A Simulation Study of Optimal Design Strategies for Cluster Randomized Trials with Budgetary Limitations

Yingxi Kong

Abstract

Background: Cluster Randomized Trials (CRTs) are widely used in public health research to evaluate interventions. Designing CRTs involves balancing the number of clusters (G) and the number of observations per cluster (R) under budget constraints to optimize the precision of treatment effect estimates. This study explores optimal experimental designs for CRTs with normally distributed and Poisson-distributed outcomes, focusing on the effects of design parameters and variance components.

Method: Using the ADEMP framework, we design a comprehensive simulation study to evaluate various combinations of G, R, and cost ratio (c_1/c_2) under a fixed budget of \$10,000. Performance metrics, including variance, bias, mean squared error (MSE), and coverage of the treatment effect estimate (β), were assessed for 100 iterations per scenario. We also examined the impact of underlying data generation parameters on model precision.

Results: For both outcome types, increasing the number of clusters improved precision by reducing the variance of treatment effect estimates, particularly for smaller G. Optimal designs included $G = 30$, $R = 79$, and $c_1/c_2 = 5$ for normally distributed outcomes, and $G = 30$, $R = 314$, and $c_1/c_2 = 20$ for Poisson outcomes. Higher between-cluster variance γ^2 were associated with increased variance and reduced coverage for both cases, indicating lower precision in treatment effect estimates. Moreover, for Poisson-distributed outcomes, a higher baseline mean also increased variance and lowered coverage.

Conclusion: Optimal experimental designs for CRTs require careful balancing of clusters, within-cluster observations, and cost ratios to maximize precision under budget constraints. Higher between-cluster variance and baseline means reduce model precision, underscoring the importance of accounting for these factors when designing CRTs. Future research should explore dynamic cost structures, expanded parameter ranges, and larger simulation iterations to enhance generalizability.

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, a balanced trade-off between the number of clusters and the number of observations within each cluster 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 aim 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.

In designing our trial, we identify budget, B in dollars, as the given resource limit where B might also represent 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 refer to the repeated measurements from the same biological sample which are highly correlated and different clusters refer 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 aim 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 cluster (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 (https://github.com/nikivivi9/Practical-Data-Analysis-Portfolio.git). 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. The experiment function 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, and Cost Ratio

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 table 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.

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
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

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 generally decreased 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 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. 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.

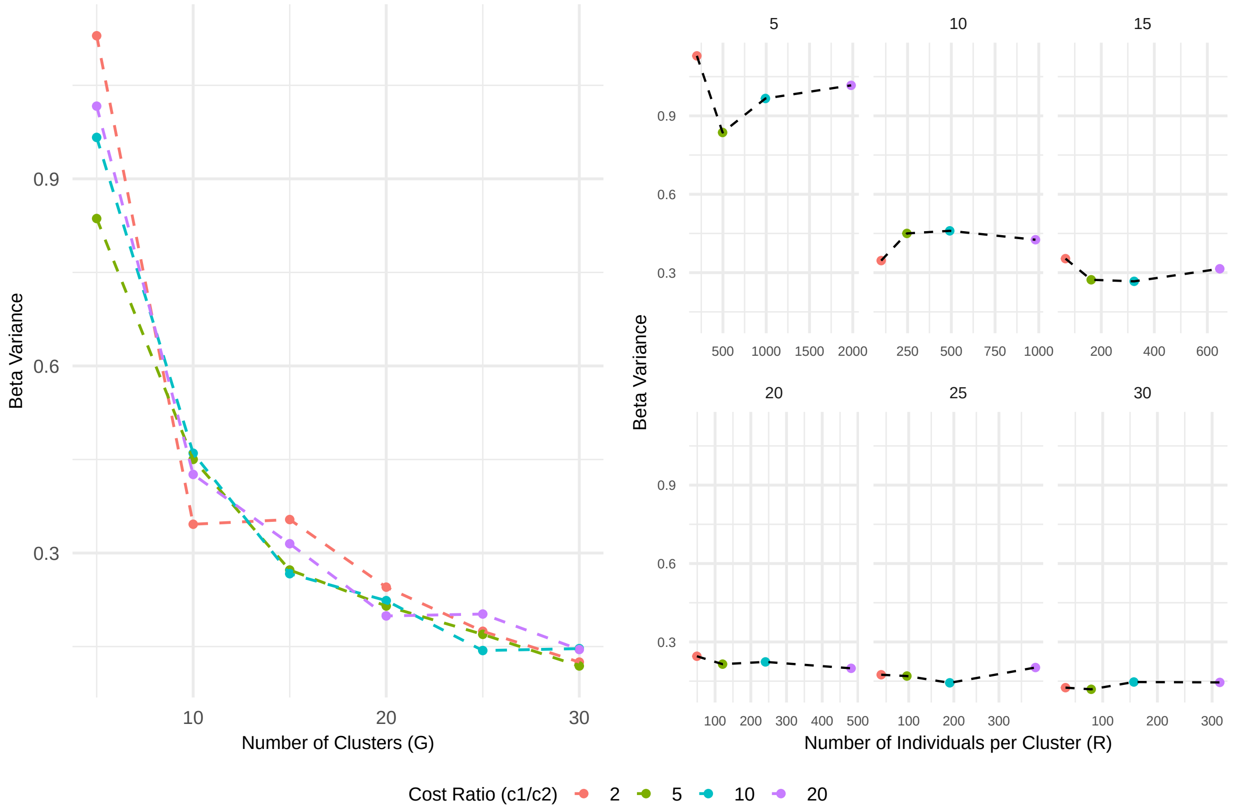

Figure 1: Beta Variance by Number of Clusters (G)

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 it has great performance with other metrics, as shown in `Table 3`. The following variation on underlying data generation parameters would be based on this design scenario.

Table 3: Optimal Design (Outcome Normal Distribution)

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 the 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 to high levels. 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: $\gamma^2 = 9$, $ICC = \frac{9}{9+1} = 0.9$

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	1.491135	0.0158661	0.0088652	0.0157861	94
1.000	1	1.483679	0.1188668	0.0163207	0.1179445	95
9.000	1	1.696329	1.5700444	0.1963288	1.5928889	90

In `Table 4`, fixing within-cluster variance at 1, the estimated treatment effect remains close to the true value with lower γ^2 and the estimate more deviated from the true value at $\gamma^2 = 9$. 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 0.1963 at $\gamma^2 = 9$. Additionally, the MSE increases with higher γ^2 and the coverage remains around 94%-95% for lower γ^2 while decreases slightly to 90% at $\gamma^2 = 9$.

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 vary σ^2 to present low, medium, and high level of ICC as follows:

- 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	1.494616	0.1464453	0.0053837	0.1450099	95
1	1.000	1.481067	0.1250949	0.0189327	0.1242024	95
1	0.111	1.502974	0.1214301	0.0029743	0.1202247	96

In `Table 5`, we observed that the estimated treatment effect is consistently close to the true value while the variance of β exhibits a slight increase as the within-cluster variance increase, same for the MSE. However, the impact of within-cluster variances on estimation precision is much less notable compared to that of between-cluster variances. Bias does not exhibit a clear trend and the coverage remains consistent as σ^2 varies. 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- and within-cluster variability reduces precision. We did not observe obvious trend on estimate 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. These insights underscore the importance of carefully balancing design parameters and variance components to achieve reliable and precise treatment effect estimates.

Poisson Distributed Outcome

Vary G, R, and Cost Ratio

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 ($\alpha = 0$, $\beta = 1.5$, and $\gamma^2 = 1$), with a budget limit of B = \$10,000 and the cost of the first sample in each cluster set at \$c_1 = \$\$20. After the data simulation process, hierarchical model was applied to each simulated data using the `glmer()` function from the `lmerTest` package and performance metrics were calculated based on the model results. Performing same steps as normally distributed outcomes, all simulated data are saved in the `Data_possion` folder and the performance metrics results tables 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 for most cases.

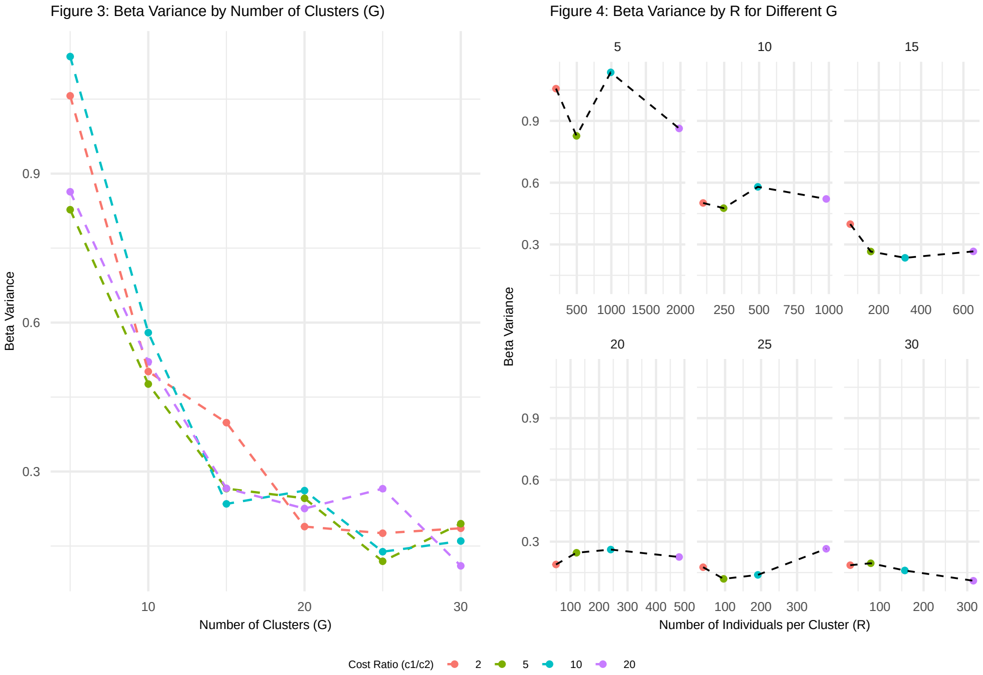

Figure 3: Beta Variance by Number of Clusters (G)

Figure 4: Beta Variance by R for Different G

Comparing the results across all combinations of design parameters, the design with G = 30, R = 314, and cost ratio = 20 exhibits to be the optimal. This design presents the lowest variance of β and relatively great performance for other metrics, as shown in `Table 6`. The following variation on underlying data generation parameters would be based on this scenario. The full table including all comparing scenarios locates within the `Table Results` folder, named as `summary_metrics_df_poisson.csv`.

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	314	20	0.1100694	0.0007004	0.1089692	97

Vary Between-cluster Variance

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 individual-level variability 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: $\gamma^2 = 10$, $ICC = \frac{10}{10+1} = 0.909$

Table 7: Metrics Summary Varying Between-cluster Variance

Between-cluster Variance	Beta Estimate	Beta Variance	Beta Bias	MSE	Coverage
0.111	1.495584	0.0155644	0.0044158	0.0154282	93
1.000	1.540230	0.1286061	0.0402299	0.1289385	96
10.000	1.627900	1.5120583	0.1279004	1.5132962	91

`Table 7` summarizes the impact of varying the between-cluster variance on the treatment effect estimate and associated performance metrics. The estimated treatment effect remains close to the true value of 1.5 with lower γ^2 but shows motr deviation as γ^2 increases. In addition, the variance of β increases substantially as γ^2 grows, indicating reduced precision. Both bias and MSE exhibits an increasing trend while coverage decreases as the between-cluster variability increases.

Vary Baseline Mean

Maintaining other parameters constant ($\beta = 1.5$ and $\sigma^2 = 1$), we also varied baseline mean value ($\alpha = 0, 2, 4, 6, 8, 10$) to investigate its impact on the model precision shown in `Table 8`. The estimated value of the treatment effect remain close to the true value across cases. For smaller α, the variance of treatment effect estimate is relatively low while the variance increases for higher α values ($\alpha = 8, 10$), suggesting reduced precision in the estimates when the baseline mean becomes higher. Moreover, the coverage decreases for higher α values, suggesting less reliable confidence intervals for the estimate.

Table 8: Metrics Summary Varying Between-cluster Variance

Between-cluster Variance	Beta Estimate	Beta Variance	Beta Bias	MSE	Coverage
0	1.519431	0.1238002	0.0194306	0.1229397	93
2	1.528565	0.1093212	0.0285653	0.1090440	92
4	1.541733	0.1345461	0.0417327	0.1349423	93
6	1.571183	0.1277670	0.0711830	0.1315564	95
8	1.512625	0.1632526	0.0126253	0.1617795	87
10	1.573386	0.1659348	0.0733862	0.1696610	91

In conclusion, Poisson-distributed outcomes exhibits similar insights as what we observed with normally distributed outcomes. Increasing the number of clusters significantly improves estimate precision by reducing the variance of treatment effect estimates, emphasizing the importance of designing studies with sufficient clusters to ensure robust results. Additionally, higher between-cluster variance and baseline mean values were associated with increased variance in treatment effect estimates, indicating reduced model precision.

Discussion

This study, in collaboration with Dr. Zhijin Wu, explored optimal experimental design for cluster randomized trials (CRTs) under budget constraints. Using simulation methods, we investigated how underlying data generation parameters influence the selection of an optimal design for both normally distributed and Poisson-distributed outcomes, providing comprehensive insights into designing strategy and their implications.

Our findings with normally distributed outcomes highlight that increasing the number of cluster substantially improves the precision of the treatment effect estimate. However, this improvement comes at the cost of reduced cluster size within a fixed budget, which could limit power for within-cluster analyses. The cost ratio does present a slight decreasing trend shown in `Table 2` which is less pronounced compared to the number of clusters. For Poisson outcomes, similar trends were observed. Increasing the number of clusters reduced the variance of β, particularly for smaller G values. From our simulations, the design configuration with G = 30, R = 79, and a cost ratio of $c_1/c_2 = 5$ shown as optimal for the normally distributed outcomes and the design with G = 30, R = 314, and a cost ratio of $c_1/c_2 = 20$ shown as optimal for the Poisson distributed outcomes. These designs consistently achieved low variance and bias, along with high coverage, making them suitable for resource-constrained settings.

Implementing the optimal design, for normally distributed outcomes, higher between-cluster variance (γ^2) reduced precision, increased bias, and lower coverage, particularly at higher levels of γ^2. The within-cluster variance (σ^2) had a less dramatic effect on precision, smaller within-cluster variance resulted in improved precision and reduced noise. The baseline mean value (α) and the treatment effect value (β) does not directly impact the variance of our treatment effect estimation. For Poisson outcomes, similar trend for γ^2 were observed while higher baseline mean value α presents to reduce precision in the poisson case.

Although our findings provide insights of study design strategy, we have limitations. Our fixed budget value, fixed sample cost, and predefined parameter ranges may not comprehensively capture the variability of real world scenarios. Further investigation into dynamic cost scaling and an expanded parameter comparison, which would better reflect real-world cases, could provide additional meaningful insights. Moreover, to reduce computational demands, our simulations were performed with only 100 iterations per scenario. While this provides preliminary insights, the limited number of iterations may not fully capture the variability across scenarios. Simulations with a larger number of iterations is recommended for more robust and generalizable conclusions.

```r
knitr::opts_chunk$set(echo = FALSE, warning = FALSE, message = FALSE)

library(tidyverse)
library(lmerTest)
library(ggplot2)
library(kableExtra)
library(ggpubr)
summary_metrics_df <- read.csv("C:/Users/yingx/OneDrive/Desktop/Fall 2024/PHP 2550/Practical-Data-Analys
varyGR_summary <- summary_metrics_df %>%
  dplyr::select(G, R, c1c2_ratio, beta_variance, beta_bias, mse, coverage)

varyGR_summary %>%
  filter(c1c2_ratio == 10) %>%
  kable(col.names = c("Number of Cluster (G)", "Number of Observations per Cluster (R)",
                      "Cost Ratio (c1/c2)", "Beta Variance", "Beta Bias", "MSE", "Coverage"),
        booktabs = TRUE, caption = "Metrics Summary Varying Number of Clusters") %>%
  kable_styling(font_size = 7, latex_options = c("repeat_header", "HOLD_position", "scale_down")) %>%
  column_spec(1, width = "3cm") %>%
  column_spec(2, width = "3cm") %>%
  column_spec(3, width = "2cm") %>%
  column_spec(4, width = "1.5cm") %>%
  column_spec(5, width = "1cm") %>%
  column_spec(6, width = "1cm") %>%
  column_spec(7, width = "1cm")
varyGR_summary %>%
  filter(G == 20) %>%
  kable(col.names = c("Number of Cluster (G)", "Number of Observations per Cluster (R)",
                      "Cost Ratio (c1/c2)", "Beta Variance", "Beta Bias", "MSE", "Coverage"),
        booktabs = TRUE, caption = "Metrics Summary Varying Cost Ratio") %>%
  kable_styling(font_size = 7, latex_options = c("repeat_header", "HOLD_position", "scale_down")) %>%
  column_spec(1, width = "3cm") %>%
  column_spec(2, width = "3cm") %>%
  column_spec(3, width = "2cm") %>%
  column_spec(4, width = "1.5cm") %>%
  column_spec(5, width = "1cm") %>%
  column_spec(6, width = "1cm") %>%
  column_spec(7, width = "1cm")
# plot of G vs. beta variance
plot1 <- ggplot(summary_metrics_df, aes(x = G, y = beta_variance, color = as.factor(c1c2_ratio))) +
  geom_point(size = 1) +
  geom_line(size = 0.5, linetype = "dashed") +
  labs(title = "Figure 1: Beta Variance by Number of Clusters (G)",
       x = "Number of Clusters (G)",
       y = "Beta Variance",
       color = "Cost Ratio (c1/c2)") +
  theme_minimal() +
  theme(axis.title = element_text(size = 7),
        title = element_text(size = 7),
        axis.text = element_text(size = 7),
        legend.title = element_text(size = 7),
        legend.text = element_text(size = 7),
        legend.key.size = unit(0.3, "cm"),
        strip.text = element_text(size = 6),
        legend.position = "bottom")
# plot of R vs. beta variance
plot2 <- ggplot(summary_metrics_df) +
  facet_wrap(~ G, scales = "free_x") +
  geom_point(aes(x = R, y = beta_variance, color = as.factor(c1c2_ratio)), size = 1) +
  geom_line(aes(x = R, y = beta_variance), size = 0.4, linetype = "dashed") +
  labs(title = "Figure 2: Beta Variance vs. R by Different G",
       x = "Number of Individuals per Cluster (R)",
       y = "Beta Variance",
       color = "Cost Ratio (c1/c2)") +
  theme_minimal() +
  theme(axis.title = element_text(size = 7),
        title = element_text(size = 7),
        axis.text = element_text(size = 5),
        legend.title = element_text(size = 7),
        legend.text = element_text(size = 7),
        legend.key.size = unit(0.3, "cm"),
        strip.text = element_text(size = 6),
        legend.position = "bottom")

ggarrange(plot1, plot2, common.legend = TRUE, legend = "bottom")

varyGR_summary %>%
  filter(G == 30 & c1c2_ratio == 5) %>%
  kable(col.names = c("Number of Cluster (G)", "Number of Observations per Cluster (R)",
                      "Cost Ratio (c1/c2)", "Beta Variance", "Beta Bias", "MSE", "Coverage"),
        booktabs = TRUE, caption = "Optimal Design (Outcome Normal Distribution)") %>%
  kable_styling(font_size = 7, latex_options = c("repeat_header", "HOLD_position", "scale_down")) %>%
  column_spec(1, width = "3cm") %>%
  column_spec(2, width = "3cm") %>%
  column_spec(3, width = "2cm") %>%
  column_spec(4, width = "1.5cm") %>%
  column_spec(5, width = "1cm") %>%
  column_spec(6, width = "1cm") %>%
  column_spec(7, width = "1cm")
summary_metrics_df_gamma <- read.csv("C:/Users/yingx/OneDrive/Desktop/Fall 2024/PHP 2550/Practical-Data
summary_metrics_df_gamma <- summary_metrics_df_gamma %>%
  dplyr::select(-c(G, R, c1, c2, c1c2_ratio, total_cost,alpha, beta))

summary_metrics_df_gamma %>%
  kable(col.names = c("Between-cluster Variance", "Within-cluster Variance",
                      "Beta Estimate", "Beta Variance", "Beta Bias", "MSE", "Coverage"),
        booktabs = TRUE, caption = "Metrics Summary Varying Between-cluster Variance") %>%
  kable_styling(font_size = 7, latex_options = c("repeat_header", "HOLD_position", "scale_down")) %>%
  column_spec(1, width = "3cm") %>%
  column_spec(2, width = "3cm") %>%
  column_spec(3, width = "2cm") %>%
  column_spec(4, width = "1.5cm") %>%
  column_spec(5, width = "1cm") %>%
  column_spec(6, width = "1cm") %>%
  column_spec(7, width = "1cm")
summary_metrics_df_sigma <- read.csv("C:/Users/yingx/OneDrive/Desktop/Fall 2024/PHP 2550/Practical-Data
summary_metrics_df_sigma <- summary_metrics_df_sigma %>%
  dplyr::select(-c(G, R, c1, c2, c1c2_ratio, total_cost, alpha, beta))


summary_metrics_df_sigma %>%
  kable(col.names = c("Between-cluster Variance", "Within-cluster Variance",
                      "Beta Estimate", "Beta Variance", "Beta Bias", "MSE", "Coverage"),
        booktabs = TRUE, caption = "Metrics Summary Varying Within-cluster Variance") %>%
  kable_styling(font_size = 7, latex_options = c("repeat_header", "HOLD_position", "scale_down")) %>%
  column_spec(1, width = "3cm") %>%
  column_spec(2, width = "3cm") %>%
  column_spec(3, width = "2cm") %>%
  column_spec(4, width = "1.5cm") %>%
  column_spec(5, width = "1cm") %>%
  column_spec(6, width = "1cm") %>%
  column_spec(7, width = "1cm")
summary_metrics_df_poisson <- read.csv("C:/Users/yingx/OneDrive/Desktop/Fall 2024/PHP 2550/Practical-Dat
# plot of G vs. beta variance
plot3 <- ggplot(summary_metrics_df_poisson, aes(x = G, y = beta_variance, color = as.factor(c1c2_ratio)
  geom_point(size = 1) +
  geom_line(size = 0.5, linetype = "dashed") +
  labs(title = "Figure 3: Beta Variance by Number of Clusters (G)",
       x = "Number of Clusters (G)",
       y = "Beta Variance",
       color = "Cost Ratio (c1/c2)") +
  theme_minimal() +
  theme(axis.title = element_text(size = 6),
        title = element_text(size = 6),
        axis.text = element_text(size = 6),
        legend.title = element_text(size = 5),
        legend.text = element_text(size = 5),
        legend.key.size = unit(0.3, "cm"),
        strip.text = element_text(size = 6))
# plot of R vs. beta variance
plot4 <- ggplot(summary_metrics_df_poisson) +
  facet_wrap(~ G, scales = "free_x") +
  geom_point(aes(x = R, y = beta_variance, color = as.factor(c1c2_ratio)), size = 1) +
  geom_line(aes(x = R, y = beta_variance), size = 0.4, linetype = "dashed") +
  labs(title = "Figure 4: Beta Variance by R for Different G",
       x = "Number of Individuals per Cluster (R)",
       y = "Beta Variance",
       color = "Cost Ratio (c1/c2)") +
  theme_minimal() +
  theme(axis.title = element_text(size = 6),
        title = element_text(size = 6),
        axis.text = element_text(size = 6),
        legend.title = element_text(size = 5),
        legend.text = element_text(size = 5),
        legend.key.size = unit(0.3, "cm"),
        strip.text = element_text(size = 6),
        legend.position = "bottom")

ggarrange(plot3, plot4, common.legend = TRUE, legend = "bottom")
varyGR_summary_poisson <- summary_metrics_df_poisson %>%
  dplyr::select(G, R, c1c2_ratio, beta_variance, beta_bias, mse, coverage)

varyGR_summary_poisson %>%
  filter(G == 30 & c1c2_ratio == 20) %>%
  kable(col.names = c("Number of Cluster (G)", "Number of Observations per Cluster (R)",
                      "Cost Ratio (c1/c2)", "Beta Variance", "Beta Bias", "MSE", "Coverage"),
        booktabs = TRUE, caption = "Metrics Summary Varying Cost Ratio") %>%
  kable_styling(font_size = 7, latex_options = c("repeat_header", "HOLD_position", "scale_down")) %>%
  column_spec(1, width = "3cm") %>%
  column_spec(2, width = "3cm") %>%
  column_spec(3, width = "2cm") %>%
  column_spec(4, width = "1.5cm") %>%
  column_spec(5, width = "1cm") %>%
  column_spec(6, width = "1cm") %>%
  column_spec(7, width = "1cm")
summary_metrics_df_poisson_gamma <- read.csv("C:/Users/yingx/OneDrive/Desktop/Fall 2024/PHP 2550/Practi
summary_metrics_df_poisson_gamma <- summary_metrics_df_poisson_gamma %>%
  dplyr::select(-c(G, R, c1, c2, c1c2_ratio, total_cost, alpha, beta))

summary_metrics_df_poisson_gamma %>%
  kable(col.names = c("Between-cluster Variance", "Beta Estimate", "Beta Variance", "Beta Bias", "MSE",
        booktabs = TRUE, caption = "Metrics Summary Varying Between-cluster Variance") %>%
  kable_styling(font_size = 7, latex_options = c("repeat_header", "HOLD_position", "scale_down")) %>%
  column_spec(1, width = "3cm") %>%
  column_spec(2, width = "3cm") %>%
  column_spec(3, width = "2cm") %>%
  column_spec(4, width = "1.5cm") %>%
  column_spec(5, width = "1cm") %>%
  column_spec(6, width = "1cm")
summary_metrics_df_poisson_alpha <- read.csv("C:/Users/yingx/OneDrive/Desktop/Fall 2024/PHP 2550/Practi

summary_metrics_df_poisson_alpha <- summary_metrics_df_poisson_alpha %>%
  dplyr::select(-c(G, R, c1, c2, c1c2_ratio, total_cost, gamma_sq, beta))

summary_metrics_df_poisson_alpha %>%
  kable(col.names = c("Between-cluster Variance", "Beta Estimate", "Beta Variance", "Beta Bias", "MSE",
        booktabs = TRUE, caption = "Metrics Summary Varying Between-cluster Variance") %>%
  kable_styling(font_size = 7, latex_options = c("repeat_header", "HOLD_position", "scale_down")) %>%
  column_spec(1, width = "3cm") %>%
  column_spec(2, width = "3cm") %>%
  column_spec(3, width = "2cm") %>%
  column_spec(4, width = "1.5cm") %>%
  column_spec(5, width = "1cm") %>%
  column_spec(6, width = "1cm")
```

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