MedicalXpress

Calculating genetic links between diseases, without the genetic data

December 10 2019

Credit: CC0 Public Domain

Physicians use standard disease classifications based on symptoms or location in the body to help make diagnoses. These classifications, called nosologies, can help doctors understand which diseases are closely related, and thus may be caused by the same underlying issues or respond to the same treatments.

An important part of understanding disease is estimating its heritability,

that is, what percentage of disease variation in individuals is due to inherited genetic variants versus environmental causes like exposure to pollution, infections or trauma. Traditionally, to calculate the heritability of a given disease, researchers needed expensive data sets containing all kinds of medical and genetic data plus detailed knowledge of family relationships. In a new study, data scientists from the University of Chicago estimated heritability and mapped out relationships among thousands of diseases using data from electronic health records.

The study, published December 3, 2019 in *Nature Communications*, calculated statistical curves of each disease's prevalence over an average lifetime, showing which tend to strike earlier or later in life. The researchers also created "disease embeddings," or groupings of diseases that show how closely they are related to each other based on diagnostic codes and notes in the health record. Using similarities in these curves and patterns revealed by the disease embeddings, researchers could then estimate heritability and genetic correlations between diseases.

"It used to be that every new estimate of heritability or genetic and environmental correlations between diseases was a big deal," said Andrey Rzhetsky, Ph.D., a data scientist at UChicago who is the paper's senior author. "Here we were able to estimate thousands of heritability values and hundreds of thousands of correlations, doing what used to be very expensive and slow at a very large scale."

Early onset vs late onset

To build the team's statistical models, postdoctoral researcher Gengjie Jia, Ph.D., the paper's first author, used data from Truven MarketScan, a database of de-identified health claims of 151 million people in the United States over 11 years. They also included data from the Danish National Patient Registry (5.6 million people over 21 years) and the Swedish National Health Registry (9.4 million people over 44 years).

They then created disease prevalence curves that plot the percentage of people who have a disease at each age.

The curves document statistically significant changes in a condition's prevalence over the average lifespan. Different extremes and shapes of the curves show whether a disease is more prevalent at younger (early onset) or older (late onset) ages. The researchers can also identify dips or spikes in the curve that may be a sign of environmental trigger events that can influence disease, such as puberty, changes in diet, trauma or exposure to infections.

The team also built "disease embeddings," or relationships between diseases, using a neural network model to analyze several different factors around when a disease appears in a medical record. This analysis was modeled after natural language processing that defines a word's underlying semantics by analyzing its surrounding words. In a health record, a disease is like a word, and the historical record of conditions they develop over a lifetime form a sentence. For example, "headaches" might later be followed by "migraines" as physicians narrow down a diagnosis. Therefore, when you plot them on a two-dimensional map, headaches would appear closer to migraines than, say, stomach cramps.

"The system is learning from real sequences in the patient data by optimizing 20 parameters for each disease," Rzhetsky said. "From that context, given a patient's past health history, the network is trying to anticipate what comes next. You can think about it like what happens in the doctor's mind as they make a diagnosis."

Identifying new patterns

As they studied the data, several patterns started to emerge. In the U.S. data, early onset diseases outnumbered late onset conditions, but were less prevalent in the population. This could be because routine newborn

screening and monitoring of children tends to identify more diseases, or because diseases with a strong genetic component tend to strike earlier and cause more deaths.

When two diseases are closely correlated by genetics alone, the shapes of their prevalence curves are likely to be very different. If they are linked only by environmental factors, they are much more similar, but the curves are most similar when both environmental and genetic correlations are high.

The researchers also saw that some diseases that would appear to be closely related, like psychiatric conditions, clustered into different groups based on mean onset age. Attention deficit hyperactivity disorder and autism, for example, are early onset, whereas schizophrenia, bipolar disorder and depression tend to be late onset.

Jia said that this initial run with such large health datasets validates their approach to classifying diseases based on similarity of the shapes of the curves. While at a high level, the result matched commonly accepted classifications and associations between groups of diseases, it did identify some surprises. For example, parasitic infections were found to align with an array of noninfectious diseases, such as neurofibromatosis, tympanic membrane disorders of the ear, osteogenesis imperfecta (brittle bone disease) and congenital eye anomalies.

The disease prevalence curves, standardized across age and sex, have never previously been systematically compared like this study does (click here to see a searchable database of sex-and-country-stratified prevalence curves for over 500 diseases). Now, the team hopes to refine these tools and use them to help fill in the gaps for understudied conditions.

"Our estimates can be used for deciding where to allocate research

resources," Rzhetsky said. "Does this disease have a stronger genetic or environmental component? We did this through a whole spectrum of diseases, so it's a general tool that can be applied to other conditions as they arise."

More information: Gengjie Jia et al, Estimating heritability and genetic correlations from large health datasets in the absence of genetic data, *Nature Communications* (2019). DOI: 10.1038/s41467-019-13455-0

Provided by University of Chicago