Literature for further reading:

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3668100

A little more description on what is covered in this reading + some extras (no examples though). Gives some tips on how to plan good data collection.

http://dept.stat.lsa.umich.edu/~jerrick/courses/stat701/notes/mi.html

Multiple imputation in R using MICE, preceded by an intro to MCAR, MAR and MNAR. Imputation is explained below.

Missing data in survival analysis

Missing data are a common problem in research. The conclusions of analyses where the data are complete can be very different from analyses with incomplete data. How do you make sure your analysis yields the correct conclusion, even though the data are not complete?

First, you need to understand **why** some data are missing. This is important, because the techniques you decide to apply depend on the reason some data are missing. Be aware that there is no statistical test telling us why the data are missing. This is done by combining reason and knowledge on how the data were collected. Something I've emphasised throughout this course and the previous ones in the series is that there is no substitute for getting to know your data. Part of this is by tabulation and histograms etc, but another key part of it comes before any descriptive analysis – knowing how the data were generated and the potential for missing or invalid values in each data field. Let's now recap patterns of missingness.

We say that data are 'missing completely at random' (MCAR) when the complete cases (patients without any missing values for a given data item) are a random sample of the whole dataset (all patients). One patient is just as likely to have missing values as any other patient: males just as likely as females, older patients just as likely as younger ones etc. This can happen when a participant didn't have time to fill out the questionnaire or some information was lost or misplaced - and none of these things happened in a systematic way. This is the easiest situation to deal with, though sadly it's often rather an unrealistic assumption.

More often, you'll have to deal with data that are 'missing at random' (MAR). In this case, missingness can be explained by other variables for which there is full information. For example, if people with a higher education are less likely to disclose their income, then income is MAR because the chance of income values being missing depends on the patient's education. In this situation, which is pretty common, you can "fill in" the missing values on the basis of another variable, so if you

know their education you can predict their income well. Statistical methods exist to deal with this that are beyond the scope of this course, though I'll list them briefly below.

Finally, data that are 'missing not at random' (MNAR) are neither MAR nor MCAR. For example, you could be missing medical information on the severity of diabetes when they are too ill to see a doctor and provide that information; missingness depends partly on the diabetes status, as is the case for MAR, but it also depends on the severity of illness, which can't always be captured. In general, data are MNAR when the missingness is specifically related to what's missing and so the probability of the value being missing depends on unobserved variables, i.e., variables not in your data set. This is generally the most problematic type.

Now that we know what we are talking about when we say missing data, we can have a look at different methods for dealing with incomplete data. Luckily, you only need to understand the general idea and pick the right tool, as the computer will do rest of the work. Here are some of the most used techniques for handling missing data.

Complete case analysis (or available case analysis, or listwise deletion):

In this approach, the cases with missing data are simply omitted from the analysis. If the data are MCAR, this will produce unbiased estimates as long as the sample size is still sufficiently large. If the data are MAR or MNAR, the estimates will be biased. That's a good reason why you need to understand the reason for the missing values. It's tempting to just hope they're completely random, but you need to think through the problem, run some descriptive analyses and ask the data provider if necessary and possible.

Mean substitution (or mean imputation):

Replace ("impute") the missing values of a variable, with the mean of the available values of the same variable. For example, if some male patients are missing values, then just assign them the overall mean value for the male patients who do have values. This has the advantage of not changing the overall mean for that variable. However, it artificially decreases the estimated variation. It also makes it difficult to detect correlations between the imputed variable and other variables. Hence mean substitution always gives biased results and is not recommended.

Multiple imputation:

Missing variables are assumed to be MAR (or MCAR) and are imputed by drawing from a distribution. This is done multiple times and yields multiple different completed datasets. Each of these datasets is analysed, and the results are combined into a single overall result. Multiple imputation has been shown to yield unbiased results for MAR or MCAR data. It can be done in R.

Maximum likelihood:

This approach also gives unbiased results for MAR (or MCAR) data. Data are assumed to be normally distributed with a certain (multivariate) mean and variance. Observed data are used to compute the mean and variance, and missing data are drawn from the resulting normal distribution. We draw many times from the distribution until the mean and variance of the completed data are as close as they can get to that of the observed data. Fortunately, you don't have to do that yourself. There are many packages that can do that for you in R!

You may have noticed that I've not suggested any approach for MNAR data. This is because MNAR data need to be handled on a case-by-case basis. Basically, it's more complicated.

Source: survival analysis r public health