**Missing data background**

To recap, the definition of missing data is the NA values appearing within datasets showing an unobserved value. A systematic report from 2021 showed nearly half the articles covering 62 studies didn’t report missing data, and over half neglected reporting how they dealt with any missingness.

**Missingness Mechanisms**

There are 3 missingness mechanisms to consider before handling missing data.

**Missing Completely At Random** assumes that the missingness has nothing to do with the observed or the missing values. Follow-up could be terminated because the study ended early.

**Missing At Random** assumes that the probability of missingness depends on the observed values. Baseline characteristics of patients may not be controllable in the study, so they could be removed.

**Missing Not At Random** assumes that the probability of missingness depends on the missing data itself. People may not fill out survey questions if they find them too sensitive or inappropriate.

**Approach to Missing Data**

**Complete case analysis** involves analysing data where the patients with missing data are removed and only patients with complete data throughout the study are analysed. If you attempt to apply simple linear regression in R on a dataset with missing data, it automatically conducts complete case analysis and removes them.

While this is a “handy” way of dealing with it, we are introducing bias into our results as we are selecting what to analyse. It also reduces statistical power as we are removing patients from the dataset instead of filling in the gaps.