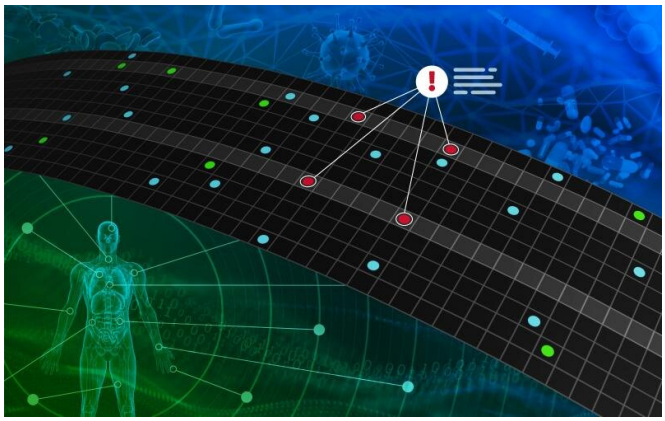
Researchers develop novel method to identify complex medical relationships

28 April 2022



ORNL, VA and Harvard researchers developed a sparse matrix full of anonymized information on what is thought to be the largest cohort of healthcare data used for this type of research in the U.S. The matrix can be probed with different methods, such as KESER, to gain new insights into human health. Credit: Nathan Armistead/ORNL, U.S. Dept. of Energy

A team of researchers from the Department of Veterans Affairs, Oak Ridge National Laboratory, Harvard's T.H. Chan School of Public Health, Harvard Medical School and Brigham and Women's Hospital has developed a novel, machine learning-based technique to explore and identify relationships among medical concepts using electronic health record data across multiple health care providers.

The method, called Knowledge Extraction via Sparse Embedding Regression, or KESER, was published recently in *npj Digital Medicine*. The process integrates electronic health record data from two large institutions—the VA and Boston-based Partners health care—and provides automated feature selection that leads to phenotype identification algorithms and knowledge discovery.

"KESER provides a high-level view of the relationships between clinical knowledge that we can't always see when caring for patients at the individual or group level," said Dr. Katherine Liao, a principal investigator of KESER at VA Boston and associate professor of medicine at Harvard Medical School. "We look forward to translating the study's methods and results from applications in clinical research to advancements in clinical care."

The project is part of the phenomics core work directed by Drs. Kelly Cho and Mike Gaziano from VA Boston and Harvard under the VA's Million Veteran Program, or MVP, a "national research program to learn how genes, lifestyle, and military exposures affect health and illness," according to the VA Office of Research and Development MVP website.

In 2016, ORNL began collaborating with the VA on MVP-CHAMPION, a big-data initiative under the MVP program, to create a large, precision-medicine platform to host the VA's vast medical record dataset, consisting of records for some 24 million veterans. In efforts to strengthen crosscutting innovation in support of numerous research projects under this joint VA-DOE program, ORNL worked closely with MVP Data Core from VA Boston and Harvard to identify specific research areas to pursue. Among those was an effort to answer the question: What elements do we need to find within electronic health records to correctly identify a given phenotype?

Working with what they think is the largest cohort of health care data used for this type of research in the U.S., the team set out to automate the identification of phenotypic relationships while providing visibility into the underlying machine learning assumptions and decision processes.

To do that, they designed and built the four-step KESER methodology: converting data into a structured format, constructing a low dimensional

vector representation of each medical code, selecting features to attribute importance and mapping attributed relationships as a network.

Data processing and representation learning

ORNL played a key role in the tedious yet essential work of processing and structuring a variety of medical data—patient procedures, diagnoses and measurements, as well as physician notes, prescription information and more—from millions of patients across the VA and Partners health care.

"There is a lot of unstructured data processing that goes on before you end up with a structured piece of information that can be put into statistical methods," said Edmon Begoli, ORNL AI Systems section head and principal investigator on the MVP-CHAMPION project. "The team spent years laboring through the data to get it into a state where we could start using it for research."

With the processed data, the team built a co-occurrence matrix, consisting of more than 100,000 types of events, or health care codes—essentially a massive, but sparse, data table with a row and column for every possible health care code. Each co-occurrence in time between two events helps create a clearer, more detailed picture of a given phenotype.

Leveraging ORNL's big data infrastructure and expertise in scientific computing—essential when working at this scale of data—the team worked to automate the data pre-processing and make the process publicly available.

"A researcher or institution can download the code, store their data in the correct format and our process will do all the steps needed to integrate their data with everyone else's," said Everett Rush, ORNL research scientist and lead data engineer on the project.

The research team has taken great care to protect patient privacy throughout the project. The team processed all the VA's data inside ORNL's secure protected health data infrastructure. After crunching it into an anonymous summary level, they shared it with Harvard and other collaborators. The resulting KESER matrix retains no links to individual patients.

"There's no way to trace from the end results back to an individual patient because these are aggregates," said Dallas Sacca, an ORNL senior solutions engineer. Sacca manages the protected health data enclave at ORNL and reviews each piece of data to ensure it meets the HIPAA guidelines for de-identification before allowing it to leave the enclave.

Knowledge extraction

The matrix is full of anonymized information on this immense cohort of patients that can be probed with different methods, such as KESER, to gain new insights into human health. Using a series of modern statistical methods, the team transformed summary data into vectors, tuned a model that encodes the relatedness of each vector and extracted the most important features and feature weights for each phenotype.

"These statistical methods, which include Gaussian graphical models for sparse modeling of covariance structures, are particularly capable in attribution of importance that exposes potential causal relationships, a concept with which classical AI technology, such as deep learning, tends to struggle," said George Ostrouchov, ORNL senior research scientist and lead statistician on the MVP-CHAMPION project.

After running the KESER method, the team selected eight phenotypes—including depression, rheumatoid arthritis and ulcerative colitis—to explore. Using the features selected by KESER, they trained models to identify the phenotypes of interest.

Future research

The possibilities enabled by KESER's novel ability to anonymize, integrate and analyze data from multiple health care institutions seem limitless.

Tianxi Cai, professor of Biomedical Informatics at Harvard Medical School and a principal investigator of KESER, said, "We are excited to have a highly scalable approach that can handle matrices an

order of magnitude bigger than what we are working with now."

The team is already incorporating more clinical descriptors into the knowledge graphs. Additionally, the team has started exploring the knowledge graphs to better understand emerging diseases.

"In a situation like COVID, for example, where everybody needs to share data and we need to start investigating all the different things that are related to this specific disease, you would potentially be able to do that with this system," said Chuan Hong, assistant professor at Duke University, who led research on the KESER project as an instructor at Harvard last year. "It's basically plug-and-play; you go to the data warehouse, follow the four-step process and directly integrate your results."

The potential for future collaboration and discovery may be the project's greatest success. "This innovation will facilitate multi-center collaborations," the team wrote in *Nature*, "and bring the field closer to the promise of creating distributed networks for learning across institutions while maintaining patient privacy."

More information: Chuan Hong et al, Clinical knowledge extraction via sparse embedding regression (KESER) with multi-center large scale electronic health record data, *npj Digital Medicine* (2021). DOI: 10.1038/s41746-021-00519-z

Provided by Oak Ridge National Laboratory