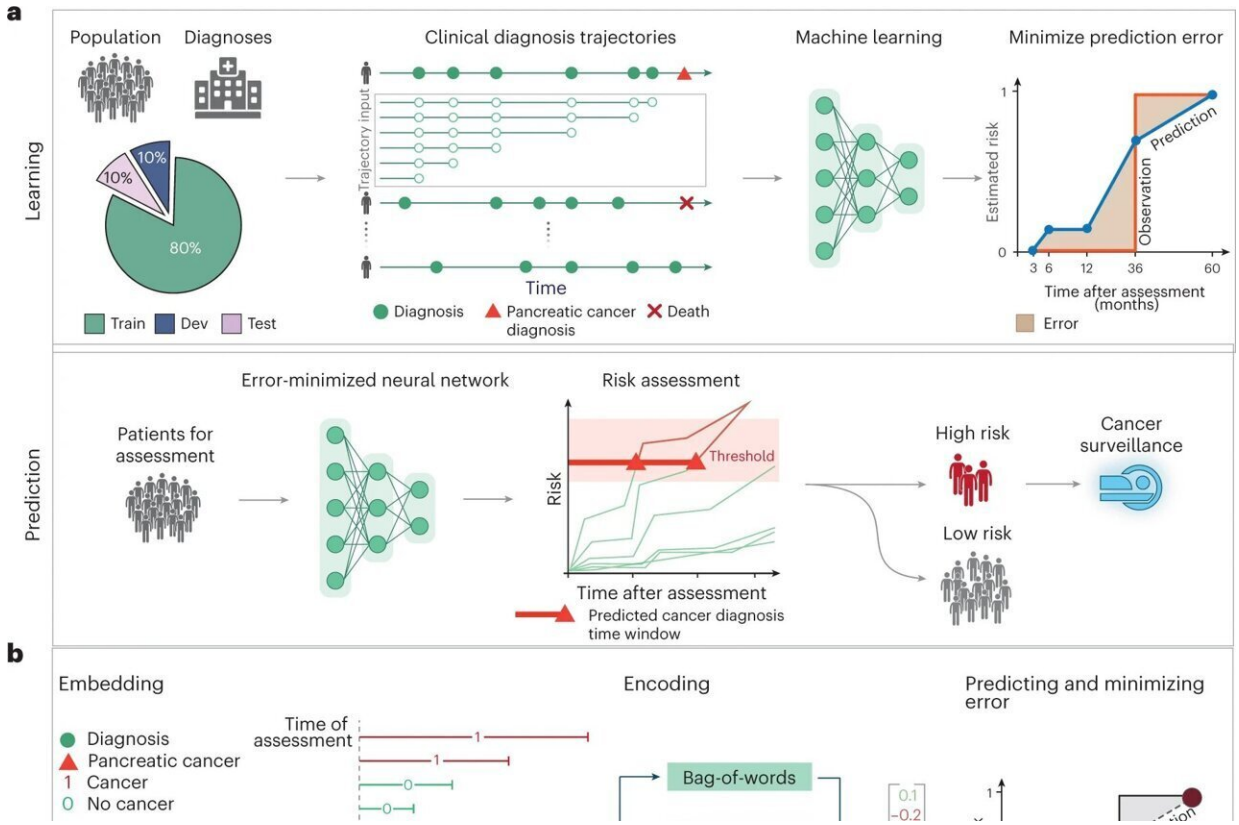
AI predicts future pancreatic cancer

May 8 2023



Training and prediction of pancreatic cancer risk from disease trajectories. **a**, Learning: The general ML workflow starts with partitioning the data into a training set (Train), a development set (Dev) and a test set (Test). The trajectories for training input are generated by sampling continuous subsequences of diagnoses for each patient's diagnosis history, each starting with the first record but with different endpoints. The training and development sets are used for training so as to minimize the prediction error—that is, the difference between a risk score function (prediction) and a step function (observation), summed over all instances. Prediction: A model's ability to accurately predict is evaluated using the withheld test set. The prediction model,

depending on the prediction threshold selected from among possible operational points, discriminates between patients at higher and lower risk of pancreatic cancer. The risk model can guide the development of surveillance initiatives. **b**, The model trained with real-world clinical data has three steps: embedding, encoding and prediction. The embedding machine transforms categorical disease codes and timestamps of these disease codes into a lower-dimensional real number continuous space. The encoding machine extracts information from a disease history and summarizes each sequence in a characteristic fingerprint in the latent space (vertical vector). The prediction machine then uses the fingerprint to generate predictions for cancer occurrence within different time intervals after the time of assessment (3, 6, 12, 36 and 60 months). The model parameters are trained by minimizing the difference between the predicted and the observed cancer occurrence. **c**, Terminology for timepoints and intervals. The last event of a disease trajectory coincides with the time of assessment. From the time of assessment, cancer risk is assessed within 3, 6, 12, 36 and 60 months. To test the influence of close-to-cancer diagnosis codes on the prediction of cancer occurrence, exclusion intervals are used to remove diagnoses in the last 3, 6 and 12 months before cancer diagnosis. Credit: *Nature Medicine* (2023). DOI: 10.1038/s41591-023-02332-5

An artificial intelligence tool has successfully identified people at the highest risk for pancreatic cancer up to three years before diagnosis using solely the patients' medical records, according to new research led by investigators at Harvard Medical School and the University of Copenhagen, in collaboration with VA Boston Healthcare System, Dana-Farber Cancer Institute, and the Harvard T.H. Chan School of Public Health.

The findings, published May 8 in *Nature Medicine,* suggest that AI-based population screening could be valuable in finding those at elevated risk for the disease and could expedite the diagnosis of a condition found all too often at advanced stages when treatment is less effective and outcomes are dismal, the researchers said. Pancreatic cancer is one of

the deadliest cancers in the world, and its toll projected to increase.

Currently, there are no population-based tools to screen broadly for pancreatic cancer. Those with a family history and certain genetic mutations that predispose them to pancreatic cancer are screened in a targeted fashion. But such targeted screenings can miss other cases that fall outside of those categories, the researchers said.

"One of the most important decisions clinicians face day to day is who is at high risk for a disease, and who would benefit from further testing, which can also mean more invasive and more expensive procedures that carry their own risks," said study co-senior investigator Chris Sander, faculty member in the Department of Systems Biology in the Blavatnik Institute at HMS. "An AI tool that can zero in on those at highest risk for pancreatic cancer who stand to benefit most from further tests could go a long way toward improving clinical decision-making."

Applied at scale, Sander added, such an approach could expedite detection of pancreatic cancer, lead to earlier treatment, and improve outcomes and prolong patients' life spans.

"Many types of cancer, especially those hard to identify and treat early, exert a disproportionate toll on patients, families and the healthcare system as a whole," said study co-senior investigator Søren Brunak, professor of disease systems biology and director of research at the Novo Nordisk Foundation Center for Protein Research at the University of Copenhagen. "AI-based screening is an opportunity to alter the trajectory of pancreatic cancer, an aggressive disease that is notoriously hard to diagnose early and treat promptly when the chances for success are highest."

In the new study, the AI algorithm was trained on two separate data sets totaling 9 million patient records from Denmark and the United States.

The researchers "asked" the AI model to look for telltale signs based on the data contained in the records. Based on combinations of disease codes and the timing of their occurrence, the model was able to predict which patients are likely to develop pancreatic cancer in the future. Notably, many of the symptoms and disease codes were not directly related to or stemming from the pancreas.

The researchers tested different versions of the AI models for their ability to detect people at elevated risk for disease development within different time scales—6 months, one year, two years, and three years.

Overall, each version of the AI algorithm was substantially more accurate at predicting who would develop pancreatic cancer than current population-wide estimates of disease incidence—defined as how often a condition develops in a population over a specific period of time. The researchers said they believe the model is at least as accurate in predicting disease occurrence as are current genetic sequencing tests that are usually available only for a small subset of patients in data sets.

The 'angry organ'

Screening for certain common cancers such as those of the breast, cervix, and prostate gland relies on relatively simple and highly effective techniques—a mammogram, a Pap smear, and a blood test, respectively. These screening methods have transformed outcomes for these diseases by ensuring early detection and intervention during the most treatable stages.

By comparison, pancreatic cancer is harder and more expensive to screen and test for. Physicians look mainly at family history and the presence of genetic mutations, which, while important indicators of future risk, often miss many patients. One particular advantage of the AI tool is that it could be used on any and all patients for whom health

records and medical history are available, not just in those with known family history or genetic predisposition for the disease.

This is especially important, the researchers add, because many patients at high risk may not even be aware of their genetic predisposition or family history.

In the absence of symptoms and without a clear indication that someone is at high risk for pancreatic cancer, clinicians may be understandably cautious to recommend more sophisticated and more expensive testing, such as CT scans, MRI or endoscopic ultrasound. When these tests are used and suspicious lesions discovered, the patient must undergo a procedure to obtain a biopsy. Positioned deep inside the abdomen, the organ is hard to access and easy to provoke and inflame. Its irritability has earned it the moniker "the angry organ."

An AI tool that identifies those at the highest risk for pancreatic cancer would ensure that clinicians test the right population, while sparing others unnecessary testing and additional procedures, the researchers said.

About 44 percent of people diagnosed in the early stages of pancreatic cancer survive five years after diagnosis, but only 12 percent of cases are diagnosed that early. The survival rate drops to 2 to 9 percent in those whose tumors have grown beyond their site of origin, researchers estimate.

"That low survival rate is despite marked advances in surgical techniques, chemotherapy, and immunotherapy," Sander said. "So, in addition to sophisticated treatments, there is a clear need for better screening, more targeted testing, and earlier diagnosis, and this where the AI-based approach comes in as the first critical step in this continuum."

Previous diagnoses portend future risk

For the current study, the researchers designed several versions of the AI model and trained them on the health records of 6.2 million patients from Denmark's national health system spanning 41 years. Of those patients, 23,985 developed pancreatic cancer over time. During the training, the algorithm discerned patterns indicative of future pancreatic cancer risk based on disease trajectories, that is, whether the patient had certain conditions that occurred in a certain sequence over time.

For example, diagnoses such as gallstones, anemia, type 2 diabetes, and other GI-related problems portended greater risk for pancreatic cancer within 3 years of evaluation. Less surprisingly, inflammation of the pancreas was strongly predictive of future pancreatic cancer within an even shorter time span of two years.

The researchers caution that none of these diagnoses by themselves should be deemed indicative or causative of future pancreatic cancer. However, the pattern and sequence in which they occur over time offer clues for an AI-based surveillance model and could prompt physicians to monitor those at elevated risk more closely or test accordingly.

Next, the researchers tested the best performing algorithm on an entirely new set of patient records it had not previously encountered—a U.S. Veterans Health Administration data set of nearly 3 million records spanning 21 years and containing 3,864 individuals diagnosed with pancreatic cancer. The tool's predictive accuracy was somewhat lower on the US data set.

This was most likely because the US dataset was collected over a shorter time and contained a somewhat different patient population profiles—the entire population of Denmark in the Danish data set versus current and former military personnel in the Veterans' Affairs data set.

When the algorithm was retrained from scratch on the US dataset, its predictive accuracy improved.

This, the researchers said, underscores two important points: First, ensuring that AI models are trained on high quality and rich data. Second, the need for access to large representative datasets of clinical records aggregated nationally and internationally. In the absence of such globally valid models, AI models should be trained on local health data to ensure their training reflects the idiosyncrasies of local populations.

More information: Søren Brunak, A deep learning algorithm to predict risk of pancreatic cancer from disease trajectories, *Nature Medicine* (2023). DOI: 10.1038/s41591-023-02332-5. www.nature.com/articles/s41591-023-02332-5

Provided by Harvard Medical School