Statistical model helps predict when high blood pressure, heart disease and diabetes will occur

April 20 2021

Credit: Unsplash/CC0 Public Domain

One of the promises of new methods of personalized medicine is that individual risks for diseases can be assessed using large DNA datasets.

But many diseases are highly multifactorial, meaning that genetic risk factors are spread throughout the DNA. Finding these elusive connections and constructing a reliable and trackable statistical model from them is the goal of Matthew Robinson at the Institute of Science and Technology (IST) Austria and his international team.

A myriad of genetic factors can influence the onset of diseases like high blood pressure, heart diseases, and type 2 diabetes. If we were to know how the DNA influences the risk of developing such diseases, we could shift from reactive to more preventive care, not only improving patients' quality of living but also saving money in the health system. However, tracing the connections between the DNA and disease onset requires solid statistical models that reliably work on very large datasets of several hundred thousand patients.

Matthew Robinson, Assistant Professor at the Institute of Science and Technology (IST) Austria, together with an international team of researchers has now developed a new mathematical model that improves the predictive quality gained from large sets of patient genomic data. This method could help develop personalized predictions about health risks, similar to what a physician does when discussing a family's medical history.

Sampling from billions

Human DNA consists of several billion base pairs that encode the biological structure and functions. In their study, the scientists selected several hundred thousand genetic markers—short parts of the DNA sequence—for their investigations. Using their statistical model, they then linked these the composition of these markers to the onset of high blood pressure, heart disease or type 2 diabetes in the patients in the database. The researchers were specifically interested in the patients' age at disease onset. With this information, they can then use their model to

predict probabilities for when a disease might occur.

Yet, this statistical model cannot construct direct relations between certain genes and disease onset, but only provides an improved prediction of probabilities of disease onset. There is also an important difference between commonly used black-box models for big data studies and this method by Robinson and his colleagues: Black-box models produce predictions, but their inner workings cannot easily be understood by humans because of the many layers of abstraction they use. In contrast, the model by Robinson and his colleagues provides trackable statistical computations.

Being able to understand the inner workings of a mathematical model for producing predictions about health and disease onset is an important part of an ethical approach to using large sets of sensitive patient data. With this, the researcher can explain how the predictions were generated.

Using patient data

Harnessing the full potential of such predictive methods requires both effective models and the collection of large genomic datasets that comes with its own concerns of data security and privacy that both the researchers and the health care system have to address.

Strict measures of data security have to be obeyed when using patient data. Only with the permission of the respective ethics boards, the researchers were able to access anonymized patient data from state-funded biobanks—large collections of genetic patient data—in both the UK and Estonia. They used the data from the UK to build their model and the data from Estonia to test its predictive power. The latter even produced some first personalized risk assessments of disease onset. These then will be relayed through the Estonian health care system back

to the patients, giving them the incentive to take preventive steps.

The new statistical model by Robinson and colleagues is just one step towards using the full potential of large genomic datasets for preventive health care. Both the models and the data infrastructure of biobanks, together with a robust and secure data protection system, are needed to fulfill the promises of personalized predictive medicine.

The study is published in *Nature Communications*.

  More information: Sven E. Ojavee et al. Genomic architecture and prediction of censored time-to-event phenotypes with a Bayesian genome-wide analysis, *Nature Communications* (2021). DOI: 10.1038/s41467-021-22538-w

Provided by Institute of Science and Technology Austria