Research reveals de-identified patient data can be re-identified

December 18 2017, by Dr Vanessa Teague, Dr Chris Culnane And Dr Ben Rubinstein

The health data were published to contribute to “research, community information, policy development and policy evaluation.” Credit: iStock

In August 2016, Australia's federal Department of Health published medical billing records of about 2.9 million Australians online. These records came from the Medicare Benefits Scheme (MBS) and the

Pharmaceutical Benefits Scheme (PBS) containing 1 billion lines of historical health data from the records of around 10 per cent of the population.

These longitudinal records were de-identified, a process intended to prevent a person's identity from being connected with information, and were made public on the government's open data website as part of its policy on accessible public data.

Now, we find that patients can be re-identified, using known information about the person to find their record. This was disclosed to the Department in December 2016.

Data sharing and privacy

The health data was published to contribute to "research, community information, policy development and policy evaluation". And it's extensive – it included all publicly-reimbursed medical and pharmaceutical bills for selected patients, spanning the thirty years from 1984 to 2014.

This is the second phase of our analysis of the MBS/PBS sample dataset. In September 2016, we found that the encryption of supplier IDs was easily reversed. We immediately informed the department and the dataset was then taken offline.

Our motive is to inform government policy with a demonstration of the surprising ease with which de-identification can fail. Access to high-quality, and sometimes sensitive, data is a modern necessity for many areas of research, but we now face the challenge of how to deliver that access, while protecting the privacy of the people in those datasets.

This example highlights the risky balance between data sharing and

privacy.

But there are a range of alternative solutions that should be considered, including the use of differential privacy for published data, and secure, controlled access to sensitive data for researchers.

Re-identifying patients

We found that patients can be re-identified, without decryption, through a process of linking the unencrypted parts of the record with known information about the individual.

Our findings replicate those of similar studies of other de-identified datasets:

- A few mundane facts taken together often suffice to isolate an individual.
- Some patients can be identified by name from publicly available information.
- Decreasing the precision of the data, or perturbing it statistically, makes re-identification gradually harder at a substantial cost to utility.

The first step is examining a patient's uniqueness according to medical procedures such as childbirth. Some individuals are unique given public information, and many patients are unique given a few basic facts, such as year of birth or the date a baby was delivered.

We found unique records matching online public information about seven prominent Australians, including three (former or current) MPs and an AFL footballer.

A unique match may not be an accurate re-identification, because only

10 per cent of Australians are included in this sample data. Some apparent re-identifications might be wrong, because there is a coincidental resemblance to someone who actually isn't in that 10 per cent. However, we can cross-reference, using a second dataset of population-wide billing frequencies, which can sometimes reveal that this patient is unique in the whole population.

The second step is examining uniqueness according to the characteristics of commercial datasets we know of but cannot access directly. There are high uniqueness rates that would allow linking with a commercial pharmaceutical dataset, and with the billing data available to a bank. This means that ordinary people, not just the prominent ones, may be easily re-identifiable by their bank or insurance company.

This contributes to the debate over the relationship between re-identification, uniqueness and confidence. While uniqueness does not imply re-identification—particular data that is known to be held by certain parties, does imply the opportunity for re-identification.

What to do with data?

These de-identification methods were bound to fail, because they were trying to achieve two inconsistent aims: the protection of individual privacy and publication of detailed individual records. De-identification is very unlikely to work for other rich datasets in the government's care, like census data, tax records, mental health records, penal information and Centrelink data.

While the ambition of making more data more easily available to facilitate research, innovation and sound public policy is a good one, there is an important technical and procedural problem to solve: there is no good solution for publishing sensitive complex individual records that protects privacy without substantially degrading the usefulness of the

data.

Some data can be safely published online, such as information about government, aggregations of large collections of material, or data that is differentially private. For sensitive, complex data about individuals, a much more controlled release in a secure research environment is a better solution. The Productivity Commission recommends a "trusted user" model, and techniques like dynamic consent also give patients greater control and visibility over their personal information.

The re-identification amendment

Earlier this year, the Australian government announced plans to amend the Privacy Act to criminalise re-identification of published government data. The proposed criminal offences would apply if "the information was published on the basis that it was de-identified personal information ".

The proposed amendment has not (yet) passed, though half of the relevant Senate committee recommended that it should, despite noting, "concerns that have been expressed about aspects of this Bill by submitters, including the introduction of criminal offences, the reversed burden of proof and the retrospective application of the Bill".

A dissenting report stated, "the bill adopts a punitive approach towards information security researchers and research conducted in the public interest. In contrast, government agencies that publish poorly de-identified information do not face criminal offences and are not held responsible... the Bill discourages research conducted in the public interest as well as open discussion of issues which may have been identified."

We agree.

Algorithms for protecting online security and privacy need careful scrutiny in order to be improved and strengthened. The technical issues are challenging and complex. Legislating against re-identification will hide, not solve, mathematical problems, and have a chilling effect on both scientific research and wider public discourse.

Our hope is that this research contributes to a fair, open, scientific and constructive discussion about open data sharing and patient privacy.

Provided by University of Melbourne