Health Data Science Roles

What do health data scientists do? There isn't really a single job in the area of health data science, and even job titles are used differently in different organisations. Here I am going to describe three roles but do keep in mind that the titles I use are not unique, so you should always look carefully at the job description to get a feel for which of the three roles a particular job posting fits better with.

The three roles are data engineer, data analyst and data scientist.

Data engineers build the infrastructure to collect, store and organise the data. It's a role that requires good programming knowledge, SQL, Java or Python, command line programming, as well as understanding of data pipelines and storage systems. It is quite separate from the other two roles as it doesn't require much knowledge of the health domain – data engineers do not describe or find new insights from health data. But I'm mentioning this role because without good infrastructure, it is impossible to do the actual data science.

Moving on to the data analyst role, the focus here is to explore the data, and create visualisations and dashboards. By dashboards we mean grouping relevant information in one place to make it easier to understand and keep track of a situation described by the data. For example, a data analyst may collect information about patient admissions from different parts of a hospital and display it in such a way as to easily show how hospital bed use changes day by day.

Data analysts don't build databases (that's the data engineers' job), but they do query databases using SQL. They often use spreadsheets like Excel to perform simple analyses as well as business intelligence tools like Tableau or Power BI. These are tools that do not require programming, they are mostly based on menus and allow analysts to quite easily and quickly create engaging dashboards.

The final role, the data scientist, goes beyond what an analyst does. Data scientists have a strong background in statistics or machine learning and often have advanced degrees. They don't just describe data, they use it to answer research questions and predict future behaviours. If we go back to the example of the hospital bed use, a health data scientist would go beyond a dashboard, they might use machine learning to develop a forecasting algorithm for bed use, so that the hospital can use resources more efficiently. Business Intelligence tools are generally not sufficient for this type of work, so health data scientists use programming languages such as R or Python.

For both data analysts and data scientists, but in particular for data scientists who may be designing research studies, a deep understanding of the health domain is hugely important.

Health data is particularly challenging, it is messy and incomplete, there are many different types of health data, a lot of it unstructured, there could be biases in how data is recorded, and we could be introducing more biases in our analysis if we are not careful. So expertise in health data is crucial.

Here we've discussed three roles related to health data science: the data engineer, who designs and maintains the infrastructure, the data analyst who describes the data, and the data scientist who finds new insights. The data analyst and scientist jobs are closely related, they both need expertise in the health domain, with the scientist role requiring use of more advanced analysis methods and programming tools.