Missing data 2. Missing data mechanisms: MCAR, MAR, MNAR, and why they matter

Tra My Pham, Nikolaos Pandis, Ian R White

In the first article ***[add ref at proof stage – adjust the refs section accordingly]***, we saw that any analysis with missing data makes untestable assumptions about the missing values. The use of different statistical methods rests on different missing data assumptions, and it is important to be transparent about which assumption we are making when implementing a given method. As we will see in article 5, multiple imputation,[1] a popular approach for handling missing data, is typically performed assuming data are missing at random (described below).

Rubin[2] formally introduced the concept of the *missingness mechanism,* which describes how the chance of data being missing is associated with the values of the variables included in our analysis. Missingness is commonly categorised into Missing Completely At Random (MCAR), Missing At Random (MAR), and Missing Not At Random (MNAR). We will now illustrate these three missingness mechanisms in the example described in article 1.

The example was created using data from a randomised controlled trial comparing how probing depth on the lower anterior teeth evolves over time between two types of lingual retainers.[3] Our data set contains data on individuals' age at baseline, *age25*, which is fully observed with no missing values, and mean probing depth across six teeth at time point 1, *mean_pd1*, which has some missing values. Here, the missingness mechanism refers to how the chance of *mean_pd1* being missing depends on *age25* (<25/≥25 years old, fully observed), and the value (partially observed) of *mean_pd1*. These relationships can be presented graphically using directed acyclic graphs (DAGs), as illustrated in Figure 1. A DAG displays assumptions about the relationships between variables. The assumptions are represented by lines which connect one variable to another. These lines are directed, with a single arrowhead indicating the direction of their effects. DAGs are acyclic, meaning they cannot contain any loops in which a variable causes itself.[4]

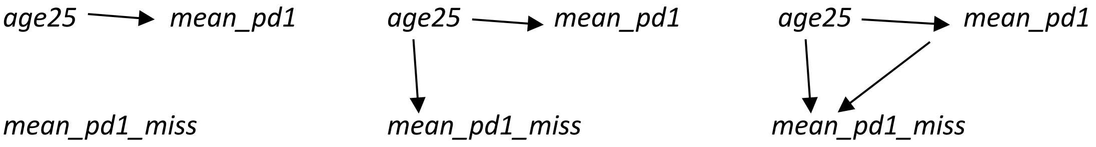
We create a variable *mean_pd1_miss* which is a binary indicator of missingnesss in *mean_pd1*, i.e. *mean_pd1_miss* takes value 1 if *mean_pd1* is observed, and 0 if *mean_pd1* is missing. We can now describe the three missingness mechanisms.

- Missing Completely At Random (MCAR): *mean_pd1* is MCAR if the chance of *mean_pd1* being missing is independent of *age25* and the (possibly missing) value of *mean_pd1*. This assumption is illustrated in Figure 1a, where there are no arrows pointing to *mean_pd1_miss* from either *age25* or *mean_pd1*. This assumption means that the missing data are fully comparable to the observed data.

- Missing At Random (MAR): *mean_pd1* is MAR conditional on *age25* if the chance of *mean_pd1* being missing is independent of the (possibly missing) value of *mean_pd1*, after controlling for *age25* (Figure 1b). This means the chance of *mean_pd1* being missing can vary with *age25*, but within each age group, the chance of *mean_pd1* being missing is the same for all individuals. In Figure 1b, both *mean_pd1* and *mean_pd1_miss* are influenced by *age25* (represented by two arrows going from *age25*), implying *mean_pd1_miss* is associated with *mean_pd1* if we do not control for *age25*. Controlling for *age25* (the common cause of *mean_pd1* and *mean_pd1_miss*) removes the association between *mean_pd1_miss* and *mean_pd1*, i.e. *mean_pd1_miss* is independent of *mean_pd1*, conditional on *age25*. This assumption means that the missing data in one age group are fully comparable to the observed data in the same age group.

- Missing Not At Random (MNAR): *mean_pd1* is MNAR given *age25* if the chance of *mean_pd1* being missing still depends on the (possibly missing) value of *mean_pd1*, even after we have controlled for *age25* (Figure 1c). This means that within each age group, the chance of *mean_pd1* being missing may still vary with the values of *mean_pd1* (e.g. *mean_pd1* might be missing more frequently for individuals with greater probing depth values compared with those with lower probing depth values who are in the same age group). As seen in Figure 1c, in addition to the arrow from *age25* to *mean_pd1_miss*, there is another arrow going directly from *mean_pd1* to *mean_pd1_miss*. Hence, *mean_pd1_miss* is not independent of *mean_pd1,* even after we have controlled for *age25*. This assumption means that the missing data in one age group are not comparable to the observed data, even in the same age group.

Figure 1. Illustration of missingness mechanisms. age25, age at baseline (≥25/<25 years old, fully observed); mean_pd1, mean probing depth at time 1; mean_pd1_miss, missingness indicator of mean probing depth at time 1; MCAR, missing completely at random; MAR, missing at random; MNAR, missing not at random.



a. MCAR b. MAR c. MNAR

Of these categorisations, MCAR is the most restrictive assumption, under which the missingness of the data does not relate to any values in the data set, whether observed or missing. This assumption is seldom plausible in practice, since there is likely other information collected in the data set that explains how values in a variable have become missing. MNAR is the least restrictive assumption but also the trickiest to deal with, as in practice we rarely know what the appropriate model for the missingness mechanism looks like. MAR is the assumption used by the standard implementation of multiple imputation (as will be seen in article 5). In the above example, we described MAR using just a single variable *age25*, but the definition extends to multiple variables. MAR can therefore be made more plausible by collecting data on additional explanatory variables that may explain the missing values.

The next article in this series will explain how we can explore the missing data in order to decide which assumption is reasonable (MCAR, MAR, or MNAR) and to plan an analysis.

References

1. Rubin DB. *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley; 1987.
2. Rubin DB. Inference and missing data. *Biometrika*. 1976;63(3):581-592. doi:10.1093/biomet/63.3.581

3. Węgrodzka E, Kornatowska K, Pandis N, Fudalej PS. A comparative assessment of failures and periodontal health between 2 mandibular lingual retainers in orthodontic patients. A 2-year follow-up, single practice-based randomized trial. *Am J Orthod Dentofac Orthop*. 2021;160(4):494-502.e1. doi:10.1016/j.ajodo.2021.02.015
4. Greenland S, Pearl J, Robins JM. Causal diagrams for epidemiologic research. *Epidemiology*. 1999;10:37-48.