Learning to predict diabetes

Machine learning could be used to identify high-risk patients and improve healthcare services.

Machine learning approaches can effectively identify patients who are likely to develop diseases such as diabetes, according to new research from KAIMRC.

A team led by Sherif Sakr evaluated how well several machine learning methods predicted diabetes incidence in patients who had taken part in the Henry Ford Exercise Testing (FIT) project, which collected metabolic, medical

BRAIN LIGHT / ALAMY STOCK PHOTO

Blocking FSH with an antibody produces a dramatic effect on body fat in mice.

and demographic data from thousands of patients undergoing exercise treadmill stress testing. Sherif's team used data from more than 32,000 of the patients who had had a five-year follow-up and no previous heart conditions or coronary artery disease. About 5,000 of the patients had developed diabetes by the time of the follow-up.

Sakr's goal was to determine which machine learning algorithm could reliably predict a patient's likelihood of developing diabetes with high accuracy. "If you can identify people who are at risk before they get sick, that would save a lot of money and resources and improve healthcare service. Treating people before they get sick is much more effective and efficient," Sakr says. He believes machine learning can achieve that goal.

The team identified 13 relevant clinical variables in the FIT data and tested several classification algorithms to determine how well they predicted a patient's odds of developing diabetes. They aimed to use as few variables as possible to increase the model's practicality. Initially, the algorithms had a performance of about 70%, with none emerging as a clear winner.

However, the FIT data suffers from an imbalance common to most healthcare datasets, with many more representatives of one class than another — far more healthy patients than diseased patients, for example. Such an imbalance hampers the algorithms and posed a major challenge in this project.

To rectify this, Sakr's team evaluated two approaches. Removing data points from the larger class (healthy patients) to reduce the difference had little impact on the outcome. However, increasing the size of the smaller class (diseased patients) by generating random new data with similar characteristics significantly improved performance, raising it to nearly 92%, significantly better than traditional statistical techniques.

The study has clearly shown that machine learning can identify high-risk patients, improving efforts at disease prevention and guiding the management of hospital resources. "There's a huge amount of data in the healthcare domain, but there hasn't yet been effective use of it," says Sakr, who plans to cross-validate these models with another cohort, and is developing models that can accurately predict the likelihood of several other significant diseases.

Alghamdi, M., Al-Mallah, M., Keteyian, S., Brawner, C., Ehrman, J. et al. Predicting diabetes mellitus using SMOTE and ensemble machine learning approach: The Henry Ford ExercIse Testing project. PLOS ONE 12, e0179805 (2017).