Published in final edited form as:

Contemp Clin Trials. 2014 September ; 39(1): 86–94. doi:10.1016/j.cct.2014.08.001.

On the design and analysis of clinical trials with correlated outcomes

Dean Follmann and Michael Proschan¹

¹Address for both authors: Biostatistics Research Branch, National Institute of Allergy and Infectious Diseases, 6700B Rockledge Drive MSC 7609, Bethesda, MD 20892
dfollmann@niaid.nih.gov

SUMMARY

The convention in clinical trials is to regard outcomes as independently distributed, but in some situations they may be correlated. For example, in infectious diseases, correlation may be induced if participants have contact with a common infectious source, or share hygienic tips that prevent infection. This paper discusses the design and analysis of randomized clinical trials that allow arbitrary correlation among all randomized volunteers. This perspective generalizes the traditional perspective of strata, where patients are exchangeable within strata, and independent across strata. For theoretical work, we focus on the test of no treatment effect $\mu_1 - \mu_0 = 0$ when the n dimensional vector of outcomes follows a Gaussian distribution with known $n \times n$ covariance matrix Σ , where the half randomized to treatment (placebo) have mean response μ_1 (μ_0). We show how the new test corresponds to familiar tests in simple situations for independent, exchangeable, paired, and clustered data. We also discuss the design of trials where Σ is known before or during randomization of patients and evaluate randomization schemes based on such knowledge. We provide two complex examples to illustrate the method, one for a study of 23 family clusters with cardiomyopathy, the other where the malaria attack rates vary within households and clusters of households in a Malian village.

Keywords

Correlated data; Clustering; Permutation test; Social Networks; Strata

1 Introduction

Clinical trials are traditionally designed and analyzed under the assumption that outcomes from the randomized participants are independent and identically distributed (iid). In practice, however, there are certain settings where it is natural to regard the outcomes as being correlated. For example, infectious diseases can be caused by a common source, so villagers using the same tainted well should tend to have more similar infection outcomes than strangers. Members of the same family may have similar outcomes based on shared

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

DNA and similar environment. Finally, one can also think of participants from the same site, or who have have similar covariates as having more correlated outcomes than participants who have dissimilar covariates.

The presence of correlation seems troubling for inference, as traditional analyses require the data to be iid on the null. However, Proschan & Follmann argued that correlation is not a concern provided the randomization process ignores the correlation, and thus conventional methods of analysis that ignore correlation are still valid [1]. Other work in this area has considered the effect of clustering, where patients in the same cluster are exchangeable but independent across clusters [2]. Examples include center or site or common baseline covariates. One can show that if randomization is stratified by cluster, the type I error rate can be conservative if the cluster is ignored in the analysis. Parzen et al showed that the usual/naive analysis that ignores clustering is appropriate if either there is no correlation within clusters, or no correlation between treatment assignment within clusters. If there is positive correlation between treatment indicators, e.g. cluster randomized trials, the naive analysis is anti-conservative, while if there is negative correlation between treatment indicators, e.g. crossover trials where each person is a cluster, then the naive analysis is conservative[3]. Proschan proved that the use of the unpaired t-test is conservative in the presence of paired data with nonnegative correlation [4].

The focus of this paper is to explore efficient ways to design and analyze clinical trials where the correlation structure is known or knowable. Unlike previous authors, we specifically allow for nonzero correlation among *all* randomized participants. Our theoretical work focuses on the Wald test based on the mle of the treatment effect $\mu_1 - \mu_0$ when the vector of outcomes follows an n -variate Gaussian distribution with known covariance Σ , where the half randomized to treatment (placebo) have mean response μ_1 (μ_0). We evaluate use of knowledge of Σ in the design and analysis of clinical trials for trials with paired or clustered data, and then evaluate two disparate and unconventional examples. We conclude with a discussion.

2 Model

Suppose we conduct a clinical trial with a continuous endpoint Y in n individuals, with $n/2$ randomized to control ($Z=0$) and $n/2$ randomized to an intervention ($Z = 1$). We allow for arbitrary correlation between the outcomes. Without loss, assume that $i = 1, \dots, n/2$ are control and the remainder treatment. We further assume that the vector of outcomes $\mathbf{Y}$ is multivariate normal with mean $\boldsymbol{\mu}' = \left(1'_{\frac{n}{2}}\mu_0, 1'_{\frac{n}{2}}\mu_1\right)$ and covariance Σ . We are interested in estimating μ_1, μ_0 and testing whether the treatment effect $\mu_1 - \mu_0$ equals zero.

We can write a model corresponding to this setup as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u},$$

where the first column of $\mathbf{X}$ is a vector of ones and the second column of $\mathbf{X}$ has the 1st $n/2$ elements 0 and the remainder 1. The parameter vector is $\boldsymbol{\beta} = (\mu_0, \mu_1)$, $E(\mathbf{u}) = 0$ and $\text{Cov}(\mathbf{u}) = \Sigma$.

Σ . Under this setup, the mle of β minimizes the quadratic form $(\mathbf{Y} - \mathbf{X}\beta)' \Sigma^{-1} (\mathbf{Y} - \mathbf{X}\beta)$ (see [5]) and we obtain

$$\hat{\beta} = (\mathbf{X}' \Sigma^{-1} \mathbf{X})^{-1} \mathbf{X}' \Sigma^{-1} \mathbf{Y},$$

with

$$\text{Cov}(\hat{\beta}) = (\mathbf{X}' \Sigma^{-1} \mathbf{X})^{-1}.$$

The estimate $\hat{\beta}$ is also known as the generalized least squares estimate (GLS). It can be derived even if u is not multivariate normal as the minimizer of the quadratic form provided $E(u) = 0$ and $\text{Cov}(u) = \Sigma$.

Comments

- Under independence or an exchangeable correlation structure, $\hat{\mu}_Z = \bar{Y}_Z$ for $Z = 0, 1$, where $\bar{Y}_Z$ is the sample mean in group Z and $\hat{\Delta} = \bar{Y}_1 - \bar{Y}_0$
- Under paired data we have $n/2 = n_d + n_c$ pairs where n_d is the number discordant with respect to treatment assignment, e.g. a pair randomized to T and C, and n_c is the number of pairs concordant with respect to treatment assignment.

Note that for the n_d discordant pairs, an estimate of Δ is the average paired difference which is

$$\bar{D}_d \stackrel{D}{=} N \left(\Delta, \frac{2\sigma^2(1-\rho)}{n_d} \right).$$

For the n_c concordant pairs, there must be an equal number of TT and CC. An estimate of Δ using the concordant pairs is the difference in sample means between TT and CC pairs:

$$\bar{D}_c \stackrel{D}{=} N \left(\Delta, \frac{2\sigma^2(1+\rho)}{n_c} \right).$$

Without knowledge of Σ we would estimate Δ with the overall average of all the data, which is

$$\hat{\Delta}_N = \bar{Y}_1 - \bar{Y}_0 = \frac{\bar{D}_d n_d + \bar{D}_c n_c}{n_d + n_c} \stackrel{D}{=} N \left(\Delta, \frac{n_d 2\sigma^2(1-\rho) + n_c 2\sigma^2(1+\rho)}{(n_d + n_c)^2} \right). \quad (1)$$

Call this the naive estimate. If $n_d = n_c = n/4$, the variance reduces to $4\sigma^2/n$.

The mle of Δ can be obtained by combining $\bar{D}_d, \bar{D}_c$ using inverse variance weights. The resultant estimate $\hat{\Delta}$ is

$$\hat{\Delta} \stackrel{D}{=} N \left(\Delta, \frac{2\sigma^2 (1 - \rho^2)}{n_c (1 - \rho) + n_d (1 + \rho)} \right). \quad (2)$$

If $n_d = n_c = n/4$ one can easily quantify the efficiency advantage of the mle based on knowledge of Σ to the estimate without such knowledge,

$$\frac{\text{var}(\hat{\Delta})}{\text{var}(\hat{\Delta}_N)} = 1 - \rho^2.$$

3 Design

At times, Σ will be known at the start of the study, and this information can be used to design an efficient trial. This is well known in simple settings. For example, if we have paired data, the optimal design is to randomize one in each pair to treatment and one to control. The paired difference $D_j, j = 1, \dots, n/2$ is normal with mean Δ and variance $2\sigma^2(1-\rho)$ and thus the average difference is

$$\bar{D} \stackrel{D}{=} N \left(\Delta, \frac{4\sigma^2 (1 - \rho)}{n} \right).$$

From (2) with $n_d = n_c = n/4$ we have that randomizing without knowledge of Σ but analyzing with knowledge of Σ gives us

$$\hat{\Delta} \stackrel{D}{=} N \left(\Delta, \frac{4\sigma^2 (1 - \rho^2)}{n} \right).$$

It follows that the relative efficiency of the two approaches, $\text{var}(\bar{D}) / \text{var}(\hat{\Delta}_N)$, is $1/(1 + \rho)$. Thus if $\rho = .5$ we can reduce the sample size by approximately 33% if Σ is known at the time of randomization. The approximation obtains because n_c will likely not exactly equal n_d .

With arbitrary Σ , the optimal randomization is difficult. In principle, one might consider all single block randomizations (i.e. with $n/2$ T and $n/2$ C), determine $\text{var}(\hat{\Delta})$ for each, order all randomizations by $\text{var}(\hat{\Delta})$, and then randomly select one of the top B designs. If a re-randomization p-value was desired, B should be large. Otherwise one could take the most efficient design and use a standard normal null distribution for the standardized test statistic. Note that there may be a large number of most efficient designs, as in the paired data example, or a large number of nearly optimal designs.

In general, the number of possible randomizations is very large, so enumeration is not really feasible. Consider the following “correlation paired randomization (CPR)” approach. We

identify the pair with the largest correlation and then randomly assign treatment and control within this pair. We exclude this pair and repeat with the remaining $n-2$ volunteers. We continue in this fashion until all pairs have been randomized. If n is odd, we can flip a coin for the last patient. For paired data with non-negative correlation, the CPR approach automatically recapitulates the usual blocked randomization within pairs.

In practice, one need only know the ranks of all pairwise correlations, not the actual values, to perform CPR. Suppose that $\Sigma = h(\theta)$ with unknown θ , but the ranks of the correlations are the same for all θ . Then one could randomize based on the correlation ranks and estimate θ after the data had been collected for analysis. Examples of such Σ s include spatial correlation, clustered data, and family structure.

At times, Σ (or the ranks of all correlations) may not be entirely known at the start of the trial but revealed sequentially. That is, prior to the treatment assignment of the m th volunteer, we may know the $m \times m$ covariance matrix for the first m volunteers Σ_m (see examples in later sections). We would also know the randomization assignments of the first $m-1$ volunteers $\mathbf{Z}_{m-1}$. Based on this information we could calculate the variance of the treatment effect estimate under assignment to treatment and control, say $\text{var} \left\{ \hat{\Delta} (Z_m=1) \right\}$ and $\text{var} \left\{ \hat{\Delta} (Z_m=0) \right\}$ respectively. We could choose the assignment that produces the smaller variance, or pick it with probability 3/4, say. The former is similar to the sequential randomization scheme of Atkinson [6] for trials where covariates are used in the design and analysis. Atkinson proposes to assign the next treatment assignment to minimize the variance of the treatment effect based on knowledge of the covariates.

Correlation paired randomization is not always the best randomization in terms of minimizing the variance of $\hat{\Delta}$. In the appendix we provide an extreme example with $n = 6$ where the variance of $\hat{\Delta}$ for the optimal $\mathbf{Z}$ approaches 0 as Σ approaches singularity, while both CPR and simple randomization have variances that remain greater than 0. Furthermore, for this extreme example one can show that simple randomization results in a smaller variance for $\hat{\Delta}$ than CPR. Thus CPR cannot be used uncritically and should be evaluated in the design stage to ensure it has acceptable properties for the problem at hand.

4 Evaluation of Simple Clusters

In this section we evaluate the performance of our estimators of θ using simple and correlation paired randomization under two different kinds of clustering.

- Paired Data. Here we have 25 pairs, each with correlation ρ but which are independent of each other. Such a study is typically analyzed using a paired t-test. It can be viewed as one kind of extreme Σ .
- Two strata. Here we have 2 strata, each with 25 patients, where outcomes within a stratum are correlated, but are independent across strata. Such a study is typically analyzed using a stratified t-test. It can be viewed as a different kind of extreme Σ .

For each scenario, 10,000 clinical trials are simulated. We evaluate the naive estimate, which ignores the correlation in the design and analysis, and we evaluate the mle under

simple randomization and also under CPR. For paired data, CPR corresponds to the randomization one uses in a crossover trial while for two clusters, CPR corresponds to stratified randomization where there are two strata, e.g. sites, of equal size. These settings are common in clinical trials and the approach of randomizing using CPR and an analysis using knowledge of Σ is what trialists generally do. So these simulations provide a reference for what would happen with a naive design and analysis, and provide some hints as to what behavior might occur in more complex settings.

For paired data and data with two clusters, we see a strong dependence of efficiency on ρ , a well known result. We see substantial gains in relative efficiency using correlation paired randomization compared to simple randomization for paired data, but no efficiency gains with data consisting of two clusters. Simple randomization with paired data results in only about 1/2 the pairs being discordant with respect to treatment assignment which is the optimal randomization that results in the minimum variance of $\hat{\Delta}$. For two clusters, the optimal randomization is to have a 50:50 split between T and C within each of the clusters. A simple randomization is quite likely to get close to this split even with only 50 patients.

5 Evaluation of Complex Structure

In this section we evaluate the performance of naive and maximum likelihood estimates of under simple and correlation paired randomization for two complex examples. We describe the data and correlation structure for the two examples in the first two subsections and then evaluate the different estimators in the last subsection.

5.1 NIH study of Familial Cardiomyopathy

A study at NIH examined the effect of different factors on a measure of heart size Y . One hundred twelve family clusters were obtained with 91 clusters of size 1 and the other 21 clusters ranging in size from 2 to 11. Within family clusters, the relationships among members were known and thus the expected amount of shared DNA could be determined. Nonidentical siblings should share about half their DNA as they get about 1/2 from father and 1/2 from mother. A Grandson-Granddaughter pair should share about one-quarter of their DNA, and so on. Figure 2 displays the degree of genetic sharing between the 21 (nonsingleton) family clusters. The smaller clusters of size 2 tend to share 1/2 or 1/4 of their genetic material while the larger clusters tend to include individuals with less sharing. The 1st and 3rd (top left to right) clusters of size 3 have pairwise correlations of 1/2, 1/2, and 1/4, while the 2nd and 4th clusters of size 3 have pairwise correlations of 1/8, 1/2, 1/4.

Suppose we want to design and evaluate a trial of an intervention on a future Y adjusted for baseline covariates and exploiting within family correlation. To fix ideas, this could be for this specific cohort, or for a different cohort where the family structure is known *a priori*. To proceed we need to have estimates of key design parameters. Towards this end, consider the model

$$Y_{ij} = \beta_0 + \beta_1 I(\text{female}_{ij}) + \beta_2 \text{age}_{ij} + \beta_3 \text{poly}_{ij} + \epsilon_{ij} \quad (3)$$

where i indexes family clusters, j indexes members within a family cluster, $I()$ is the indicator function, poly is the number of polymorphisms ($=0, 1$, or 2), and φ_{ij} is a Gaussian error term with mean 0 and variance σ^2 . We postulate a parametric correlation structure

$$\text{corr}(\epsilon_{ij}, \epsilon_{ik}) = \rho_{jk} = \frac{1}{1 + \exp(-\alpha_0 - \alpha_1 S_{jk})},$$

where S_{ij} is the expected proportion of shared DNA between persons j and k within the same cluster.

This correlation structure allows for a variety of smooth shapes that depend on S_{jk} and includes exchangeable ($\alpha_1 = 0$) and independent ($\alpha_0 = -1, \alpha_1 = 0$) structures. We further assume that volunteers from different families are independent. Based on this model, the mles for $(\alpha_0, \alpha_1) = (.47, .20)$. A likelihood ratio test of $\alpha_1 = 0$ has a p-value of .99 while a likelihood ratio test of $(\alpha_0, \alpha_1) = (\alpha_0, 0)$ versus $(0, 0)$ has a p-value $<.0001$. The within-pair correlation among pairs that share 1/64 of their DNA is estimated as $\rho = .61$, while the within-pair correlation among pairs that share 1/2 of their DNA is estimated as $\rho = .63$.

5.2 Malaria in a Malian Village

Malaria is an infectious disease spread by mosquitoes that carry the parasite *plasmodium falciparum*, which is transmitted to humans during a blood feeding. In certain areas, malaria is endemic and virtually all children are repeatedly infected over time. Infections can be cleared by drugs or a successful immune response, but infections can be severe and may lead to death.

Rural Malian villages are arranged in concessions, which are a collection of houses that are grouped together and generally walled off with areas for animals, cooking, sleeping, etc. Multiple extended families often live in concessions. Since malaria is caused by mosquito infection, it makes sense that children who live close together may tend to have similar attack rates, as they may be close to or far from the nearest source of mosquitoes. Additionally, children in the same family or concession may tend to have similar attack rates due to similar environment (e.g. bed nets), similar genes (e.g. hemoglobinopathies such as sickle cell that protect against malaria), or similar behavior (e.g. playing outside at dusk near the river or telling stories inside).

NIAID has a field site in Kenieroba, Mali and two nearby hamlets to study the natural history of malaria in children. In the future, such sites may be used to evaluate experimental malaria vaccines, so efficient designs are of great interest. The data consists of $n = 1506$ children who live in 388 households contained within 145 concessions. The largest household has 23 children while the median household has 3. The largest concession has 42 children in 8 houses with 1, 1, 1, 2, 5, 5, 9, and 18 children in each, respectively. The median concession has 8 children in two domiciles with 2 and 6, respectively. The number of cases of malaria over 4 seasons, Y , was recorded for each child and has a median of 2 with 10th and 90th percentiles equal to 0 and 7. Figure 3 presents a plot of the households and

concessions in Kenieroba. Each dot represents a child with darker dots representing larger Y s.

While Y is a count rather than a Gaussian response, the ideas of this paper can still apply with an appropriate model that accommodates correlation within concessions and households. We assume a random effects Poisson model where children, house, and concession each have random effects. Thus conditional on the random effects, the Poisson parameter for a child's count is

$$\exp(\beta_0 + b_{si} + b_{hi} + b_{ci}), \quad (4)$$

where b_{si} , b_{hi} , b_{ci} are independent mean zero Gaussian random variables with variances σ_s^2 , σ_h^2 , σ_c^2 , respectively. The random effect b_{hi} is the same for all i in the same house while b_{ci} is the same for all i in the same concession. We estimated this model with $(\sigma_s^2, \sigma_h^2, \sigma_c^2) = (.62, .21, .11)$. Likelihood ratio tests showed that all variances were significantly different from zero. About 2/3 of the extra-subject variability is due to house (.21/(.21+.11)), with concession contributing about 1/3 of this variance.

This model assumes that children within a house have a common correlation, children within a concession but from different houses have a common correlation, and children from different concessions are independent. Since $\sigma_h^2 > \sigma_c^2$ here, correlation paired randomization can be described as follows. For houses with an even number of children, we perform blocked randomization. For houses with an odd number of children we remove a child and form a blocked randomization for the remaining children. These removed children from the odd numbered houses who are all from the same concession are subject to a similar algorithm: if an even number of children are removed from the households within a concession a blocked randomization is performed. If an odd number of children are removed, one child is randomly further removed who receives the intervention with probability 1/2, the remaining children then receive blocked randomization.

5.3 Evaluation of Two Complex Examples

As before, we evaluate the naive estimate under simple randomization as well as the mle under CPR randomization. For the NIH cardiomyopathy study we restrict to the 21 clusters of Figure 2, plus a singleton cluster so there are 86 subjects. The correlation model is described in the previous section, and we evaluate the estimated correlation structure as well as perturbations from it. Overall, Σ is block diagonal with families defining the blocks. Figure 3 provides visualization of the correlation structures used in the simulations.

For the Malaria vaccine study we evaluate performance using data simulated using the model (4) using σ_p^2 , σ_h^2 and σ_c^2 as estimated in the data and perturbations from these estimates. We analyze the simulated data using a Poisson mixture model. Conditional on random effects, we assume that the mean count for a child follows a Poisson model with mean

$$\lambda_i = \exp(\beta_0 + Z_i\beta_1 + b_i + b_h + b_c),$$

where b_i, b_h, b_c are independent normals with variances $\sigma_p^2, \sigma_h^2, \sigma_c^2$, and $b_h (b_c)$ the same for all i within a common house (concession). The naive analysis ignores concession and house information by forcing $\sigma_h^2 = \sigma_c^2 = 0$.

As before, for each scenario 10,000 clinical trials are simulated and the sample variances of the different estimates calculated. Because the estimates are unbiased, we report the ratio of sampling variances with the naive randomization and analysis the benchmark.

Using the estimated correlation structure for the NIH family study, we see about a 50% reduction in the variance using the mle compared to the naive estimate, and about a 20% additional reduction with correlation paired randomization. The other correlation structures show decreasing efficiency with weakening correlation. The relative advantage of correlation paired randomization ranges from about 10% to 20% over these scenarios. Thus, use of the correlation structure here would result in a substantial improvement in the variance of the estimate of treatment effect.

For the malaria vaccine trial, we see that use of the mle compared to the naive analysis reduces the variance of the estimated treatment effect $\hat{\beta}_1$ by about 10% under simple randomization. If we additionally use Σ in randomization the variance is reduced by about 14% compared to a naive analysis using simple randomization. If the variance terms are increased, greater correlation is induced and there are greater gains in efficiency with methods that exploit Σ .

In principle, there may be randomization sequences that perform better than CPR; one pathological example with $n = 6$ is given in the appendix. To explore this issue with the NIH family study we performed an additional simulation with Σ set to its estimated value. We simulated m simple randomizations and identified the smallest variance of $\hat{\Delta}$, say

$W = \text{var}_{(1:m)}(\hat{\Delta})$. Think of this as the “best of breed” randomization sequence out of m . We

repeated this 30 times and calculated the mean ($\overline{W}$), variance (S_W^2) and 95% confidence interval for W ; i.e. $\overline{W} \pm 1.96S_W$. We varied m from 1 to 100,000 by factors of 10. We also simulated 30 randomly chosen CPR randomizations and, for each of the 30, recorded

$W = \text{var}_{CPR}(\hat{\Delta})$. We then calculated the mean, variance and 95% confidence interval for these 30 W s analogous to above. Figure 4 presents the results. We see that the expected “best of breed” decreases with m , as it must, and that with $m = 100,000$ the average

$W = \text{var}_{(1:100000)}(\hat{\Delta})$ lies above the confidence interval for the CPR estimate. CPR may be unbeatable for this setting, though proving it is complicated. In specific settings, simulations like these could be useful to help support the use of CPR.

6 Discussion

This paper discussed the design and analysis of clinical trials with arbitrary correlation between outcomes. In the simplest cases, this structure corresponds to familiar settings such as crossover trials or multi-center trials where randomization is stratified by site. Our formulation can be viewed as a smooth generalization of stratification in randomized trials. Strata in clinical trials are implicitly assumed to have exchangeable outcomes within strata and independence across strata. We generalize this concept by allowing different correlations between all pairs of volunteers within the trial. In more general settings where Σ is arbitrary, we show that efficiency gains are possible if knowledge about Σ is used. We introduce a new kind of randomization called correlation paired randomization, which can be used for general correlation structures and which recapitulates paired or stratified randomization in simple situations. This approach to randomization is *ad hoc* and not always optimal. Future research into Σ based randomization would be interesting. In practice, one should evaluate CPR or other randomization schemes to ensure they are acceptable in specific settings.

The practicality of CPR depends on the ability to estimate the correlation matrix, either in its entirety before the trial has begun, or sequentially as randomization proceeds. Both of our examples were for existing cohorts for whom the correlation could be estimated. The estimated correlation matrices could then be used to randomize these cohorts in a future trial for a future outcome. In some settings the primary endpoint Y may also be available at baseline, such as in a study of blood pressure, or viral load, say Y_0 . This would allow the sequential estimation of a parametric correlation structure based on Y_0 , as new patients are enrolled.

While this paper focused on correlation induced by clustering, family structures and spatial locations, other correlation structures may be uncovered using improved technology and awareness. With improvements in technology, it is increasingly possible to contemplate the identification and exploitation of “occult” correlation among trial participants. Examples include the use of social media, dynamic proximity of cell phone locations, or social networks see e.g. [7-9]. Specification and estimation of the correlation structures that govern outcomes in such settings could be a fruitful area for future research.

7 Appendix: An Extreme Covariance Matrix

Let $n = 6$, and suppose the highest to lowest correlations are: $\rho_{16} = \rho$, $\rho_{12} = \rho_{13} = \rho_{46} = \rho_{56} = \rho - \epsilon$, $\rho_{25} = \rho - 2\epsilon$, $\rho_{23} = \rho_{45} = \rho - 3\epsilon$, where $\rho > 0$ and all other pairs are uncorrelated. Thus, the correlation matrix is:

$$\Sigma = \begin{pmatrix} 1 & \rho - \epsilon & \rho - \epsilon & 0 & 0 & \rho \\ \rho - \epsilon & 1 & \rho - 3\epsilon & 0 & \rho - 2\epsilon & 0 \\ \rho - \epsilon & \rho - 3\epsilon & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & \rho - 3\epsilon & \rho - \epsilon \\ 0 & \rho - 2\epsilon & 0 & \rho - 3\epsilon & 1 & \rho - \epsilon \\ \rho & 0 & 0 & \rho - \epsilon & \rho - \epsilon & 1 \end{pmatrix},$$

Under CPR we sequentially randomize in pairs going from the largest to smallest correlation which here is (1, 6), (2, 5), and (3, 4).

Because ε is arbitrarily small, the correlation matrix can be made arbitrarily close to

$$\Sigma = \begin{pmatrix} 1 & \rho & \rho & 0 & 0 & \rho \\ \rho & 1 & \rho & 0 & \rho & 0 \\ \rho & \rho & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & \rho & \rho \\ 0 & \rho & 0 & \rho & 1 & \rho \\ \rho & 0 & 0 & \rho & \rho & 1 \end{pmatrix}, \quad (5)$$

The eigenvalues of (5) are 1 , $1 - 2\rho$, $1 + \rho\sqrt{2}$, $1 - 1 - \rho\sqrt{2}$, $1 + (1 + \sqrt{3})\rho$, and $1 + (1 - \sqrt{3})\rho$. Each of these must be positive because a nonsingular covariance matrix must be positive definite. This imposes the constraint:

$$\rho < 1/2.$$

The formula for $\text{var}(\hat{\Delta})$ as a function of the randomization sequence is given in Table A0 below. To use the table, divide each number in the “Randomization” column by 64, and then write its base 2 representation. For example, the numbers in the first row are 7 and 56; dividing by 64 and expressing the fractions in base 2 gives .000111 and .111000, respectively. Therefore, if after randomizing the pairs (1, 6), (2, 5), and (3, 4), we get

$$z_0 = (0, 0, 0, 1, 1, 1) \quad \text{or} \quad z_1 = (1, 1, 1, 0, 0, 0), \quad \text{then}$$

$$\text{var}_{z_0}(\hat{\Delta}) = \text{var}_{z_1}(\hat{\Delta}) = \frac{2(2\rho^2 - 1)}{4\rho - 3}. \quad (6)$$

As $\rho \rightarrow 1/2$, this variance approaches 1.

Table A0

Formula for the variance of $\hat{\Delta}$ for an extreme Σ as a function of ρ . The number under “Randomization” divided by 64 is the base 2 representation of the randomization sequence

Randomization	$\text{var}(\hat{\Delta})$	$\lim_{\rho \rightarrow 1/2} \text{var}(\hat{\Delta})$
7,56	$\frac{2(2\rho^2 - 1)}{4\rho - 3}$	1
22,26,37,41	$\frac{2(2\rho^2 - 1)}{4\rho^2 - 3}$	1/2

Randomization	$\text{var}(\hat{\cdot})$	$\lim_{\rho \rightarrow 1/2} \text{var}(\hat{\cdot})$
11,52	$\frac{2(1-2\rho^2)}{4\rho+3}$	1/5
21,25,38,42	$\frac{2(1-2\rho)(2\rho^2-1)}{4\rho^2+2\rho-3}$	0
13,14,19,28,35,44,49,50	$\frac{2(2\rho-1)(2\rho-3)(2\rho^2-1)}{4\rho^4-28\rho^3+11\rho^2+18\rho-9}$	0

The nonrandom rule

$$z_2 = (1, 0, 0, 1, 1, 0)$$

(i.e., assign patients 1,4, and 5 to treatment) corresponds to the number .100110 = 38/64, so from the fourth row of Table A0,

$$\text{var}_{z_2}(\hat{\Delta}) = \frac{2(2\rho^2-1)(1-2\rho)}{4\rho^2+2\rho-3}. \quad (7)$$

which approaches 0 as $\rho \rightarrow 1/2$. This shows that $\text{var}_{CPR}(\hat{\Delta}) / \text{var}_{z_2}(\hat{\Delta}) \rightarrow \infty$ as $\rho \rightarrow 1/2$ because $\text{var}_{CPR}(\hat{\Delta})$ is not 0.

For this setting, we can compare CPR i.e. randomizing (1, 6), (2, 5), and (3, 4) in pairs to a randomization that ignores Σ , i.e. randomizing using a single block of size 6. Table A1 reports the results. We see that for ρ approaching .5, CPR can have a larger variance than blocked randomization.

Table A1
Sampling variance for CPR and single block
randomization for an extreme Σ as a function of ρ

Assignment	$\rho = 0$	$\rho = .1$	$\rho = .25$	$\rho = .45$	$\rho = .49$	$\rho \rightarrow .5$
Random: one block	.6667	.6258	.5429	.3454	.2547	.2200
Random: CPR	.6667	.6166	.5226	.3560	.3121	.300
Z_1	.6667	.7538	.8750	.9917	.9996	1
Z_2	.6667	.5681	.3889	.0922	.0196	0

REFERENCES

1. Proschan MA, Follmann DA. Cluster without fluster: The effect of correlated outcomes on inference in randomized clinical trials. *Statistics in Medicine*. 2008; 27:795–809. [PubMed: 17594680]
2. Kahan BC, Morris TP. Improper analysis of trials randomized using stratified blocks of minimisation. *Statistics in Medicine*. 2012; 31:328–340. [PubMed: 22139891]

3. Parzen M, Lipsitz SR, Dear KBG. Does clustering affect the usual test statistics of no treatment effect in a randomized trial? *Biometrical Journal*. 1998; 40:385–402.
4. Proschan MA. On the distribution of the unpaired t-statistic with paired data. *Statistics in Medicine*. 1996; 15:1059–1063. [PubMed: 8783442]
5. Fomby, TB.; Hill, RC.; Johnson, SR. *Advanced Econometric Methods*. Springer-Verlag; New York: 1984.
6. Atkinson AC. Optimum biased coin designs for sequential clinical-trials with prognostic factors. *Biometrika*. 1982; 1982; 69:61–67.
7. Christakis NA, Fowler JH. The spread of obesity in a large social network over 32 years. *The New England Journal of Medicine*. 2007; 357:370–379. [PubMed: 17652652]
8. Christakis NA, Fowler JH. Social Network Sensors for Early Detection of Contagious Outbreaks. *PLOS One*. 2010; 5:1–8.
9. Hoff PD, Raftery AE, Handcock MS. Latent space approaches to social network analysis. *Journal of the American Statistical Association*. 2002; 97:1090–1098.

Figure 1. Genetic sharing for twenty-one family clusters from an NIH cardiomyopathy study. Line width reflects the amount of shared material, which can be $1/2$, $1/4$, $1/8$, $1/16$, $1/32$. Siblings or parent-child share $1/2$, grandfather-grandson share $1/4$, etc

Figure 2.

Malaria attack rates for children in the center of a Malian village. Each location denotes a child's house with darker coloring reflecting children with more malaria episodes. Locations are slightly perturbed to ease visualization.

Figure 3.

Correlation as a function of the proportion of shared DNA among family members under different correlation structures. Each curve corresponds to a scenario used in the simulation. Ticks on the x-axis correspond to individual family pairs. The flat bolded line corresponds to the estimated correlation.

Figure 4.

A comparison of the variance of $\hat{\Delta}$ under correlation paired randomization versus the minimum variance over m simple randomizations for the NIH cardiomyopathy study. The solid horizontal line gives the average and dashed horizontal lines the 95% confidence interval for $W = \text{var}_{CPR}(\hat{\Delta})$ based on 30 randomly chosen CPR sequences. For each m we

form an analogous average and confidence interval based on 30 replicates of $W = \text{var}_{(1:m)}(\hat{\Delta})$, the smallest variance out of m simple randomizations.

Table 2
Sampling variance ratios of the mle $\hat{\Delta}$ compared to the naive estimate $\hat{\Delta}_N$. The naive estimate ignores the correlation structure in the design and analysis. The mle uses Σ in the analysis, but not design, (Simple Randomization) or in the design and analysis (Correlation Paired Randomization)

Correlation Structure	$\text{var}(\hat{\Delta})/\text{var}(\hat{\Delta}_N)$	
	Simple Randomization	Correlation-Paired Randomization
Paired Data $\rho = .90$	.19	.10
Paired Data $\rho = .75$	.44	.25
Paired Data $\rho = .50$	.75	.50
Paired Data $\rho = .25$	.94	.75
Paired Data $\rho = .00$	1.00	1.00
Two clusters $\rho = .90$	.18	.18
Two clusters $\rho = .75$	.39	.39
Two clusters $\rho = .50$	.67	.66
Two clusters $\rho = .25$	.86	.86
Two clusters $\rho = .00$	1.00	1.00

Table 3
Sampling variance ratios of the mle $\hat{\Delta}$ for clinical trials compared to the naive estimate $\hat{\Delta}_{N\mathbf{N}}$. The naive estimate ignores the correlation structure in the design and analysis. We evaluate the mle under two different designs: simple randomization and correlation paired randomization. Covariance structures and other details are described in the test. Bolded settings correspond to the estimated parameters

Correlation Structure	$\text{var}(\hat{\Delta})/\text{var}(\hat{\Delta}_{N\mathbf{N}})$	
	Simple Randomization	Correlation-Paired Randomization
NIH family study $\alpha_0 = .47, \alpha_1 = .20$	.49	.40
NIH family study $\alpha_0 = 0, \alpha_1 = 4$	.20	.15
NIH family study $\alpha_0 = -1, \alpha_1 = 4$	.45	.34
NIH family study $\alpha_0 = -2, \alpha_1 = 4$	.75	.61
Malaria Vaccine Trial $\sigma_h^2 = .21, \sigma_c^2 = .10$	.89	.86
Malaria Vaccine Trial $\sigma_h^2 = .42, \sigma_c^2 = .20$	.75	.72
Malaria Vaccine Trial $\sigma_h^2 = .63, \sigma_c^2 = .30$	.66	.60