De-identification

Summary and discussion

The NSW Government encourages government agencies to manage and share their information assets for reuse.

The need for data de-identification arises when agencies intend to publish, share or re-use data which contains personal information. It will need to be established if the information intended for sharing will need to be de-identified, a process involving the removal or alteration of information that could identify an individual. De-identification and confidentialisation of data can help agencies balance the potentially conflicting goals of information transparency and privacy.

Using information that does not identify individuals is the surest way to protect confidentiality. As a general rule, de-identify any information asset that does not need to include personal identifiers.

Usually, de-identification includes two steps:

1. remove personal identifiers, such as an individual’s name, address, date of birth or other identifying information
2. remove or alter other information that may allow an individual to be identified, for example, because of a rare characteristic of the individual, or a combination of unique or remarkable characteristics that would enable identification.

Confidentialisation

Confidentialisation involves firstly, de-identifying data and secondly, taking the additional step of assessing and managing the risk of indirect identification occurring in the de-identified dataset.

In practice, ensuring that an information asset has been ‘de-identified’ will require agencies to adopt a risk assessment approach similar to the one involved in confidentialisation.

In de-identifying an information asset, an agency may need to retain an original, unaltered copy of the record for continued service delivery or to comply with record-keeping obligations.

Although de-identification is effective in preventing re-identification of an individual, it may not remove the risk altogether. There may be a possibility that another dataset or other information could be matched with the de-identified information. Agencies can actively assess the risk of re-identification to mitigate this risk. This should occur both before an information asset is de-identified and after disclosure of a de-identified asset.
Case Study

The Heritage Health Provider Network, Heritage Health Prize

De-identification of health data was used in the Heritage Health Prize (HHP), a competition launched by the Heritage Provider Network (HPN).

HPN, a health maintenance organization based in California, launched its Heritage Health Prize in April 2011. The objective of the competition is to construct a model to predict the number of days a patient will be hospitalized in the following year, by using the current and previous years’ claims data. The winner is the team or individual with the most accurate model, and receives a US $3 million cash prize.

The competition requires that large datasets containing private health information of patients be made publicly available. In the United States there is no legislative requirement to obtain patient consent to disclose health information if the data are deemed de-identified under the Health Insurance Portability and Accountability Act (HIPAA). Therefore, all datasets to be used in the competition were de-identified and management of the risk of re-identification of datasets was addressed.

De-Identifying Techniques

- **Pseudonyms** for Member ID, Provider ID, Vendor etc were used.
- **Top-Coding** was used. Top-coding is a process in which quantitative values that are considered uncommonly high are limited to an upper bound. This reduces the risk of individuals standing out.
- **Truncation of Claims.** The data involved a number of health claims made by patients. It is plausible that an individual could be identified if their demographic information was known, and it was known they had made an unusually large number of claims. Therefore the number of claims per patient at the 95th percentile was truncated in order to reduce the risk of re-identification of ‘extreme outliers’.
  - **Removal of High-Risk Patients and Claims.** Patients who were considered to be high risk were removed from the dataset to avoid the chance that their disease, condition, or procedure could be inferred from patterns in the data
  - **Suppression of Provider, Vendor, and PCP Identifiers.** It was possible that providers could have patterns of treatment that make them stand out. For example, a provider could be the only one with a particular specialty in a specific place of service who performed procedures on patients with a particular diagnosis. For these records the provider ID was suppressed.

* A detailed description of the de-identification techniques and the methods used to test them are outlined here.

Re-Identification Risk Assessment

A re-identification risk assessment of HHP datasets found the risk exposure was considered very low under the HIPAA Safe Harbor de-identification standard. The assessment also concluded that the HHP provides an example of, and methodology for, achieving open data principles for longitudinal health data, using good de-identification practices.

De-identifying data allowed for the disclosure data regarding individuals while simultaneously protecting those individuals from the risk of re-identification. The de-identification methods used by the Heritage Provider Network for the Heritage Health Prize are outlined here.
Useful resources

The National Statistical Service’s Confidentiality Information Series explains and advises on a range of issues about confidentialising data, including basic techniques to confidentialise data and manage risks.

The Australian National Data Services provides materials on techniques for de-identification and ethical considerations, including:

- De-identifying your data
- Ethics, consent and data-sharing

Dispelling the Myths Surrounding De-identification: Anonymization Remains a Strong Tool for Protecting Privacy, Ann Cavoukian and Khaled El Emam, Information and Privacy Commissioner Canada, 2011. This article illustrates the importance of de-identifying personal information before it is used or disclosed, with a particular focus on protecting health information. The authors conclude that if proper de-identification techniques and re-identification risk measurement procedures are used, re-identification remains a relatively difficult task.

Source: this Fact Sheet includes information from De-identification of data and information, Information Policy agency resource 1, Office of the Australian Information Commissioner, 2014. This resource provides general advice about de-identification to help Australian Government agencies to balance their transparency and privacy objectives.

Opinion 05/2014 on Anonymisation Techniques, Article 29 Data Protection Working Party, Directorate C of the European Commission (2014) is a useful analysis of the effectiveness and limits of existing anonymisation techniques (against an European Union legal background), looking at the residual risks of identification in data.


This work is licenced under the Creative Commons Attribution 4.0 International Licence. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/legalcode
Please give attribution to: © State of New South Wales [Office of Finance and Services], 2015