Train@Ed Postdoctoral Fellowship Project Summary

Project Title: Novel approaches to large-scale data analysis with applications in precision healthcare

Timothy I. Cannings (timothy.cannings@ed.ac.uk) and Catalina A. Vallejos (catalina.vallejos@igmm.ed.ac.uk)

In June 2018 Sir Alan Wilson, CEO of the Alan Turing Institute, said "It's clear that data science and artificial intelligence will revolutionise healthcare". This revolution is starting to be realised for two main reasons. First, new biomedical measurement techniques (e.g. imaging and genomic techniques) have become standardised and more affordable, leading to a large amount of data being available. Second, statistical and machine learning methods are being specifically designed to analyse these complex datasets in a way that is accessible to the end-user (e.g. well-documented open source software). There are of course many challenges remaining. This proposal focusses on developing efficient methods for the analysis of large-scale data. In some settings, large-scale data may be ultrahigh-dimensional, where each observation consists of measurements on a vast number (thousands, or even millions) of covariates. In other settings, we may have a large number of observations, in which case many existing methods are computationally intractable. Sometimes these problems occur in combination, with high-dimensional information being recorded across large cohorts.

This project aims to develop novel, scalable statistical methodology designed to deal with vast amounts of data. We have two specific large-scale data types in mind. First, modern methods such as high-throughput next-generation sequencing (NGS), produce vast amounts of ultrahigh-dimensional -omics data – we typically have measurements on more than 10^4 genes for each sample. Moreover, NGS technologies for -omics data are now increasingly focused on the characterisation of single cells. For instance, using single-cell RNA sequencing (scRNA-seq) we are able to measure expression levels across thousands of genes in hundreds of thousands of cells. The second source of data will be modern large-scale multineuronal recording methods, such as multielectrode arrays, calcium imaging, and optogenetic techniques. The challenge in this field is to collect and process large amounts of neural data and then accurately interpret the signals, which are incredibly noisy, continuously changing and travel throughout the body at incredible speed.