MedicalXpress

Two methods to de-identify large patient datasets greatly reduced risk of re-identification

July 28 2017

Two de-identification methods, k-anonymization and adding a "fuzzy factor," significantly reduced the risk of re-identification of patients in a dataset of 5 million patient records from a large cervical cancer screening program in Norway.

The study is published in the *Cancer Epidemiology, Biomarkers & Prevention*, a journal of the American Association for Cancer Research, by Giske Ursin, MD, PhD, director of Cancer Registry of Norway, Institute of Population-based Research.

"Researchers typically get access to de-identified data, that is, data without any personal identifying information, such as names, addresses, and Social Security numbers. However, this may not be sufficient to protect the privacy of individuals participating in a research study," said Ursin.

Patient datasets often have sensitive data, such as information about a person's health and disease diagnosis that an individual may not want to share publicly, and data custodians are responsible for safeguarding such information, Ursin added. "People who have the permission to access such datasets have to abide by the laws and ethical guidelines, but there is always this concern that the data might fall into the wrong hands and be misused," she added. "As a data custodian, that's my worst mightmare."

To test the strength of their de-identification technique, Ursin and colleagues used screening data containing 5,693,582 records from 911,510 women in the Norwegian Cervical Cancer Screening Program. The data included patients' dates of birth, and cervical screening dates, results, names of the labs that ran the tests, subsequent cancer diagnoses, if any, and date of death, if deceased.

The researchers used a tool called ARX to evaluate the risk of re-identification by approaching the dataset using a "prosecutor scenario," in which the tool assumes the attacker knows that some data about an individual are in the dataset. An attack is considered successful if a large portion of individuals in the dataset could be re-identified by someone who had access to some of the information about these individuals.

The team assessed the re-identification risk in three different ways: First they used the original data to create a realistic dataset that contained all the abovementioned patient information (D1). Next, they "k-anonymized" the data by changing all the dates in the records to the 15th of the month (D2). Third, they fuzzied the data by adding a random factor between -4 to +4 months (except zero) to each month in the dataset (D3).

By adding a fuzzy factor to each patient's records, the months of birth, screening, and other events are changed; however, the intervals between the procedures and the sequence of the procedures are retained, which ensures that the dataset is still usable for research purposes.

"We found that changing the dates using the standard procedure of k-anonymization drastically reduced the chances of re-identifiying most individuals in the dataset," Ursin noted.

In D1, the average risk of a prosecutor identifying a person was 97.1 percent. More than 94 percent of the patient records were unique, and

therefore those patients ran the risk of being re-identified. In D2, the average risk of a prosecutor identifying a person dropped to 9.7 percent; however, 6 percent of the records were still unique and ran the risk of being re-identified. Adding a fuzzy factor, in D3, did not lower the risk of re-identification further: The average risk of a prosecutor identifying a person was 9.8 percent, and 6 percent of the records ran the risk of being re-identified.

This meant that there were as many unique records in D3 as in D2. However, scrambling the months of all records in a dataset by adding a fuzzy factor makes it more difficult for a prosecutor to link a record from this dataset to the records in other datasets and re-identify an individual, Ursin explained.

"Every time a research group requests permission to access a dataset, data custodians should ask the question, 'What information do they really need and what are the details that are not required to answer their research question,' and make every effort to collapse and fuzzy the data to ensure protection of patients' privacy," Ursin said.

Patient data are in general very well safeguarded and re-identification is not yet a major threat, Ursin added. "However, given the recent trend in sharing data and combining datasets for big-data analyses—which is a good development—there is always a chance of information falling into the hands of someone with malicious intent. Data custodians are, therefore, rightly concerned about potential future challenges and continue to test preventive measures."

According to Ursin, the main limitation of the study is that the approaches to anonymize data in this study are specific to the dataset used; such approaches are unique for each dataset and should be designed based on the nature of the data.

Ursin declares no conflicts of interest.