Researchers use novel technique to analyze 53 million points of clinical data

April 6 2023

Credit: Pixabay/CC0 Public Domain

Researchers from Children's Hospital of Philadelphia (CHOP) and

Drexel University were able to analyze 53 million patient notes from more than 1.5 million individual patients to identify similarities in their medical histories that can help pinpoint potential risks for developing future diseases and the trajectory of those conditions.

This method of identifying phenotypic similarities exceeds the capacity any other current computational models. The findings were recently published in the journal *Artificial Intelligence in Medicine*.

While many technological advances have made it possible to analyze genetic data on a large scale, there are challenges with applying the same techniques to clinical data because it is often not standardized, meaning researchers may describe or record the same symptoms differently. The Human Phenotype Ontology (HPO) was established to serve as a dictionary of more than 15,000 clinical terms to help standardize how clinical information for patients can be analyzed, accelerating the integration of precision medicine into clinical practice.

Prior studies from the Epilepsy Neurogenetics Initiative (ENGIN) at CHOP have analyzed clinical data from tens of thousands of patient notes that have helped reveal novel genetic targets in patients with different genetic childhood epilepsies. To expand on this success, researchers from CHOP's Department of Biomedical Health Informatics (DBHi), Drexel's College of Computing & Informatics, and ENGIN utilized Arcus, a suite of tools and services developed at CHOP that links biological, clinical, research and environmental data for the purpose of conducting innovative, data-driven research. The Arcus platform is designed to help reveal how sets of data overlap at a massive scale.

"This work follows trends in the artificial intelligence and machine learning field where complex, disparate information is transformed into something we can represent in a standardized way, thereby allowing

machines to classify medical conditions and predict future disease risk," said senior study author Scott Haag, Ph.D., Assistant Research Professor in the Department of Computer Science at Drexel's College of Computing & Informatics and the Supervisor of the Arcus Data Science Team at CHOP.

In this study, researchers analyzed data from 1,504,582 patients with a variety of diagnoses and syndromes with 53,955,360 electronic notes in the Arcus data repository. As a result, they identified 9,477 distinct phenotypes. This technique demonstrated a high degree of agreement with the judgment of experts in the various clinical fields represented in this data.

By transforming complex and multidimensional phenotypes from the HPO format into arrays, the method developed by the researchers in this study will enable efficient representation of these phenotypes for downstream tasks that require deep phenotyping. HPO is used in both common and rare diseases. For example, this study provides arrays of clinical data that can be used to distinguish a condition like autism spectrum disorder from rare neurodevelopmental disorders.

"This study demonstrated that utilizing an array made up of the clinical terms we identified could exceed the capacity of other much more computationally complex methods of analyzing phenotypes," said first author Maryam Daniali, a Ph.D. candidate in Computer Science at Drexel University. "This allowed us to map similarities between phenotypes using millions of points of data, significantly surpassing previous methods that relied on thousands of data points."

"The algorithm we developed in this study has the potential to be utilized in finding similarities between clinical trajectories and identifying novel genetic causes of diseases," said Ingo Helbig, MD, a pediatric neurologist in CHOP's ENGIN program and Scientific Director of the

Arcus Omics program. "This will allow us to use machine learning in tandem with existing methods to analyze risks and patient prognoses in a more efficient manner at large scale."

More information: Maryam Daniali et al, Enriching representation learning using 53 million patient notes through human phenotype ontology embedding, *Artificial Intelligence in Medicine* (2023). DOI: 10.1016/j.artmed.2023.102523

Provided by Children's Hospital of Philadelphia