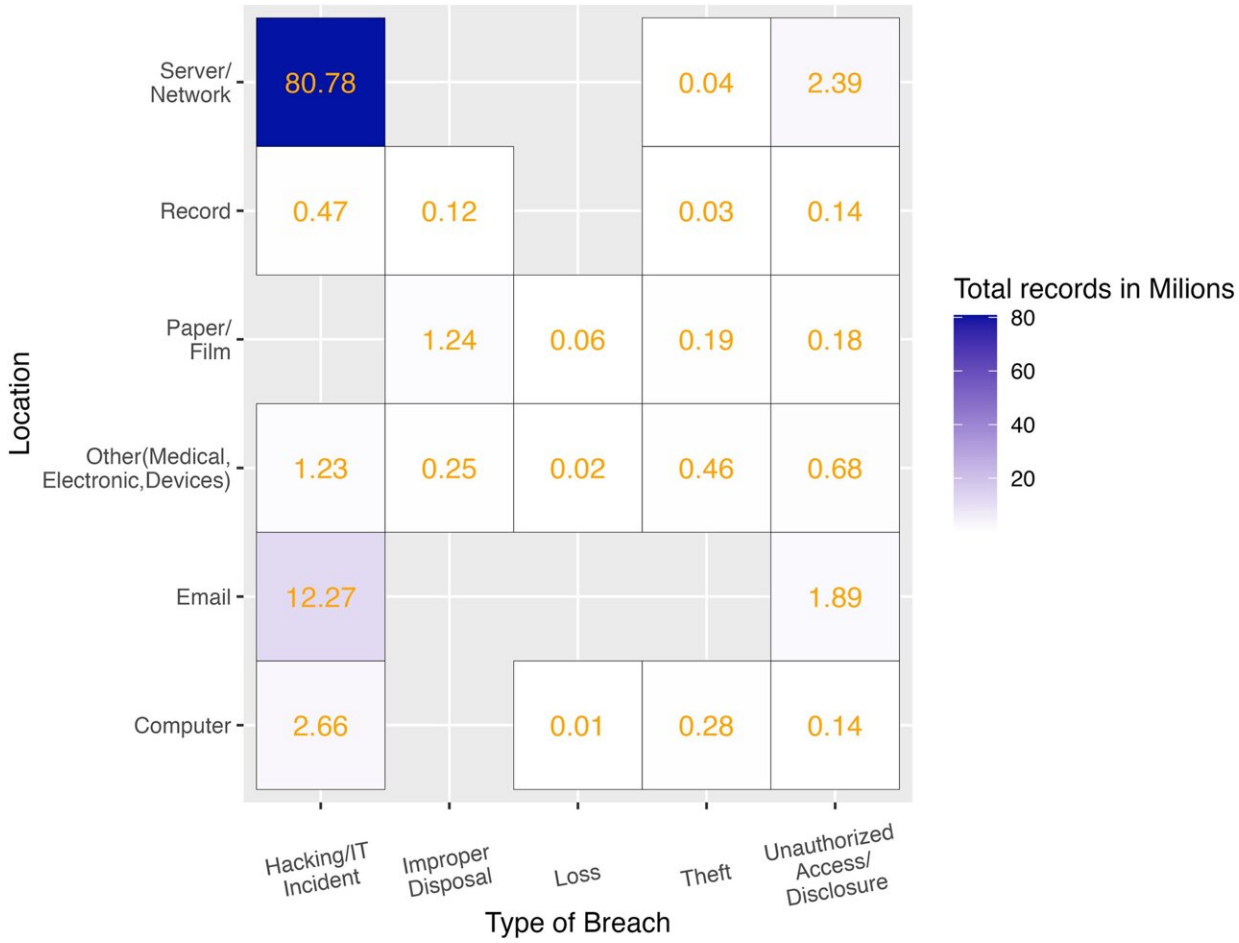
Study finds the risks of sharing health care data are low

October 7 2022, by Anne Trafton



Magnitude of data breaches considering total number of records affected by type of breach and location in the U.S. between November 2019 and November 2021. Elaborated from raw data. Credit: *PLOS Digital Health* (2022). DOI: 10.1371/journal.pdig.0000102

In recent years, scientists have made great strides in their ability to develop artificial intelligence algorithms that can analyze patient data and come up with new ways to diagnose disease or predict which treatments work best for different patients.

The success of those algorithms depends on access to patient health data, which has been stripped of personal information that could be used to identify individuals from the dataset. However, the possibility that individuals could be identified through other means has raised concerns among privacy advocates.

In a new study, a team of researchers led by MIT Principal Research Scientist Leo Anthony Celi has quantified the potential risk of this kind of patient re-identification and found that it is currently extremely low relative to the risk of data breach. In fact, between 2016 and 2021, the period examined in the study, there were no reports of patient re-identification through publicly available health data.

The findings suggest that the potential risk to patient privacy is greatly outweighed by the gains for patients, who benefit from better diagnosis and treatment, says Celi. He hopes that in the near future, these datasets will become more widely available and include a more diverse group of patients.

"We agree that there is some risk to patient privacy, but there is also a risk of not sharing data," he says. "There is harm when data is not shared, and that needs to be factored into the equation."

Celi, who is also an instructor at the Harvard T.H. Chan School of Public Health and an attending physician with the Division of Pulmonary, Critical Care and Sleep Medicine at the Beth Israel Deaconess Medical Center, is the senior author of the new study. Kenneth Seastedt, a thoracic surgery fellow at Beth Israel Deaconess Medical Center, is the

lead author of the paper, which appears today in *PLOS Digital Health.*

Risk-benefit analysis

Large health record databases created by hospitals and other institutions contain a wealth of information on diseases such as heart disease, cancer, macular degeneration, and COVID-19, which researchers use to try to discover new ways to diagnose and treat disease.

Celi and others at MIT's Laboratory for Computational Physiology have created several publicly available databases, including the Medical Information Mart for Intensive Care (MIMIC), which they recently used to develop algorithms that can help doctors make better medical decisions. Many other research groups have also used the data, and others have created similar databases in countries around the world.

Typically, when patient data is entered into this kind of database, certain types of identifying information are removed, including patients' names, addresses, and phone numbers. This is intended to prevent patients from being re-identified and having information about their medical conditions made public.

However, concerns about privacy have slowed the development of more publicly available databases with this kind of information, Celi says. In the new study, he and his colleagues set out to ask what the actual risk of patient re-identification is. First, they searched PubMed, a database of scientific papers, for any reports of patient re-identification from publicly available health data, but found none.

To expand the search, the researchers then examined media reports from September 2016 to September 2021, using Media Cloud, an open-source global news database and analysis tool. In a search of more than 10,000 U.S. media publications during that time, they did not find a single

instance of patient re-identification from publicly available health data.

In contrast, they found that during the same time period, health records of nearly 100 million people were stolen through data breaches of information that was supposed to be securely stored.

"Of course, it's good to be concerned about patient privacy and the risk of re-identification, but that risk, although it's not zero, is minuscule compared to the issue of cyber security," Celi says.

Better representation

More widespread sharing of de-identified health data is necessary, Celi says, to help expand the representation of minority groups in the United States, who have traditionally been underrepresented in medical studies. He is also working to encourage the development of more such databases in low- and middle-income countries.

"We cannot move forward with AI unless we address the biases that lurk in our datasets," he says. "When we have this debate over privacy, no one hears the voice of the people who are not represented. People are deciding for them that their data need to be protected and should not be shared. But they are the ones whose health is at stake; they're the ones who would most likely benefit from data-sharing."

Instead of asking for patient consent to share data, which he says may exacerbate the exclusion of many people who are now underrepresented in publicly available health data, Celi recommends enhancing the existing safeguards that are in place to protect such datasets. One new strategy that he and his colleagues have begun using is to share the data in a way that it can't be downloaded, and all queries run on it can be monitored by the administrators of the database. This allows them to flag any user inquiry that seems like it might not be for legitimate research

purposes, Celi says.

"What we are advocating for is performing data analysis in a very secure environment so that we weed out any nefarious players trying to use the data for some other reasons apart from improving population health," he says. "We're not saying that we should disregard patient privacy. What we're saying is that we have to also balance that with the value of data sharing."

More information: Kenneth P. Seastedt et al, Global healthcare fairness: We should be sharing more, not less, data, *PLOS Digital Health* (2022). DOI: 10.1371/journal.pdig.0000102

This story is republished courtesy of MIT News (web.mit.edu/newsoffice/), a popular site that covers news about MIT research, innovation and teaching.

Provided by Massachusetts Institute of Technology