The Health Data Science Workflow

Health Data Science involves using data to address healthcare problems. How does that work in practice? Here we are going to talk about the data science workflow: four general steps that we follow in any data science project.

First of all, we need the actual data. There are different ways to get health data. We may be collecting it ourselves. For example, imagine that we are interested in finding out whether a novel exercise programme has an effect on wellbeing. We would recruit our participants and then have them fill in a questionnaire at the start of the study, enrol them into the exercise programme and monitor their activity levels perhaps using a wearable device, and then have them fill in another questionnaire at the end of the study. The activity data, and the two questionnaire responses, baseline and post-experiment, are the datasets for our study. We then need to make sure we store this data in a safe way, and get on with the rest of the study. This is an example of a primary data research project.

But we could also be dealing with secondary data. This is data that already exists, for example it could be routinely collected healthcare data which is available for research under strict conditions. Let's say we have the question: In Scotland, do patients with diabetes who take their medications regularly have fewer heart attacks? To answer this question, we need to identify patients with diabetes who are on medication and information on heart attacks. There are two national datasets held by Public Health Scotland that contain this data.

1. PIS: The Prescribing Information System, that contains data about diabetes medication, and
2. SMR: Scottish Morbidity Record. If we assume that most heart attacks are treated in hospital, we can identify heart attacks by the clinical diagnosis, stored in the SMR dataset.

Before we can answer our question, we need to link these two datasets, so that we get heart attack information about the same people we've identified as people with diabetes on medication. In this example, we haven't had to collect our own data, but we've had to access and link existing datasets.

The first step in our data science workflow is to get the data, whether it's new or existing. The second step is to prepare or clean the data. This is a very important step, especially in secondary analysis of data that was not originally collected for research, so it may not be in the correct format for what we want to do with it. So we may need to organise it differently, and deal with missing values, or values that don't seem quite right.

Now, before we move on. It is important to really understand the data and the definitions of all the variables in our dataset. For example, clinical diagnoses that I mentioned earlier, in the SMR dataset, are stored using ICD codes (which stands for International Statistical Classification of Diseases and Related Health Problems). So there's a particular code for "heart attack". This sounds simple enough, but there are actually six diagnosis fields in SMR: One Main Diagnosis field (the primary diagnosis) and five 'Other Diagnosis' fields (for any accompanying / secondary diagnoses).

For our study, to remind you, we want to find out whether patients with diabetes who take their medications regularly have fewer heart attacks. So do we want to include these other diagnosis fields, to pick up everyone with a record of a heart attack – even if it was not their main diagnosis? What if this secondary diagnosis was used by doctors to refer to a heart attack that actually happened in the past? Do we want this in our dataset? This is just to make you aware of some of the challenges of working with health data. We need to make sure we have the appropriate data to answer our research question, otherwise even the most sophisticated analysis is not going to give us a sensible answer. Now this is really something we want to address at step 1, when we get our data, make sure it's the appropriate data. Unfortunately, sometimes problems do not become apparent until step 2, when we start looking into the data we have.

So after spending quite a bit of time cleaning our data, the next step is to analyse it. There's different tools we can use here, statistical hypothesis testing or prediction modelling using machine learning. The thing to remember is that all tools make assumptions that we need to be aware of, so that we apply them appropriately.

Finally, the last step in the data science workflow is communication. A very important step, because no matter how great our study design and data analysis has been, it won't be useful, unless we can communicate our results to others. How we do this depends on our target audience. Are we preparing a report for clinicians, who want to see numbers in tables? Are we presenting our results to policy makers or managers with limited time who just want a clear and short message? Are we putting together an infographic with engaging graphs for the general public? When we are deciding how to communicate our results, we need to keep in mind the needs of our target audience.

So this is the data science workflow: First, we get the data, either by collecting it ourselves, or linking existing datasets. Then, we clean our data, analyse it, and communicate our results.