1 Numpy and Pandas

For much of healthcare data analytics and modelling we will use NumPy and Pandas as the key containers of our data. These libraries allow efficient manipulation and analysis of very large data sets. They are very considerably faster than using a 'pure Python' approach.

NumPy and Pandas are distributed with all main scientific Python distributions.

1.1 NumPy

NumPy is a library for supporting work with large multi-dimensional arrays, with many mathematical functions. NumPy is compatible with many other libraries such as Pandas (see below), MatPlotLib (for plotting), and many Python maths, stats and optimisation libraries.

When we import NumPy as as library it is standard practice to use the *as* statement which allows it to be referenced with a shorter name. We will import as *np*:

```
import numpy as np
```

1.2 Pandas

Pandas is a library that allows manipulation of large arrays of data. Data may be indexed and manipulated based on index. Data is readily pivoted, reshaped, grouped, merged.

We will use pandas alongside numpy. Generally, NumPy is faster for mathemetical functions, but Pandas is more powerful for data manipulation.

As with numpy, we import with a shortened name:

```
import pandas as pd
```