Artificial intelligence may fall short when analyzing data across multiple health systems

November 6 2018

Artificial intelligence (AI) tools trained to detect pneumonia on chest X-rays suffered significant decreases in performance when tested on data from outside health systems, according to a study conducted at the Icahn School of Medicine at Mount and published in a special issue of *PLOS Medicine* on machine learning and health care. These findings suggest that artificial intelligence in the medical space must be carefully tested for performance across a wide range of populations; otherwise, the deep learning models may not perform as accurately as expected.

As interest in the use of computer system frameworks called convolutional neural networks (CNN) to analyze medical imaging and provide a computer-aided diagnosis grows, recent studies have suggested that AI image classification may not generalize to new data as well as commonly portrayed.

Researchers at the Icahn School of Medicine at Mount Sinai assessed how AI models identified pneumonia in 158,000 chest X-rays across three medical institutions: the National Institutes of Health; The Mount Sinai Hospital; and Indiana University Hospital. Researchers chose to study the diagnosis of pneumonia on chest X-rays for its common occurrence, clinical significance, and prevalence in the research community.

In three out of five comparisons, CNNs' performance in diagnosing diseases on X-rays from hospitals outside of its own network was significantly lower than on X-rays from the original health system.

However, CNNs were able to detect the hospital system where an X-ray was acquired with a high-degree of accuracy, and cheated at their predictive task based on the prevalence of pneumonia at the training institution. Researchers found that the difficulty of using deep learning models in medicine is that they use a massive number of parameters, making it challenging to identify specific variables driving predictions, such as the types of CT scanners used at a hospital and the resolution quality of imaging.

"Our findings should give pause to those considering rapid deployment of artificial intelligence platforms without rigorously assessing their performance in real-world clinical settings reflective of where they are being deployed," says senior author Eric Oermann, MD, Instructor in Neurosurgery at the Icahn School of Medicine at Mount Sinai. "Deep learning models trained to perform medical diagnosis can generalize well, but this cannot be taken for granted since patient populations and imaging techniques differ significantly across institutions."

"If CNN systems are to be used for medical diagnosis, they must be tailored to carefully consider clinical questions, tested for a variety of real-world scenarios, and carefully assessed to determine how they impact accurate diagnosis," says first author John Zech, a medical student at the Icahn School of Medicine at Mount Sinai.

This research builds on papers published earlier this year in the journals Radiology and Nature Medicine, which laid the framework for applying computer vision and deep learning techniques, including natural language processing algorithms, for identifying clinical concepts in radiology reports for CT scans.

Provided by The Mount Sinai Hospital