Estimating Missing Data's Effects on Causal Inference with Diff-in-Diff and IP-weighting

Nathen Byford

2024-12-16

Introduction

Missing data is a common problem among statistical analyses. Data can be missing due to a variety of reasons, form a subject not answering a question, to a subject leaving a study for one reason or another. Sometimes missing data is numerous and other times a study can have no missing data. Often times when a study has plentiful missing values classical statistical methods using the complete cases will be biased and something is needed to be done.

In causal inference the issue of missing data is no different, there can be unintended bias introduced based on values that are missing. Causal inference methods might have more or less bias introduced by missing data due to the fact that we are trying to estimate counter factual outcomes, outcomes that don't exist in the first place. These estimates for the counter factual outcomes are based on the data observed in the study and if values are missing, information about the counter factuals is also being lost. Because of this I aim to investigate how causal inference estimates differ when there is missing data.

Missingness in Data

It is important to understand the different types of missing data that can emerge in studies. Missing data can be classified into three main types: Missing Completely at Random (MCAR), Missing at Random (MAR), and Missing Not at Random (MNAR), each with distinct implications for analysis Little and Rubin (2019). MCAR occurs when the missingness is entirely unrelated to both observed and unobserved data, meaning the data is missing purely by chance. In this case, traditional statistical techniques like complete case analysis remain valid, as the missing data introduces minimal bias. MAR arises when the probability of missing data is related to observed data but not to the missing values themselves. While this scenario introduces bias, it can often be addressed through techniques like multiple imputation that account for the relationship between observed variables and missingness. MNAR, on the other hand, is the most challenging type of missing data, where the missingness is directly related to the unobserved data. For example, patients may drop out of a study because their condition worsens. MNAR often introduces significant bias that cannot be addressed using standard techniques without strong assumptions. Specialized methods, such as machine learning-based imputation, are typically required to mitigate the impact of MNAR data.

Methods

This study looks into the differences in the estimated treatment effect for compete case analysis in MNAR data compared to the imputed data estimated treatment effects. The following subsections go into detail about the methods of data imputation and causal inference to estimate the treatment effect.

Data Imputation

Using Machine Learning techniques Haliduola, Bretz, and Mansmann (2022) are able to impute MNAR data from a clinical trial for anxiety medication. The first step of the imputation was to cluster the data by tox response. This response curve is used to group better understand the differences between subjects based on their initial and continued response to the drug. Due to the time component of the data a recursive neural network was utilized in the data imputation. In addition due to the small sample size of some cluster over sampling was used in the training dataset. Because of this method, the data that are MNAR can be imputed with minimal loss of information and induced bias.

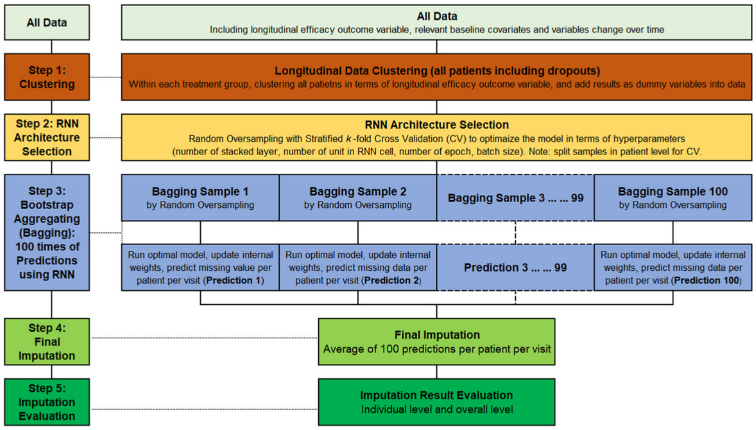
All Data
All Data
Including longitudinal efficacy outcome variable, relevant baseline covariates and variables change over time
Step 1: Clustering
Longitudinal Data Clustering (all patients including dropouts)
Within each treatment group, clustering all patietns in terms of longitudinal efficacy outcome variable, and add results as dummy variables into data
Step 2: RNN Architecture Selection
RNN Architecture Selection
Random Oversampling with Stratified k-fold Cross Validation (CV) to optimaize the model in terms of hyperparameters (number of stacked layer, number of unit in RNN cell, number of epoch, batch size). Note: split samples in patient level for CV.
Step 3: Bootstrap Aggregating (Bagging): 100 times of Predictions using RNN
Bagging Sample 1 by Random Oversampling
Bagging Sample 2 by Random Oversampling
Bagging Sample 3 … … 99
Bagging Sample 100 by Random Oversampling
Run optimal model, update internal weights, predict missing value per patient per visit (Prediction 1)
Run optimal model, update internal weights, predict missing data per patient per visit (Prediction 2)
Prediction 3 … … 99
Run optimal model, update internal weights, predict missing data per patient per visit (Prediction 100)
Step 4: Final Imputation
Final Imputation
Average of 100 predictions per patient per visit
Step 5: Imputation Evaluation
Imputation Result Evaluation
Individual level and overall level

Figure 1: Data Imputation Process

Causal Inference Methods

Two causal inference methods will be used to estimate the treatment effect of the anxiety drug from the baseline checkup to the final checkup. Difference-in-Difference (Diff-in-Diff) and Inverse Probability Weighting (IP-weighting) are two widely used methods for estimating the average treatment effect (ATE) in causal inference. Diff-in-Diff compares changes in outcomes over time between treated and untreated groups, leveraging the assumption that trends would have been parallel in the absence of treatment. This method is particularly useful when pre-treatment data are available and helps account for time-invariant confounding. IP-weighting, on the other hand, uses propensity scores to create a pseudo-population where the treatment assignment is independent of observed covariates, thereby adjusting for potential confounding. Combining these methods can provide complementary insights: Diff-in-Diff focuses on changes over time while IP-weighting ensures balance in baseline characteristics, offering a robust approach to estimating the ATE in the presence of complex data structures or confounding variables.

Analysis

Exploratory Data Analysis

The response variable is based on the Hamilton Anxiety Rating Scale (HAMA), therefore a lower score represents a better response. These scores where observed at a baseline at week 0 and then after treatment at week 1, 2, 4, and 6. Below in figure 2 we can see the scores of each subject at each checkup, untreated is on the left and treated subjects are on the right. The observations shown as a red “X” are ones that contain missing score values. There are 80 missing scores in the dataset.

In the following tables we can see how these missing scores are distributed. We can see in table 3 the split of the missing values between treated and untreated. The split is fairly even with 38 missing scores in the

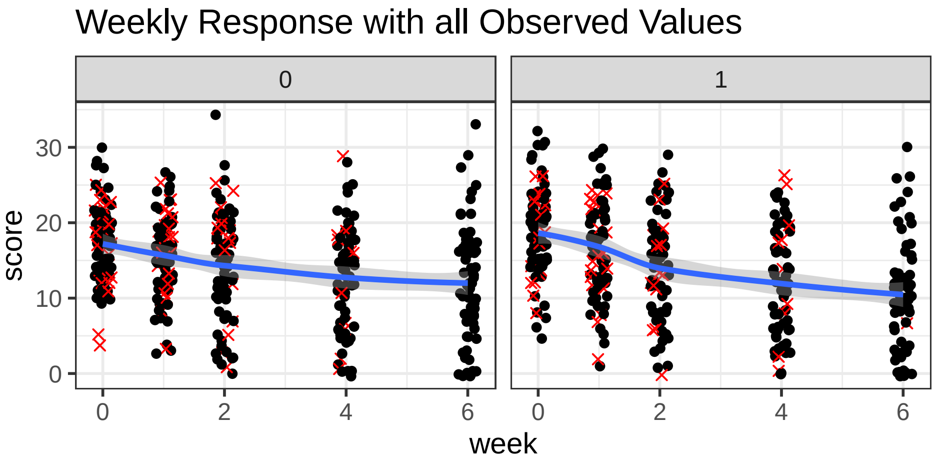

Figure 2: Observed responses by Week

treatment group and 42 missing scores in the untreated group. We can see that there is slightly more missing values in the untreated group. Looking in table 4 the missing values incease as time goes on in the study. This is possibly due to the drug potentially not working or even the oposite. This time related missingness is a sign that there can be a reason the data are missing and a pattern that could be causing the missing.

Complete Case analysis

Using the complete cases in the dataset 80 observations are lost due to missing response values. Using complete case analysis will most likely result in biased estimated due to the previously mentioned MNAR nature of the data. The goal of causal inerence is to estiamte the average treatement effect and doing so often includes estimating counterfactual outcomes. The estimated values of the counterfactual outcome may be biased by the missing information in the missing data. More sophisticated causal inference methods account for the fact that the counterfactuals are missing themeselves and may have less bias estimates.

The first method used is Diff-in-Diff, this method relies on the assumptions that the treatment and control are similar with parallel trends in the outcome. In figure 3 the trend lines for the complete case analysis can be seen. Most importantly we can see that the trend lines from baseline to week 1 are parallel, if there is some period before the drug takes effect this provides evidence that the parallel trends assumption is correct. Additionally Knowing that the data comes from a clinical trial we can say that the the treatment and control were most likely well randomized.

Fitting the Diff-in-Diff model the results are shown in table 1 are not significant at the 5% significance level. What we do see is a a slight decrease in the score for the treated group compared to the control group. The 95% confidence interval for the ATE, shown as DID, is between -1.1 and 0.003. This interval includes 0 so we cannot state that the ATE is not 0 at this significance level.

The next method tested was IP-weighting, this method relies on the additional assumption of positivity. Calculating the weights results all positive values between 0 and 1 satisfying positivity. The results for IP-weighting are shown in table 2. In this table we have the treatment variable, week/time variable, and interaction term between the two. This interaction term is what provides an estimate of the ATE.

Table 1: Difference in difference estimates week 0 to 6

Characteristic	**Beta**	**95% CI**[1]	**p-value**
treat	1.6	-0.69, 3.9	0.2
week	-0.86	-1.2, -0.48	<0.001
DID	-0.51	-1.1, 0.03	0.063

[1] CI = Confidence Interval

Table 2: IP-weighting Estimates week 0 to 6

Characteristic	**Beta**	**95% CI**[1]	**p-value**
treat	3.1	0.86, 5.4	0.007
week	-0.86	-1.2, -0.49	<0.001
treat * week	-0.58	-1.1, -0.05	0.033

[1] CI = Confidence Interval

Imputed Values analysis

After imputing the missing values using the machine learning method, the same causal inference methods are applied to the imputed dataset to determine how missing data effect these methods. These imputed values can be seen in figure 4.

Table 3: Difference in difference estimates with imputed values weeks 0 to 6

Characteristic	**Beta**	**95% CI**[1]	**p-value**
treat	1.4	-0.56, 3.4	0.2
week	-0.66	-0.99, -0.33	<0.001
DID	-0.54	-1.0, -0.07	0.025

[1] CI = Confidence Interval

Table 4: IP-weighting estimates with imputed values weeks 0 to 6

Characteristic	**Beta**	**95% CI**[1]	**p-value**
treat	2.8	0.83, 4.8	0.006
week	-0.66	-0.98, -0.34	<0.001
treat * week	-0.61	-1.1, -0.14	0.011

[1] CI = Confidence Interval

Conclusion and Discussion

The analysis highlights the impact of missing data on causal inference estimates and the benefits of using imputation methods to address MNAR data. When comparing the results from complete case analysis (Tables 1 and 2) with those from the imputed dataset (Tables 3 and 4), several key differences emerge. For both Difference-in-Difference (Diff-in-Diff) and IP-weighting, the imputed dataset yields slightly different point estimates. For instance, the estimated treatment effect from Diff-in-Diff changed marginally from -0.51 (95% CI: -1.1, 0.03) to -0.54 (95% CI: -1.0, -0.07). Similarly, for IP-weighting, the interaction term's estimate shifted from -0.58 (95% CI: -1.1, -0.05) to -0.61 (95% CI: -1.1, -0.14).

More notably, the confidence intervals in the imputed dataset are narrower across both methods. This suggests that imputing the missing data reduces the variability in the estimates, likely by leveraging the additional information provided by the machine learning imputation process. For example, the confidence interval for the Diff-in-Diff treatment effect became tighter, providing greater precision and allowing significance to be detected at the 5% level (p = 0.025). These narrower intervals indicate increased statistical power when accounting for missing data, underscoring the value of sophisticated imputation techniques in handling MNAR scenarios.

Overall, the results demonstrate that complete case analysis risks bias and reduced precision due to the exclusion of incomplete observations. By contrast, imputation improves both the validity and reliability of causal effect estimates. These findings emphasize the importance of employing advanced data imputation strategies in studies with substantial missingness, particularly when the missing data mechanism is not random. Future research should explore the generalizability of these methods across different datasets and consider alternative imputation approaches to further refine causal inference under MNAR conditions.

References

Haliduola, Halimu N., Frank Bretz, and Ulrich Mansmann. 2022. "Missing Data Imputation in Clinical Trials Using Recurrent Neural Network Facilitated by Clustering and Oversampling." *Biometrical Journal* 64 (5): 863–82. https://doi.org/10.1002/bimj.202000393.

Little, Roderick J. A., and Donald B. Rubin. 2019. *Statistical Analysis with Missing Data*. John Wiley & Sons.

Appendix

A1: EDA

Table 5: Missing Values by Treatment

Treatment	NA count
1	38
0	42

Table 6: Missing Values by Week

Week	NA count
0	0
1	0
2	14
4	23
6	43

A2: Complete case analysis

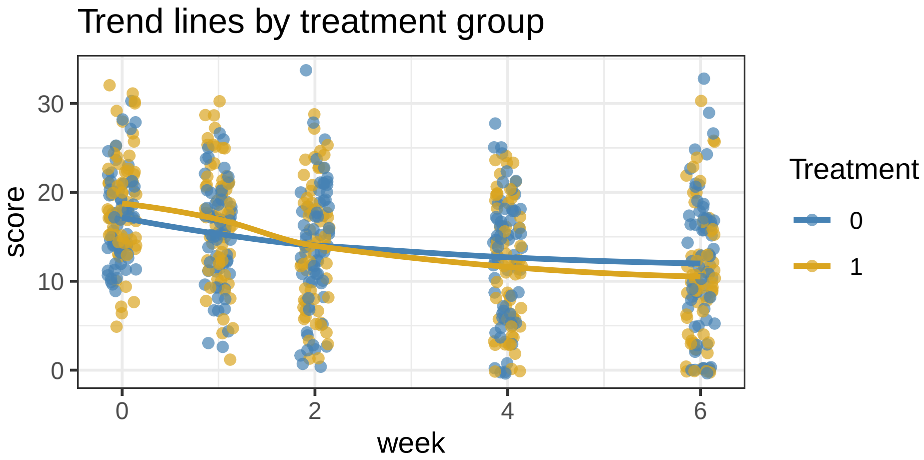

Figure 3: Treatment and Controll Trend Lines by Week

A3: Imputed Data

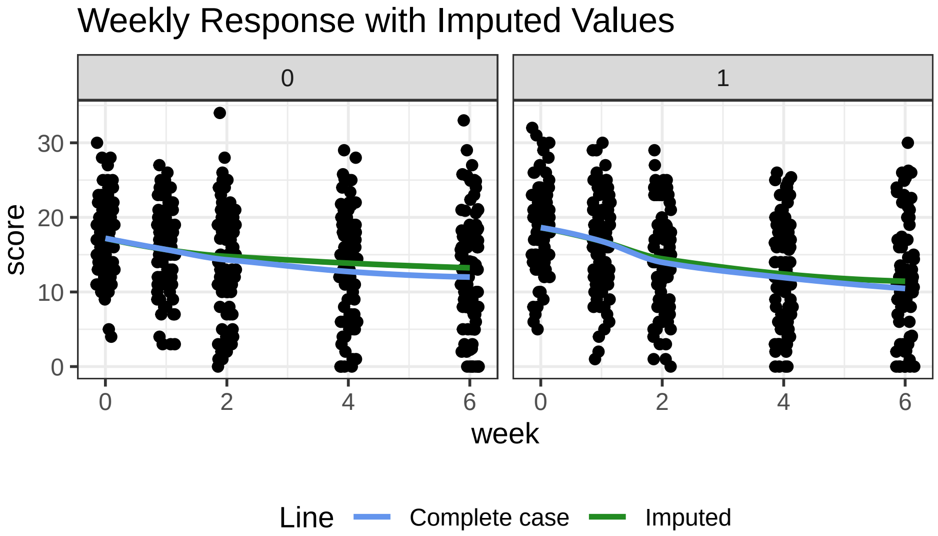

Figure 4: Plot of imputed data values and trend