

June 21, 2011

Software tool determines odds of breaching patient privacy

In the age of Facebook and Twitter, privacy may not seem as important as it once was, with people publicly sharing everything from the status of their romantic relationships to pictures of themselves in various stages of inebriation. But when it comes to medical information, privacy is still very much a priority, at least for health institutions that possess that data. So when health information is shared for secondary purposes, such as medical research, considerable effort is made to ensure people can't be identified from those data sets.

The problem is that assessing the reidentification risk of health data is time-consuming and somewhat mysterious. It seems to be based more often on opinions than empirical evidence. Research ethics boards (REBs) in Canada and institutional research boards (IRBs) in the United States determine if health data has been sufficiently "anonymized."

Two new software tools may make the work of these boards easier and, in the process, also make the lives of medical researchers easier. At least, that is the hope of the person behind the development of the tools.

"My hope is to encourage more availability of data for research," says Khaled El Emam, Canada Research Chair in electronic health information at the University of Ottawa in Ontario and chief executive officer of Privacy Analytics, an Ottawa-based company that creates software to protect individual privacy in sensitive data. "To get access to existing data now, there is so much uncertainty. The research ethics board review process can be long and painful, and can take many months of back and forth."

Reidentification risk assessment has long been performed in an ad hoc manner. There are no rules or international standards. Decisions are based on the experiences and perceptions of ethic boards' members, some of whom are far more conservative than others. As a result, there is much variability in how different REBs and IRBs estimate privacy risks.

"They do what they think is best, but they haven't been informed by objective evidence," says El Emam. "We have provided that evidence."

The free tools, the product of three to four years of research, are called REB Wizard (www.ehealthinformation.ca/rebwizard/ca/) and IRB Wizard (www.ehealthinformation.ca/irbwizard/). The software calculates reidentification risk percentages in different parts of Canada or the US, based on an analysis of census data. Users can change certain variables, such as the number of digits in a postal code, to adjust the risk percentage until it is in an acceptable range.

According to El Emam, deciding how much risk is acceptable is still up to individual institutions, and it typically ranges from a 33% risk of reidentification all the way down to a 1% risk, though somewhere around 20% is most common. The software

does have limitations, says El Emam, as it is based on statistical models and therefore produces estimates, which are never 100% accurate.

Still, “it’s a very good starting point to start having a discussion about risk,” says El Emam.

Anything that helps REBs do a better job of determining risk assessments is a welcome development, says Carole Gentile, chair of the Children’s Hospital of Eastern Ontario’s REB. “Khaled’s work has taken it to the next step,” she says. “It gives us an empirical tool around which to make decisions.”

Why is reidentification risk assessment so important?

According to Gentile, protecting patients’ privacy is a basic tenet of medical research. “The principle on which it rests is that people have proprietary rights on their health information. ... Institutions have to protect people against unintended disclosures.”

Bradley Malin, director of the Health Information Privacy Lab at Vanderbilt University in Nashville, Tennessee, also welcomes the introduction of technology into the field of data deidentification. “There has been a lot of research that has provided guidelines in terms of how you should protect data,” says Malin. “But there is not that much in terms of software available to the general public or groups that would be sharing data. In the past, it’s been very ad hoc in terms of how people anonymize or temper the identifiability of data sets.”

That’s not to say that challenges don’t remain. Institutions must still decide how much protection is enough. They must also define the type of risk they intend to minimize. There is, for example, there is “prosecutor risk,” which is risk of someone obtaining the health information about an individual they know is in a data set. Then there is “journalist risk,” which is the risk of someone picking a random person from a geographical area and identifying them in a data set. There is also “marketer risk,” or the average risk of identifying anybody in a dataset.

Organizations also put themselves at risk, both legally and in terms of reputation, if a medical record that it discloses for research is found to be identifiable. As well, there is the risk of harm to an identified patient. Identify theft is a possibility, as this type of fraud is easier to commit when you have access to someone’s medical history. And there is also the possibility of psychological harm caused by having medical conditions made known to others.

“Certain types of information are more stigmatizing than others,” says Malin. “If someone is reidentified and found to have a broken toenail, that would be much less of a concern than if that someone is positive for Chlamydia.”

Traditionally, there have been two ways to satisfy the deidentification requirements of an IRB. One is called “safe harbour.” This is when a long list of items — including names, dates and social security numbers — is removed from the data. The problem with this approach is that the data set loses important details and becomes “kind of neutered,” Malin says.

The other way to meet requirements is to find an expert to convince an IRB that reidentification will not be a problem, but that route is seldom chosen because there is no clear indication of who is an expert in this area.

Automating risk assessment via software provides another way, hopefully a better one, says Malin, who notes that El Emam has been very transparent in how he created

REB Wizard and IRB Wizard, publishing information that shows “what’s going on underneath the hood.”

“Now IRBs don’t have to individually reason about these problems or find experts,” says Malin. “Automating this whole process and making it transparent is definitely a service to the research community.” — Roger Collier, *CMAJ*

DOI:10.1503/cmaj.109-3843