Data Cleaning

Why is data cleaning such as important part of the health data science workflow?

Healthcare is one of the most complex domains as there are many different sources of health data and many different types of health data. And what we want to use data for is to improve care, help with diagnosis, make better treatment decisions. If our data is not accurate, if it's messy and incomplete, then any analysis we do with this data, any models we build will also be inaccurate. And in this domain more than any other, this can have detrimental consequences for people's health.

What causes messy data in healthcare? There could be errors or missing information when data is entered into our healthcare administrative systems. Or lack of standardisation during data collection means that the same information may be recorded differently at different places, or by different people, or there may be changes over time.

Let's look at some examples of messy data you might encounter in a health dataset. I should say that this is a completely fake dataset, based on the format of typical electronic health records.

This dataset is clearly about some kind of operation. It includes a patient ID number, the patient date of birth, the date of the operation, and some kind of score measured before and after the operation.

The first thing I want to discuss is the idea of tidy data. This is a particular structure of a dataset that makes it easier to analyse. This comes down to standardisation. If datasets have a consistent data structure, then we can develop tools that work with this structure and that makes certain common data science tasks easy because we can use these existing tools.

What does a tidy dataset look like?

There are just three rules:

1. Every variable forms a column
2. Every observation forms a row
3. Each type of observational unit forms a table

Let's start with the third rule. In our table, what is the observational unit? What is the data about? Well, it's called operation_data, so it must be about this certain type of operation. But then why does the table also contain the date of birth of patients? This is demographic information, it's not information about the operation. You can see the problem that this

creates. If a single patient has multiple operations, they would appear in this table in multiple rows, and their date of birth would appear multiple times, which might lead to errors. There is in fact one PatientID that appears twice, 1002. But their date of birth is different. Was it typed in wrong? Is it actually two different patients? This is a separate issue that we would need to solve by talking to the people who created the dataset. But in terms of a tidy dataset, this really should be two tables: a table with demographic data, which only gets updated when a new patient enters the system, and a table with data about the operations, which is updated every time a patient has an operation, whether it's an existing patient or not. Now, when we analyse the data, we might want to join these two datasets, but they should be stored separately.

Next, let's look at the other two rules. We want variables in columns, and observations in rows. This is fine for the demographics table. Our variables are patient ID and date of birth, and there is one column for each, and our observations are patients, and we have one row for each patient. But what about the operation data table? Patient ID and operation date are variables, but score pre-op and score post-op are combinations of variables and values: score is a variable, but whether this score was measure before or after the operation, that's two different values of a variable we could call date of score. Let's turn this into a tidy dataset and have a look at it. You are probably thinking: is this an improvement? The new format is much longer. It's true that the wider and shorter format is quite efficient, so it's not useless, it might be a good way to display the information. But remember, the main aim of tidy datasets is standardisation, so even though it's longer, the tidy version of the dataset is better suited for data analysis.

Our datasets are now tidy, what else about them might require some cleaning?

Missing values. Here, one of our pre-operation scores is missing. It's also possible that one of the values for date of birth is missing, as first of January 1900 is not really a valid birthdate. For this we would again need to look at what information we have about this dataset, or ask the dataset creators whether 1st of January 1900 is the code they use for missing data for this variable.

What do we do about missing data? We are not going to go into a lot of detail here, but there is basically three things we can do. We could just remove these observations from our dataset, we could in some cases leave them as they are, as some types of algorithms can work with missing data, or we can impute them, which means assign a value to them based on the rest of the data we have. It is important to check whether there is any pattern in the missing data. For example, let's say that a score was more likely to be missing if a patient was older. Then if we just removed all the operations for which the score was missing, we would be removing many more older than younger patients, and if age is associated with

whatever it is we are investigating, then that could have consequences for the results of our study.

A couple of other things to consider: are all variables within the range we expect? Here, all our operations are from 2020, expect for one that's from 2002. Is this a real date, or is it a typo? We need to check. Are all our variables measured in the same units? If, for example, our dataset contained patients' height, is that all in metres, or is it sometimes in cm, or even in feet?

Here we've discussed how to prepare our health data for analysis. Hopefully one thing that has become very clear is that before we do anything else, we really need to understand our data, be confident about what each variable represents and how missing values are coded.