Data Analysis

In the Health Data Science workflow, after we get the data and we clean it, by which we mean make it tidy, check that all values make sense, deal with missing data, at this state we are ready to start our data analysis. Of course the type of data analysis we do, depends on the aims of our project, which we would have defined at the start, before we even got our data together.

No matter what we're trying to achieve, it's always a good idea to start with some descriptive statistics. I would argue that this could be part of the data cleaning stage as it allows you to notice possible errors, like unexpected values, and deal with them before further analysis. What do we mean by descriptive statistics? If you have categorical data, which is a variable that can take a limited number of possible values/categories, like smoking vs no smoking, or cancer stage, you could look at frequencies: how often does each of these categories occur in the data? If you have continuous data, such as blood pressure or life expectancy, you could look at typical values in the data, like the mean or the median, you could look at how spread-out the data is from the typical value, using the standard deviation or the inter-quartile range. You could also make graphs, like histograms to look at the shape of the data, and box plots that allow you to identify outliers, values that are significantly different from the rest of the data, and they could be due to the variability in the measurement, or they could indicate an error.

Following this, there are a few different things you could do. You may just want to explore and understand the data without asking a particular question. For example, imagine you have data on children's vaccinations in Scotland. There's a few different vaccines given at different ages, so you may want to look at vaccination rates for each of these vaccines, in different parts of Scotland, maybe across time - look at how the picture has changed over the years. For this, you need to put together appropriate visualisations, and maybe combine them all in a dashboard. This could be a very useful tool for a team of public health experts who want to monitor vaccine uptake.

Using this dashboard would allow you to notice things that need further investigation, for example a sudden decrease in vaccine uptake in a certain location. You'd start looking at the data more closely, and maybe develop a hypothesis about the reason for this decrease. This has now led you to an actual question you want to answer, so you'll need to design a specific study and collect data or get access to the data you need to answer the question.

Visualisations and dashboards are a great tool for exploratory data analysis, and they can lead to the generation of hypotheses.

If you do have a hypothesis you want to test, you would use the appropriate statistical tool. We won't go into details here, but what statistical test you do depends on a few things: are you working with continuous data or categorical data? Is your continuous data normally distributed? Are you looking into one or multiple variables to explain an outcome? If you don't want to make any assumptions about the relationship between your input variables and your outcome, but you have a very large dataset and want to find patterns in it, you could use machine learning methods. The most common type of machine learning used in healthcare to date is supervised machine learning, which learns to map variables to labels by looking at large amounts of pre-labelled data. For example, machine learning algorithms trained on hundreds of thousands of images have been used to detect breast cancer from mammography examinations.

So when it comes to the actual data analysis, there are many options, depending on your aims and the types of data that you have. And yet, I want to finish by reminding you of the importance of the previous step in the data science workflow: data cleaning. You could be using the most state-of-the-art analysis method, but if you have not spent the time ensuring the quality of your data: garbage in, garbage out.