Predicting cancer risks on the basis of national health data

Scientists from the German Cancer Research Center (DKFZ) and the European Bioinformatics Institute EMBL-EBI, Hinxton, UK, are using the Danish health registers to predict individual risks for 20 different types of cancer with a high degree of accuracy. The prediction model can also be transferred to other healthcare systems. It could help to identify people with a high risk of cancer, for whom individualized early detection programs could be tested in studies.

If cancers are detected early, the chances of cure are generally higher and patients require less intensive treatment. However, screening programs for the early detection of cancer only exist for a few tumor diseases - and not all people take advantage of these offers.

If there were a simple way to filter out people with a very high risk of developing cancer, screening programs could be developed specifically for those at risk. Researchers led by Moritz Gerstung from the DKFZ and the European Bioinformatics Institute EMBL-EBI, Hinxton, UK, have now published a feasibility study on this topic. The data scientists used the comprehensive data from the Danish health register, in which all clinical diagnoses of the population are stored, to quantify the individual disease risks for 20 different types of cancer.

The researchers first trained a prediction model on the data of 6.7 million adult Danes from 1995 to 2014. The training data set included more than 1,000 different previous diagnoses, as well as cancer in family members, age and - where available - basic body data and risk factors such as tobacco consumption or obesity.

The model was then validated on the data sets from 2015 to 2018, covering 4.7 million Danes, and delivered a high level of predictive accuracy. The model enables predictions about the individual risks of developing 20 different types of cancers. Over the course of a lifetime, the model achieved an accuracy of 81 percent. Taking age and gender effects into account, the accuracy was 59 percent. The model achieved the highest accuracy for cancers of the digestive system, as well as for thyroid, kidney and uterine cancer.

In order to check whether this predictive performance was also confirmed in health data from other countries, the researchers also validated their model using data from the UK Biobank and achieved comparable accuracy. The analyses do not allow an exact prediction of which person will develop cancer. However, they do determine the individual risk and enable a comparison with people of a similar age.

"With this study, we wanted to demonstrate that it is essentially possible to model individual cancer risks on the basis of national health data," explains Moritz Gerstung. Such risk stratification could help to offer further early detection tests to those people who would benefit most. In addition to established early detection methods, these could in future include blood-based cancer tests, for example, which are the subject of intensive research worldwide and are already being tested in clinical trials in some cases. The underlying hope is that in future, a certain number of tests could detect more cancers following risk stratification, with people at low risk could be avoided unnecessary tests and false-positive results and overdiagnosis could be prevented.

However, as Moritz Gerstung says, a reliable database is essential for this. "The Danish health data is unique because it covers a large period of time and can be linked with each other. Only a few European countries offer something equivalent, such as Finland and Sweden or special research cohorts in the UK.

Efforts are also underway in Germany to establish national digital health infrastructures. "It would make sense to consider which type of data is best suited for assessing cancer risk at the planning stage," says Gerstung. In his current work, the ICD-10 diagnosis codes, which are also used in other European healthcare systems, have proven to be useful.

Since basic information on body measurements and known risk factors such as tobacco consumption also provided important information, it seems advisable to facilitate the collection of such data at population level. "If this data had been available across the board in the Danish health registers, our prediction model would probably have been even more accurate," summarizes Gerstung.

Publication:

Jung, Alexander W., et al. "Multi-cancer risk stratification based on national health data: a retrospective modelling and validation study." The Lancet Digital Health 2024, DOI: 10.1016/S2589-7500(24)00062-1

Press release

23-May-2024
Source: German Cancer Research Center

Further information

- German Cancer Research Center