng Z, Diehr P, Peterson A, McLerran D. Selected Statistical Issues in Group Randomized Trials. *Annu Rev Public Health.* 2001;22:167-187.
26. Raudenbush SW, Martinez A, Spybrook J. Strategies for improving precision in group-randomized experiments. *Educ Eval Policy Anal.* 2007;29(1):5-29.
27. Rhodes W. Pairwise Cluster Randomization: An Exposition. *Eval Rev.* 2014;38(3):217-250.
28. Ivers NM, Halperin IJ, Barnsley J, et al. Allocation techniques for balance at baseline in cluster randomized trials: a methodological review. *Trials.* 2012;13(1):120-128.
29. Imbens GW. Experimental design for unit and cluster randomised trials. Paper presented at: International Initiative for Impact Evaluation; June 15–17, 2011; Cuernavaca, Mexico.
30. DeLong E, Li E, Cook A. Pair-Matching vs Stratification in ClusterRandomized Trials. 2014; https://dcricollab.dcri.duke.edu/sites/NIHKR/KR/Pairing-vs-stratification_V1.0.pdf Accessed 21 January 2021.
31. Turner EL, Fan L, Gallis JA, Prague M, Murray DM. Review of Recent Methodological Developments in Group-Randomized Trials: Part 1--Design. *Am J Public Health.* 2017;107(6):907-915.
32. Grosskurth H, Mosha F, Todd J, et al. Impact of improved treatment of sexually transmitted diseases on HIV infection in rural Tanzania: randomised controlled trial. *Lancet.* 1995;346(8974):530-536.
33. Campbell M, Grimshaw J, Steen N. Sample size calculations for cluster randomised trials. Changing Professional Practice in Europe Group (EU BIOMED II Concerted Action). *J Health Serv Res Policy.* 2000;5(1):12-16.
34. Campbell MK, Mollison J, Grimshaw JM. Cluster trials in implementation research: estimation of intracluster correlation coefficients and sample size. *Stat Med.* 2001;20(3):391-399.
35. Adams G, Gulliford MC, Ukoumunne OC, Eldridge S, Chinn S, Campbell MJ. Patterns of intra-cluster correlation from primary care research to inform study design and analysis. *J Clin Epidemiol.* 2004;57(8):785-794.
36. Elley CR, Kerse N, Chondros P, Robinson E. Intraclass correlation coefficients from three cluster randomised controlled trials in primary and residential health care. *Aust N Z J Public Health.* 2005;29(5):461-467.
37. Knox SA, Chondros P. Observed intra-cluster correlation coefficients in a cluster survey sample of patient encounters in general practice in Australia. *BMC Med Res Methodol.* 2004;4(1):30.

38. Littenberg B, Maclean CD. Intra-cluster correlation coefficients in adults with diabetes in primary care practices: The Vermont Diabetes Information System Field Survey. *BMC Med Res Methodol.* 2006;6(1):20.
39. Gail MH, Mark SD, Carroll RJ, Green SB, Pee D. On design considerations and randomization-based inference for community intervention trials. *Stat Med.* 1996;15(11):1069-1092.
40. Eldridge SM, Ukoumunne OC, Carlin JB. The Intra-Cluster Correlation Coefficient in Cluster Randomized Trials: A Review of Definitions. *Int Stat Rev.* 2009;77(3):378-394.
41. Hayes RJ, Moulton LH. *Cluster randomised trials.* Boca Raton: CRC Press; 2009.
42. Collins LM, Schafer JL, Kam CM. A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychol Methods.* 2001;6(4):330-351.
43. Ukoumunne OC, Gulliford MC, Chinn S, Sterne JA, Burney PG. Methods for evaluating area-wide and organisational-based interventions in health and health care: a systematic review. *Health Technol Assess.* 1999;3(5).
44. Thompson SG, Pyke SDM, Hardy RJ. The design and analysis of paired cluster randomized trials: an application of meta-analysis techniques. *Stat Med.* 1997;16(18):2063-2079.
45. Duncan C, Jones K, Moon G. Context, composition and heterogeneity: Using multilevel models in health research. *Soc Sci Med.* 1998;46(1):97-117.
46. Crespi CM. Improved Designs for Cluster Randomized Trials. *Annu Rev Public Health.* 2016;37(1):1-16.
47. Kahan BC, Morris TP. Improper analysis of trials randomised using stratified blocks or minimisation. *Stat Med.* 2012(4):328.
48. King G, Gakidou E, Ravishankar N, et al. A "politically robust" experimental design for public policy evaluation, with application to the Mexican universal health insurance program. *J Policy Anal Manag.* 2007;26(3):479-506.

Table 1: Number of strata and clusters per stratum for the stratified design and number of paired clusters for the matched-pair design evaluated for each combination of design parameter values

Scenario	Simple design	Stratified design		Matched-pair design
	Total number of clusters *(2pq)*	**Number of strata** *(p)*	**Number of clusters per stratum** *(2q)*	**Number of paired clusters** *(q)*
Small number of large clusters $(\boldsymbol{n = 20, 50, 100})$				
1	8	2	4	4
2	20	2	10	10
3	20	5	4	10
Large number of small clusters $(\boldsymbol{n = 5, 20, 50})$				
4	40	2	20	20
5	40	5	8	20
6	40	10	4	20

n is the number of individuals sampled per cluster

Table 2: Maximum possible value of the matching correlation (ρ_M) for given values of the intra-cluster correlation coefficient of the outcome (ρ_y) and the cluster size (n)

	Cluster size (n)				
ρ_y	**5**	**20**	**50**	**100**	**1000**
0.001	0.005	0.020	0.048	0.091	0.500
0.005	0.025	0.091	0.201	0.334	0.834
0.01	0.048	0.168	0.336	0.503	0.910
0.05	0.208	0.513	0.725	0.840	0.981
0.1	0.357	0.690	0.847	0.917	0.991
0.3	0.682	0.896	0.955	0.977	0.998

Table 3: Summary of the standardised bias (as percentage of the empirical standard error) of the estimator for the matched-pair, stratified and simple designs with the respective analytical methods by the matching correlation and the six study design scenarios given in Table 1. Standardised bias is summarised across the parameter value combinations for the intra-cluster correlation, sample cluster size, and ratio of the correlation between unmatched clusters within strata to the correlation between matched clusters (N=1098 simulation design parameter combinations altogether)

Matching correlation	**0**			**0.05**			**0.1**			**0.3**			**0.5**		
Study design	**Median**	**Min**	**Max**	**Median**	**Min**	**Max**	**Median**	**Min**	**Max**	**Median**	**Min**	**Max**	**Median**	**Min**	**Max**
Scenario 1: Number of clusters = 8 with 2 strata and 4 matched pairs (N=199)															
	N= 15			**N= 52**			**N= 48**			**N= 44**			**N= 40**		
Pair/M and **Pair/U**†	-0.088	-3.837	3.606	0.029	-2.918	2.818	-0.209	-4.322	2.610	0.010	-2.141	2.332	-0.028	-3.578	3.700
Stratified	-0.508	-2.037	3.051	0.327	-5.722	4.611	0.184	-5.430	4.586	0.604	-6.310	8.016	-0.086	-5.312	6.703
Simple	0.193	-3.653	4.614	-0.208	-4.988	4.160	0.151	-5.333	2.842	0.213	-5.884	3.478	-0.385	-5.791	4.520
Simple/PS	0.323	-3.235	5.425	0.066	-5.287	3.495	0.109	-4.603	3.034	-0.120	-5.789	4.470	-0.609	-4.100	6.326
Scenario 2: Number of clusters = 20 with 2 strata and 10 matched pairs (N=199)															
Pair/M and **Pair/U**†	0.717	-1.056	2.104	0.074	-3.783	3.584	0.118	-2.279	2.365	-0.107	-2.067	2.276	0.273	-2.302	2.625
Stratified	-0.877	-2.786	2.542	-0.295	-6.451	9.119	0.007	-7.115	5.400	0.900	-2.794	6.975	0.151	-4.120	6.579
Simple	0.946	-1.570	3.636	0.404	-4.516	3.830	0.583	-4.552	6.932	0.323	-7.271	7.201	-0.545	-5.100	4.993
Simple/PS	1.315	-2.134	3.691	0.439	-5.071	4.572	0.588	-5.262	7.379	0.340	-4.802	7.209	-0.653	-5.947	3.372
Scenario 3: Number of clusters = 20 with 5 strata and 10 matched pairs (N=199)															
Pair/M and **Pair/U**†	-0.259	-3.426	1.366	0.062	-2.180	2.050	0.118	-3.085	2.109	-0.057	-2.292	2.260	-0.289	-3.188	2.354
Stratified	0.778	-4.812	4.527	-0.127	-4.661	5.927	0.002	-4.392	6.046	-0.110	-3.804	4.714	-0.090	-4.579	4.867
Simple	-0.245	-5.256	2.609	-1.120	-4.472	4.899	0.070	-5.336	5.901	-0.034	-4.155	4.397	0.202	-5.738	7.448
Simple/PS	-0.969	-5.649	2.893	-0.495	-6.320	5.060	0.206	-6.109	5.134	-0.314	-3.838	5.159	0.796	-5.918	5.284

Matching correlation	0			0.05			0.1			0.3			0.5		
Study design	**Median**	**Min**	**Max**	**Median**	**Min**	**Max**	**Median**	**Min**	**Max**	**Median**	**Min**	**Max**	**Median**	**Min**	**Max**
Scenario 4: Number of clusters = 40 with 2 strata and 20 matched pairs (N=167)															
	N= 15			**N= 44**			**N= 44**			**N= 36**			**N= 28**		
Pair/M and **Pair/U**†	-0.031	-5.961	4.159	-0.117	-2.666	3.533	-0.340	-3.756	5.932	-0.076	-3.265	3.769	0.305	-2.362	2.793
Stratified	-0.122	-2.733	3.983	-0.170	-4.074	9.533	0.874	-6.368	4.600	0.353	-8.569	5.501	-0.546	-3.681	4.238
Simple	-0.900	-4.829	4.339	0.436	-7.269	6.589	0.405	-5.997	3.736	0.689	-4.056	5.090	-1.001	-4.008	4.320
Simple/PS	-0.586	-4.553	4.427	0.281	-7.630	6.578	0.204	-5.418	3.895	0.691	-5.404	4.672	-0.980	-4.247	3.784
Scenario 5: Number of clusters = 40 with 5 strata and 20 matched pairs (N=167)															
Pair/M and **Pair/U**†	-0.081	-2.730	4.294	-0.164	-3.236	3.577	0.183	-3.670	2.796	0.373	-4.219	5.422	0.478	-4.497	3.477
Stratified	0.482	-1.859	5.539	0.497	-4.096	3.412	-0.213	-4.156	3.994	-0.008	-2.917	4.320	-0.173	-3.101	4.471
Simple	-0.137	-6.437	3.757	-0.682	-3.968	3.609	0.111	-4.184	3.288	0.745	-4.841	3.666	0.715	-3.813	4.009
Simple/PS	0.627	-6.407	4.178	-0.404	-4.236	3.955	-0.325	-4.588	3.909	0.908	-3.766	4.609	0.746	-3.984	4.631
Scenario 6: Number of clusters = 40 with 10 strata and 20 matched pairs (N=167)															
Pair/M and **Pair/U**†	0.719	-2.699	2.734	0.021	-4.037	4.756	0.185	-4.351	4.906	0.074	-3.376	2.927	0.371	-2.111	2.715
Stratified	0.731	-3.746	5.999	0.259	-4.177	4.262	0.114	-5.409	3.261	0.059	-5.480	8.110	-0.357	-3.687	4.084
Simple	-0.691	-4.572	3.897	-0.320	-5.205	5.249	0.062	-3.387	5.091	0.609	-3.475	4.745	0.821	-3.116	5.014
Simple/PS	-0.643	-5.657	3.201	-0.501	-5.164	5.541	0.319	-4.970	4.153	1.188	-4.128	4.131	1.137	-3.772	5.620

N = Number of design parameter combinations

† Matched-pair design with the unmatched analysis (**Pair/U**) produced the same estimates for the intervention as the matched-pair analysis for this study design (**Pair/M**)

Scenarios 1 to 3: Small number of large clusters

Scenarios 4 to 6: Large number of small clusters

Table 4: Guidance for choosing between the matched-pair, stratified and simple randomised designs for cluster randomised trials

Clusters per arm	**Weak matching correlations*** **(0.05, 0.1)**	**Strong matching correlations*** **(0.3, 0.5)**
< 4	Not recommended [22,44] †	
4 to 9	Simple Matched-pair with unmatched analysis‡	Stratified Simple without§ or with‖ post-stratification adjustment
10 to 20	Simple¶	Matched-pair Stratified** Simple with post-stratification adjustment‖

* Refer to Supplement 1 for determining the strength of the matching correlation

† Under powered to detect realistic effect sizes and not generalisable

‡ Stated *a priori* at the design stage

§ Simple design without adjustment when the matching correlation is 0.3

‖ When the number of clusters is large relative to the number of strata

¶ No efficiency gains in using restricted randomisation compared to simple randomisation

**Stratified design and matched-pair design are equally efficient when correlation between unmatched clusters within strata is as great as that between matched clusters

Figure 1: Mean width of the 95 percent confidence intervals (CI) of each study design with respective analytical methods* (*y*-axis) by the ratio of the correlation between unmatched clusters within strata to the correlation between matched clusters (*f*) § (*x*-axis) when the intra-cluster correlation is 0.3 and the sample cluster size is 50. The rows in the figure correspond to the six study design scenarios presented in Table 1 and the columns correspond to different values of the matching correlations (denoted as MC in the figure).

* See Methods section for the key to abbreviations for the study designs and respective analytical methods indicated by the different plot symbols given in the legend.

Matched-pair design with unmatched analysis (**Pair/U**) was not evaluated when the matching correlation was 0.3 and 0.5 because coverage was greater than the nominal level.

§ $f = \frac{\rho_U}{\rho_M}$

Note: The values on the *x*-axis are jittered slightly for each study design to avoid overlapping plotted lines and the scale on the *y*-axis changes for the different number of clusters (that is, 8, 20 and 40).

Figure 1

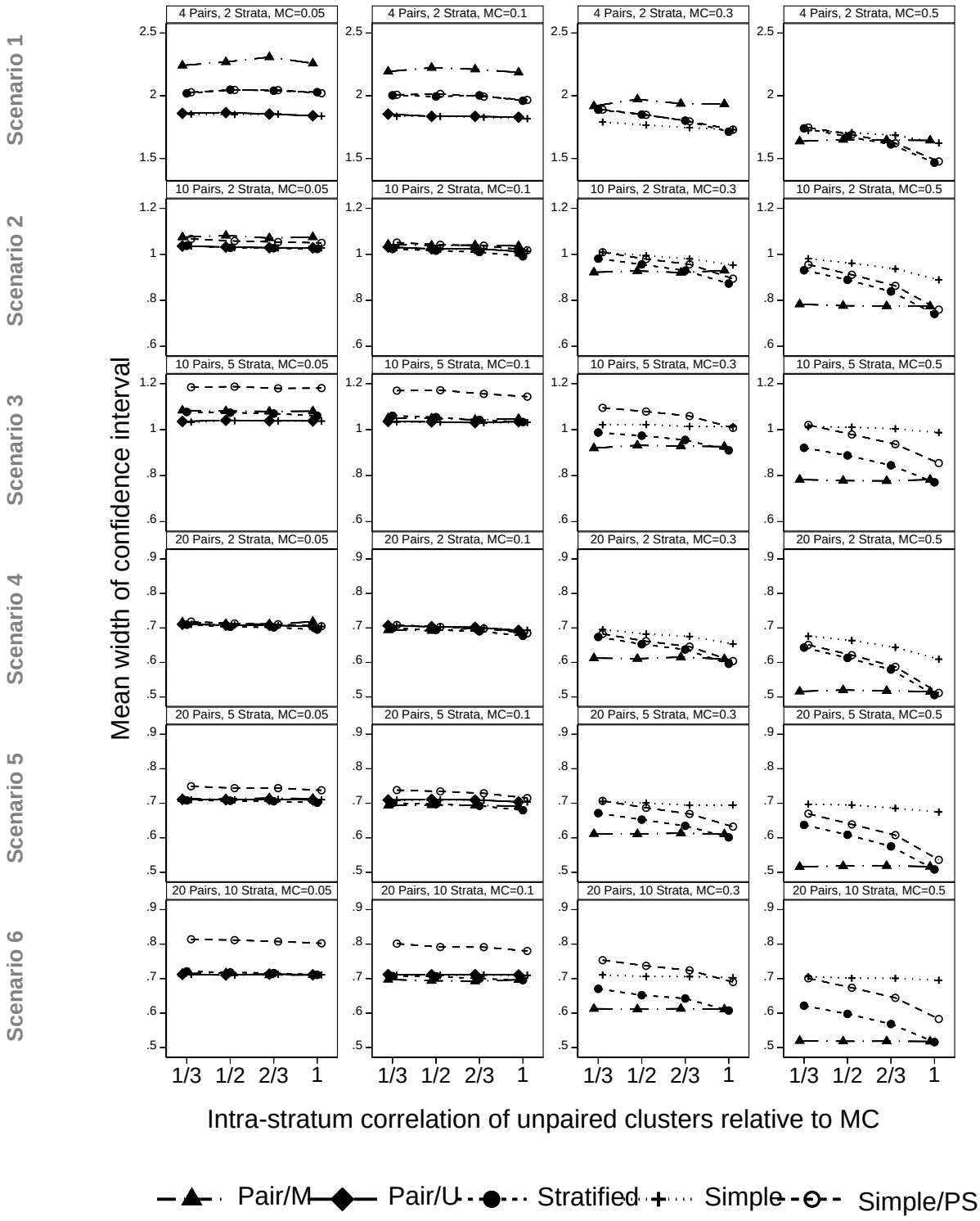


Figure 2: Relative efficiency* of each study design with respective analytical methods§ compared to the matched-pair design (*y*-axis) by the ratio of the correlation between unmatched clusters within strata to the correlation between matched clusters (*f*)† (*x*-axis) when the intra-cluster correlation is 0.3 and the sample cluster size is 50. The rows in the figure correspond to the six study design scenarios presented in Table 1 and the columns correspond to different values of the matching correlations (denoted as MC in the figure).

* Relative efficiency is the ratio of the mean confidence interval width of each study design with the respective analytical methods§ to the mean confidence interval width of the matched-pair design with the matched analysis. A ratio of one indicates that the study design provides the same mean confidence interval width (i.e., is as efficient) as the matched pair design, a ratio below one indicates the study design is more efficient than the matched-pair design, and a ratio above one indicates the study design is less efficient than the matched pair design.

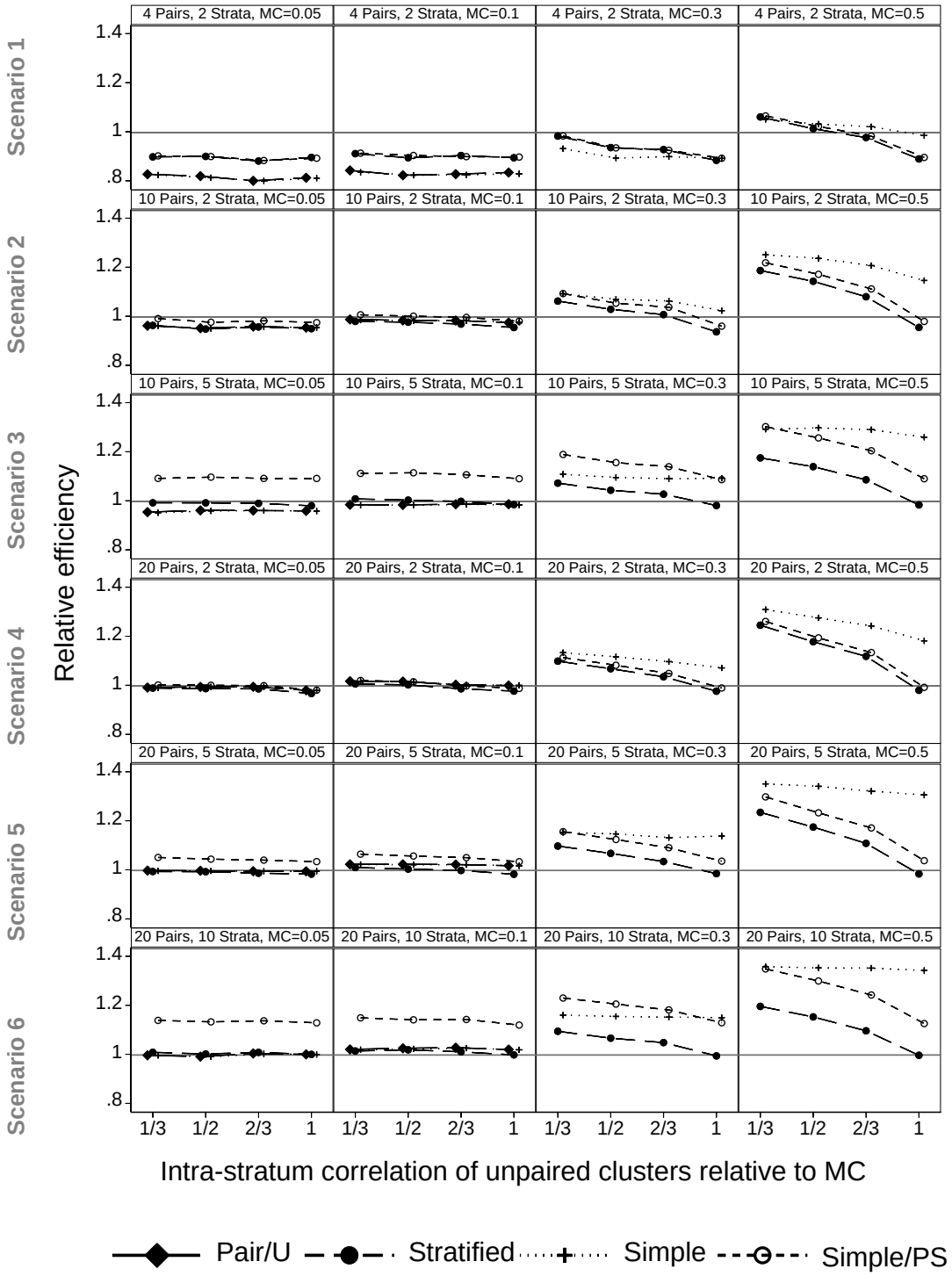
§ See Methods section for the key to abbreviations for the study designs and respective analytical methods indicated by the different plot symbols given in the legend.

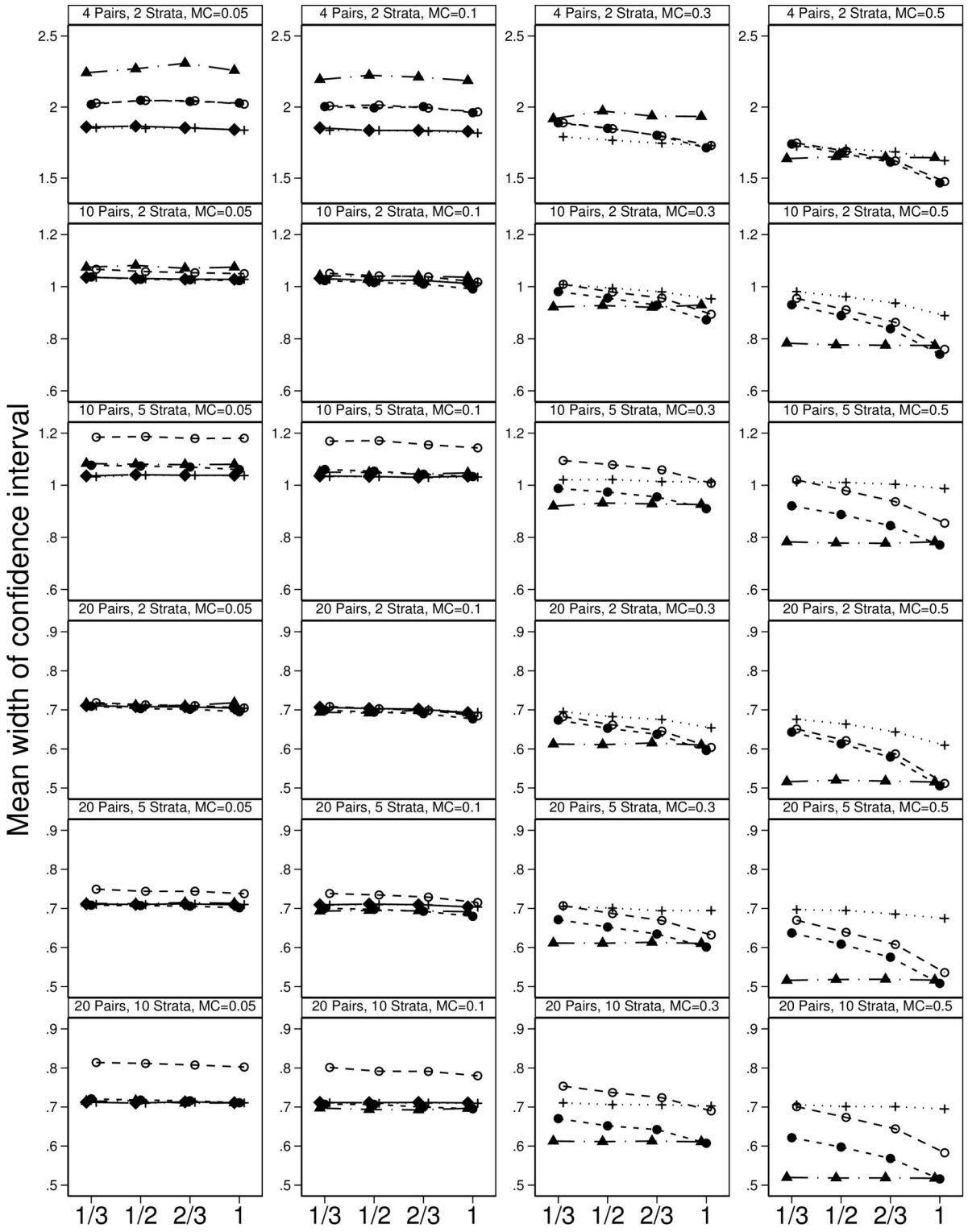
Matched-pair design with unmatched analysis (**Pair/U**) was not evaluated when the matching correlation was 0.3 and 0.5 because coverage was greater than the nominal level

† $f = \frac{\rho_U}{\rho_M}$

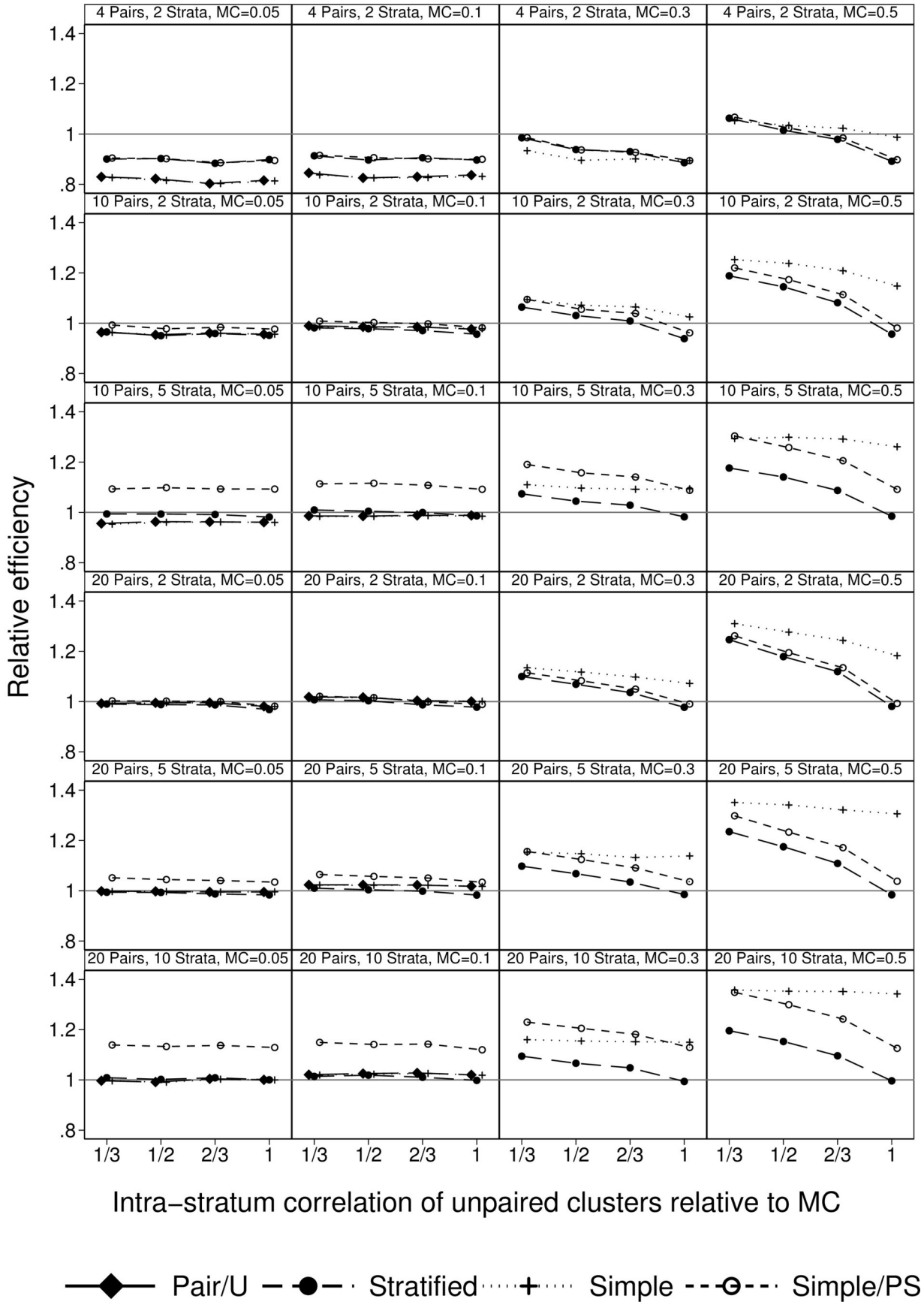
Note: The values on the *x*-axis are jittered slightly for each study design to avoid overlapping plotted line.

Figure 2





sim_9152_chondros_figure2.eps



sim_9152_chondros_figure3.eps