Assessment 3: Privacy Case Study Report (Re-identification of Medicare/Pharmaceutical Benefits Scheme data)

CP5806: Data and Information – Management, Security, Privacy and Ethics

Nikki Fitzherbert 13848336

Section 1: Case study description

In August 2016, the Department of Health (Health) published a de-identified longitudinal dataset containing individual claims records for a random sample of 10 per cent of the Australian population (approximately 2.5 million people) that had made a Medicare Benefits Scheme (MBS) or Pharmaceutical Benefits Scheme (PBS) claim between 1984 and 2014 (Office of the Australian Information Commissioner [OAIC], 2018).

The dataset was published on data.gov.au, which holds anonymised public sector data published by Australian federal, state and local government agencies plus datasets deemed to be of national interest (Digital Transformation Agency, n.d.). Health had released the dataset to further the Australian government's open data policy, which stated that public sector data should be made available to the community for research and evaluative purposes wherever possible (Culnane, Rubinstein, & Teague, 2017; Department of Health, 2018).

About a month and 1,500 downloads later, researchers from the University of Melbourne notified Health that they were able to re-identify medical service providers by reverse-engineering the encryption process. Health subsequently removed the dataset from the website in order to maintain its security and integrity, despite there being little evidence thus far of any personal information having being compromised (Anderson, 2016). In addition, Health and the OAIC also launched separate investigations and a bill was introduced into parliament to make it a criminal office to re-identify public sector datasets.

In December 2017, the same researchers notified Health that they had been able to cross-reference records in the dataset with other online information to identify specific individuals (Culnane et al., 2017).

Whilst the proposal to criminalise re-identification of public sector data was never passed into law, the events prompted a revision of the Australian government's public data release policies. Furthermore, Health was found to have breached sections of the *Privacy Act 1988* (the Privacy Act) relating to inadequate removal of sensitive personal information prior to the dataset's release (OAIC, 2018).

Being able to re-identify individual records was seen as an issue for two main reasons. Firstly, Australians are increasingly concerned about the privacy and security of their personal information, and perceived data breaches can damage community trust in policy makers and data custodians (Grubb, 2017). Secondly, it highlighted the problems with the public release of de-identified individual data records as part of an open data agenda. That is, there may be no easy way to completely negate the risk of record re-identification without significantly degrading a dataset's scientific or business utility (Culnane et al., 2017; Varghese, 2016).

Whilst OAIC reports indicated that very few notifiable data breaches in Australia are due to data sharing, this was not the first time the privacy of individuals in a de-identified dataset has been at risk (for example Narayanan & Shmatikov, 2008, 2010). Furthermore, any potential data breach involving personal information is

considered a serious matter in Australia as most are attempted with malicious or criminal motives (OAIC, 2020).

Finally, the risk of data breaches through re-identification of personal information affects multiple parties in the Australian community. For example,

- Australian government agencies may become more risk-averse in releasing similar datasets into the public domain, which would reduce the general public benefit gained from analysis undertaken by academic and private-sector institutions (Productivity Commission [PC], 2017);
- Individuals may become more concerned about unauthorised access to their personal information as data breaches can lead to emotional distress, embarrassment, discrimination and a risk to personal safety (Prictor, 2018) (Prictor, 2018); and
- Researchers and industry may be prevented from assisting government to improve data security and privacy protection techniques, which would ultimately hinder innovation and evidence-based policy development (Culnane, Rubinstein, & Teague, 2016).

Section 2: Arguments supporting the causal party

The following section briefly summarises why Health released the MBS/PBS dataset into the public domain and the actions taken after becoming aware that medical service provider numbers could be re-identified.

As an Australian government agency, Health was committed to promoting and adhering to the (government's) public data policy, which states that member agencies must take actions designed to optimise and re-use public data, by default openly release non-sensitive data, and collaborate with organisations in the private and research sectors to extend the value of public data for the benefit of the Australian community (Department of the Prime Minister and Cabinet, n.d.).

However, Health also understood that it had obligations regarding Australian privacy legislation, which protects the personal information of individuals. De-identification of the dataset via one or a combination of techniques would enable it to be shared and published freely (PC, 2017). At the time of the release, Health had considered its obligations to the Privacy Act and judged that the anonymisation techniques applied to the data were sufficient to prevent re-identification of individuals or medical service providers (OAIC, 2018).

Health also acted quickly after being made aware of the dataset's vulnerability in 2016 with the provider numbers. Not only did the organisation remove the dataset from the Australian government's open data portal, but it commenced an internal investigation into what had occurred and how to remedy the situation in order to ensure that further re-identification of records did not occur and to retain the integrity of the dataset (Easton, 2016).

Despite no personal information having yet been comprised, Health and other involved parties took the situation very seriously. The Health Minister made a public apology to doctors in Perth and the Attorney-General introduced a bill that proposed to make it illegal to re-identify any public sector data such as the MBS/PBS dataset (Anderson, 2016).

Section 3: Arguments supporting the affected parties

There was ample critique of Health's actions during and after the incident. Five of those are briefly summarised here.

The OAIC found that the risk assessment and de-identification processes undertaken by Health prior to the dataset's release were insufficient given the internal expertise and resources available to them at the time. This contributed to their determination that Health had breached the Privacy Act three times (OAIC, 2018; Pash, 2018).

The Office of the Victorian Information Commissioner (OVIC) examined the incident as part of a broader report on the use of de-identification for public data. They concluded that prior to its release, insufficient consideration had been given to the amount of auxiliary information available that could be used to re-identify individual records in the dataset (Office of the Victorian Information Commissioner, 2018).

Several people argued that since it would be just about impossible to de-identify any high-dimensional dataset to the point that the organisation could be assured that the privacy of medical providers and patients is protected, Health should have re-considered whether to release it on data.gov.au at all (for example Grubb, 2017; Varghese, 2016).

Culnane, Rubinstein and Teague (2017) stated that not only had others been aware of how simple it would be to re-identify medical providers in the dataset based on social media conversations, but that there was a very simple solution for the issue. For example, Health could have used a standard encryption algorithm or simply a randomly chosen unique number for each service provider and patient (Culnane et al., 2017).

They also argued that criminalising the re-identification of public data was a mis- guided reaction to the situation. It would have done nothing to prevent malicious attempts to gain authorised access to personal and sensitive information, and only inhibited legitimate research into improving privacy protection techniques (Culnane et al., 2016).

Section 4: The author's standpoint

After consideration of the arguments presented by the parties involved (such as Culnane et al (2017), Health, and the OAIC (2018)), the author supports the affected parties as their arguments were considered stronger and more convincing.

The final section works through the five-step decision making process based on that stance.

1. Develop problem statement(s)
    a. The privacy of service providers and up to 2.5 million Australians was put at risk due to the ability to reverse-engineer the de-identification techniques applied to the MBS/PBS claims dataset by Health and identify individuals using auxiliary information.
2. Identify alternatives
    a. Do not release details of de-identification techniques used.
    b. Revise internal risk assessment procedures for releasing sensitive data.
    c. Improve privacy-protection algorithms applied to datasets.
    d. Do not release sensitive datasets at all or only at a higher level of data aggregation.
    e. Restrict access to sensitive datasets to approved organisations.
3. Choose alternative(s)
    a. In the short-term, the best solution would be option c.
4. Implement decision
    a. This is a simple and effective solution (Culnane et al., 2017; Fang, Wen, Zheng, & Zhou, 2017) that would not require immediate broad-scale organisational change, which takes time in a large agency such as Health. Furthermore, it would maintain the scientific and business value of the dataset to external parties that value having access to high-dimensional public data for research and evaluative purposes (PC, 2017).
5. Evaluate results
    a. If successful, there would be minimum future reports in the media about data breaches from the release of Health's datasets on data.gov.au.
    b. Other parties would not be able to find vulnerabilities that could lead to re-identification of personal information in datasets.
    c. There would be no further determinations from the OAIC that Health has breached the Privacy Act via data sharing on data.gov.au.

Total word count: 1,528 words

- Section 1: 602 words
- Section 2: 312 words
- Section 3: 310 words
- Section 4: 304 words

References

Anderson, S. (2016, September 29). Medicare dataset pulled after academics find breach of doctor details posible. *ABC News*. Retrieved from https://www.abc.net.au/news/2016-09-29/medicare-pbs-dataset-pulled-over-encryption-concerns/7888686

Culnane, C., Rubinstein, B., & Teague, V. (2016). Crime and privacy in open data. Retrieved from https://pursuit.unimelb.edu.au/articles/crime-and-privacy-in-open-data

Culnane, C., Rubinstein, B., & Teague, V. (2017). *Health data in an open world.* Retrieved from https://arxiv.org/ftp/arxiv/papers/1712/1712.05627.pdf

Department of Health. (2018). Data access and release policy. Retrieved from https://www1.health.gov.au/internet/main/publishing.nsf/Content/Data-Access-Release-Policy

Department of the Prime Minister and Cabinet. (n.d.). Public data policy. Retrieved from https://pmc.gov.au/public-data/public-data-policy

Digital Transformation Agency. (n.d.). data.gov.au. Retrieved from https://data.gov.au/page/about

Easton, S. (2016). Service provider IDs unmasked in open health data, investigation underway. Retrieved from https://www.themandarin.com.au/70905-health-service-provider-ids-unmasked-in-medicare-and-pbs-open-data/

Fang, W., Wen, X. Z., Zheng, Y., & Zhou, M. (2017). A survey of big data security and privacy preserving. *IETE Technical Review, 34*(5), 544-560. doi:10.1080/02564602.2016.1215269

Grubb, B. (2017, December 18). Health record details exposed as 'de-identification' of data fails. *The Sydney Morning Herald*. Retrieved from https://www.smh.com.au/technology/australians-health-records-unwittingly-exposed-20171218-p4yxt2.html

Narayanan, A., & Shmatikov, V. (2008). Robust de-anonymization of large sparse datasets. *2008 IEEE Symposium on Security and Privacy (sp 2008)*, 111-125. doi:10.1109/SP.2008.33

Narayanan, A., & Shmatikov, V. (2010). Myths and fallacies of "personally identifiable information". *Communications of the ACM, 53*, 24-26. doi:10.1145/1743546.1743558

Office of the Australian Information Commissioner. (2018). *Publication of MBS/PBS data*. Retrieved from https://www.oaic.gov.au/assets/privacy/privacy-decisions/investigation-reports/publication-of-mbs-pbs-data.pdf

Office of the Australian Information Commissioner. (2020). Notifiable data breaches report: January-June 2020. Retrieved from https://www.oaic.gov.au/privacy/notifiable-data-breaches/notifiable-data-breaches-statistics/notifiable-data-breaches-report-january-june-2020/#comparison-of-top-five-industry-sectors

Office of the Victorian Information Commissioner. (2018). *Protecting unit-record level personal information: The limitations of de-identification and the implications for the Privacy and Data Protection Act 2014*. Retrieved from https://ovic.vic.gov.au/wp-content/uploads/2018/07/Protecting-unit-record-level-personal-information.pdf

Pash, C. (2018, March 29). A data bungle put at risk the private health details of millions of Australians. Retrieved from

https://www.businessinsider.com.au/data-breach-private-health-details-medicare-2018-3
Prictor, M. (2018). Patients and the data breach notification maze. Retrieved from https://pursuit.unimelb.edu.au/articles/patients-and-the-data-breach-notification-maze
Productivity Commission. (2017). *Data availability and use*. Retrieved from https://www.pc.gov.au/inquiries/completed/data-access/report/data-access.pdf
Varghese, S. (2016). Personal data de-identification practically impossible: Expert. Retrieved from https://www.itwire.com/business-it-news/data/personal-data-de-identification-practically-impossible-expert.html