

How Machine Learning Can Identify Patients at Risk of Diabetes

Samara Rosenfeld

In the U.S., more than 7 million patients have undiagnosed Type 2 diabetes mellitus. But a recent [study](#) found that by using machine learning on data that already exist in a patient's electronic health record (EHR), large populations of patients at high-risk of the condition can be predicted with 88% sensitivity.

What's more, the machine learning model had a positive predictive value of 68.6%.

Chaitanya Mamillapalli, M.D., endocrinologist at Springfield Clinic in Illinois, and Shaun Tonstad, principal and software architect at Clarion Group in Illinois, along with their research team, aimed to evaluate a machine learning model to screen EHRs and identify potential patients with undiagnosed Type 2 diabetes mellitus.

Mamillapalli told *Inside Digital Health*[™] that the team extracted data from an EHR at the Springfield Clinic. The data extracted was based on non-glucose parameters, including age, gender, race, body mass index, blood pressure, creatinine, triglycerides, family history of diabetes and tobacco use.

The team had an initial sample size of 618,022 subjects, but only 85,719 subjects had complete records.

After extracting the data, the subjects were equally split into training and validation datasets.

In the training group, the machine learning model was trained using the decision jungle binary classifier algorithm based on the parameters to learn if a subject is at-risk of diabetes.

The validation set classified the risk of the disease from the extracted non-glycemic parameters.

The validation subject probabilities were then compared to how the team defined Type 2 diabetes mellitus — random glucose greater than 140 mg/dL and/or HbA1c greater than 6.5%.

The predictive accuracy was also measured with area under the curve for the receiver operating characteristic curve and F1-score.

In the dataset, the model identified more than 23,000 true positives and 3,250 false negatives.

If the machine learning model is deployed in the back end of the EHR, physicians will be prompted if a patient's health data shows that the patient is at high-risk of diabetes and should be screened, Mamillapalli said.

Mamillapalli said that patients generally go undiagnosed for four to six years before formally knowing that they have Type 2 diabetes mellitus. He told us that because of this, the patient is exposed to complications, which could cost up to \$33 billion per year in the U.S.

But identifying the condition as early as possible could decrease the risk of complications.

However, screening rates for diabetes is still only at 50%.

In a written statement to *Inside Digital Health*[™] from Mamillapalli, he wrote that "using an automated, scalable electronic model, we can deploy this tool to screen large chunks of the population."

Mamillapalli said that the second phase of the team's research is to change the algorithm slightly to diagnose prediabetes, which affects 90 million people, but is only diagnosed in 10% of that population.

"As the predictive accuracy is improved, this machine learning model may become a valuable tool to screen large populations of patients for undiagnosed (Type 2 diabetes mellitus)," the authors wrote.

The findings of the study, titled "Development and validation of a machine learning model to predict diabetes mellitus diagnoses in a multi-specialty clinical setting," were presented at the American Association of Clinical Endocrinologists in California.