Project 3

Yingxi Kong

Abstract

Introduction

Cluster Randomized Trials (CRTs) are a common design approach in public health research. In CRTs, clusters are randomly assigned to either a treatment or a control group, and the treatment effect is estimated by comparing outcomes between those groups. This method is well suited to testing differences in a method or approach to patient care (as opposed to evaluating the physiological effects of an intervention)[1]. However, balanced trade-off between the number of clusters and the number of observations within each clusters becomes a challenge in designing CRTs. People argues that increasing the number of clusters enhances the ability to detect treatment effects by reducing variability

This study, conducted in collaboration with Dr.Zhijin Wu, aims to explore optimal experimental design under certain budget constraint to maximize the precision of the treatment effect estimation, specifically focusing on how varying parameters such as sampling costs, between- and within-cluster variability, and underlying outcome distributions impact the precision of treatment effect estimates. By employing a comprehensive simulation framework, we will investigate the trade-offs between the number of clusters and the number of observations per cluster. We consider scenarios with normally distributed outcomes and extend the analysis to Poisson-distributed outcomes, reflecting a range of potential real-world application and providing more applicable insights for designing efficient and cost-effective CRTs.

Methods

This study aims to investigate how to optimally select the number of clusters and the number of observations within each cluster under budget constraints in CRTs. In addition, we aims to explore how the optimal design changes as varying parameters like cost structures, levels of within- and between-cluster variability, and treatment effect. Specifically, this study would address two key objectives: (1). determining the optimal allocation of resources under a fixed budget limit and (2). understand how variations in underlying data generation mechanisms affect the precision of treatment effect estimation. To achieve these, we performed a comprehensive simulation framework guided by the ADEMP framework, ensuring the consistency and clarity when evaluating different design scenarios.

Background

In designing our trial, we identify budget, B in dollars, as the given resource limit where B might also represents time or other resources required for the data collection process. In each cluster, the first sample cost c_1 and other additional samples in the same cluster cost c_2 where $c_2 < c_1$. Additionally, as we mentioned earlier, samples in the same cluster must be assigned to the same treatment or control group and their measurement might be correlated. Similarly, in sequencing data, samples in clusters refers to the repeated measurements from the same biological sample which are highly correlated and different clusters refers to different samples.

We would consider settings where Y is normally distributed and where Y follows a Poisson distribution. G $(i = 1, ..., G)$ represents the number of clusters where R $(j = 1, ..., R)$ represents the number of observations

within each cluster. Additionally, X_i is a binary indicator of the treatment assignment for cluster i where $X_i = 0$ for control and $X_i = 1$ for treatment, and Y_{ij} represents the observed outcome.

ADEMP Framework

- **Aims:** The primary aim of this simulation study is to identify optimal combination of cluster size and the number of observations per cluster under budget constraint, maximizing the precision of treatment effect estimation. In addition, we aims to explore how other data generation parameters, like cost ratio, variance, baseline mean, and treatment effect, would impact the optimal design.
- **Data-Generating Mechanism:**
  - Normally Distributed Outcomes:

    For observation j (j = 1, ..., R) in cluster i (i = 1, ..., G), we have:

    $$\mu_i = \alpha + \beta X_i + \epsilon_i \text{ with } \epsilon_i \sim N(0, \gamma^2), \text{ that is } \mu_i \sim N(\mu_{i0}, \gamma^2)$$

    $$Y_{ij} = \mu_i + e_{ij} \text{ with } e_{ij} \sim iid\ N(0, \sigma^2), \text{ that is } Y_{ij}|\mu_i \sim N(\mu_i, \sigma^2)$$

    Here, α represents the baseline mean value while β represents the treatment effect. γ^2 indicates the between-cluster variability and σ^2 captures the within-cluster variability.
  - Poisson Distributed Outcomes:

    For observation j (j = 1, ..., R) in cluster i (i = 1, ..., G), we have

    $$log(\mu_i) = \alpha + \beta X_i + \epsilon_i \text{ with } \epsilon_i \sim N(0, \gamma^2)$$

    $$Y_{ij}|\mu_i \sim Poisson(\mu_i)$$
- **Estimand:** The primary estimand for both outcome types is the treatment effect β, the difference in the expected outcome between the treatment and the control groups.
- **Methods:** We identify a fixed budget limit value $B = \$10,000$ for all simulation scenarios, and the cost for the first sample in each cluster fixed at $c_1 = \$20$. To evaluate the trade-offs between G and R, we varied G and the cost ratio c_1/c_2. The total cost for each design was calculated using the following formula:

  $$Total\ Cost = G \times c_1 + G \times (R-1) \times c_2$$

  To satisfy the budget constraint, the total cost should be restricted to:

  $$\text{Total cost } \leq B$$

  Using these criteria, we computed the maximum feasible number of observations per clusters (R) for each pair of G and c_1/c_2. For each combination of G, R, and cost ratio, we generated data for 100 iterations and fit a hierarchical model on each to identify the optimal design which maximizes precision within the budget constraints.

  Additionally, using the identified optimal design, we further varied key data generation parameters (γ^2, σ^2, α, and β) independently to explore their influence on model precision. Data were generated for 100 iterations under each parameter variation, and a hierarchical model was applied to investigate the impact. Similar steps are performed for Poisson-distributed outcomes. The Results section will detail these steps further and provide insights from the simulations.

- **Performance Measures:** The primary measures of performance is the variance of the treatment effect estimate, which reflects the precision of the estimate. We also incorporate metrics like bias, MSE, and coverage to provide a more comprehensive evaluation of our models.

Results

To facilitate data simulation under various scenarios, we developed custom data simulation functions for both normally distributed and the Poisson-distributed outcomes which allow flexible specification of key parameters, such as G, β, γ^2, and σ^2. These functions are located in the `Data Simulation.R` script within the `R` folder of the project directory. In addition, we developed custom experiment functions for each outcome distribution located in the `Experiment.R` script within the `R` folder of the project directory. These functions take the simulated data, fit the appropriate model, and calculate performance metrics, such as variance, bias, mean squared error (MSE), and coverage, across scenarios. This approach ensures a reproducible process for simulation, modeling, and model performance evaluation across diverse experimental designs.

Normally Distributed Outcome

Vary G, R, c1/c2

Start with normally distributed outcome, we explore a range of design scenarios by varying number of clusters (G = 5, 10, 15, 20, 25, 30) and cost ratios (c_1/c_2 = 2, 5, 10, 20). The underlying data generation parameters were fixed ($\alpha = 0$, $\beta = 1.5$, $\gamma^2 = 1$, and $\sigma^2 = 1$), with a budget limit of B = \$10,000 and the cost of the first sample in each cluster set at c_1 =\$20.

For each scenario, utilizing the data simulation function, the maximum feasible R was calculated based on the budget and the cost structure of our design. Clusters were randomly assigned to treatment or control groups in each iteration, ensuring representation of both groups, and outcomes were simulated using the hierarchical model for normally distributed data. The simulated datasets were saved as CSV files in the `Data_normal` folder for further analysis. Each data was then processed through our experiment function, where a hierarchical model was fit using the `lmer()` function from the `lmerTest` package. All performance metrics, including variance, bias, MSE, and coverage, were calculated and saved in the `Results_normal` folder for further evaluation.

From the range of design scenarios explored, we selected key results to present our key findings shown in `Table 1` and `Table 2`. The full results located within the `Table Results` folder, denoted as `summart_metrics_df.csv`. B, R, and the cost ratio are relatively interact with each other under fixed budget limit. With the same cost ratio, larger number of clusters groups results less number of observations in each cluster. Similarly, with fixed number of cluster, higher cost ratio allows for more observations in each cluster.

Observing `Table 1`, keeping the cost ratio constant, increasing the number of clusters decreases the number of observations in each cluster. The variance of β decreased substantially with higher G, dropping from 0.996 at G = 5 to 0.147 at G = 30, indicating improved precision with more clusters. Bias remains relative low across scenarios, with a maximum of 0.172 at G = 10 and a minimum value of 0.027 at G = 25. MSE decreased consistently as G increased while coverage gradually improved with higher G.

Additionally, in `Table 2`, keeping the number of cluster fixed at 20, the number of observations in each cluster as the cost ratio becomes higher. The variance of β and MSE slightly decreases as the cost ratio increases, indicating improved precision. Bias remains low without a clear trend while the coverage improved consistently as the cost ratio increases.

Table 1: Metrics Summary Varying Number of Clusters

Number of Cluster (G)	Number of Observations per Cluster (R)	Cost Ratio (c1/c2)	Beta Variance	Beta Bias	MSE	Coverage
5	991	10	0.9664506	0.0813709	0.9634073	84
10	491	10	0.4600079	0.1723095	0.4850983	92
15	324	10	0.2668408	0.0489170	0.2665652	91

Table 1: Metrics Summary Varying Number of Clusters *(continued)*

Number of Cluster (G)	Number of Observations per Cluster (R)	Cost Ratio (c1/c2)	Beta Variance	Beta Bias	MSE	Coverage
20	241	10	0.2236300	0.1111983	0.2337588	93
25	191	10	0.1436221	0.0272516	0.1429285	95
30	157	10	0.1467200	0.0584557	0.1486699	92

Table 2: Metrics Summary Varying Cost Ratio

Number of Cluster (G)	Number of Observations per Cluster (R)	Cost Ratio (c1/c2)	Beta Variance	Beta Bias	MSE	Coverage
20	49	2	0.2452454	0.0135748	0.2429772	90
20	121	5	0.2150198	0.0171457	0.2131636	94
20	241	10	0.2236300	0.1111983	0.2337588	93
20	481	20	0.1991018	0.0305745	0.1980456	96

Similar findings can be observed from `Figure 1` and `Figure 2`. The variance of β decreases sharply with increasing G, specifically for smaller number of clusters with $G \leq 15$. Variance trends are consistent across different cost ratios, with minor differences at higher cluster counts. In addition, observing `Figure 2`, for certain values of G (G = 5, 15, 25), the variance of β initially decreases as R increases. However, once R surpasses a certain threshold, a slight increase in variance is observed.

Figure 1: Beta Variance by Number of Clusters (G)

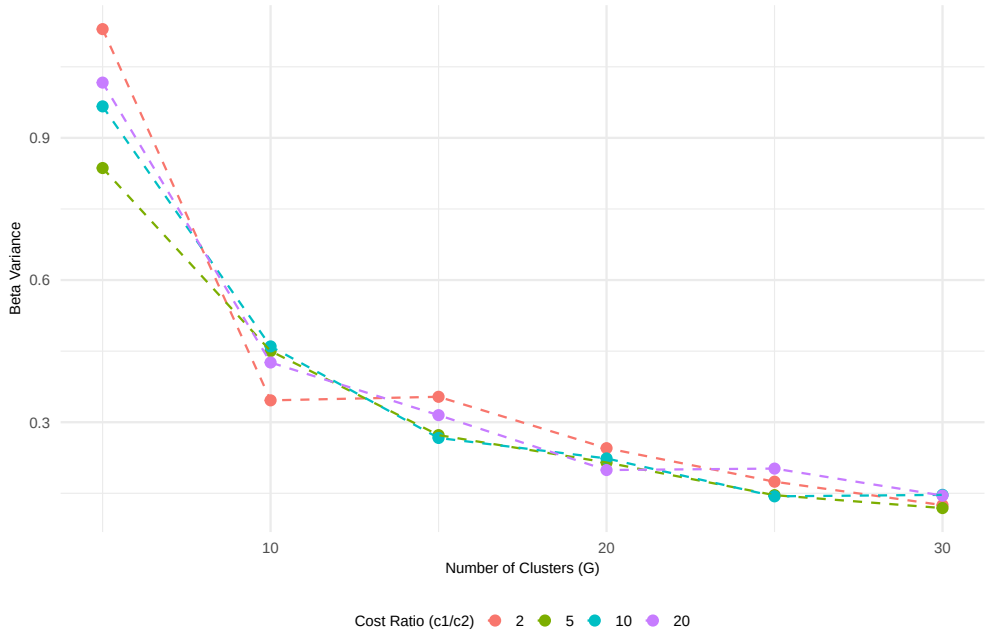
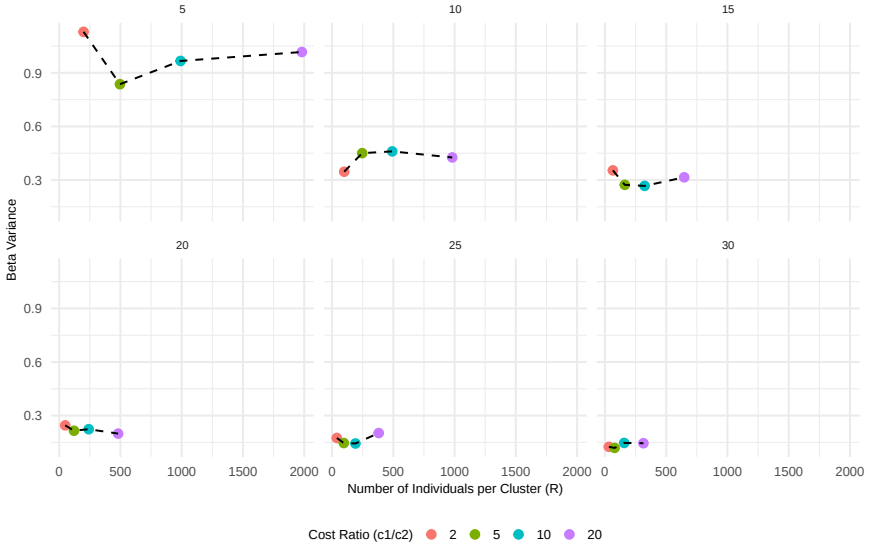


Figure 2: Beta Variance vs. R by Different G



Comparing the results across all combinations of design parameters, we identified the design with G = 30, R = 79, and cost ratio = 5 as the optimal configuration. This design exhibits the lowest variance of β and relatively lower values for other performance metrics, as shown in `Table 3`. The following variation on underlying data generation parameters would be based on this scenario.

Table 3: Metrics Summary Varying Cost Ratio

Number of Cluster (G)	Number of Observations per Cluster (R)	Cost Ratio (c1/c2)	Beta Variance	Beta Bias	MSE	Coverage
30	79	5	0.1188668	0.0163207	0.1179445	95

Vary Between-cluster Variance

Starting from varying between-cluster variance (γ^2), we keep other parameters constant ($\alpha = 0$, $\beta = 1.5$, and $\sigma^2 = 1$). Considering instra-cluster correlation (ICC), defined as:

$$ICC = \frac{\gamma^2}{\gamma^2 + \sigma^2}$$

We create scenarios with difference ICC values to investigate the impact of between-cluster variability on performance metrics. By varying γ^2, we systematically adjusted the ICC to represent low, moderate, and high levels of between-cluster correlation. This allowed us to explore how increased clustering affects the precision and bias of the treatment effect estimate (β) across various design configurations.

- Low ICC: $\gamma^2 = 0.111$, $ICC = \frac{0.111}{0.111+1} = 0.1$
- Medium ICC: $\gamma^2 = 1$, $ICC = \frac{1}{1+1} = 0.5$
- High ICC: $\gamma^2 = 10$, $ICC = \frac{10}{10+1} = 0.9090$

Table 4: Metrics Summary Varying Between-cluster Variance

Between-cluster Variance	Within-cluster Variance	Beta Estimate	Beta Variance	Beta Bias	MSE	Coverage
0.111	1	1.476395	0.0184819	0.0236049	0.0188543	95
1.000	1	1.484544	0.1433764	0.0154559	0.1421815	96

Table 4: Metrics Summary Varying Between-cluster Variance *(continued)*

Between-cluster Variance	Within-cluster Variance	Beta Estimate	Beta Variance	Beta Bias	MSE	Coverage
10.000	1	1.350199	1.5012129	0.1498011	1.5086411	91

In `Table 4`, fixing within-cluster variance, the estimated treatment effect remains close to the true value, 1.5. The variance of β increases substantially as γ^2 increases, indicating reduced precision with higher between-cluster variability. The bias increases slightly as γ^2 increase as well, with a maximum value of 0.1498 at $\gamma^2 = 10$. Additionally, the MSE increases with higher γ^2 and the coverage remains around 95%-96% for lower γ^2 while decreases slightly to 91% at $\gamma^2 = 10$. A larger γ^2 indicates greater variability between clusters, meaning clusters differ more significantly from each other. This increased variability makes it harder to isolate the treatment effect from the random effects introduced by the cluster differences, thereby reducing the precision of the treatment effect estimate.

Vary Within-cluster Variance

To vary the within-cluster Variance (σ^2), we remain other parameter constant ($\alpha = 0$, $\beta = 1.5$, and $\gamma^2 = 1$). Again, we decided the variance value based on the ICC levels:

- Low ICC: $\sigma^2 = 9$, $ICC = \frac{1}{1+9} = 0.1$
- Medium ICC: $\sigma^2 = 1$, $ICC = \frac{1}{1+1} = 0.5$
- High ICC: $\sigma^2 = 0.111$, $ICC = \frac{1}{1+0.111} = 0.9001$

Table 5: Metrics Summary Varying Within-cluster Variance

Between-cluster Variance	Within-cluster Variance	Beta Estimate	Beta Variance	Beta Bias	MSE	Coverage
1	9.000	1.436137	0.1642759	0.0638627	0.1667116	92
1	1.000	1.517234	0.1636513	0.0172340	0.1623118	93
1	0.111	1.487674	0.1420927	0.0123261	0.1408237	93

In `Table 5`, we observed that as the within-cluster variance increase, the estimated treatment effect is still close to the true value while the variance of β exhibits a slight increase, same for the MSE. Bias presents a increasing tread as σ^2 increases while the coverage remains consistent. A lower within-cluster variance suggests that members within a cluster are more similar to each other, leading to reduced noise in the data and more precise estimates of the treatment effect.

In conclusion, with normally distributed outcome, our findings demonstrate the effect of design parameters and variance components in cluster randomized trials. Increasing the number of cluster improves the precision of the treatment effect estimate. Additionally, higher between-cluster variance reduces precision. Conversely, lower within-cluster variability enhances precision. We did not observe obvious trend on variance as we varied the baseline mean value (α) and the treatment effect (β). The baseline mean value represents the overall average outcome. Changing in this would shift the outcome distribution but would not affect the variability in the data. Similarly, although β represents the magnitude of the treatment effect, it would not directly affect the variance of β. These insights underscore the importance of carefully balancing design parameters and variance components to achieve reliable and precise treatment effect estimates.

Poisson Distributed Y

Extending our analysis to the Poisson Distributed outcome, we explore a range of design scenarios by varying number of clusters (G = 5, 10, 15, 20, 25, 30) and cost ratios ($c_1/c_2 = 2, 5, 10, 20$). The underlying data generation parameters were fixed ($\alpha = 0$, $\beta = 1.5$, $\gamma^2 = 1$, and $\sigma^2 = 1$), with a budget limit of B = \$10,000

and the cost of the first sample in each cluster set at c_1 =\$20. Performing same steps as normally distributed outcomes, simulated data are located in the `Data_possion` folder and the performance metrics results were stored in the `Results_poisson` folder.

`Figure 3` exhibits similar findings as we observed for the normal distributed outcome. Increasing G substantially reduces the treatment effect variance, particularly for smaller G values. Additionally, various cost ratio does not present obvious effect on the model precision. In `Figure 4`, when we fixed the number of clusters, we observed that the treatment effect variance shows a slight decreasing trend as R increases.

Figure 3: Beta Variance by Number of Clusters (G)

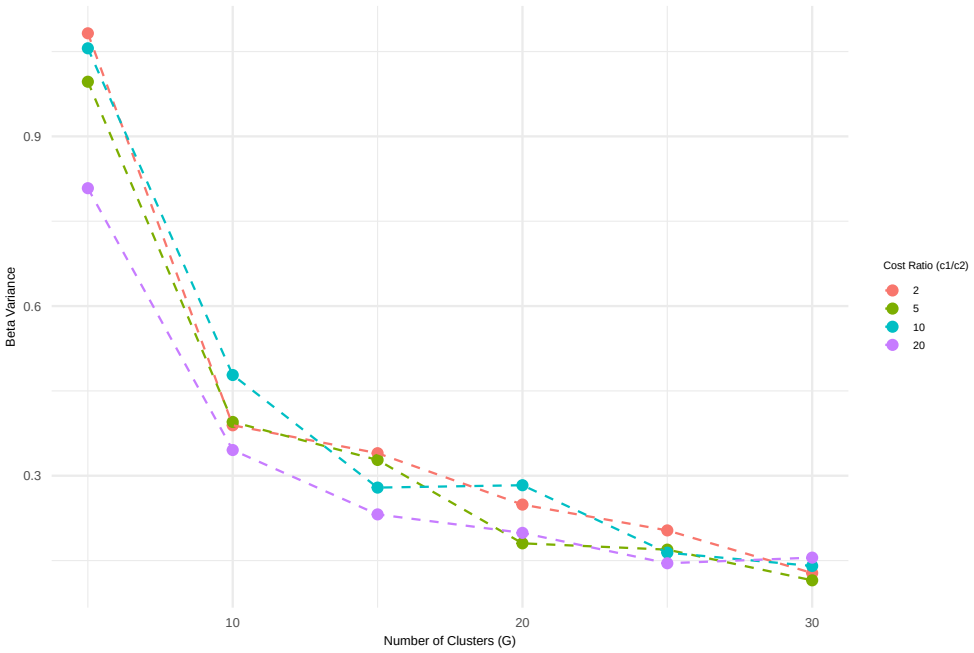
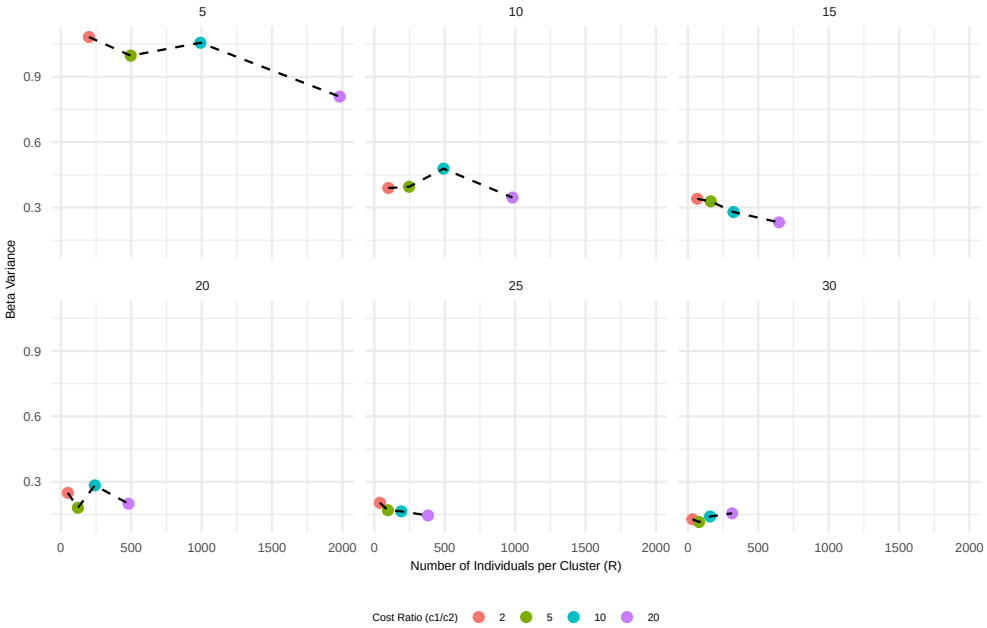


Figure 4: Beta Variance by R for Different G



Comparing the results across all combinations of design parameters, the design with G = 30, R = 79, and

cost ratio = 5 still exhibits to be the optimal. This design exhibits the lowest variance of β and relatively lower values for other performance metrics, as shown in `Table 6`. The following variation on underlying data generation parameters would be based on this scenario.

Table 6: Metrics Summary Varying Cost Ratio

Number of Cluster (G)	Number of Observations per Cluster (R)	Cost Ratio (c1/c2)	Beta Variance	Beta Bias	MSE	Coverage
30	79	5	0.1150802	0.0029277	0.113938	95

Vary gamma_sq

To explore the impact of varying between-cluster variances γ^2 on model performance, we kept other parameters consistent ($\alpha = 0$, $\beta = 1.5$). Considering the intra-cluster correlation (ICC), defined as:

$$ICC = \frac{\gamma^2}{\gamma^2 + \mu}$$

μ represents the baseline mean outcome where given $\alpha = 0$, $\mu = exp(\alpha) = 1$. We adjust γ^2 to represents different levels of ICC:

- Low ICC: $\gamma^2 = 0.111$, $ICC = \frac{0.111}{0.111+1} = 0.1$
- Medium ICC: $\gamma^2 = 1$, $ICC = \frac{1}{1+1} = 0.5$
- High ICC: $\gamma^2 = 10$, $ICC = \frac{10}{10+1} = 0.909$

The performance metrics results shown in`Table 7`

Table 7: Metrics Summary Varying Between-cluster Variance

Between-cluster Variance	Beta Estimate	Beta Variance	Beta Bias	MSE	Coverage
0.111	1.497346	0.0126839	0.0026536	0.0125641	96
1.000	1.529952	0.1736862	0.0299520	0.1728465	90
10.000	1.375978	1.2239380	0.1240220	1.2270801	91

Vary alpha

Vary beta

Reference

Discussion

Reference

1. Heagerty, P. J., Biostatistics, N. P. T. C., & Core, S. D. (2024). *Experimental designs and randomization schemes: Cluster randomized trials.* In *Rethinking Clinical Trials: A Living Textbook of Pragmatic Clinical Trials*; NIH Pragmatic Trials Collaboratory. https://doi.org/10.28929/204