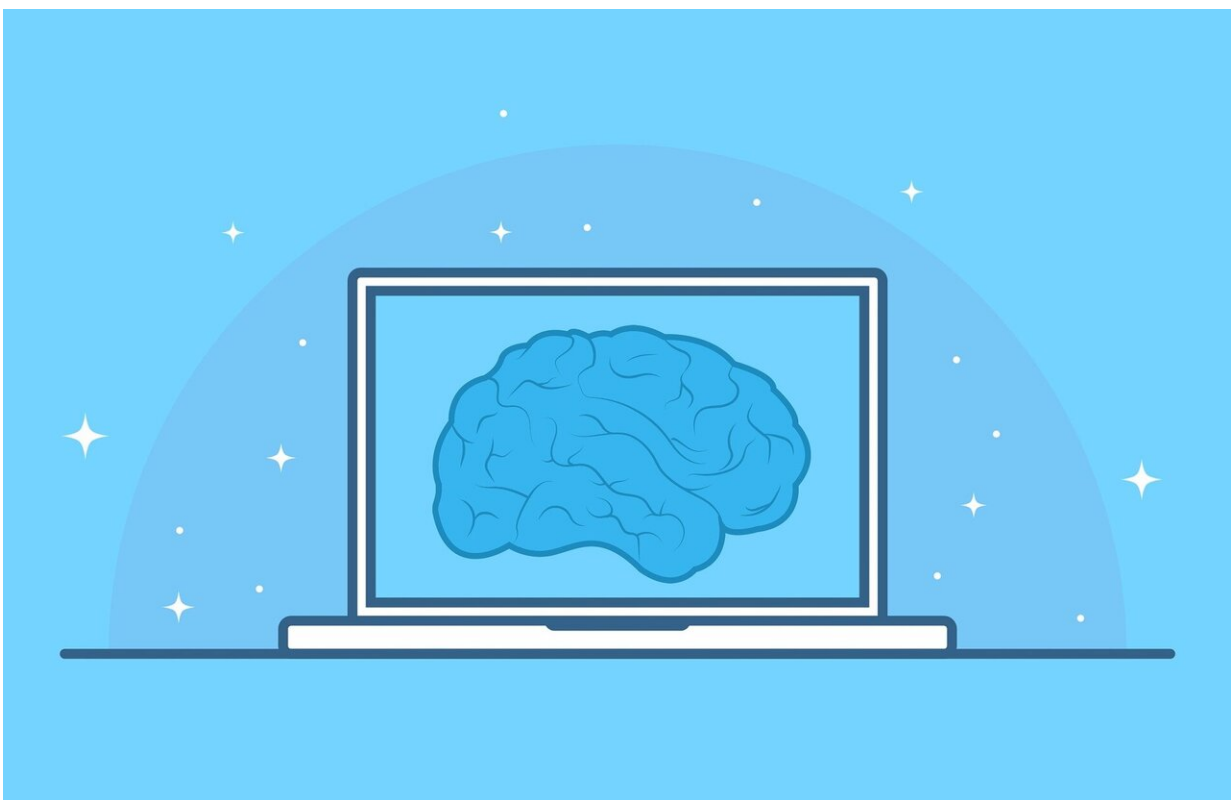# The whisper of schizophrenia: Machine learning finds 'sound' words predict psychosis

June 13 2019, by Carol Clark

Credit: CC0 Public Domain

A machine-learning method discovered a hidden clue in people's language predictive of the later emergence of psychosis—the frequent

use of words associated with sound. A paper published by the journal *npj Schizophrenia* published the findings by scientists at Emory University and Harvard University.

The researchers also developed a new machine-learning method to more precisely quantify the semantic richness of people's conversational language, a known indicator for psychosis.

Their results show that automated analysis of the two language variables—more frequent use of words associated with sound and speaking with low semantic density, or vagueness—can predict whether an at-risk person will later develop psychosis with 93 percent accuracy.

Even trained clinicians had not noticed how people at risk for psychosis use more words associated with sound than the average, although abnormal auditory perception is a pre-clinical symptom.

"Trying to hear these subtleties in conversations with people is like trying to see microscopic germs with your eyes," says Neguine Rezaii, first author of the paper. "The automated technique we've developed is a really sensitive tool to detect these hidden patterns. It's like a microscope for warning signs of psychosis."

Rezaii began work on the paper while she was a resident at Emory School of Medicine's Department of Psychiatry and Behavioral Sciences. She is now at fellow in Harvard Medical School's Department of Neurology.

"It was previously known that subtle features of future psychosis are present in people's language, but we've used machine learning to actually uncover hidden details about those features," says senior author Phillip Wolff, a professor of psychology at Emory. Wolff's lab focuses on language semantics and machine learning to predict decision-making and

mental health.

"Our finding is novel and adds to the evidence showing the potential for using machine learning to identify linguistic abnormalities associated with mental illness," says co-author Elaine Walker, an Emory professor of psychology and neuroscience who researches how schizophrenia and other psychotic disorders develop.

The onset of schizophrenia and other psychotic disorders typically occurs in the early 20s, with warning signs—known as prodromal syndrome—beginning around age 17. About 25 to 30 percent of youth who meet criteria for a prodromal syndrome will develop schizophrenia or another psychotic disorder.

Using structured interviews and cognitive tests, trained clinicians can predict psychosis with about 80 percent accuracy in those with a prodromal syndrome. Machine-learning research is among the many ongoing efforts to streamline diagnostic methods, identify new variables, and improve the accuracy of predictions.

Currently, there is no cure for psychosis.

"If we can identify individuals who are at risk earlier and use preventive interventions, we might be able to reverse the deficits," Walker says. "There are good data showing that treatments like cognitive-behavioral therapy can delay onset, and perhaps even reduce the occurrence of psychosis."

For the current paper, the researchers first used machine learning to establish "norms" for conversational language. They fed a computer software program the online conversations of 30,000 users of Reddit, a social media platform where people have informal discussions about a range of topics. The software program, known as Word2Vec, uses an

algorithm to change individual words to vectors, assigning each one a location in a semantic space based on its meaning. Those with similar meanings are positioned closer together than those with far different meanings.

The Wolff lab also developed a computer program to perform what the researchers dubbed "vector unpacking," or analysis of the semantic density of word usage. Previous work has measured semantic coherence between sentences. Vector unpacking allowed the researchers to quantify how much information was packed into each sentence.

After generating a baseline of "normal" data, the researchers applied the same techniques to diagnostic interviews of 40 participants that had been conducted by trained clinicians, as part of the multi-site North American Prodrome Longitudinal Study (NAPLS), funded by the National Institutes of Health. NAPLS is focused on young people at clinical high risk for psychosis. Walker is the principal investigator for NAPLS at Emory, one of nine universities involved in the 14-year project.

The automated analyses of the participant samples were then compared to the normal baseline sample and the longitudinal data on whether the participants converted to psychosis.

The results showed that higher than normal usage of words related to sound, combined with a higher rate of using words with similar meaning, meant that psychosis was likely on the horizon.

Strengths of the study include the simplicity of using just two variables—both of which have a strong theoretical foundation—the replication of the results in a holdout dataset, and the high accuracy of its predictions, at above 90 percent.

"In the clinical realm, we often lack precision," Rezaii says. "We need

more quantified, objective ways to measure subtle variables, such as those hidden within language usage."

Rezaii and Wolff are now gathering larger data sets and testing the application of their methods on a variety of neuropsychiatric diseases, including dementia.

"This research is interesting not just for its potential to reveal more about mental illness, but for understanding how the mind works—how it puts ideas together," Wolff says. "Machine learning technology is advancing so rapidly that it's giving us tools to data mine the human mind."

Provided by Emory University