POTENTIAL DEVIATION FROM MAR

SHUYI TAN

1. **Introduction**

In clinical studies, subjects sometimes are assessed by either questionnaires or interviews on a voluntary basis. Missing data as a result drop-outs is a common problem in statistical analysis of clinical data. In addition to drop-outs, some subjects can be missing at one follow-up time and then measured again at one of the next, which leads to complicated and challenging data missing patterns.[1] It is widely acknowledge that a reckless complete-case analysis, in which incomplete records are dropped, may causes selection bias and therefore results in imprecise estimates. Therefore, the analysis of data with missing values is in need of careful planning and attention.

To illustrate the concepts in a more detailed way, it would be helpful to clarify the mechanisms of data missingness. There are three typical mechanisms of missing data specified by Little and Rubin: missing complete at random (MCAR), missing at random (MAR), and missing not at random (MNAR).[2] If the data missingness occurs for reasons unrelated to the analysis question, and hence independent of the variables of interest, the data is said to be MCAR. [3] Under MCAR, the incomplete dataset is representative for the complete dataset.Therefore, an analysis that solely make use of complete cases require the assumption of MCAR to ensure unbiased reulsts. A more common situation is MAR, in which the missing data is conditionally independent of missing data given the observed value. In other words, it allows estimates of the missing data based on the subjects with complete data.[4] Otherwise, if the missing data depends on missing data, and the dependency continues even given the observed data, then the data is said to be MNAR.

Under the prevalent assumption of MAR, researchers have developed various plausible methods such as the full information direct maximum likelihood and multiple imputation to yield unbiased results. However, the assumption of MAR sometimes may be not appropriate because the MAR and MNAR cannot be distinguished based on the observed data. It is easy to understand since the unobserved value is unknown and cannot be assessed if unobserved can be predicted based on observed data. [4] The main objective of this project is explore the outcome of data analysis with missing values that is suspect of MNAR. A dataset that are potentially MNAR will be modelled based on the MAR assumption and different MNAR scenarios receptively. Then, the results will be compared with sensitivity analysis to explore whether the deviation from MAR is the case and the its impact.

2. **Sensitivity Analysis**

2.1. **Dataset Overview.** The illustrative data is a subset of dataset collected from a longitudinal cohort study of people living with HIV in the New York City. Subjects was recruited in 1994 from a a large number of medical care and social service agencies serving HIV. The

Date: April 20, 2021.

subset of data is restricted to the first sixth round of interviews.[5] There are 532 observations with 8 variables in the dataset as described in Table 1:

Vairable	**Type**	**Description**
log_virus	Numeric	Log of self reported viral load level
age	Numeric	Age at time of the interview
mental	Factor	Binary measure of poor mental health (1=Yes, 0=No)
damage	Numeric	Ordered interval for the CD4 count
income	Factor	Annual family income in 10 intervals
treatment	Factor	A continuous scale of physical health with a theoretical range between 0 and 100
healthy	Numeric	Continuous scale of physical health w

Table 1. Dataset Overview

Regarding data missingness, a missing value in the log virus load level was assigned to individuals who either could not recall their viral load level, did not have a viral load test in the six month preceding the interview, or reported their viral loads as ”good” or ”bad” [6]. Table 2 summarized the proportion of missing values in each variables. It shows that there exist a large number of missing data in the main variable of interest, which exceeds the most commonly acceptable percentage 25%.

Variable	**log_virus**	**age**	**income**	**healthy**
	33.6%	4.6%	7.1%	4.6%
Variable	**mental**	**damage**	**treatment**	
	4.6%	11.8%	4.6%	

Table 2. Data Missingness of Each Variable

Considering the nature of clinical studies, nonignorable drop-out is the most frequently encountered situation. [7] In addition, subjects who lost to follow-up have a much higher predicted $CD4^+$ T cell count than other subjects.[8] Usually patients with higher poor health may be less likely to complete questionnaires or interviews, even if the extent to which this occurs cannot be ascertained from the observed data. Therefore, we assume that non-responders tend to have high viral load than the responders, which means that the data is potentially MNAR. We considered that the missing viral load data may be MNAR, while the MAR assumption is also likely to hold.

2.2. **Complete Case Analysis.** To acquire a baseline result for the following compassion, we will start from a complete-case analysis, in which all drop-out subjects are excluded. To model the association between mental health (as measured by a binary variable) to the self-reported viral load (transformed into a binary variable) with five covariates (age, damage, income, and treatment), a logistic regression model is fitted to the complete data. The results are summarized in Table 3, which presents the estimated regression coefficient and standard error of viral load adjusted on all other covariates, the count of each level of viral load, and the percentage of patients with a given modality of viral load respectively. In the complete-case analysis, the association between mental health and viral load is statistically insignificant with a large p-value.

	Estimate	SE	Count	% of Viral Load	p-value
$<$ 500 c/ml	-	-	179	53.6%	-
$\geq$ 500 c/ml	0.434	0.291	156	46.6%	0.1353

Table 3. Results of Complete Case Analysis

2.3. **MAR Analysis.** When the mechanism of missing data is MAR, valid results can be obtained by using appropriate methods, among which a multiple imputation is an efficient approach. The idea of multiple imputation follows from regression imputation, it uses the observed data to predict the missing values, but appropriately takes into account the uncertainty in the imputed values. To achieve this, latent observations are replaced by proper values acquired from an appropriate predictive distribution of the unobserved data given the observed data. Parameters of the imputation model under the assumption of MAR are estimated using the 'mice' function in R. Specifically, the imputation method to be used for each variable in data is based on the characteristic of each variable respectively. For example, the missing values in the column of binary viral load will be imputed by logistic regression. The process is repeated several times to create five complete datasets. The analysis model is then fitted to each 'complete' dataset, and the results are combined for inference, which incorporate the uncertainty both sampling uncertainty and uncertainty due to missing data.[3]

The results obtained from the procedure described above under MAR with $M = 5$ imputations are presented in Table 4.

	Estimate	SE	Count	% of Viral Load	p-value
$<$ 500 c/ml	-	-	253	50.0%	-
$\geq$ 500 c/ml	0.484	0.281	253	50.0%	0.057

Table 4. Results of MAR Analysis

Compare the result of MAR analysis to that of complete case analysis, it can be observed that the estimated regression coefficients of the viral load increases and the association between poor mental health and the viral load becomes more statistically significant based on the most commonly used level of significance ($\alpha = 5\%$).

2.4. **MNAR Analysis.** Having reviewed some of the methods for conducting sensitivity analysis under MNAR, I will focus on one particularly accessible approach that reflects possible deviation from the MAR assumption, where the data are multiply-imputed and then modified to reflect plausible MNAR scenarios. Following Noemie Resseguier's method of multiple imputation by chained function, the imputation model obtained from the previous MAR analysis will be modified by specifying supplementary parameters θ that allows us to specify that the distribution of the variable of interest is different among subjects with missing value and among subjects without missing value, conditionally on all variables included in the imputation model.[9] This is inspired by Rubin's pattern-mixture models and be expressed by a simple equation:

$$x_{missing}|\mathbf{x}_{\text{other variables}} = x_{missing}|\mathbf{x}_{\text{other variables}} \times \theta \tag{1}$$

As mentioned above, the variable of interest: viral load has been transformed into a binary variable, and it will be imputed by logistic regression. In this case, the supplementary parameter will be expressed as the odds ratios, which corresponds to the comparing the modality of interest, viral load in our case, among subjects with missing values with those who are without missing values. Since our variable is binary, θ takes only one value in our case. It is postulated that non-responders tend to have higher viral load than the responders, hence different MNAR scenarios will be explored by constructing models under $\theta = 1.2$, 1.5, and 1.8, with results being summarized in Table 5.

	Estimate	**SE**	**Count**	**% of Viral Load**	**p-value**
θ **= 1.2**					
$<$ **500 c/ml**	-	-	250	49.2%	-
$\geq$ **500 c/ml**	0.474	0.281	258	50.8%	0.078
θ **= 1.5**					
$<$ **500 c/ml**	-	-	248	48.8%	-
$\geq$ **500 c/ml**	0.472	0.261	260	51.2%	0.048
θ**= 1.8**					
$<$ **500 c/ml**	-	-	238	46.9%	-
$\geq$ **500 c/ml**	0.473	0.261	270	53.1%	0.041

Table 5. Results of MNAR Analysis

Comparing these results with MAR analysis (Table 4), the resulting regression estimates of the viral load decrease, but gain greater statistical significance with smaller larger values. If we solely interpret the results in terms of statistical significance, this may lead to completely opposite conclusion that the association between mental health and viral load is not substantial. However, this phenomenon also happens among models under different assumed MNAR scenarios, which can be observed from the p-values in Table 5. In addition, if solely observing from regression estimates, the variation between the MAR estimates and MNAR was smaller than our postulate. Therefore, it would be described that the estimates can be considered robust to the MNAR assumption.

3. Conclusion and Discussion

In summary, there is no concrete method to distinguish whether a data is MAR or MNAR because it is technically challenging to estimate unobserved data from unobserved values. However, sensitivity analyses can be conducted in order to assess whether conclusions are robust to plausible departures from the MAR assumption. By analyzing a potentially MNAR dataset, we have observed that the estimates of the variable of interest are roughly robust. However, there do exist visible impacts of the MARN assumptions. To be specific, when adopting a larger θ, which represents a stronger departure from MAR,the regression estimates decrease, but the magnitude of variations is very small, which may reveal that in this specific dataset the deviation from MAR has little impact on the estimates. Otherwise, by comparing the p-values across the complete-case analysis, MAR analysis, and MNAR analysis, it is obvious that the p-values are shifting to the direction of statistical significance, which may be signals of the impact of the MAR departure. Connecting to the data analysis in clinical studies, similar behaviors may result in misleading conclusions on whether a treatment effect is significant, which usually is of particular interest other than the magnitude of coefficient estimates.

There exist some limitations in this project that prevents us from drawing a solid conclusion on the impact of deviation from the MAR assumption. One of the biggest causes is that the deviation from the MAR in the illustrative data is small. As a result, the damage caused by the violation of the MAR assumption is relatively not obvious. Some future work may utilize a simulated dataset that artificially includes more intended and not random data missingness. For example, it could be a longitudinal dataset that involves MNAR drop-out mechanisms. The results of sensitivity analysis will be expected to demonstrate larger variations between models under the assumption of MAR and MNAR respectively. In addition, the choice of the supplementary parameters in our case relies on a pre-determined postulate that non-responders are more likely to have higher viral load than responders and then three seemly plausible values of θ are chosen. However, it would be more secure to choose a wider range of supplementary parameters to explore more MNAR scenarios.

References

[1] Ibrahim, Joseph G. and Molenberghs, Geert. *Missing data methods in longitudinal studies: a review.* Test (Madrid, Spain), 18(1), 1–43. https://doi.org/10.1007/s11749-009-0138-x

[2] Little, Roderick and Rubin, Donald. *Statistical Analysis with Missing Data, Third Edition* Wiley Series in Probability and Statistics. New York: Wiley. 1987

[3] Leurent, B., Gomes, M., Faria, R. et al. *Sensitivity Analysis for Not-at-Random Missing Data in Trial-Based Cost-Effectiveness Analysis: A Tutorial.* PharmacoEconomics 36, 889–901. 2018 https://doi.org/10.1007/s40273-018-0650-5

[4] Sterne, J. A., White, I. R., Carlin, J. B., Spratt, M., Royston, P., Kenward, M. G., Wood, A. M., & Carpenter, J. R. *Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls.* BMJ (Clinical research ed.), 338, b2393. https://doi.org/10.1136/bmj.b2393. 2009

[5] Wong, T., Chiasson, M. A., Reggy, A., Simonds, R. J., Heffess, J., Loo, V. *Antiretroviral therapy and declining AIDS mortality in New York City.* Journal of urban health : bulletin of the New York Academy of Medicine, 77(3), 492–500. 2000 Received from https://doi.org/10.1007/BF02386756

[6] Gelman,Andrew. Hill, Jennifer. Su, Yu-Sung. Yajima, Massanao. Pittau, Maria. Goodrish, Ben. Si, Yajuan. Kropkom Jon. *MI: Missing Data Imputation and Model Checking.* R package version 0.09-11; 2010.

[7] Wu, J., Ibrahim, J., Chen, M., Schifano, E., Fisher, J. *BAYESIAN MODELING AND INFERENCE FOR NONIGNORABLY MISSING LONGITUDINAL BINARY RESPONSE DATA WITH APPLICATIONS TO HIV PREVENTION TRIALS.* Statistica Sinica, 28(4), 1929-1963. 2010.

[8] Moore, C. M., MaWhinney, S., Forster, J. E., Carlson, N. E., Allshouse, A., Wang, X., Routy, J. P., Conway, B., & Connick, E. *Accounting for dropout reason in longitudinal studies with nonignorable dropout.* Statistical methods in medical research, 26(4), 1854–1866. https://doi.org/10.1177/0962280215590432. 2017.

[9] Resseguier,Noemie. *Multivariate Imputation by Chained Equations (Iteration Step for sensitivity analysis).* Package ‘SensMice’ October 04, 2010.

Appendix A. Codes

The codes are stored in github, and here is the link: codes

Department of Statistics, University of British Columbia