Assessing Causal Effects of Genetically-Guided Therapy on Hospital Outcomes in Severe Depressive Disorders: A Quantitative Analysis of the CYP-GUIDES Data

Submitted by: Ziyue (Julia) Wang, Anusha Kuppahally, Malavi Ravindran

Abstract

This paper delves into the clinical trial described in the paper of Gualberto Ruaño et al. in 2020 on the effect of CYP2D6 gene-guided therapy on patients with major depression. In contrast to the original study's qualitative findings, we performed quantitative analysis on the causal impact of genetically-guided therapy on hospital length of stay (LOS) and 30-day readmission rates (RAR). Having employed Fisherian and Neymanian test statistics and utilized various variance reduction techniques, we revealed no significant causal relationship amongst the genetically guided therapy and LOS and RAR. Relevant code can be found in this Github repository: https://github.com/jzywaaaaa/causal_effect_analysis_on_CYP_GUIDES_data.git.

I. Introduction

Our project will focus on the CYP-GUIDES data from the paper of Ruaño et al. (2020) which records a randomized controlled trial (RCT) on the effect of genetically-guided therapy on 1,500 patients with major depressive disorder or severe depression. 477 patients were assigned to standard therapy (Group S) and 982 to medications primarily metabolized by the CYP2D6 enzyme (Group G). There are 39 covariates including gender, age, race/ethnicity, psychotropic medication (and number of medications), and number of administrations.

In the original study, no distinct difference between the treated group and the control group was discovered. However, the approach taken by Ruaño et al. was relatively qualitative. We supplemented this paper's analysis by quantifying the causal effect, if any, between the genetically guided treatment and the two responses – hospital length of stay (LOS) and readmission rate (RAR) in hospitalized patients with severe depressive disorders. By doing so, we hoped to help assess resource allocation and the optimal therapy treatment for psychiatric hospitalizations.

We found neither the Fisherian nor the Neymanian test statistics for completely randomized experiment (CRE) to be significant. Implementing variance reduction techniques like post-stratification and regression adjustment (via simple linear regression, Lin's estimator, and random forest) was able to reduce the variance of our test statistic by as much as 98%. However, all but one of the 95% Neymannian confidence intervals (for both responses) included zero. Hence, we concluded that there was no significant causal relationship between the genetically guided therapy to LOS or RAR in the general population. Nonetheless, given the drastic decrease in variance in post-stratification analyses, this new treatment could be effective to a specific subgroup.

II. Related Work

This project primarily referenced "Results of the CYP-GUIDES randomized controlled trial: Total cohort and primary endpoints." This paper discussed the design of the RCT, the hypothesis, results across

covariates, and significance of outcomes (Ruaño et al. 2020). Data was collected as a completely randomized experiment (CRE), meaning that treatment was randomly assigned across the entire group of patients. The study hypothesized that genetically-guided treatment would decrease LOS and RAR compared to standard treatment. This paper did not conduct any causal inference, and instead simply calculated the mean and standard deviations for both response variables across a few covariates, such as demographics, number of prescriptions, and metabolic reserve (MR). While the paper recognized some differences in average LOS across some covariates, the average treatment effect was not quantified. There was also no evaluation of the treatment effect, as there is no mention of significance or robust hypothesis testing. The paper concluded that there was no significant difference in LOS or RAR between the groups, but recommended further steps, such as stratification by certain covariates, controlling for confounders, and new methods for collecting data. We incorporated these recommendations in our project. This type of research is important, as new treatments can both enhance patient care and optimize resource usage in hospitals. Our project aimed to go beyond this study and truly establish whether there is a causal relationship between the treatment and improved patient outcomes.

III. Data

Our data was collected from the paper mentioned above. The .csv files can be found at this GitHub repo. Among the 1,500 patients, 477 were assigned to standard therapy (Group S) and 982 were assigned to genetically-guided therapy (Group G) (see Appendix Figure 1 for the flowchart of the process of the RCT). For the **treatment** column, we used 1 to denote the genetically-guided therapy and 0 to denote the standard therapy. We also removed 41 rows without a treatment.

The primary response variable is **LOS**, which measures the amount of time in days that each patient stayed in the hospital while being treated for a severe depressive disorder. The secondary outcome is **RAR**, which is a boolean variable indicating if the patient was readmitted to the hospital 30 days after being discharged. After examining a histogram of LOS, we found the distribution to be severely right-skewed. Based on procedures that the paper also employed, we decided to exclude outliers and transform LOS to log(LOS) to normalize the distribution.

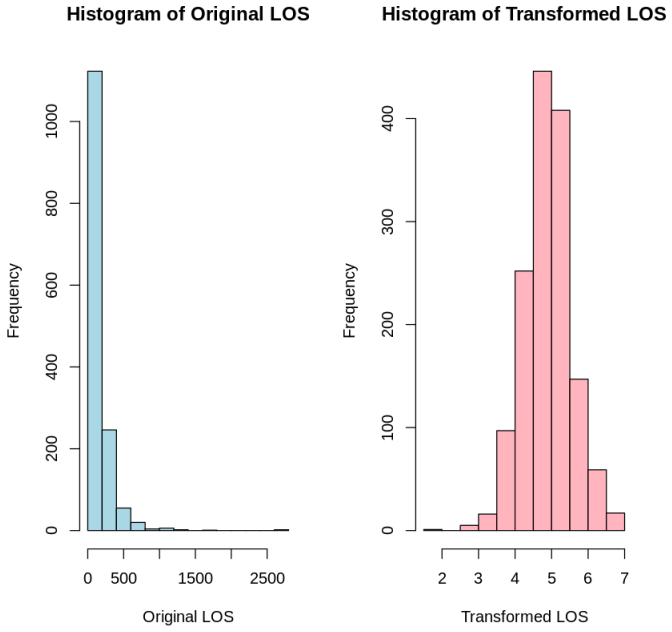

Figure 1. LOS before and after log transformation

Although this was a clinical trial and it was assumed that all extreme values of LOS were not mistakes, the basis for excluding outliers was that patients with extremely high LOS could introduce confounders, as those patients likely had additional health issues—this means that their LOS would likely not be affected by the treatment as many of these issues are not related to psychiatric medications. While the paper did not exclude outliers, the issue of confounders was mentioned later on in the recommendations for future analysis; thus, we decided to exclude patients that had a LOS greater than 1,000 days and log-transformed the data. Since there were only 11 patients that had a LOS greater than 1,000 days, there was no significant data loss once we completed this transformation.

In addition to the treatment and the two responses, our data included 39 covariates. While most columns were factors or numbers, there were a few columns that required additional cleaning. The **diagnosis** column originally consisted of long strings of the diagnosis of the patient (e.g., "MDD, Single Episode, Severe Without Psychotic Features"). We extracted the key information from this column and broke it down into three columns: condition (with levels "MDD", "depressive disorder NOS", and "other"), recurrence (with levels "single episode", "recurrent", and "unspecified"), and severity (with levels "severe", "moderate", and "unspecified"). After plotting the count plots of all the categorical variables, we noticed that the **physician** column, which indicates which physician was responsible for the patients, was quite imbalanced. In particular, out of the 23 levels, 8 had fewer than 20 records. Hence, we merged these 8 levels into one level called "other".

By plotting the covariate balances across treatment and control, we noticed that the distributions of **race**, **condition**, the amount of **trazodone** administered to patients, as well as the number of psychotropic medications (**n_pt_med**) were imbalanced in the treatment and the control groups (Figure 2). Hence, it would be beneficial to perform post-stratification and re-analyze the CRE as a stratified randomized experiment while controlling these four covariates respectively.

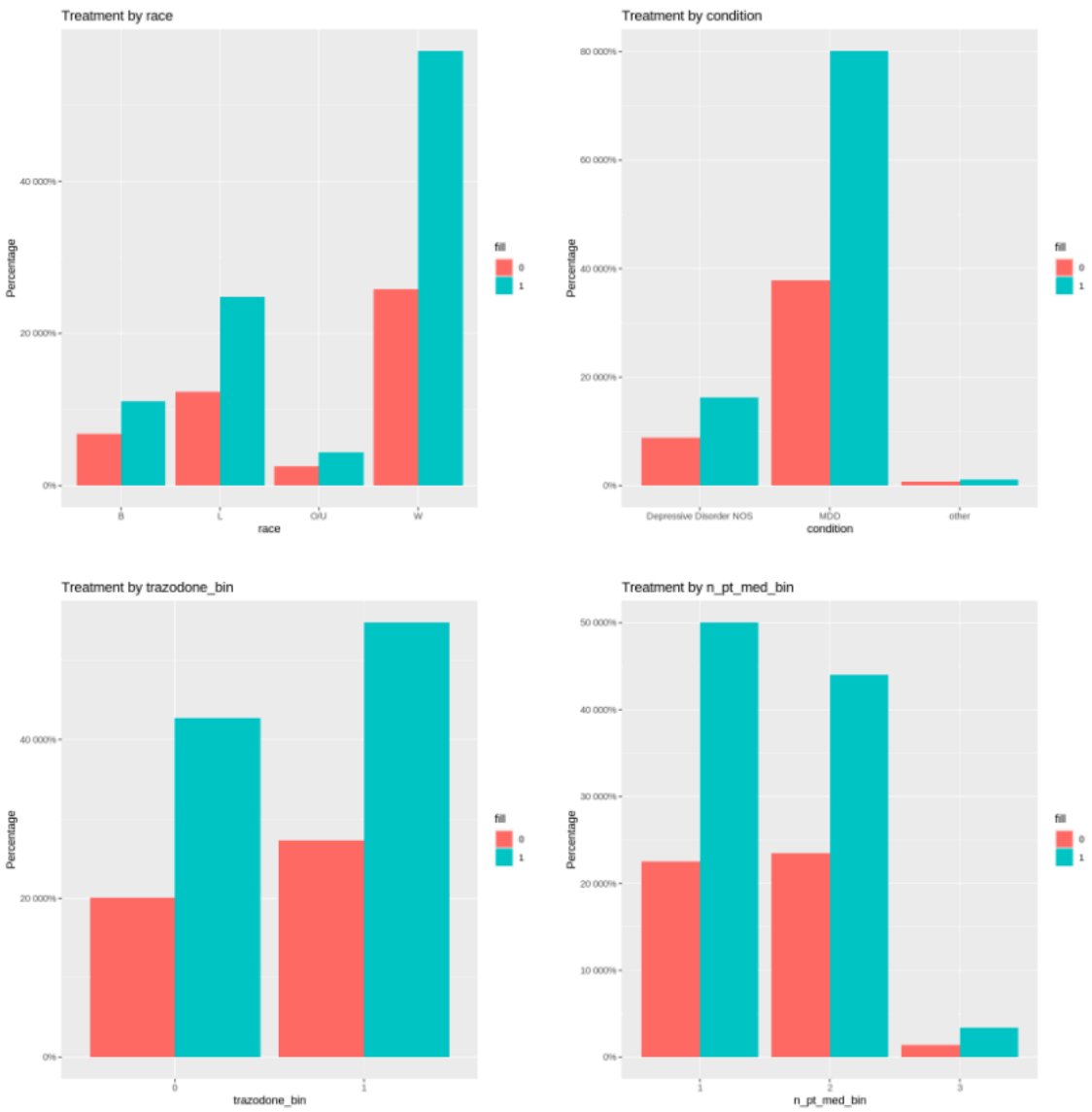

Figure 2. Covariate Balance across Treatment and Control

IV. Methods

Our objective is to determine if a causal relationship exists between genetically-guided treatment and two key outcomes: (i) the log of the patient's length of stay in the hospital (log(LOS)) and (ii) whether or not the patient is readmitted to the hospital 30 days after being discharged (RAR). For both of these outcomes we employ the following inferences:

I. **Fisher Randomization Test (FRT):** Fisherian inference allowed us to test whether the effect of genetically-guided treatment is zero for any individual unit. As the data was collected via a completely randomized experiment, we first computed the simplest difference in means estimator, $\hat{\tau}$, for each outcome. To compute the p-value from our FRT, we applied a Monte-Carlo procedure in which we randomly permuted the treatment assignments and recomputed the difference in means over 10,000

iterations. The p-value was then the proportion of permutations where the simulated difference in means was equal to or greater than the observed difference in means.

II. **Neyman's Repeated Sampling Inference:** Neymenian inference allowed us to test whether the *average* effect of genetically-guided treatment was zero. For both the log(LOS) and RAR outcomes, we estimated the variance of $\hat{\tau}$ as the sum of overall variation and the variation due to sampling and obtained 95% confidence intervals identically to the classical two-sample setting (assuming approximate normality).

III. **Variance Reduction Techniques:** We also sought to reduce variance by the introduction of relevant covariates from the data. Variance reduction was explored in two primary ways: (i) Post-Stratification and (ii) Regression Adjustment

A. Post Stratification: Although our data was collected as a CRE, we identified imbalanced covariates and then assumed the data was collected as a SRE according to those covariates to reduce the variance. The variables we identified were: race, condition, trazodone, and the number of psychotropic medications (Figure 2). We then found the estimated average treatment effect, variance, and 95% confidence interval when stratifying by each covariate.

B. Regression adjustment: We also regressed our two outcomes of interest on relevant pre-treatment covariates. Prior to constructing various regression models, we employed elastic net regularization to select a subset of features to use for each outcome. The elastic net method is a hybrid of Lasso (L1) and Ridge (L2) regularization, where parameter α controls the tradeoff between feature selection and coefficient shrinkage. For log(LOS), we set $\alpha = 1$ (which corresponds to LASSO) in order to reduce our dimension from 41 to 15. For RAR, we set $\alpha = 0.01$ in order to reduce our dimension from 41 to 12. We performed elastic net regularization separately for each outcome to avoid assuming an identical set of influential features across both outcomes. We altered α as a part of our sensitivity analysis to understand how robust our regression adjustment methodologies were to the selection of inputs.

1. **Fisher Randomization Test with covariates:** We first conducted simple linear regressions for each of our outcomes, log(LOS) and RAR. Each was regressed on the treatment indicator and covariates. In this setting, our estimate of causal effect for each outcome was selected as the regression coefficient on the treatment indicator. Associated p-values were extracted from the model outputs.
2. **Lin's estimator:** Lin's estimator regresses the outcome of interest on the treatment, mean-centered covariates, and the interaction between treatment and mean-centered covariates, in order to achieve variance reduction and unbiasedness. We applied Lin's estimator for log(LOS) and RAR by modifying our original linear regressions to mean-center covariates and include interaction terms. For both outcomes, we obtained the estimated average treatment effect, robust standard errors (and variance), and 95% confidence intervals.
3. **Generalized Lin's Estimator– Random Forest:** We also generalized Lin's estimator to a flexible Machine Learning approach in the form of a Random Forest. Random

Forest is a versatile Machine Learning algorithm that operates by constructing multiple decision trees during the training phase and outputting the average prediction of the individual trees. Our high-level approach was the following:

a) Generate $\hat{\mu}_0(x)$ and $\hat{\mu}_1(x)$, which were prediction models on the control and treatment units respectively.
b) Calibrate $\hat{\mu}_0(x)$ and $\hat{\mu}_1(x)$ by their respective biases.
c) Use calibrated predictive models to infer missing potential outcomes.
d) Generate an estimate of causal effect with the difference in means, where missing potential outcomes were replaced by inferred values.

We employed the cross-fitting procedure discussed in class to obtain valid 95% confidence intervals and variance estimates in the generalized Lin's setting.

V. Experiments & Results

I. Primary Response Variable: log(LOS)

Method	$\hat{\tau}$	**p-value (Fisher's null)**	**Variance**	**95% CI**
Fisher Randomization Test	0.0196	0.2976	–	–
Neyman's Repeated Sampling	0.0196	–	0.0014	(-0.054, 0.0929)
Post-SRE: race	-0.0038	–	7.19e-05	(-0.0204, 0.013)
Post-SRE: condition	-0.0052	–	1.5188e-05	(-0.013, 0.0024)
Post-SRE: trazodone	0.0403	–	0.0007	(-0.010, 0.091)
Post-SRE: # PT meds	0.0546	–	0.0006	(0.0049, 0.104)
Fisher Randomization Test with covariates	0.031	0.271	–	–
Lin's Estimator	0.031	–	0.00078	(-0.024, 0.086)
Generalized Lin's Estimator with Random Forest	0.019	–	0.00054	(-0.026, 0.064)

Table 1. Estimates and Significance by Methodology for log(LOS)

When treating the data as a simple CRE we did not find a significant difference in log(LOS) by treatment. Using Fisher's Randomization Test, we found an estimated average treatment effect of 0.0196 and a

p-value of 0.2976, greater than our selected significance level of 0.05. Thus, we failed to reject Fisher's null hypothesis that the treatment has no effect. Using Neyman's null hypothesis, which states that the average treatment is 0, we found an estimated variance of 0.0014 and a confidence interval of (-0.054, 0.093). Because this interval contained 0, it did not provide evidence against Neyman's null hypothesis, so we failed to reject Neyman's null hypothesis. These results follow logically, as if Fisher's null hypothesis stands, so does Neyman's null hypothesis.

After conducting post-stratification across 4 covariates, we determined that stratification by each covariate accomplished variance reduction compared to the simple CRE value. Stratifying by condition produced the lowest overall variance estimate. However, only the confidence interval for the number of psychotropic medications did not contain 0 (as highlighted in the table)—thus, stratifying by race, condition, and trazodone did not provide evidence against Neyman's null hypothesis, but stratifying by the number of psychotropic medications did. Regression adjustment, particularly the generalized Lin's estimator, achieved variance reduction as well. However, the three regression adjustment methodologies also indicated a lack of significant causal effect, with large p-values and confidence intervals including zero.

In terms of sensitivity, our findings demonstrate considerable robustness across varied methodologies. The maximum absolute difference between any pair of estimates was 0.0455, and all but one estimate was significant. These outcomes collectively indicated a lack of discernible causal effect of genetically guided therapy on hospital length of stay. To further our sensitivity analysis, we also explored the impact of altering input features in our regression-adjusted models. To vary our input features, we simply manipulated the regularization parameter α in our elastic net regularization. For log(LOS), we utilized $\alpha = 0.02$ in order to select a broader set of features. Figure 3 depicts the estimates and confidence intervals for both Lin's and the generalized Lin's estimators when varying α, and highlights that our regression adjustment methodologies are robust to varying input features.

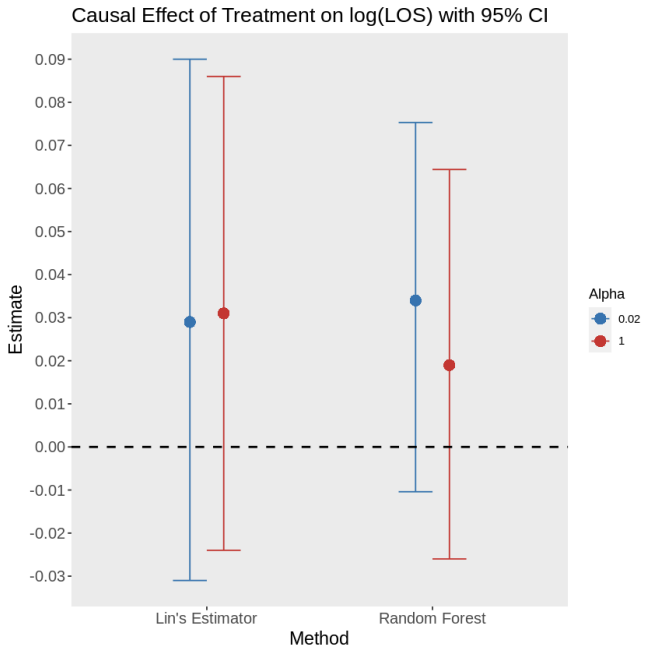

Figure 3. Causal Effect of Treatment on log(LOS) with varied α

II. Secondary Response Variable: RAR

Method	$\hat{\tau}$	**p-value (Fisher's null)**	**Variance**	**95% CI**
Fisher Randomization Test	0.0099	0.3098	–	–
Neyman's Repeated Sampling	0.0099	–	0.0003	(-0.022, 0.0419)
Post-SRE: race	-0.0016	–	9.255e-06	(-0.008, 0.0043)
Post-SRE: condition	0.0011	–	1.277e-06	(-0.001, 0.003)
Post-SRE: trazodone	-0.0016	–	0.00013	(-0.024, 0.021)
Post-SRE: # PT meds	0.0082	–	0.00012	(-0.013, 0.0297)
Fisher Randomization Test with covariates	0.005	0.978	–	–
Lin's Estimator	0.003	–	0.0003	(-0.032, 0.039)
Generalized Lin's Estimator with Random Forest	0.005	–	0.0003	(-0.029, 0.040)

Table 2. Estimates and Significance by Methodology for RAR

We repeated the same procedure of treating the data as a simple CRE, and found that there was also no significant difference in RAR by treatment. After conducting Fisher's Randomization Test, we found an estimated average treatment effect of 0.0099 and a p-value of 0.3098. As this p-value was greater than our selected significance level, we failed to reject Fisher's null hypothesis that the treatment has no effect. As stated previously, it then follows that Neyman's null hypothesis should also stand. This was supported in our calculations when using Neyman's null hypothesis, for which we found an estimated variance of 0.0003 and a confidence interval of (-0.022, 0.042). Since this confidence interval contained 0, it did not provide evidence against Neyman's null hypothesis that the average treatment effect is 0—so, we failed to reject Neyman's null hypothesis.

After repeating our post-stratification procedure for RAR across the same covariates, we found that stratifying by each covariate reduced variance compared to the simple CRE result. Similarly to log(LOS), stratifying by condition also produced the lowest overall variance estimate. However, in this case, stratifying by each of these covariates resulted in confidence intervals containing 0. So, for RAR, post-stratification did reduce variance but did not provide evidence against Neyman's null hypothesis. Thus, we failed to reject the null hypothesis. Both Lin's and the generalized Lin's estimators failed to achieve variance reduction over the simple

CRE variance and estimates derived from these regression adjustment methodologies also indicated a lack of significant causal effect, with large p-values and confidence intervals including zero.

In terms of sensitivity, our findings again demonstrated robustness across varied methodologies. While some estimates varied in their directionality, the maximum absolute difference between any pair of estimates was 0.0155 and none were significant. These outcomes collectively indicate a lack of discernible causal effect of genetically guided therapy on 30-day readmission. To further our sensitivity analysis, we again explored the impact of altering input features in our regression-adjusted models. For RAR, we utilized $\alpha = 0.005$ in order to select a broader set of features. Figure 4 depicts the estimates and confidence intervals for both Lin's and the generalized Lin's estimators when varying α, and highlights that our regression adjustment methodologies are robust to varying inputs.

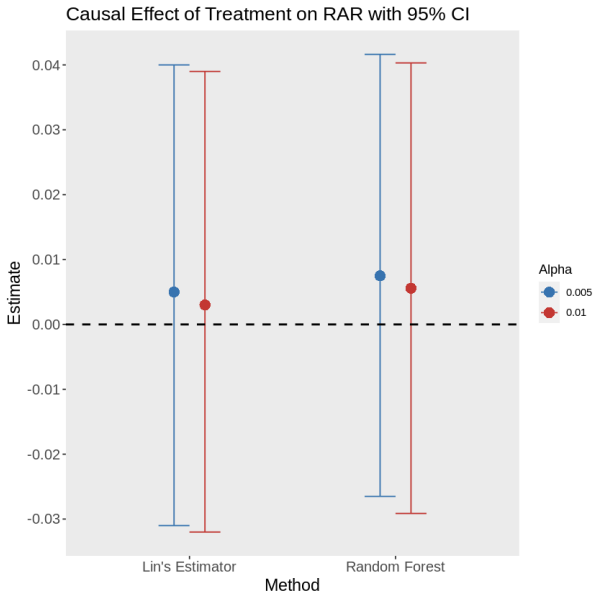

Figure 4. Causal Effect of Treatment on RAR with varied α

VI. Conclusion

After conducting these methods, we did not find robust evidence of a difference in LOS and RAR based on genetically guided therapy. While this conclusion was also stated in the paper we referenced, there is now causal inference to substantiate that claim. However, we recommend that additional analyses should occur to fully support this conclusion. As stated previously, improving treatments is important both for improving patient outcomes and resource allocation—as there is scientific backing to support genetically-guided therapy, it is important to fully test whether this treatment produces meaningful results. We recommend that future studies stratify by scientifically relevant variables (e.g. metabolic reserve, medication), perform sensitivity analysis by physician, and control for the type of electronic medical record (EMR) used to record data.

References

Ruaño, Gualberto, et al. “Results of the CYP-GUIDES randomized controlled trial: Total cohort and primary endpoints.” *Contemporary Clinical Trials*, vol.89, article 105910, Feb. 2020. Accessible from: https://www.clinicalkey.com/#!/content/playContent/1-s2.0-S1551714419306263?returnurl=null&referrer=null.

Shashwat Tiwari. “Clinical Dataset of the CYP-GUIDES Trial: Cytochrome Psychotropic Genotyping Under Investigation for Decision Support Data.” *Kaggle*, 2020. Accessible from: https://www.kaggle.com/datasets/shashwatwork/clinical-dataset-of-the-cypguides-trial/.

Tang, Wan, and Tu, Xin M. *Modern Clinical Trial Analysis*. Springer: 2013. Accessible from: https://link.springer.com/book/10.1007/978-1-4614-4322-3.

Goetghebeur, Els, and Molenberghs, Geert. “Causal Inference in a Placebo-Controlled Clinical Trial with Binary Outcome and Ordered Compliance.” *Applications and Case Studies* pp.928-934, February 27, 2012. Accessible from: https://www.tandfonline.com/doi/abs/10.1080/01621459.1996.10476962.

Appendix

Figure [[]]: Flow Diagram of Randomized Assignment.

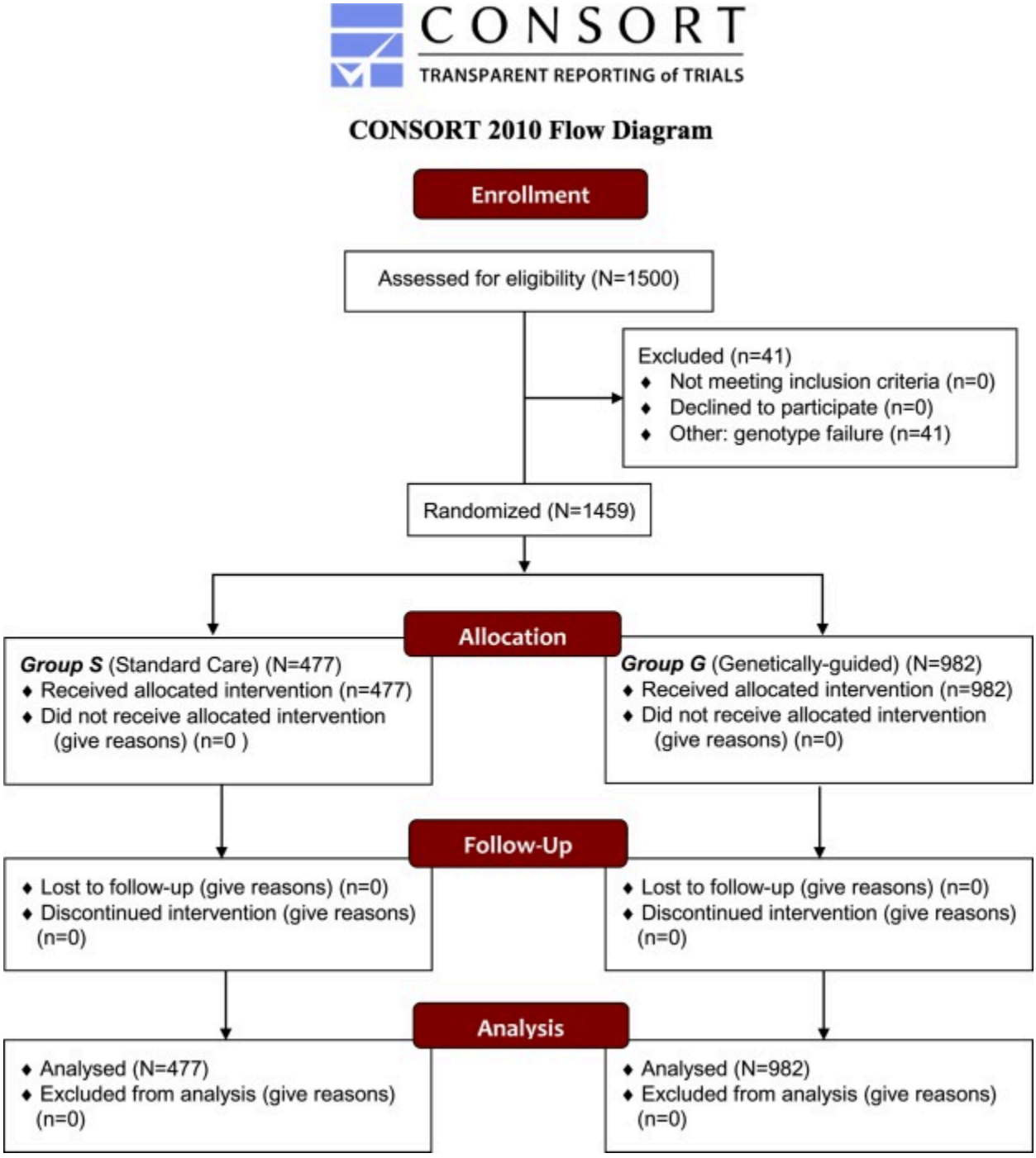

Figure [x]: Histograms of LOS and Transformed LOS

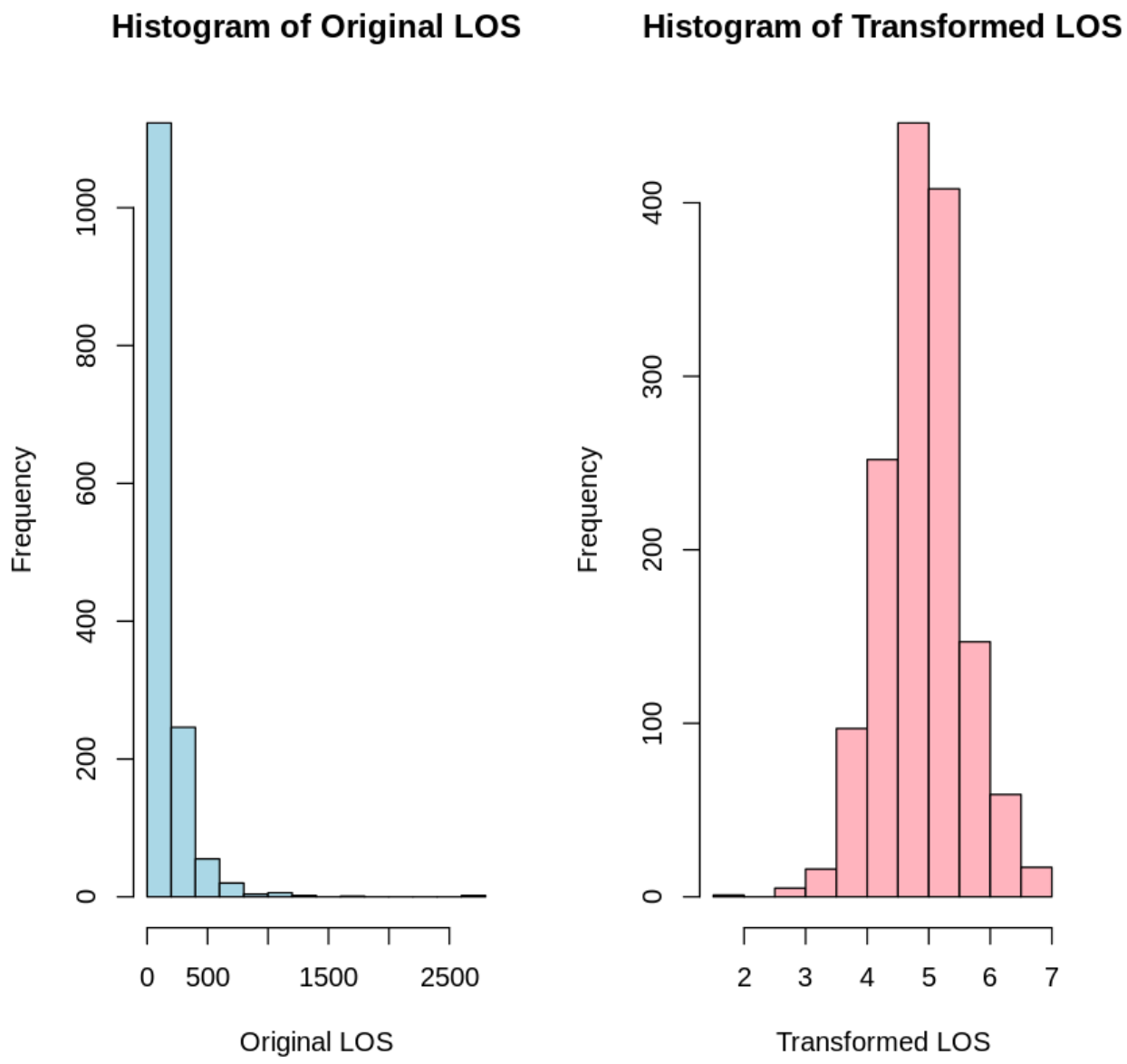

Figure [x]: Imbalanced Covariates