# Section 4 Assignment

**Question 1:** You have been provided with a dataset from an observational cohort study to evaluate the effect of daily supplementation with 2000 mg of Omega-3 Fatty Acids (1200 mg EPA + 800 mg DHA) on the number of severe asthma attacks over a three month follow-up period. The outcome is a binary [0,1] indicator of severe asthma attacks (any versus none). The exposure is an indicator of whether the person took their supplement or not. There is also information on four strong baseline confounding variables: age (in years), body mass index (in kg/m^2), asthma medication use (regular versus irregular use), and a measure of overall diet quality. All of the variables in the dataset are complete (i.e., no missing data). You've been asked to demonstrate to your clinical colleagues the types of missingness, and how they might affect the confounder adjusted association between Omega-3 supplementation and asthma attacks. Specifically, you've been asked to set a proportion of the observations for the Omega-3 supplementation exposure variable to missing under MCAR, MAR, and MNAR. The exact proportion of missingness doesn't matter. Describe three different mechanism you would use to set a proportion of the exposure variable to missing if the missingness was MCAR (mechanism 1), MAR (mechanism 2), and MNAR (mechanism 3).

RQ: association of daily supplements on asthma attacks Exposure: Omega-3 binary indicator exposure Outcome: Binary indicator of severe asthma attacks (any/none) Confounders: age (in years), body mass index (in kg/m^2), asthma medication use (regular versus irregular use), and a measure of overall diet quality

**MCAR:** Data are considered missing completely at random (MCAR) if the probability of missingness does not depend on any other variables, either measured or unmeasured. Mechanism: -randomly assign a probability of missingness and apply it to the exposure variable

**MAR:** Data are considered missing at random (MAR) if the probability of missingness depends on other variables, AND these other variables are measured and available in the data Mechanism: - figure out the probability of missingness for another measured variable in our dataset & apply that probability of missingness to the exposure variable.

**MNAR:** Data are considered not missing at random if the probability of missingness depends on other variables, AND these other variables are NOT measured, and thus not available in the data. - generate probability of missingness based on another variable, apply that probability of missigness to the exposure

variable, then delete that variable from the dataset ie: base it on bmi, then delete bmi in order to properly simulate that data not being in the dataset

**Question 2:**

**Define non-monotone missingness.** Nonmonotone missing data arises when it is not possible to arrange the data in order of increasing missingness. A typical example of non-monotone missing data is in a followup study, if various individuals miss some study visits then rejoin later on, the data is considered non-monotone missing (compared to if the individual drops off and just doesn't return)

**Describe the main complication introduced by the non-monotone missingness pattern.** The imputation process is more complicated than if the data is monotone missingness pattern. The missing data for nonmonotone in a given variable has to be imputed conditionally on another variable in the dataset that contains missingness as well.

**What type of imputation techniques can you use with non-monotone missing data.** We must use iterative imputation techniques (such as MICE), thus missing data is imputed using previously imputed data until the data converges to a stable solution (with the missing values as a function of other features). Multiple imputations involves creating multiple plausible imputations for the missing data. These imputations are based on a model that incorporates all of the data in the dataset including the missing data. The results from each of the imputed datasets are combined to obtain estimates that account for the uncertainty introduced by the imputation process.

**Question 3:** You are asked to estimate the ATE, and the outcome variable is missing 17% of it's observations. You and your colleagues are confident that this missingness is MAR. Your clinical colleague states that they want you to do a complete case analysis because you "should never impute the outcome."

Complete case analyses proceed by removing all the observations with any missing data. This approach is generally only valid if missing data are MCAR. Data are considered missing at random (MAR) if the probability of missingness depends on other variables that are measured and available in the dataset. While generally, CCA are only to be used if the missing data are MCAR, there can be a scenario where a CCA will still be valid, even though the missing outcome data are MAR.

**Describe an analytic scenario where a complete case analysis will still be valid, even though your missing outcome data are MAR.** If the missing outcome data is missing at random, you can still use complete case analysis if the variable that the outcome data missingness is dependent on is not a variable that is being used in the analysis and is independent of the outcome, and you are using a g computation or conditional regression model.

**Describe an analytic scenario where a complete case analysis will NOT be valid when outcome is MAR.** When the variable that the outcome data missingness is dependent on is related to other variables that are in the analysis (such as a confounding variable), then it is not appropriate to use a complete case analysis.

**Question 4:** Use the `section4_q4_data.csv` data to estimate the average treatment effect between quitting smoking and the difference in weight between 1971 and 1982. Use marginal standardization AND inverse probability of exposure weighting to estimate this effect, and adjust for sex, age, and race using main effects terms only (i.e., no splines, no polynomials, no interactions). Use the `mice` package to impute any missing data.

Table 1: Average Treatement Effect and the 95% CIs using Standard Deviation for the relationships between quitting smoking and change in weight between 1971-82

| Method | Average Treatment Effect | Standard Error | 95% CI |
|---|---|---|---|
| Inverse Probability Weighting | 2.97 | 0.59 | 1.81-4.13 |
| Marginal Standardization | 3.03 | 0.60 | 1.85-4.20 |