P9185 Statistical Practices and Research for Interdisciplinary Sciences (SPRIS)

Project I Report

Effects of DAR-0100A on cognitive functioning in individuals with schizophrenia

Jungang Zou, jz3183

Department of Biostatistics,

Mailman School of Public Health, Columbia University

April 9, 2023

Contents

1 Introduction

Abnormal interpretation of reality characterizes schizophrenia (SCZ), which is a grave mental disorder. To evaluate the efficacy of different dosages (15mg, 0.5mg) of the selective D1R agonist, DAR-0100A, in improving cognitive deficits in schizophrenia (SCZ) patients, a randomized trial was conducted in which 47 clinically stable individuals with SCZ were randomly assigned to one of three groups. The study was conducted over 19 days, during which participants were admitted to an inpatient clinic and administered the study drug from day 0 to day 5, and from day 15 to day 19. Cognitive function was measured using a composite memory score, and four cognitive assessments were conducted on day 0, day 5, day 19, and day 90. The assessments were used to compare the impact of DAR-0100A on cognitive deficits in SCZ patients compared to a placebo (normal saline), with higher scores indicating better memory.

This report aimed to deal with the missingness problem in data and examine the impact of different treatments at various time intervals.

2 Exploratory Data Analysis

The dataset analyzed in this report was longitudinal and in a long format. It consisted of 47 clinically stable individuals with SCZ who were randomly assigned to one of three treatments: placebo (17 individuals), low dose DAR-0100A (14 individuals), or high dose DAR-0100A (16 individuals). We recorded the composite memory score on Days 0, 5, 19, and 90 as a numeric variable named `MEM_Comp`. Additionally, we collected age and

gender data, represented by `Age` and `Gender`, respectively. While `Age` is a numeric variable, `Gender` is a factor variable with females as the reference group. To code the treatment and time information, we used two categorical variables: `Treatment` and `Time`. The three levels of `Treatment` included the placebo group, the high-dose treatment group, and the low-dose treatment group. For `Time`, the four levels were days 0, 5, 19, and 90. The summary statistics can be found in Figure 1.

The distributions of `MEM_Comp` stratified by `Treatment` and `Time` are plotted in Figure 2. The measured memory assessment scores from the high-dose treatment group are better than the other two compared groups. The trajectory of each patient stratified by `Treatment` and `Time` is also graphed as a spaghetti plot in Figure 3.

The missingness pattern in this data is drop-out missingness. Totally, there are 24 individuals who have complete observations at all four time points. If stratified by treatment groups, there are 9 complete cases in the high-dose group, 7 complete cases in the low-dose group, and 8 complete cases in the placebo. Out of 47 subjects, 2 are missing on day 5, 12 are missing on day 19, and 16 are missing on day 90.

3 Methods

3.1 Notation

Consider a dataset consisting of n individuals, each with 4 repeated measures denoted by $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{i4})$. Additionally, at each time point j, each individual has p covariates represented by the vector $\mathbf{X}_{ij} = (X_{ij1}, \ldots, X_{ijp})^\top$.

We denote $\mathbf{Y} = (\mathbf{Y}_1, \ldots, \mathbf{Y}_n)$ and $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_n)$ and are interested in modeling the conditional distribution $f(\mathbf{Y}|\mathbf{X}, \theta)$ with some unknown parameters θ using statistical methods. To distinguish between the observed and missing components of the outcome, we use $\mathbf{Y}^{obs}$ and $\mathbf{Y}^{mis}$ to represent the observed and missing parts, respectively.

3.2 Missingness Assumptions

Generally, there are three types of missingness: Missing Completely at Random(MCAR), Missing at Random(MAR), and Missing not at Random(MNAR). We mainly focus on MCAR and MAR in this analysis.

MCAR assumes missingness is independent of other variables. Little's test [1] is used to test if the missingness in the dataset is MCAR or not. we ran Little's test on the missing data and got a p-value as 0.0002942496, so we have strong evidence to reject the null hypothesis and concluded the data is not from the MCAR assumption.

MAR assumes missingness is dependent on observed part $\mathbf{Y}^{obs}$. MAR is a commonly used assumption to deal with missing data in real life. In this analysis, we focus on the MAR assumption in this project.

3.3 Linear Mixed Effects Model with Multiple Imputation

Multiple imputation [2], is the widely used tool to deal with missing data with MAR assumption. The idea of multiple imputation (MI) is to generate several imputed datasets and pool the results from these "full" datasets into a single analysis.

In this project, we used a famous machine learning-based longitudinal

imputation tool MixRF [3] to generate 20 imputed datasets. MixRF has the ability to model the complex longitudinal surface by using the random forest to capture the non-linear relationship between covariates and outcome.

After MI, we ran a linear mixed-effects model (LMM) on each separate imputed dataset. LMM imposes random effects for each individual on the regression surface, where individuals get assessed over time, and it is immediately obvious that data records from the same individual would exhibit similarities over time. The LMM with random intercept model is defined as follows:

$$Y_{ij} = \mathbf{X}_{ij}\beta + \alpha_i + \epsilon_{ij}, i = 1, 2, ..., n, j = 1, 2, 3, 4. \quad (1)$$

where ϵ_{ij} is independent and identical distributed as $N(0, \sigma^2)$. α_i stands for the individual effect and is distributed as a normal distribution $N(0, \sigma_\tau^2)$. More details about the analysis model can be found in Appendix.1.

Finally, we use Rubin's Rule [4] to pool the parameters from each separate model.

4 Results

The pooled estimates for the LMM model are displayed in Figure 4. As we see `Baseline` is one of the most important variables to account for the longitudinal trajectory. Every unit increases in `Baseline` will lead to about 0.929 unit increase in memory scores. On the other hand, the treatment effect of the high-dose group is another significant variable ($p-value < 0.1$). The individuals assigned to the high-dose group will have a 0.36 increase in memory scores, compared with the ones who are assigned to the low-dose

group. On the other hand, the treatment effect from the low-dose group is not significant enough ($p-value = 0.804$), which means there is no evidence to find a significant difference between the low-dose group and the placebo group. Figure 5 shows the pooled estimates and 95% confidence interval for the high-dose group and the low-dose group across the longitudinal time points. We can also observe the significant difference between the treatments from the high-dose group and the low-dose group.

If considering the effect of time trajectory, we may find the main effects of different time points are not significant enough. Even if we consider the interaction effect for time and different treatments, the effects are also not significant. This result also indicates the treatment effects for all treatments stay stable over time. Figure 5 also reveals the same conclusion that the treatment effects unchanged over time.

5 Conclusion

The focus of this report is to investigate the effects of three treatments (high-dose, low-dose, placebo) on cognitive deficits among individuals with schizophrenia. To account for missing data and adjust for baseline covariates and demographic information such as age and gender, we used a linear mixed-effect model with Multiple Imputation. Our analysis suggests that the high-dose DAR-0100A regimen may be effective in improving cognitive abilities in schizophrenia patients. However, we found limited evidence to support the efficacy of the low-dose regimen of DAR-0100A. Furthermore, we observed that the treatment effects were unchanged over time.

Appendices

The appendix includes all supplementary formulas and figures that are referred to in this report.

.1 Analysis model

Specifically, our covariates consist of the baseline measurement of `MEM_Comp0`, `Time`, `Treatment` and the interaction between `Time` and `Treatment`. Since `MEM_Comp0` is included in the covariates so we exclude all baseline measurements in the outcome variable. For `Time`, we choose day 5 as the reference level and placebo group for `Treatment`. These variables are used for both imputation and the analysis model. The analysis model is:

$$
\begin{aligned}
Y_{ij} =& \beta_0 + \beta_1 * MEM\_Comp0_{ij} + \beta_2 * I(Time_{ij} = 19) + \beta_3 * I(Time_{ij} = 90) + \\
& \beta_4 * I(Treatment_{ij} = high - dose) + \beta_5 * I(Treatment_{ij} = low - dose) + \\
& \beta_6 * I(Time_{ij} = 19) * I(Treatment_{ij} = high - dose) + \beta_7 * I(Time_{ij} = 90) * \\
& I(Treatment_{ij} = high - dose) + \beta_8 * I(Time_{ij} = 19) * I(Treatment_{ij} = low - dose) \\
& + \beta_9 * I(Time_{ij} = 90) * I(Treatment_{ij} = low - dose) + \alpha_i + \epsilon_{ij}, \\
& i = 1, 2, ..., n, j = 1, 2, 3, 4.
\end{aligned}
\tag{2}
$$

.2 Figures

```
   Subject_ID         day              Age        Gender    Treatment_Group     MEM_comp           baseline       
 Min.   :1021   Min.   : 0.00   Min.   :20.00   F: 88   High Dose:64     Min.   :-3.18587   Min.   :-1.5467  
 1st Qu.:1034   1st Qu.: 3.75   1st Qu.:33.00   M:100   Low Dose :56     1st Qu.:-0.72381   1st Qu.:-0.7093  
 Median :1049   Median :12.00   Median :40.00           Placebo  :68     Median : 0.11963   Median :-0.0339  
 Mean   :1051   Mean   :28.50   Mean   :39.23                            Mean   :-0.01291   Mean   : 0.0000  
 3rd Qu.:1068   3rd Qu.:36.75   3rd Qu.:46.00                            3rd Qu.: 0.66273   3rd Qu.: 0.6630  
 Max.   :1088   Max.   :90.00   Max.   :54.00                            Max.   : 1.73880   Max.   : 1.6359  
                                                                         NA's   :30
```

Figure 1: Summary statistics for each variable

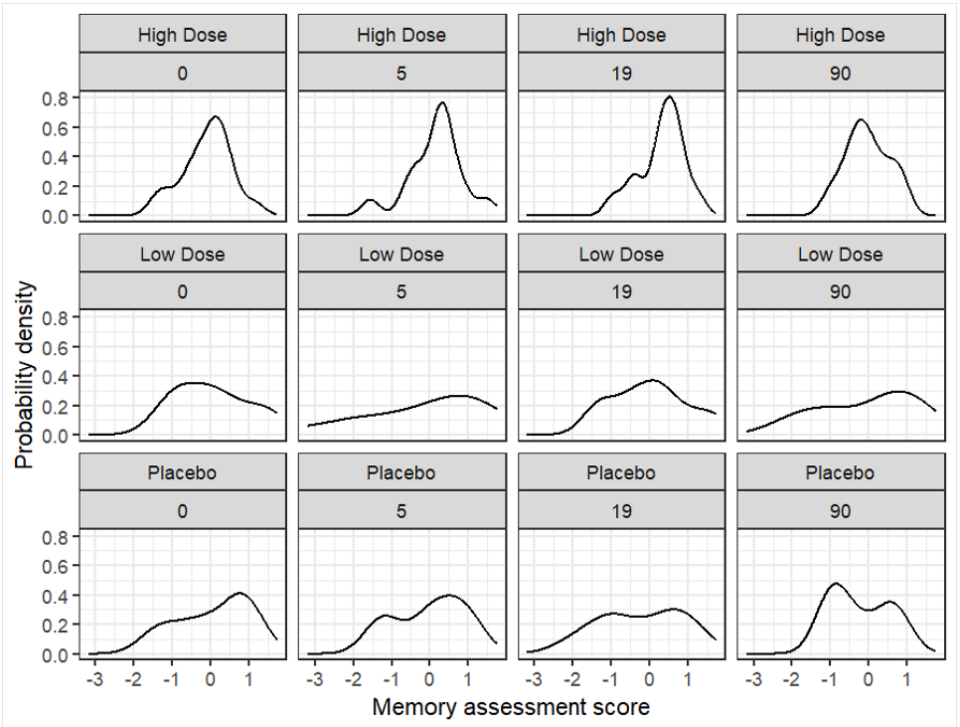

Figure 2: Distribution of outcome variable

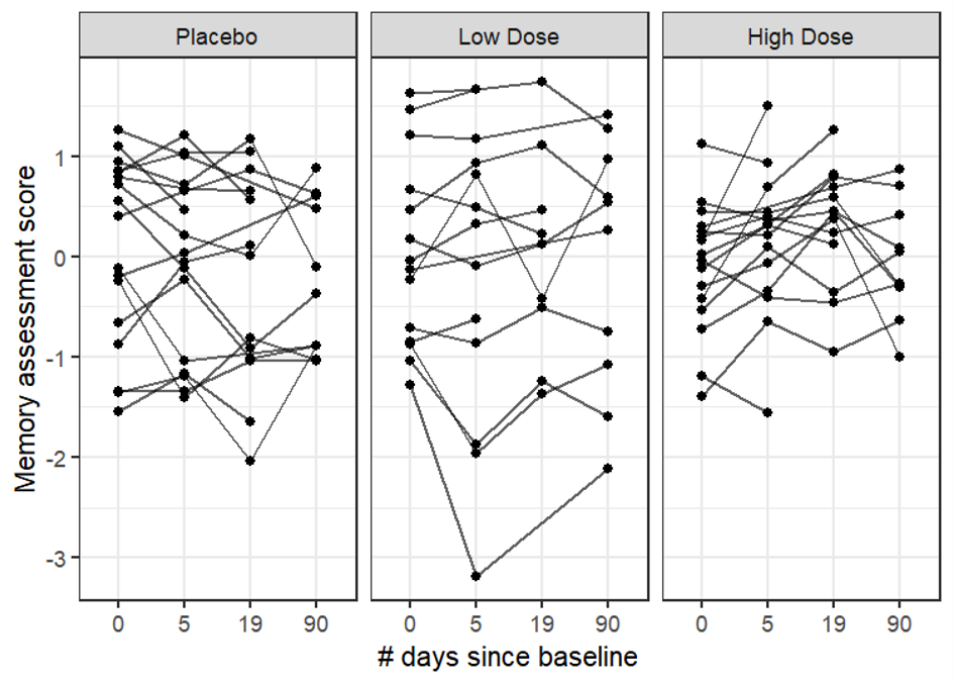

Figure 3: Spaghetti plot for longitudinal trajectory

	Estimated Value		
Variable	Coefficient	Standard error	p-value
Intercept	-0.091	0.137	0.506
Baseline	0.929	0.088	0.000**
Day 19	-0.103	0.110	0.352
Day 90	-0.025	0.110	0.818
High dose	0.360	0.197	0.068*
Low dose	-0.051	0.203	0.804
Day 19*High Dose	0.221	0.158	0.162
Day 90*High Dose	-0.039	0.158	0.806
Day 19*Low Dose	0.211	0.164	0.198
Day 90*Low dose	0.213	0.164	0.194

Table 1: Fixed effect coefficient estimates using multiply imputed data. * = $p < .1$, ** = $p < .05$.

Figure 4: Summary of pooled estimates

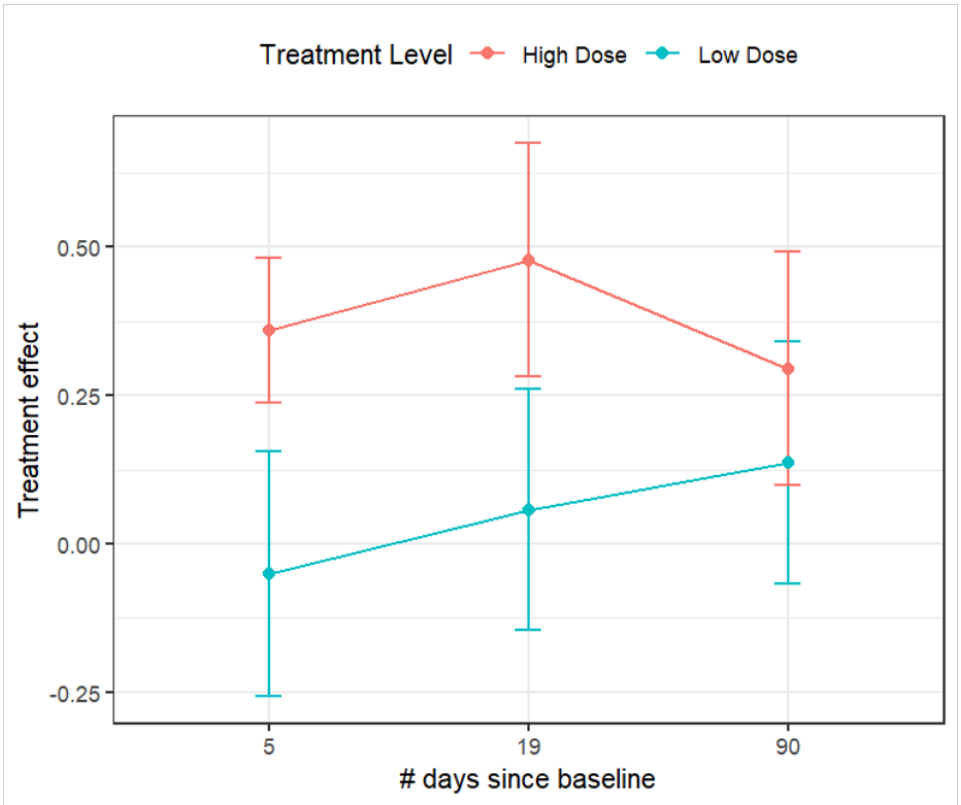

Figure 5: Estimates of treatment effects

References

[1] Roderick JA Little. "A test of missing completely at random for multivariate data with missing values". In: *Journal of the American statistical Association* 83.404 (1988), pp. 1198–1202.

[2] Donald B Rubin and Nathaniel Schenker. "Multiple imputation in healthare databases: An overview and some applications". In: *Statistics in medicine* 10.4 (1991), pp. 585–598.

[3] Wang and etc. "Imputing gene expression in uncollected tissues within and beyond GTEx". In: *The American Journal of Human Genetics* 98.4 (2016), pp. 697–708.

[4] John Barnard and Donald B Rubin. "Miscellanea. Small-sample degrees of freedom with multiple imputation". In: *Biometrika* 86.4 (1999), pp. 948–955.