Big data – a solution to the (mis)diagnosis of depression?

Wesley Kai-Xian McLoughlin
Medical School, University of Edinburgh

What is the problem?

Depression is a common mental health problem that affects 1 in 4 of us in our lifetime and is a leading cause of global disability. Mental illness currently costs the UK £95 billion per year, in support, economy losses and treatment. Currently, depression is diagnosable clinically but there are no predictive tests. Additionally, it is sometimes misdiagnosed or missed.

Physical injuries like a broken arm are easy to spot and often easy to treat, but mental health problems are often more subtle, complex in nature and treatment.

What did I find?

I found that of the 62 biomarkers I selected to analyse from the UK Biobank data, not all were usable in the model because data was not available for all individuals in the Biobank. 47 of the biomarkers measured were retained in the final model, including Vitamin D, triglycerides (a measurement of fat in the blood) and C-reactive protein (a measure of inflammation).

The final model proved to be 65% accurate at predicting depression. Whilst this sounds impressive, unfortunately this means it had a 35% error rate. But that might be slightly different from the true value as it is fairly likely that some of the individuals in the Biobank may develop depression in the future or be depressive now and have been mis- or undiagnosed.

This represents the nature of research sometimes, but to use the old adage, "if at first you don't succeed, try again".

What am I interested in?

I wanted to investigate whether I could develop a way to predict which individuals are likely to get depression using a 'big data approach'. 'Big data' research involves using databases (which contain vast amounts of information stored electronically in an organised and accessible way) and statistical computer software to look for trends in populations of patients with regards to health and disease.

What does this mean?

As big data, data science and machine learning (use of artificial intelligence to predict and model) become more important in medicine, we can learn more about diseases easier. By using large sample sizes, like the UK biobank, we can avoid a lot of bias, confounding factors (factors that influence both the independent (measured) variable and the (outcome) variable) and generate more research output from one source.

Genetic studies and screening large populations for diseases and/or disease markers means we are collecting more research data then we can analyse, understand and process. We can use techniques like machine-learning and data science to begin to try and understand it all.

I have shown as a proof of concept that biomarkers and large-scale data can predict depressive status, however more information, larger samples and a more diverse panel of biomarkers need to be used to provide a more accurate prediction.

Who am I?

I am a 3rd year Medical Student at the University of Edinburgh. I completed this project in the summer before starting my intercalated BMedSci in Anatomy and Development. As part of the 6-year Medical course at Edinburgh, students complete a 1-year Honours degree in one of 20 subjects. As a result, students get to graduate with a BMedSci and MBChB degree! I am particularly interested in research and enjoy data science. I hope to incorporate a PhD in my medical training in the near future with the aim to becoming a clinician-scientist.

What did I do?

I had access to patient data stored in the UK Biobank, which contains information on 500,000 volunteers whose health and well-being is followed over time. The data includes basic details like age and sex, blood test results along with clinical and genetic information of the volunteers (see Figure 1).

I focussed my research on biomarkers, which are biological molecules or characteristics that can be objectively measured to show an indication of a biological process or a response to something like a drug. For example, blood glucose levels are a biomarker which give us insight into how good or bad someone's insulin control is.

I chose to analyse 62 biomarkers measured from blood samples to see if a panel of suitable biomarkers could be identified as predictive of depression.

To study this 'big data', I used R – a statistical programming language to make a 'model' which analysed and controlled for a number of factors and then predicted an outcome. In this case, the model predicted whether individuals were depressed and controlled for things like age and sex. It analysed the biobank's biomarkers and excluded biomarkers that were not associated with depression.

Once the model was produced, I tested it on a subset of the UK Biobank data, measuring how good the model was at predicting depression and comparing my findings with the actual diagnosis status of the individuals in the Biobank. I plotted graphs showing the relationship of different biomarkers and depression while evaluating the model. Figure 2 shows some of the interesting separation in the results we observed for some measurements when plotting depressed against non-depressed individuals. It shows that using the two measurements, depressed and non-depressed individuals could be distinguished. Figure 3 shows how the model was evaluated with a ROC curve (a standard way of evaluating diagnostic and predictive tests).

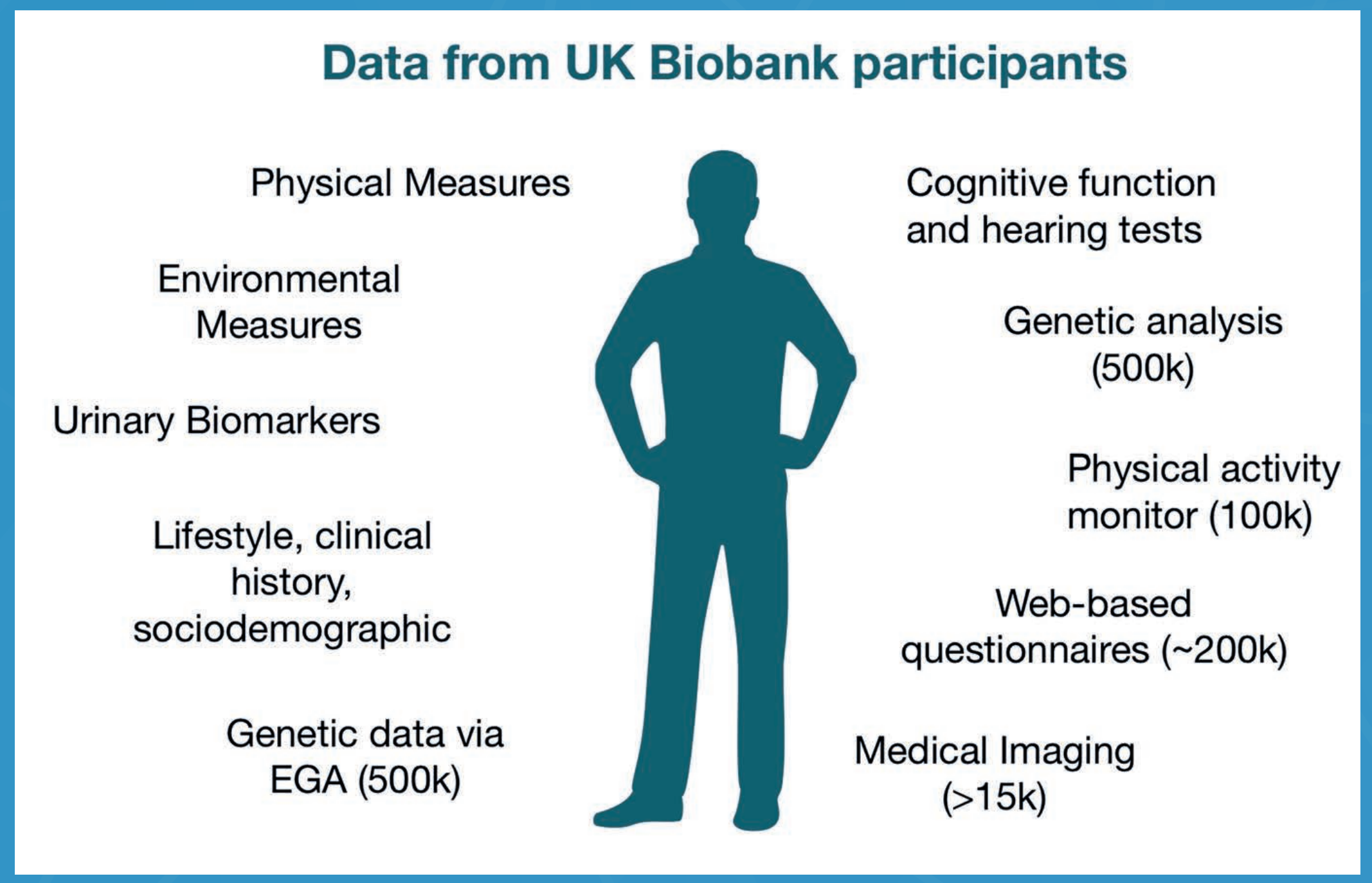
Data from UK Biobank participants
Physical Measures
Environmental Measures
Urinary Biomarkers
Lifestyle, clinical history, sociodemographic
Genetic data via EGA (500k)
Cognitive function and hearing tests
Genetic analysis (500k)
Physical activity monitor (100k)
Web-based questionnaires (~200k)
Medical Imaging (>15k)

Figure 1 Overview of the UK Biobank's collected data (adapted from figure available ebi.ac.uk), EGA: European Genome-Phenome Archive, Medical Imaging: detailed MRI scans.

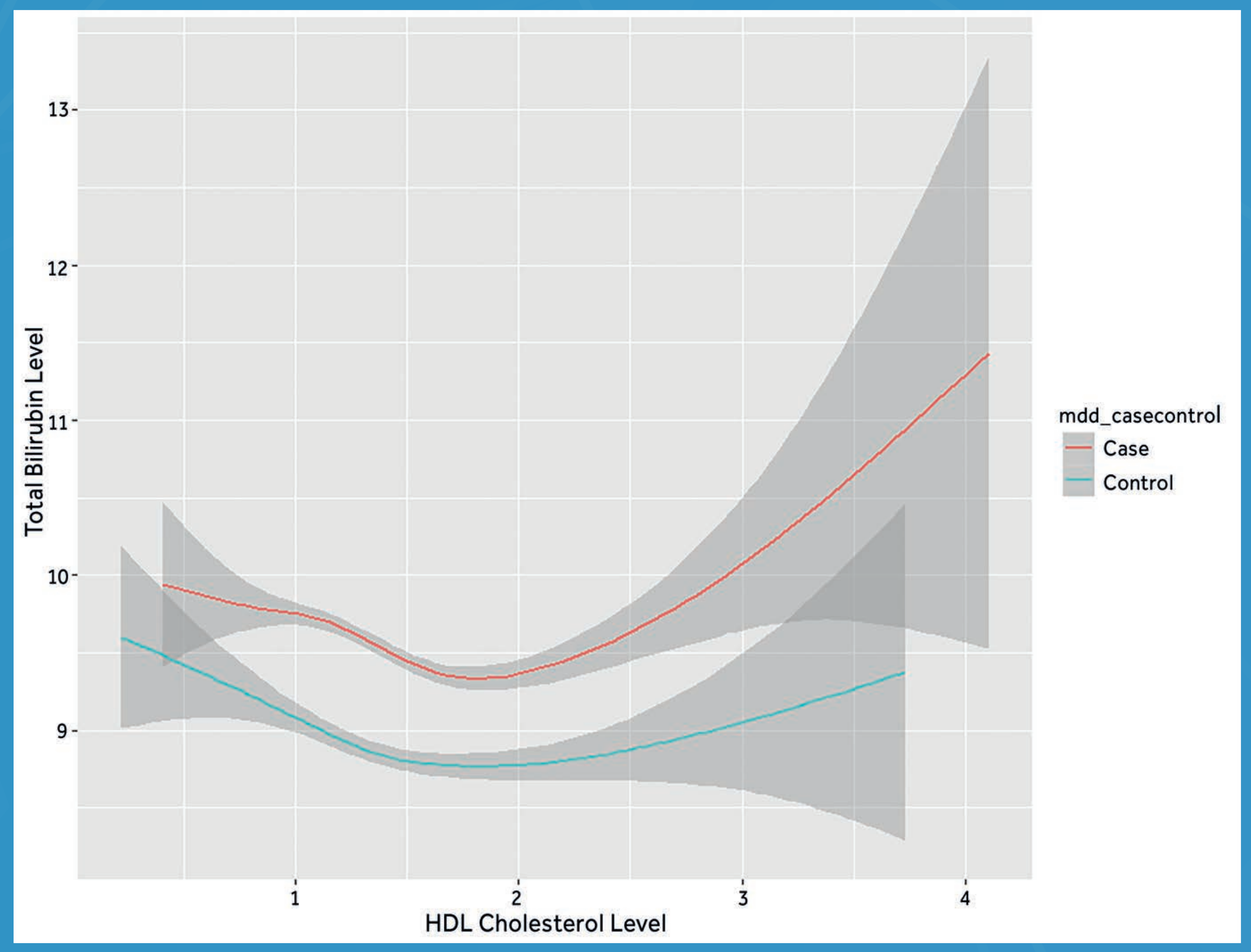

Figure 2 Graph showing measurements for two of the biomarkers analysed, HDL (high density lipoprotein) cholesterol against bilirubin, among depressed (cases) and non-depressed individuals (controls); The dark bands represent the range of uncertainty or error in the measurements.

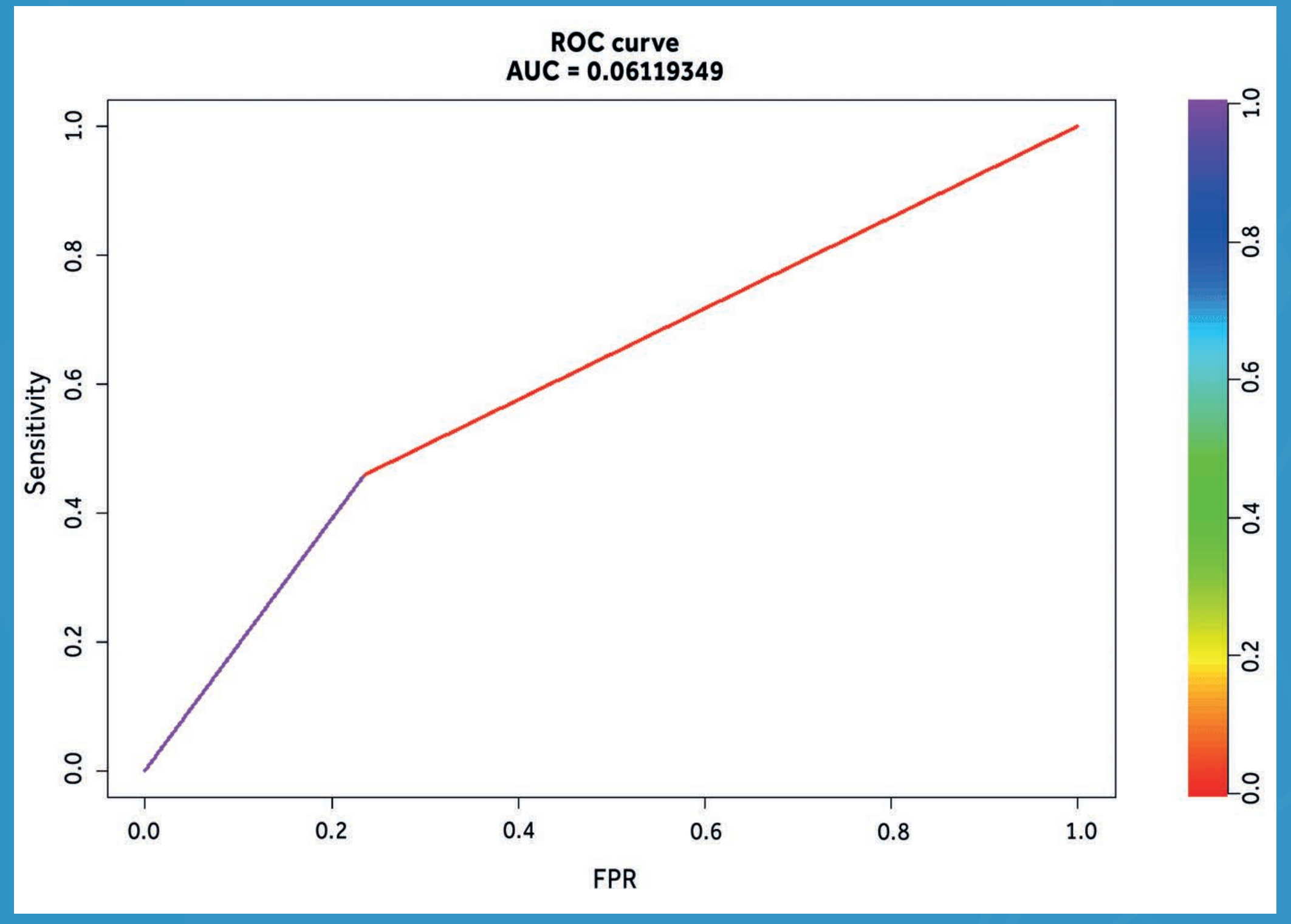

Figure 3 Receiving-Operator Characteristic (ROC) curve evaluating the model (FPR: False prediction rate), ROC is a standard method of evaluating the diagnostic or predictive value of a test or model. It compares the false and true predictive rate, AUC is a numerical representation of this. Traditionally, an AUC of 0.7 or greater is decent and 0.8 and greater is good. However, it is hard to compare this to this predictive model as it isn't a traditional diagnostic test, and is predicting the probability of an outcome as opposed to diagnosing what is already present.

Wesley was supervised during his Vacation Scholarship project, *"Analysis of large-scale UK BioBank data to establish a panel of possible predictive biomarkers for depression"*, by Professor Andrew McIntosh at the University of Edinburgh.