Machine learning models rank predictive risks for Alzheimer's disease
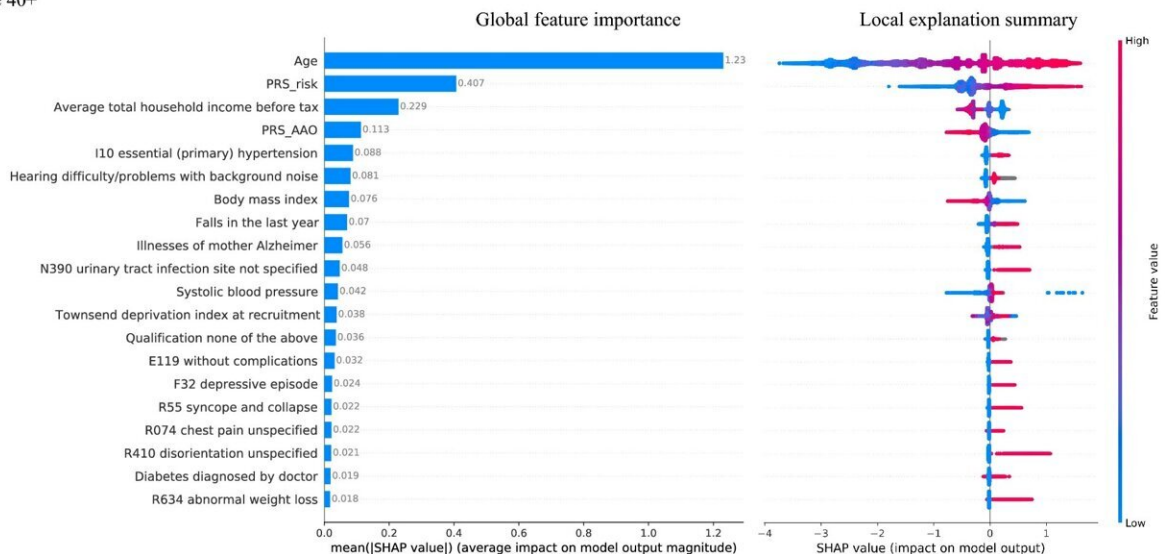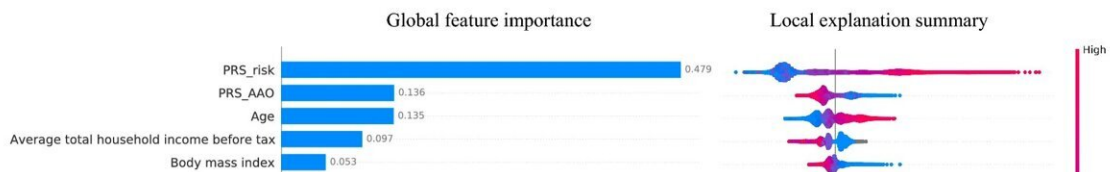
March 30 2023, by Emily Caldwell

Feature importance and impact of top 20 most important features for different age groups. SHAP global feature importance bar chart (left) and SHAP local explanation summary plot (right) for two age groups (A) Age 40 + , and (B) Age 65 + . Red and blue in the summary plot represent high and low feature values, respectively. The gray numerical values in the left bar plots show the mean (|SHAP value|), which shows the average impact on model output magnitude. The long right tails in the local explanation summary plot indicate rare but large-effect-size risk factors. Credit: *Scientific Reports* (2023). DOI:

Once adults reach age 65, the threshold age for the onset of Alzheimer's disease, the extent of their genetic risk may outweigh age as a predictor of whether they will develop the fatal brain disorder, a new study suggests.

The study, published recently in the journal *Scientific Reports*, is the first to construct machine learning models with genetic risk scores, non-genetic information and electronic health record data from nearly half a million individuals to rank risk factors in order of how strong their association is with eventual development of Alzheimer's disease.

Researchers used the models to rank predictive risk factors for two populations from the UK Biobank: White individuals aged 40 and older, and a subset of those adults who were 65 or older.

Results showed that age—which constitutes one-third of total risk by age 85, according to the Alzheimer's Association—was the biggest risk factor for Alzheimer's in the entire population, but for the older adults, genetic risk as determined by a polygenic risk score was more predictive.

"We all know Alzheimer's disease is a later-onset disease, so we know age is an important risk factor. But when we consider risk only for people age 65 or older, then genetic information captured by a polygenic risk score ranks higher than age," said lead study author Xiaoyi Raymond Gao, associate professor of ophthalmology and visual sciences and of biomedical informatics in The Ohio State University College of Medicine. "That means it's really important to consider genetic information when we work on Alzheimer's disease."

A low household income also emerged as an important risk factor, ranking either third or fourth after the effects of age and genetics.

"The finding related to income is very, very interesting," said Gao, also a member of Ohio State's Division of Human Genetics faculty, whose lab uses biomedical big data and artificial intelligence to study the genetics behind Alzheimer's and ocular diseases. "We all want to have a healthy life, and income can be such an important factor to decide what you can afford to eat, where you can afford to live, education level, access to care—and all of these possibly contribute to Alzheimer's disease."

Of the 457,936 UK Biobank participants in the sample, 2,177 individuals had developed Alzheimer's disease and 455,759 had not, and 88,309 were 65 or older.

A few non-genetic risk factors that differed between people with and without Alzheimer's disease (AD) stood out: Results showed that in people with AD, higher systolic and lower diastolic blood pressure were more common, diabetes was more prevalent, household income and education were lower, and recent falls, hearing difficulty and a mother's history of having AD were higher.

The top-20 list of risk factors for the full sample of adults also included diagnoses of high blood pressure, urinary tract infection, depressive episodes, fainting, unspecified chest pain, disorientation and abnormal weight loss. Other risk factors in the top 20 for people 65 and older included high cholesterol and gait abnormalities. These findings showed the power of adding condition codes from electronic health records to the models.

"Machine learning can explore relationships among all of those features, or variables, pick the important features and rank certain features at the top that contribute much more to Alzheimer's disease risk than the rest

of the features," Gao said. "Typically, it's not good to be highly obese, but we also see here that a lower body mass index is not good. High blood pressure is typically not good, but here we see lower diastolic blood pressure is not good. The models revealed some interesting patterns."

Building the models was a two-step process. The team first conducted genome-wide association studies using data from the Alzheimer's Disease Genetics Consortium to identify genetic variants linked to overall risk of developing Alzheimer's disease and to development of the disease after a specific age. The separate collections of variants were used to establish two polygenic risk scores, which aggregate genetic effects across the genome into a single measure of risk for each individual.

Those scores were applied to DNA data from the UK Biobank participants and combined with biobank information on conventional risk factors such as sex, education, body mass index and blood pressure, and more than 11,000 electronic health record condition codes that had been cited in individuals' records.

The team also used an algorithm in interpreting the model's output to ensure risk factor variables were weighted objectively in the analysis.

We are born with our genetic risk for disease already established, but information about how other health and socioeconomic factors affect our risk for Alzheimer's—as well as glaucoma, which Gao also studies—gives us power to take preventive measures, he said.

"If people know more about risk factors, they can possibly adjust their lifestyle. For both Alzheimer's and glaucoma, there is no cure, so prevention can help a lot," Gao said. "I also hope constructing models to make these predictions could help with drug development and effective

and low-cost screening programs."

More information: Xiaoyi Raymond Gao et al, Explainable machine learning aggregates polygenic risk scores and electronic health records for Alzheimer's disease prediction, *Scientific Reports* (2023). DOI: 10.1038/s41598-023-27551-1

Provided by The Ohio State University