Resource Optimization in a Cluster Randomized Control Trial:
Simulation Study

Thomas B. Arnold

Abstract

Researchers conducting cluster randomized control trials (cRCTs), such as genetic sequencing studies where a sample is tested multiple times, must navigate a trade-off between the number of clusters and their size to optimize their budgets. As such, cost optimization is critical. In this simulation study, we used the ADEMP framework to explore cost optimization in a cRCT of the treatment effect on outcome Y. Borrowing the framework of a genetic sequencing study, we conceptualized clusters as individuals and samples within clusters as technical replicates. We examined the effect of changing the cost of adding individuals relative to the cost of adding technical replicates, as well as the effect of changing variance between individuals and variance of a given technical replicate from an individual's true value. We performed simulations using both normal and Poisson distributions, first varying one parameter at a time, and then through factorial designs varying multiple parameters. We found that, while the optimal number of individuals and technical replicates was determined more by these parameters' relative costs than by variance, they depended on variance in a much more complex way than expected, with specific values of variance changing the optimal number of individuals and technical replicates in a non-uniform way. This latter finding underscores that estimating variance between and within clusters prior to beginning a cRCT is critical to minimizing standard error.

Introduction

In a cluster randomized control trial (cRCT), experimental units are grouped into clusters, and each cluster is assigned to either the control or treatment group. Under certain circumstances, a cRCT may be more feasible and appropriate than a fully randomized control trial (RCT). A typical example would be a nutrition intervention implemented through a school's cafeteria; in this case, a cluster could be all students in a given school, with each school randomized to either the treatment or control group. Even if randomization of all students within the school were possible, doing so would pose a risk of treatment contamination (Breukelen and Candel (2018)).

Our study addresses a fundamental cRCT design question: given constrained resources, how many clusters, and what size of cluster, will be most cost-effective? Adding clusters will usually be more costly than increasing the size of a cluster. However, observations within a cluster may be correlated, which introduces a breadth-depth trade-off. The proportion of the total variance, explained by within cluster variance, is measured as the intracluster correlation coefficient (ICC). The greater the correlation within clusters, the more clusters will be required to detect the same effect size as a comparable RCT (Elley et al. (2005)).

This cost-optimization problem has been discussed in the literature (Liu et al. (2023), Breukelen and Candel (2018), Breukelen and Candel (2012), Raudenbush (1997)), including in the context of genetic sequencing studies (Zhang, Ntranos, and Tse (2020), Moriel, Memet, and Nitzan (2024), Kim, Kubal, and Schiebinger (2024), Rashkin et al. (2017), Rashkin et al. (2017)). In a genetic sequencing study, such as the one discussed by Dr. Wu, a cluster would be a biological replicate, such as blood collected from one individual, and the samples within that cluster would be technical replicates — repeated measurements of that same biological sample — such as reads of a genomic region. The breadth-depth trade-off in this context is between the number of biological replicates (breadth) and the number of technical replicates (depth) (Moriel, Memet, and Nitzan (2024)). Optimizing this balance is crucial, as

it has significant implications for the study's ability to produce meaningful results. While increasing the number of biological replicates provides a broader view of variability between samples (Zhang, Ntranos, and Tse (2020)), adding technical replicates (such as "deep sequencing" (Brug, Nalls, and Cookson (2010))) can help identify more rare variants per individual (Rashkin et al. (2017)), which is critical in genetic studies aiming to identify less common mutations, and reduce measurement error (Moriel, Memet, and Nitzan (2024), Wu (2024)).

In this simulation study, we used the ADEMP framework to explore cost optimization in an cRCT of the treatment effect (β) on outcome variable Y, examining which optimal combinations of number of clusters and number of units within a cluster minimize standard error (SE) of β. While the findings of this study are applicable to cRCT's across domains, we borrowed the framework of a sequencing study, conceptualizing clusters as "individuals" and units within clusters as "technical replicates."

We performed simulations to optimally select the number of individuals (G) and the number of technical replicates in each cluster (r), given a budget of B. To optimize these parameters, we determined the number of individuals and number of technical replicates that produce the lowest standard error (SE) of (β), where β=0.5. Our simulations varied costs of an additional individual ((c_1) and costs of an additional technical replicate (c_2), where $c_1 > c_2$. We also modulated variation between individuals (γ) and variation in a single measurement's deviation from an individual's true value (σ) (Wu (2024)). In addition to simulations varying one parameter at a time, we performed factorial designs varying multiple parameters.

We hypothesized that, where γ is low relative to σ, the optimal number of individuals will be smaller than the optimal number of technical replicates, and that as γ increases relative to σ, the optimal number of individuals will increase and the optimal number of technical replicates will decrease.

As for cost, we expected that, where c_2 is closer to c_1, the optimal number of individuals will be larger than the optimal number of technical replicates, unless γ is much smaller than σ. We also expected that, as c_2 decreases relative to c_1, the optimal number of individuals would decrease and the optimal number of technical replicates would increase.

Methods

Defining the Problem Space

To inform our simulation study, we first seek to understand the theoretical foundations of our problem space. We use the result from Raudenbush (1997) to examine the expected behavior of our estimator. Note that we select standard error (SE) as our estimand to avoid confusion when discussing variance as there is some overlap in terminology and the variances γ and σ are frequently referenced.

Consider a cluster randomized trial with G clusters and R observations per cluster. Let X_i be the treatment indicator for cluster i, and Y_{ij} be the jth observation from cluster i. In the normal case we have α as the overall mean, β as the treatment effect, γ^2 as the between-cluster variance, and σ^2 is the within-cluster variance.

In the Poisson case, α is the log baseline rate, β is the log rate ratio for treatment, and γ^2 is the between-cluster variance on log scale. $\sum_{j=1}^{R} Y_{ij}|\mu_i \sim \text{Poisson}(R\mu_i)$ is consequential because it tells us that when we take multiple measurements within a cluster in the Poisson case, we can analyze the sum of counts rather than individual measurements.

Normal Hierarchical Model

$$
\begin{aligned}
Y_{ij} &= \mu_i + e_{ij} \\
\mu_i &= \alpha + \beta X_i + \epsilon_i \\
\epsilon_i &\sim N(0, \gamma^2) \\
e_{ij} &\sim N(0, \sigma^2) \\
i &= 1, \ldots, G \\
j &= 1, \ldots, R
\end{aligned}
$$

Poisson Hierarchical Model

$$
\begin{aligned}
\sum_{j=1}^{R} Y_{ij} | \mu_i &\sim \text{Poisson}(R\mu_i) \\
Y_{ij} | \mu_i &\sim \text{Poisson}(\mu_i) \\
\log(\mu_i) &= \alpha + \beta X_i + \epsilon_i \\
\epsilon_i &\sim N(0, \gamma^2) \\
i &= 1, \ldots, G \\
j &= 1, \ldots, R
\end{aligned}
$$

Next, we derive the heroically standard error using the normal hierarchical case. Consider a cluster randomized trial with the hierarchical model $Y_{ij} = \mu + \beta X_i + \epsilon_i + e_{ij}$ where $\epsilon_i \sim N(0, \gamma^2)$ and $e_{ij} \sim N(0, \sigma^2)$ are independent. From Wu (2024) we know that our budget constraint can be defined as $B = G(c_1 + c_2(R-1))$. This can be solved for G as $G = \frac{B}{c_1 + c_2(R-1)}$. Following the derivation in Raudenbush (1997) we see that this leads to a standard error of the treatment effect estimator $\hat{\beta}$ of:

$$\text{SE}(\hat{\beta}) = \sqrt{\frac{4(\gamma^2 + \sigma^2/R) \times (c_1 + c_2(R-1))}{B}}$$

Also from Raudenbush (1997) we can see that $R_{optimal} = \sqrt{\frac{c_1 \sigma^2}{c_2 \gamma^2}} - 1$ and therefore $G_{optimal} = \frac{B}{c_1 + c_2(R-1)}$. Substituting this into our expression for the standard error we find that:

$$\text{SE}(\hat{\beta}) = \sqrt{\frac{4(\gamma^2 + \sigma^2/R)}{G}}$$

From this we see that larger R is optimal when σ/γ or c_1/c_2 is large, when $c_2 \ll c_1$, it becomes more efficient to take multiple measurements R within clusters, and finally that $\sigma \ll \gamma$ favors scenarios with many clusters G relative to R. Additionally, Raudenbush (1997) notes this is an unbiased estimator, which has an impact on the design of our simulation study insofar as we should choose an estimand which directly measures variance rather than a combination of variance and bias.

Finally, we use the theoretical definitions above to inform our parameters of interest. We can see from our theoretical derivations that α, β, while consequential to the estimation of our treatment effect $\hat{\beta}$, will not alter the SE($\hat{\beta}$). We do not present the results of varying B because B affects SE($\hat{\beta}$) in a very predictable way that is not meaningful and does not require a simulation to be understood. A larger budget produces a lower SE($\hat{\beta}$), due to the larger allowable sample. Based on our derivations, we hypothesize that c_1, c_2, σ, and γ will be the most consequential parameters to determining $R_{optimal}$ and $G_{optimal}$, and therefore we will focus on exploring these parameters.

ADEMP Framework

A. Aims

The primary objectives of this simulation study are to determine the optimal allocation between the number of clusters (G) and replicates per cluster (R) under budget constraints, minimize the standard error of treatment effect estimation across various cost and variance scenarios, and compare optimization

strategies between normal and Poisson distributions. Secondary aims include understanding the relationship between variance components (γ^2, σ^2) and optimal design choices, evaluating how cost ratios (c_1/c_2) affect allocation decisions, and assessing the robustness of optimization strategies across different parameter combinations.

D. Data-Generating Mechanisms

Two hierarchical models form the basis of our data generation. The Normal Hierarchical Model is structured as $Y_{ij} = \mu_i + e_{ij}$ where $\mu_i = \alpha + \beta X_i + \epsilon_i$, with $\epsilon_i \sim N(0, \gamma^2)$ and $e_{ij} \sim N(0, \sigma^2)$ for $i = 1, \ldots, G$ and $j = 1, \ldots, R$. The Poisson Hierarchical Model follows $Y_{ij}|\mu_i \sim \text{Poisson}(\mu_i)$ with $\log(\mu_i) = \alpha + \beta X_i + \epsilon_i$ where $\epsilon_i \sim N(0, \gamma^2)$.

The parameter space includes fixed parameters $\alpha = 1$ (intercept) and $\beta = 0.5$ (treatment effect), while varying parameters span ranges for between-cluster standard deviation $\gamma \in [0.5, 3]$ by 0.5, within-cluster standard deviation $\sigma \in [0.2, 2]$ by 0.3, first sample cost $c_1 \in \{20, 50, 100\}$, and additional sample cost $c_2 \in \{1, 10, 19\}$. These are constrained by a fixed total budget $B = 1000$ following $B = G(c_1 + c_2(R - 1))$.

E. Estimands

The study focuses on a single primary estimand: the standard error (SE) of the treatment effect estimate $\hat{\beta}$. This choice maintains interpretable units and emphasizes precision rather than bias, supported by theoretical work demonstrating unbiased estimation. The standard error provides a direct measure of the precision with which we can estimate the treatment effect under various design configurations.

M. Methods

Our methodology encompasses two key stages. First, we probe the design space using univariate modulation of our selected parameters. Second, we explore the design space in a factorial framework. We do so by systematically generating feasible combinations of G and R given the budget constraint, maintaining a minimum of $G = 7$ clusters for valid inference and maximum $R = 200$ replicates, and subsequently testing each of these valid combinations of G and R across multiple parameter values of c_1, c_2, σ, and γ. We do so by simulating data $n_{sims} = 1000$ times for each (G, R) combination, fitting appropriate models, calculating empirical SE, and selecting the (G, R) combination that minimizes SE. This process is repeated across the full parameter space to build a comprehensive understanding of optimal design choices.

We implement model fitting using linear mixed effects models (lmer) for the normal case and generalized linear mixed models (glmer) for the Poisson case, incorporating random effects for cluster-level variation and fixed effects for treatment. The implementation leverages parallel processing for simulation replicates with appropriate error handling for non-convergent models. Validation checks verify budget constraints, minimum cluster requirements, and monitor convergence issues. Special attention is paid to boundary conditions, tracking cases where optimal solutions occur at boundaries and identifying parameter combinations that lead to unstable solutions.

P. Performance Measures

Performance evaluation occurs across multiple dimensions. Primary measures include the empirical standard error of $\hat{\beta}$, model-based standard error estimates, and coverage probability of 95% confidence intervals. Secondary measures which are calculated and tracked but not reported or discussed include

coverage, total cost utilization, and model convergence. Our metrics assess the sensitivity of optimal designs to parameter changes and stability of solutions across simulation replicates.

Results

Aim 2: The Normal Case

Univariate Case

Factorial Design

Aim 3: The Poisson Case

We skip the Poisson univariate case because our conclusions are subsumed within the discussion of the factorial design below.

Factorial Design

Comparison Between Poisson and Normal

Limitations

We identify two [*change to three if necessary*] major limitations. First, we recognize that generalized linear mixed models can be volatile. While we attempted to solve this issue by simulating many times, it might be beneficial to try a Bayesian modeling approach, which could potentially fit more reliable models with fewer repeated simulations. This is a potential area for future work.

The second major limitation is boundary effects. Optimal r and optimal G in each simulation were chosen by minimizing the SE of the estimator. However, there are many cases where we see boundary effects, i.e., the optimal r and G jump in a discontinuous way, between numbers that are not near each other, when moving from the best to the second-best answer. The reason for tthis is that we are close to a boundary, such that there is not actually that much of a penalty for choosing a different combination of r and G. A more comprehensive analysis would consider these boundayr conditions, without focusing on only one optimal r and G combination minimzing SE, but rather including those combinations that come very close to doing so. This would better quanttify risk implicit in their study design, in terms of doing the best job possible estimating the treatment effect, to researchers. In sum, because our analysis focuses only on optimal r and optimal G, and not potential very close cases, it may inflate the perceived risk of choosing the wrong r and G from σ and γ. [*note - not sure the sigma/gamma part of this last sentence makes sense*]

Optional limitation* The third major limitation is the range of σ and γ values that we test. As discussed below, one of our conclusions is that cost ratio seems to be more consequential than ICC for the range of σ and γ values that we test. However, we do not have a biological or other justification for that range.

Conclusions

Note - i am not sure that everything captures the difference between univariate and factorial results/conclusions

Overall, univariate cases yield generally expected results but factorial design revealed unexpectedly complicated ones.

The directionality of the effects of varying parameters on optimal r and optimal G were as expected. Where γ was low relative to σ, optimal G was smaller than optimal r, and as γ increased relative to σ, optimal G increased and optimal r decreased. Where c_2 was closer to c_1, optimal G was larger than optimal r, unless γ was much smaller than σ. As c_2 decreased relative to c_1, optimal G decreased and optimal r increased.

Also, varying the underlying parameters for the data generating process (i.e., all parameters other than c_1, c_2, and B) had predictable effects on our ability to estimate β. Specifically, increasing σ and/or γ increased SE, making it harder to make a confident estimate of treatment effect.

note - add any other ways data generating parameters affected SE?

When cost ratio was high (when c_1 was much higher than c_2), the specific value of σ and γ changes optimal G and optimal r more than when cost ratio was low. Of note, in high cost ratio scenarios, the optimal r and G to minimize SE of the estimate changes in a non-linear way when varying (σ and γ). Cost ratio seems to be more consequential than ICC (σ and γ) on optimal G and optimal r, at least within the values of σ and γ that we test. When the cost ratio is low, σ and γ behave in a more linear fashion. When cost ratio is high, we observe more boundary conditions: the volatility of the results in terms of optimal r and G in a high cost ratio scenario is higher than in a low cost ratio.

Optimal G and optimal r depended on ICC in a much more complex way than expected, with specific values of σ and γ changing optimal G and optimal r in a non-uniform way. We observed a non-linear relationship, with peaks and valleys, where there are different answers very near each other. [*note - the sentence right before this note needs to be clarified*] The unexpected complexity of the relationship between variance and optimal G and optimal r suggests that estimating σ and γ prior to beginning a cRCT is critical to minimizing standard error.

In sum, although the effect of c_1 and c_2 on optimal r and G was easier to understand, understanding variance was more consequential. This underscores the importance of performing pilot studies before large-scale, high-budget studies: knowing γ and σ, which are much harder to estimate in advance than c_1 and c_2, can be critically important. Even if one knows c_1 and c_2 in advance of the study, if one estimates σ and γ incorrectly, this may produce a configuration of r and G that appears to optimize the budget butt does not minimize SE.

References

Breukelen, Gerard JP van, and Math JJM Candel. 2012. "Calculating Sample Sizes for Cluster Randomized Trials: We Can Keep It Simple and Efficient!" *Journal of Clinical Epidemiology* 65 (11): 1212–18.

———. 2018. "Efficient Design of Cluster Randomized Trials with Treatment-Dependent Costs and Treatment-Dependent Unknown Variances." *Statistics in Medicine* 37 (21): 3027–46.

Brug, Marcel van der, Michael A Nalls, and Mark R Cookson. 2010. "Deep Sequencing of Coding and Non-Coding RNA in the CNS." *Brain Research* 1338: 146–54.

Elley, C Raina, Ngaire Kerse, Patty Chondros, and Elizabeth Robinson. 2005. "Intraclass Correlation Coefficients from Three Cluster Randomised Controlled Trials in Primary and Residential Health Care." *Australian and New Zealand Journal of Public Health* 29 (5): 461–67.

Kim, Jakwang, Sharvaj Kubal, and Geoffrey Schiebinger. 2024. "Optimal Sequencing Depth for Single-Cell RNA-Sequencing in Wasserstein Space." *arXiv Preprint arXiv:2409.14326*.

Liu, Jingxia, Lei Liu, Aimee S James, and Graham A Colditz. 2023. "An Overview of Optimal Designs Under a Given Budget in Cluster Randomized Trials with a Binary Outcome." *Statistical Methods in Medical Research* 32 (7): 1420–41.

Moriel, Noa, Edvin Memet, and Mor Nitzan. 2024. "Optimal Sequencing Budget Allocation for Trajectory Reconstruction of Single Cells." *Bioinformatics* 40 (Supplement_1): i446–52.

Rashkin, Sara, Goo Jun, Sai Chen, Genetics, Epidemiology of Colorectal Cancer Consortium (GECCO), and Goncalo R Abecasis. 2017. "Optimal Sequencing Strategies for Identifying Disease-Associated Singletons." *PLoS Genetics* 13 (6): e1006811.

Raudenbush, Stephen W. 1997. "Statistical Analysis and Optimal Design for Cluster Randomized Trials." *Psychological Methods* 2 (2): 173.

Wu, Zhijin. 2024. "November 12, 2024 Presentation to PHP 2550."

Zhang, Martin Jinye, Vasilis Ntranos, and David Tse. 2020. "Determining Sequencing Depth in a Single-Cell RNA-Seq Experiment." *Nature Communications* 11 (1): 774.

Appendix I: Script Code

```
# --- Preamble ---
# Date of last update: Dec. 8, 2024
# R Version: 4.3.1
# Package Versions:
#   tidyverse: 2.0.0
#   knitr: 1.45
#   kableExtra: 1.3.4
#   ggplot2: 3.4.3
#   lme4 1.1-34
#   ggrepel 0.9.6
#   latex2exp 0.9.6
#   parallel 4.3.1
#   doParallel 1.0.17


setwd("~/GitHub/cluster_RCT_sim")

# Knitr Engine Setup
knitr::opts_chunk$set(message=F,
                      warning=F,
                      error=F,
                      echo=F,
                      fig.pos = "H" ,
                      fig.align = 'center')

# Packages
options(kableExtra.latex.load_packages = FALSE) # Required to avoid floatrow error
library(knitr)
library(kableExtra)
library(ggplot2)
library(ggrepel)
library(latex2exp)
library(lme4)
library(tidyverse)
library(parallel)
library(doParallel)

# seed
set.seed(42)
```