Australian Government
Office of the Australian Information Commissioner

Publication of MBS/PBS data

Commissioner initiated investigation report

oaic.gov.au

23 March 2018

OAIC

Contents

Executive summary

On 1 August 2016, the Department of Health published on data.gov.au a collection of Medicare Benefits Schedule and Pharmaceutical Benefits Schedule related data. The data consisted of claims information for a 10% sample of people who had made a claim for payment of Medicare Benefits since 1984, or for payment of Pharmaceutical Benefits since 2003.

The Department of Health knew that this data would be extremely valuable for medical research and policy development purposes. It appears that the Department's decision to release the data was made in good faith for the public interest, on an understanding that the privacy interests of all relevant individuals had been protected.

A range of steps were taken by the Department of Health to de-identify the dataset before its public release. However, one month after the dataset was published, Drs Chris Culnane, Benjamin Rubinstein and Vanessa Teague of the University of Melbourne identified a weakness in the technique used to encrypt Medicare service provider numbers in the dataset, allowing the encryption to be reversed. This provided a potential for Medicare service providers referred to in the data to be identified. After this discovery, the dataset was further examined by an interagency taskforce involving experts from the Australian Bureau of Statistics and Data61, and separately by Drs Culnane, Rubinstein and Teague. This analysis identified that the detailed nature of the information in the dataset created a risk that some individuals may be identified by linking the dataset with other information sources.

Once the Department of Health became aware of the possible decryption of Medicare service provider numbers, the Department's response was quick and comprehensive. It acted to rapidly remove the dataset from public access, and it enhanced its processes to ensure that such an error would not be repeated. However, there were flaws in the process followed by the Department in de-identifying the dataset, assessing the risk of re-identification and deciding to publish it. The Commissioner accepts that decryption of Medicare service provider numbers per se does not mean that a provider is identified, however the result of the decryption meant that there was potential to re-identify providers. The Commissioner considered that the Department breached Australian Privacy Principle (**APP**) 1, APP 6 (in relation to health providers), and APP 11 in the course of publishing the dataset.

To assure the Commissioner and the Australian community that the Department of Health would continue to address the issues identified in the investigation, the Department offered, and the Commissioner accepted, an enforceable undertaking under s 33E of the Privacy Act on 23 March 2018. The acceptance of this enforceable undertaking ended the Commissioner's investigation.

Personal information in the dataset

Information is 'personal information', and subject to the requirements of the *Privacy Act 1988* (Cth), where it is 'about' an 'identified' individual or an individual who is 'reasonably identifiable'. Whether an individual is reasonably identifiable depends on the nature of the information in issue, and the context in which the information is held or released. Information that has been de-identified will no longer be personal information and the APPs will not apply to activities involving that information.

While the Department of Health took steps to de-identify the personal information of Medicare service providers, these measures were ultimately not sufficient. The encryption method for provider numbers was flawed, allowing Medicare service providers to potentially be re-identified from the information.

Under the Privacy Act, an individual will be ‘identifiable’ where it is possible to identify the individual from available information, including, but not limited to, the information in issue. An individual will be ‘reasonably’ identifiable where the process or steps for that individual to be identifiable are reasonable to achieve. Each of these elements of the term ‘reasonably identifiable’ inform whether the risk of re-identification under the Privacy Act will be sufficiently low such that the information will remain de-identified and will not constitute ‘personal information’. At this point in time, the Commissioner does not consider the evidence of the processes or steps for re-identifying patients in the dataset is such that it can be concluded that such processes or steps are reasonable to achieve. Thus the Commissioner does not consider this was such that the dataset contained the personal information of patients.

Breaches of the Privacy Act

As the Department of Health was required to protect personal information in the dataset of Medicare service providers, the Commissioner’s view is that the Department breached APP 6 by disclosing such personal information for a purpose other than that for which it was collected, where no relevant exception applied.

The Commissioner is also of the view that the Department of Health breached APPs 1 and 11 in the course of preparing the dataset as a whole for publication. The steps taken by the Department of Health to confirm personal information was removed from the dataset prior to its publication were inadequate relative to the sensitivity of the information and the context of its release.

The Commissioner acknowledges that these breaches were unintentional, and that the Department of Health did endeavour to protect the information it held. The Commissioner considers that the steps the Department has taken in response to the incident to date, together with the further steps it has committed to taking under the enforceable undertaking, constitute an appropriate response to the incident.

Lessons for the Department and other personal information custodians

There are important lessons from this matter, both for the Department of Health and for other custodians of valuable repositories of personal information.

The first is that the de-identification of large and rich datasets for publication to the world at large is extremely difficult. Deciding whether information has been de-identified to an extent suitable for public release requires careful, expert, and likely independent evaluation. Any such evaluation must take into account the circumstances of the particular data release. At this time, it is uncertain whether de-identification of a unit level dataset of this size and detail is possible to an extent that would permit full public release, while still maintaining the utility of the data. As part of attempting to de-identify data proposed for release, APP entities must consider the context of release. For example, they should consider limiting release of unit level data about individuals to trusted recipients, rather than to the world at large.

Second, appropriate processes and expertise should sit behind any decision to release de-identified personal information. The decision making process the Department of Health followed before releasing this data did not involve a clear and documented approval process, rigorous risk management processes, or a significant degree of cross-government coordination. This incident offers an opportunity for the Australian Government to strengthen its approach to publishing data that is based on personal information. The Commissioner notes that since this incident, the Australian Government has published guidance on the *Process for Publishing Sensitive Unit Record Level Public Data as Open Data*, which provides guidance on releasing datasets related to sensitive information. Had these measures been in place before the publication of this dataset and been followed, it appears likely that this incident would not have occurred. The Commissioner also notes the *Privacy (Australian Government Agencies – Governance) APP Code 2017*, which will take effect in July 2018, will provide additional standards of privacy protection for Australian Government agencies.

Background

Publication of the dataset

On 1 August 2016, the Department of Health published a compilation of Medicare Benefits Schedule (**MBS**) and Pharmaceutical Benefits Schedule (**PBS**) information (**MBS/PBS dataset** or **dataset**) on the website data.gov.au.[1] The dataset was publicly accessible online, and was downloaded approximately 1,500 times in the one month period it was publicly available.

The Department of Health has explained the reasons for its decision to publish the data as follows:

- Data held by Government is an important strategic national resource that holds considerable value for growing the economy, improving service delivery and transforming policy outcomes for all Australians.
- Determinations regarding the release of the data by the Department must consider the risk of identification against the benefits to the health system.
- Better evidence gives us better capacity to develop and design innovative, clinically-effective and cost-effective health services, which will lead to better patient outcomes and improve the standard of healthcare in Australia.
- Using the 10% datasets can deliver significant benefits to the health system, including:
  - better access to evidence to inform the Government's decisions about which health policies and programs to invest in;
  - a better understanding of what works, how well, for what cost, and in what circumstances; and
  - a more efficient health system, by supporting the most cost-effective treatments, strategies and interventions on broad-based independent evidence.
- The 10% data release was widely welcomed by health researchers and consumer interests, and was seen as an important initiative in health systems research.[2]

The Commissioner acknowledges the Department of Health's motivation for releasing the data was to deliver benefits to the health system. The decision to release the data was clearly made in good faith.

[1] Data.gov.au is the Australian Government's central catalogue of public data, and includes datasets available for access and download by the public at large, as well as information about datasets held by agencies to which access can be requested.

[2] Department of Health's submission to the OAIC, November 2016, p 1.

The dataset

The dataset comprised two distinct datasets – an MBS dataset and a PBS dataset – that were linked by common identifiers. The dataset contained unit level claims information for a random 10% sample of the population in the Medicare system (in effect, almost all Australian residents).

The dataset included details for each MBS and PBS claim made by the individuals in the sample population. The MBS data included almost all Medicare services accessed by the sample population between 1984 and 2014, and the PBS data included almost all uses of the PBS system by the sample population between 2003 and 2014.[3] This amounted to approximately 3 billion lines of data for approximately 2.5 million Australians.

The MBS dataset

The MBS dataset included numerous pieces of information relating to individuals within the 10% sample of the population in the Medicare system (**patients**), the Medicare services provided to those patients, and the providers of those services. For each individual, the data included a unique patient identification number (**patient PIN**); sex; year of birth; and which of four broad geographic regions the patient enrolled with Medicare.

For every MBS service provided to the patient, the dataset contained the MBS item number for the service;[4] the date the service was provided; and whether the service was provided in a hospital or elsewhere. Several pieces of information in the dataset related to billing for the service: the Medicare benefit paid for the service; the fee charged by the Medicare service provider (**service provider** or **provider**); the fee set out in the MBS for the service; and whether the service was bulk billed or patient billed.

The dataset also included a provider's unique service provider identifier; the provider's practice location; and geographic region.

The PBS dataset

The PBS dataset contained information about each prescription that was dispensed to the individuals in the sample data.

It included the patient PIN (as per the MBS data); geographic region; as well as the total number of prescriptions that have been dispensed to the patient.

For each prescription, it included the PBS item code;[5] drug type; and the date on which the prescription was dispensed. It also included information on the quantity of the item dispensed; whether the prescription was a repeat prescription; and whether the prescribing doctor had

[3] Some rarely used Medicare and PBS items were removed from the dataset as part of the de-identification process.

[4] An MBS item number is a unique identifier applied to each service or category of service listed in the Medicare Benefits Schedule. For example, item number 23 is a standard 'level B' consultation with a general practitioner of up to 20 minutes.

[5] The PBS item code is the individual number assigned to each medicine in the Pharmaceutical Benefits Schedule.

indicated that a repeat prescription should be dispensed at the same time as the original prescription.

Benefit information included the PBS benefit amount paid by the Government for the item; the contribution paid by the patient; whether the prescription was above the Safetynet and whether it was concessional; and whether the prescription was a 'standard co-payment claim' or whether it was an 'under co-payment claim'. The prescribing doctor's specialty (for example, GP, specialist, dentist) is also listed.

De-identification measures applied

The Department of Health applied a range of techniques to the dataset with the intention of ensuring that no information within the dataset could be linked back to any individual. The Department of Health also took steps to obscure the identity of any service providers the information related to by encrypting the Medicare provider identification numbers associated with each claim record.

The Department of Health applied the following measures:

- Provider identification numbers were encrypted.
- Locations were collapsed into four broad geographic regions: Australian Capital Territory and New South Wales; Victoria and Tasmania; Northern Territory and South Australia ; and Queensland and Western Australia.
- Patient PINs were encrypted.
- The patient's year of birth, rather than date of birth, was included.
- Patients aged over 100 were removed from the dataset.
- The date of service and date of supply for items were randomly perturbed up to plus or minus 14 days of the true date. Although different perturbations (between plus 14 days and minus 14 days) were applied to different patients, the same perturbation was applied to all services provided to each individual patient, thereby preserving the relative position of individual services accessed by each patient.
- Geographic aggregation was applied to the state of enrolment for Medicare (the same broad categories were applied as with provider states).
- Items with extremely low service volumes were removed from the dataset.

The methods used to encrypt and de-identify patient and medical provider information were based on an approach that had been used in a 10% PBS data release coordinated by the Department of Human Services (**DHS**) since 2005. In this data release, a quarterly 10% PBS sample dataset (the PBS dataset) had been provided to a group of approved research organisations by DHS on a fee-for-service basis. A review conducted on behalf of the Department of Health indicated that the method adopted for the encryption or scrambling of the PBS dataset was based on a method used by the former Health Insurance Commission, which dated back to the mid-1990s.

Discovery of the vulnerability and response

On 8 September 2016, Drs Culnane, Rubinstein and Teague of the University of Melbourne advised the Australian Bureau of Statistics (**ABS**) that they had been able to decrypt provider numbers contained in the MBS part of the MBS/PBS dataset on data.gov.au. The Department of Health was notified by the ABS of the dataset vulnerability on the same day.

The Department of Health took immediate action in response to notification of the vulnerability, by suspending the dataset from data.gov.au the same day. The Department of Health then contacted the University of Melbourne researchers and confirmed that the researchers did not intend to publish the decrypted data or the decryption method until the vulnerability was properly understood by the Australian Government. The Department of Health also notified the Office of the Australian Information Commissioner (**OAIC**) of the vulnerability on 9 September 2016.

The Department of Health then collaborated with the Department of the Prime Minister and Cabinet (**PMC**), specialists from the Commonwealth Scientific and Industrial Research Organisation's data innovation group Data61, the ABS, and the Australian Signals Directorate to undertake a risk assessment of this matter. The Department of Health also engaged a business management consultant to review the decision making process leading to the publication of the dataset and related matters.

As this issue raised concerns not just for the Department of Health, but likely for the whole of Government's approach to the release of public sector data, the Department of Health continued its engagement across Government agencies to initiate a broad review of, and enhancements to, the existing data release policy and process.

In the interim, the Department of Health also established revised arrangements for the MBS and PBS data to be released to approved research organisations only. These arrangements required: all data releases to be approved by the Deputy Secretary of the Health Benefits Group; no linkable MBS and PBS data to be made publically available; all requests from researchers and other non-Government entities to be first directed to DHS and subject to consideration by the External Request Evaluation Committee; the Department of Health to consider the Five Safes Framework (Safe People, Safe Projects, Safe Settings, Safe Data, Safe Output) in assessing new requests; and all data releases to be made under a formal agreement or contract with either the Department of Health or DHS and to be subject to arrangements specified in a public interest certificate.[6] In addition the Department implemented its updated Data Governance and Release Framework that had as its purpose, to increase the prominence of good corporate practice concerning data handling and use in the Department.

On 7 December 2016, PMC published a procedural note, *Process for Publishing Sensitive Unit Record Level Public Data as Open Data*.[7] This process requires a range of steps to be taken before sensitive unit record level data is published with the objective of ensuring that the dataset has been protected to the highest standard.

[6] Department of Health's submission to the OAIC, November 2016, p 6.

[7] Department of the Prime Minister and Cabinet, *Process for Publishing Sensitive Unit Record Level Public Data as Open Data* (December 2016), available online at https://blog.data.gov.au/news-media/blog/publishing-sensitive-unit-record-level-public-data (accessed March 2018).

The Commissioner acknowledges that the Department of Health took a rapid, proactive approach to both addressing the immediate issues surrounding this dataset's vulnerabilities, as well as taking initiative to address the broader whole of Government issue of large scale data publication.

Commissioner's investigation

Section 40(2) of the *Privacy Act 1988* (Cth) (the **Privacy Act**) permits the Australian Privacy Commissioner (**Commissioner**), on his own initiative, to investigate an act or practice if the act or practice may be an interference with the privacy of an individual or a breach of Australian Privacy Principle (**APP**) 1.

On 9 September 2016, the Commissioner wrote to the Secretary of the Department of Health to advise that the Commissioner was commencing an investigation under s 40(2). The Commissioner's investigation considered a range of issues, including whether the Department of Health:

1. in the course of publishing the MBS/PBS dataset, disclosed personal information of either healthcare providers or patients or both
2. had appropriate practices, procedures and systems in place to ensure that personal information proposed for de-identified release was not inappropriately used or disclosed, and was appropriately secured
3. took reasonable steps to secure the personal information used to form the dataset prior to its publication.

Relevant provisions of the Privacy Act

Agencies covered by the Privacy Act must comply with the APPs contained in Schedule 1 of the Act. The APPs apply to the handling of 'personal information', which the Privacy Act defines as:

> information or an opinion about an identified individual, or an individual who is reasonably identifiable:
>
> (a) whether the information or opinion is true or not; and
> (b) whether the information or opinion is recorded in a material form or not.[8]

APP 1 requires an APP entity to take reasonable steps to implement practices, procedures and systems that will ensure compliance with the APPs.

APP 6 prohibits an APP entity from using or disclosing personal information that it collected for a particular purpose for a secondary purpose, unless a relevant exception set out in APP 6 applies.

APP 11 requires an APP entity to take reasonable steps to protect the personal information it holds from misuse, interference and loss, and unauthorised access, modification or disclosure.

[8] Privacy Act, s 6(1).

Whether the dataset contained personal information

Whether or not the published dataset contained personal information is central to considering whether the release of the dataset was a 'disclosure' for the purpose of the Privacy Act.

The information handling obligations in the APPs attach to 'personal information'. 'Personal information' is defined in s 6(1) of the Privacy Act as information or an opinion about an identified individual, or an individual who is reasonably identifiable – whether or not the information or opinion is true, and whether or not the information or opinion is recorded in a material form. For information or an opinion to be personal information, it must be 'about' an individual who is 'identified' or 'reasonably identifiable'.[9]

Information will be 'about' an individual where the individual is a subject matter of the information.[10]

An individual is 'identified' when, as a question of fact, a link can be established between the information and the particular person.

Whether an individual is 'reasonably identifiable' from particular information is also a question of fact. An individual will be 'identifiable' where it is possible to identify the individual from available information, which includes, but is not limited to, the particular information in question. An individual is 'reasonably' identifiable where the process or steps for that individual to be identifiable are reasonable to achieve. Whether a person is 'reasonably' identifiable will depend on factors including the nature and amount of information and the context of the release of the information. The 'context of release' refers to matters such as who will hold and have access to the information, other information available to those recipients, and the practicality of identifying an individual using that information.[11] Where the processes or steps for individuals to become identifiable from available information are reasonable, those individuals will be reasonably identifiable under the Privacy Act.

The OAIC's *Australian Privacy Principles Guidelines* (**APP Guidelines**) note that where it is unclear whether an individual is reasonably identifiable, an entity should err on the side of caution and treat the information as personal information.[12]

Information will no longer be personal information where it has been de-identified, as it will no longer be information or an opinion about an identified individual or an individual who is

[9] Privacy Act, s 6(1); *Privacy Commissioner v Telstra Corporation Limited* [2017] FCAFC 4, para 60, (this case considered the definition of 'personal information' that applied prior to the commencement of the *Privacy Amendment (Enhancing Privacy Protection) Bill 2012*).

[10] *Privacy Commissioner v Telstra Corporation Limited* [2017] FCAFC 4, para 63.

[11] OAIC, *What is personal information?* (May 2017), available online at https://www.oaic.gov.au/agencies-and-organisations/guides/what-is-personal-information (accessed March 2018). See also, OAIC, *Australian Privacy Principles Guidelines* (March 2018), available online at https://www.oaic.gov.au/agencies-and-organisations/app-guidelines/ (accessed March 2018) ('**APP Guidelines**'), paras B.91-B.93.

[12] APP Guidelines, para B.94.

reasonably identifiable, in accordance with the Privacy Act definition of 'personal information'.[13] If an entity successfully de-identifies information that contains personal information, the Privacy Act will not apply to activities involving that information. However, any information which is re-identified from an apparently de-identified dataset will be personal information, and must be handled in accordance with the APPs. Similarly, a dataset that has had certain identifying information removed, but that still contains information about individuals who can be reasonably identified (for example, by combining the dataset with other available information), also contains personal information and must be handled in accordance with the APPs.

A number of steps may be required to effectively de-identify personal information: such as removal or alteration of all personal identifiers from the information; and removal or alteration of any other information that may, alone or in combination with other information, allow an individual to be reasonably identifiable. Context is a critical factor in determining whether information is sufficiently de-identified.

It is likely there will always be some risk of re-identification of personal information for complex unit level datasets, but in order to effectively de-identify personal information for the purposes of the Privacy Act, it must be the case that the individuals the information is based upon are no longer 'reasonably identifiable', considering the factors outlined above.

Information contained in the dataset

The Department of Health considered that at the time of publication of the dataset, it did not contain 'personal information', as defined in the Privacy Act. The Department of Health took into account the OAIC's APP Guidelines and considered that, given the de-identification and sampling perturbation that was employed, a person's identity would not be capable of becoming known, including by reason of information that might be expected to be available to a person to whom the disclosure was made.

Before deciding to release the dataset, the Department of Health took into account: the nature of the information or data itself; the actual or likely knowledge of the proposed recipient of the data; the information available to the recipient from both public and non-public sources; and legal and other restrictions applying to the recipient that would render identification attempts unlawful. The Department of Health considered it extremely unlikely that a person with decryption expertise would apply that expertise to decrypt the provider number.

Information about Medicare service providers

Information in the dataset about a service provided to a patient was linked to the Service Provider Number (**provider number**) of the health professional who provided the service.

The provider number is a unique number used to identify a health professional at their practice location. The number consists of a three to six digit number identifying an individual and two additional characters that identify the practice location. It is used for Medicare and Department of

[13] While 'de-identified' is a defined term in s 6(1) of the Privacy Act, that definition clarifies a specific obligation to de-identify under APP 11.2 and is relevant to certain other specific obligations. It does not set the threshold at which the APPs apply; that threshold is set by the definition of 'personal information'.

Veterans' Affairs claims processing and for determining the provider's eligibility for incentive payments. The number enables the service provider to claim Medicare benefits for services provided at a particular location, and to refer patients to another practice where they can claim a Medicare benefit.[14]

Although there is no searchable public database of provider numbers, the provider number of a particular provider would in many instances be ascertainable by, for example, patients of the service provider, a pharmacy where prescriptions from the service provider are filled, or the service provider's insurer.

The Commissioner notes that certain entities with access to provider numbers may have restrictions on their ability to use those numbers – for example, contractual obligations tied to the means by which they have access to provider numbers, or obligations under the Privacy Act.

The provider number was encrypted through conversion into an apparently meaningless number using a deterministic encryption methodology. However, the encryption method used was reversible, meaning that the encrypted numbers could be transformed back into the original provider numbers on which they were based.

At the time of publication of the dataset, technical specifications published on data.gov.au explained how the provider numbers were encrypted:

- Medicare Patient ID Numbers (PINs) were encrypted
  - Generate a random number using the original PIN as the seed.
  - Expand the encrypted PIN to 11 digits and mix with a part of the original PIN.
- Service Provider ID (SPR) encryption is similar to PIN encryption:
  - Generate a random number using the original SPR as the seed.
  - Expand the encrypted SPR to 8 digits and mix with part of the original SPR.

After the dataset was published, the encryption applied to the service provider numbers was reversed by researchers from the University of Melbourne. The researchers described the manner in which they reversed the encryption as follows:

> When we read the description of the method of encrypting supplier IDs, it suggested to us the use of pseudorandom number generation, which was insecure in that setting. We could then guess the algorithm, reverse it, and notify the government, who promptly removed the data from the website. At the time we said that there was a risk that someone else would discover the same weakness.
>
> …

[14] Department of Human Services, *Medicare provider number for health professionals*, available online at https://www.humanservices.gov.au/health-professionals/services/medicare/medicare-provider-number-health-professionals (accessed March 2018).

> The specific issue is easy to correct, using any standard encryption algorithm such as RSA or AES. Indeed, encryption was not necessary – a randomly chosen unique number for each person would have worked.[15]

Whether the dataset contained information 'about' Medicare service providers

The dataset contained various pieces of information that could, if linked to a particular provider, constitute the personal information of that service provider. For example, practice location identifier, specialty, and region are all capable of constituting personal information where a provider is an individual who is identified or who is reasonably identifiable. The MBS dataset also contained information about services provided and fees charged by individual service providers.

Each of these types of information is information 'about' a service provider: it is information about the provider and services delivered by the provider.

Whether Medicare service providers were 'reasonably identifiable'

The encryption applied to the provider numbers was able to be broken because of the manner in which the data was encrypted – by use of a pseudorandom number generator, using a seed derived from an identifier (in this case, provider numbers) rather than through use of a recognised encryption algorithm.

The encryption was broken within a month of the dataset's public release. In addition, the researchers who broke the encryption advised of an online discussion on social media that occurred before they broke the encryption, in which an unrelated group discussed the data release and the possibility of breaking the encryption by the method that was ultimately used by the researchers.[16]

The dataset was released online to the public, with no restrictions on access, and was accompanied by efforts by the Department of Health to publicise its disclosure to obtain the maximum public benefit from its release. The entire dataset was in fact downloaded more than 1,500 times during the month in which it was publicly available. At the time the dataset was released, the Department of Health indicated that the dataset was designed to enable linking with other datasets in future, such as hospital data and immunisation data.

Given the general release of the information and the scope of actual and potential recipients, no clear limit can be placed on other information that would be available to, or known by, these recipients.

[15] Culnane, Rubinstein, Teague, 'Health Data in an Open World: A report on re-identifying the MBS/PBS dataset and the implications for future releases of Australian government data' (arXiv preprint arXiv:1712.05627, 2017) ('**Health Data in an Open World**'), pp 4-5.

[16] Culnane, Rubinstein, Teague, 'Health Data in an Open World', p 5.

In light of the ease of decryption of the provider numbers made possible by the use of a pseudorandom number generator in lieu of a recognised encryption standard,[17] for the purpose of determining whether the data contains personal information of providers, given the release of the dataset to the world at large, it can be assumed that some of the people who the dataset was made available to would be able to obtain access to the decrypted provider number.

Because of this, the risk of re-identification of individual providers, via decrypted provider numbers, was not sufficiently low. In relation to the range of pieces of information relating to service providers contained in the dataset, once the provider number was decrypted, this information could be used, as part of a process, to identify particular providers. Providers were 'reasonably identifiable'.

As such, the Commissioner's view is that the dataset, in the form it was published, contained personal information of providers.

Information about patients

The dataset contained a near complete record of MBS and PBS services that were received by patients during the two sample periods. Much of the information, if it were to constitute 'personal information', would also constitute 'health information about an individual' and thus be 'sensitive information' for the purposes of the Privacy Act.[18]

Each row of information contained in the dataset is linked to a patient identifier: an apparently random number associated with an individual patient. This common identifier permits all information within the dataset relating to a particular patient to be linked. The patient identifier for each individual referred to in the dataset was created by the Department of Health by encrypting that individual's 'patient PIN'. A patient PIN is a separate numerical identifier, which is associated with the relevant individual's Medicare number. However, the connection between patient PINs and particular Medicare numbers is known only to DHS.

Unlike provider numbers, no decryption of patient PINs has occurred to date. Even if the PINs were decrypted, they would be meaningless to anyone outside DHS, the only agency able to link patient PINs to individuals' Medicare numbers.

In addition to using effectively meaningless identifiers, a number of de-identification processes were applied by the Department of Health to the dataset. For instance, the data was randomly perturbed, and some rare events were removed.

Nevertheless, even after these measures were applied, the dataset still contained a rich medical history for each individual in the sample population. Although the medical history was incomplete (certain low frequency events were removed, and medical services not linked to the PBS or MBS

[17] To effectively secure information, it is generally recommended that a recognised cryptographic algorithm be used rather than a custom written algorithm. A number of publications list algorithms that are widely recognised as secure. See, for example, Department of Defence, *2017 Australian Government Information Security Manual: Controls* (November 2017), available online at https://www.asd.gov.au/publications/Information_Security_Manual_2017_Controls.pdf (accessed March 2018), in particular, chapter on Cryptography, and sections titled 'Cryptographic Fundamentals', 'ASD Approved Cryptograph Algorithms' and 'ASD Approved Cryptographic Protocols' pp 238 – 249.

[18] Privacy Act, s 6(1).

would not be included), a history of medical treatments (as identified by MBS and PBS item numbers) was available for each patient.

Whether the dataset contained information 'about' patients

The range of information about each patient in the data included the MBS/PBS item codes the patient had claimed, the dates on which the patients received medical treatments or had prescriptions filled, and the fees paid for appointments, procedures, and prescriptions. This information is 'about' the patients the information relates to, because information such as the medical history and identifying details for an individual is information of which that individual is the subject matter.

However, the patients are not identifiable on the face of the data, because they are referred to only by the patient PIN, an effectively meaningless number. Whether the dataset contains personal information of patients depends on whether they are 'reasonably identifiable'.

Whether patients were 'reasonably identifiable'

As noted above, the information was publically released on data.gov.au on 1 August 2016. The information was available for approximately one month before it was removed on 8 September 2016. During this period, it was downloaded approximately 1,500 times. As noted above also, the dataset contained a near complete record of MBS and PBS services received by patients in the sample periods.

In determining whether patients are 'reasonably identifiable', confirmed identification of particular patients is not required, although confirmed identification may in certain circumstances be reasonably inferred.

Drs Culnane, Rubinstein and Teague reported having a high level of confidence that they had re-identified a small number of patients in the dataset by linking the dataset with information that was available online.[19] They identified a number of scenarios in which individuals might be identifiable. The report compiled by Data61 also outlined a number of re-identification scenarios. In contrast, the Department submitted that there had not been confirmed identification of these individuals, and that the possibility of a false match had not been ruled out.

The Department also noted that the individuals at risk of re-identification in the dataset had unique or rare attributes, such as individuals having undergone a series of unusual procedures, or having undergone certain procedures at a highly atypical age.

There is a marked difference between the processes and steps required to re-identify patients and providers. A unique identifier for providers, which can be linked to their identity, can be derived from the dataset following a process of decryption. In contrast, patients could only be identified by relying on unusual features of those patients, or a very full knowledge of their medical history, sufficient to distinguish them from everyone else in the dataset. While there is some risk of re-identification of patients by a sufficiently informed and skilled person, this risk is extremely low.

[19] Although it appears likely that re-identification of a small number of individuals occurred, this cannot be verified with complete certainty. It is possible that other individuals not included in the 10% sample share the same characteristics that were used to perform the re-identification.

Further, in the event that a possible match between a known person and a patient in the dataset occurs, it would be extremely difficult to confirm whether the match is correct. Because the dataset only covers a 10% sample of the population, it is possible that the known person is in fact in the 90% of the population not included in the dataset, and merely shares certain attributes with a person (the possible match) who is within the 10% sample. It may be possible to look to external sources of information to rule out this possibility for some individuals,[20] but the need for this further step lowers the risk of re-identification further.

The Commissioner's view is that the processes or steps for achieving re-identification are so extensive, and the risk of identification for any given patient so low, that the patients in the dataset are not reasonably identifiable for the purposes of the Privacy Act.

On the information before him, the Commissioner's view is that the published dataset does not contain the personal information of patients, as defined in the Privacy Act, as patients cannot be reasonably identified.

[20] See discussion in Culnane, Rubinstein and Teague, 'Health Data in an Open World'.

Disclosure of personal information (APP 6)

APP 6 regulates the use and disclosure of personal information and provides that APP entities may only use or disclose personal information for the primary purpose of collection, unless an exception applies.

The purpose for which the personal information was collected, or primary purpose, refers to the specific function or activity for which the entity collects the information.[21] Any other purpose for which an APP entity uses or discloses the personal information is a secondary purpose. An agency or organisation 'discloses' personal information when it makes the information accessible to others outside the organisation and releases the subsequent handling of the personal information from its effective control.

An APP entity may use or disclose personal information for a secondary purpose if the individual would reasonably expect the APP entity to do so, and the secondary purpose is related to the primary purpose, or another exception, set out in APP 6.2, applies.

Conclusion

As noted above, the Commissioner's view is that the dataset contains personal information of providers. There was a disclosure of this personal information through the public release of the data on data.gov.au.

The primary purpose of collection of the personal information appears to have been the administration of claims made against the MBS and PBS.

The Department of Health's publication of the dataset was for a secondary purpose of providing longitudinal health information to researchers and the public at large. It appears unlikely that the individuals the information is about would reasonably expect that their personal information would be disclosed by the Department for this secondary purpose. None of the other exceptions in APP 6 apply.

The Commissioner considers the disclosure of the personal information of service providers breached APP 6.

In contrast, the Commissioner is not satisfied that the dataset contained the personal information of patients, for the reasons outlined above. The Commissioner is therefore of the view that the personal information of patients was not disclosed through the publication of the dataset.

[21] APP Guidelines, para B.98.

Governance and protection of personal information (APPs 1 and 11)

APP 1.2 requires APP entities to take such steps as are reasonable in the circumstances to implement practices, procedures and systems that will ensure the entity complies with the APPs in the handling of personal information. The reasonable steps that an APP entity should take will depend upon circumstances including:

- the nature of the personal information held, with more rigorous steps required as the amount and sensitivity of personal information handled by an APP entity increases
- the possible adverse consequences for an individual if their personal information is not handled as required by the APPs, with more rigorous steps required as the risk of adverse consequences increases
- the nature of the APP entity. Relevant considerations include an entity's size and resources
- the practicability, including time and cost involved.[22]

APP 11.1 of the Privacy Act requires APP entities to take such steps as are reasonable in the circumstances to protect personal information they hold from misuse, interference and loss, as well as unauthorised access, modification or disclosure. The APP Guidelines explain that the reasonable steps to be taken by an entity to ensure the security of personal information depend on many factors, such as:

- the nature and size of the entity
- the volume and sensitivity of personal information
- possible adverse consequences
- practical implications of security measure implementation
- the invasiveness of the security measure itself.[23]

Relevance of APP 1 and APP 11 to the MBS/PBS de-identification process

Although the published dataset had identifiers removed and was modified to prevent re-identification, it was based on personal information (that is, information that was about an individual who was identifiable or reasonably identifiable by the entity that held the information).

De-identified information does not need to be handled in accordance with the APPs.[24] However, where an APP entity proposes to de-identify information to allow it to be released or otherwise

[22] APP Guidelines, para 1.6.

[23] APP Guidelines, para 11.7.

[24] Although the APPs do not need to be complied with when handling de-identified data, the OAIC recommends that agencies and organisations take a risk management approach when using de-identified information. Whether

used in a manner that personal information could not be used, the entity must ensure that it does so appropriately, and that the subsequent use or disclosure does not breach APP 6. APP 1.2 requires the APP entity to have in place appropriate practices, procedures and systems to ensure this occurs. Similarly, APP 11.1 requires the APP entity to take reasonable steps to protect the personal information that the dataset is based upon, using appropriate controls such as effective de-identification techniques and encryption.

The Commissioner's investigation considered whether the Department met the requirements of APP 1.2 and APP 11.1 in the course of preparing the MBS/PBS dataset for publication.

Governance measures and protections applied by the Department

The Department of Health had a range of measures in place at the time of the incident to protect the personal information it held.

General policies regarding the publication of information

The Department of Health had various policies, procedures and other guidance in place relevant to the data and personal information it held, and to the decision and process for publishing the dataset, particularly:

- Department of Health's *Data management policies*, in particular, *DM10: Data Access and Release*
- *Enterprise Data Management High Level Principles*, June 2012
- *Australian Government Public Data Policy Statement*.

A number of modifications to the Department of Health's data release and analysis strategies were being implemented at the time of the incident, including centralisation of its data stewardship functions.

Technical measures to de-identify personal information

The Department of Health utilised a number of technical measures to de-identify the dataset, as detailed in the background section of this report. The Department advised that this suite of measures was based on a methodology used for previous releases of PBS data to select researchers.

Processes and systems for approval of data release

The Data Governance Council (**DGC**) had Department-wide authority for data management at the Department of Health. On 14 October 2015, the proposed dataset was considered by the DGC, which agreed to 'the public release of linkable MBS, PBS and public hospital 10% sample dataset, noting the release proposed will be presented to the Executive for final approval.'

information is personal information or de-identified depends on context, and may change as the context of the information changes. Information that is believed to be de-identified, as in this case, may in fact not be.

On 15 April 2016, an information brief to the Deputy Secretary regarding the MBS dataset release to PharmHack, to test the linkages and identify any conflict, noted that 'a suite of confidentiality measures including encryption, perturbation, and exclusion of rare events has been applied to safeguard personal information and ensure that patients and providers cannot be identified.'

In July 2016, an information brief was sent to the Minister and Assistant Ministers notifying them of the proposed MBS/PBS dataset release, and describing the confidentiality measures applied.

The Department of Health has been unable to identify or provide any formal documented decision to release the dataset.

Risk assessment and expert advice

The Department of Health indicates that it did consider the risk of decryption and that this risk analysis was considered in line with OAIC guidance. Specifically, the Department of Health applied the 'motivated intruder' test, which tests 'whether a reasonably competent motivated person with no specialist skills would be able to identify the data or information.'

In explaining its risk assessment process to the OAIC, the Department of Health referred to a review conducted by the ABS. In 2013 the ABS was commissioned by the Department to conduct a review of the PBS data release. This review's scope covered sampling methodologies and 'confidentiality concerns.'

Sufficiency of personal information governance and protection measures

As noted above, where an APP entity is relying on de-identification processes to release what would otherwise be personal information, it must take reasonable steps to protect that information (APP 11.1) and have in place such practices, procedures and systems as are reasonable in the circumstances (APP 1.2).

Extent of reasonable steps required

As noted above, the reasonable steps required and necessary practices, procedures and systems, are determined by the circumstances, taking into account factors such as those listed in the APP Guidelines and above.

In this case, the data proposed for release was derived from highly sensitive personal information, being detailed medical records of individuals in the community, and information about the professional activities of medical providers. The Department proposed to release a very large quantity of data. It would have been apparent that a failure to de-identify this data successfully could result in serious adverse consequences for the people to whom the information related. The Department of Health is a large and complex organisation, with significant internal expertise and with resources available to permit it to access external advice.

These factors indicate that the Department should have taken very comprehensive steps to ensure the adequacy of the de-identification measures it chose to implement, and have in place rigorous practices, procedures and systems to ensure compliance with the APPs.

The Commissioner considered that, although the Department's decision to publish the dataset was taken in good faith, the Department failed to meet the high standard required by APPs 1 and 11 in the context of the dataset's publication. The Commissioner considers the Department fell short of the standard required by APP 1 and APP 11 in relation to the:

decision making process and policy framework for the dataset release

adequacy of encryption and other de-identification measures applied

selection of technical de-identification measures and assessment of re-identification risk.

Decision making process and policy framework

The Department of Health's data management policy: *Data Access and Release* provides guidelines related to ensuring the security or protection of the data that is released. The *Enterprise Data Management High Level Principles* also outline data governance principles. However, these policies do not discuss the threshold of what constitutes de-identified data, and offer minimal guidance on data release procedures.

Although these policies support the decision to publish the dataset, at the time of publication, there was no Departmental policy that provided a clear decision-making process in regard to the large-scale publication of de-identified data.

Perhaps because of this, the decision to publish the dataset was made without a formal documented decision by senior departmental officers.

The Commissioner considers that the Department of Health's failure to have sufficient practices, procedures and systems in place constituted breaches of APP 1 and APP 11.

Adequacy of encryption and other de-identification measures applied

As outlined above, the encryption applied by the Department of Health to provider numbers was reversible. The Commissioner considered that the Department of Health's reliance on reversible encryption algorithms constituted a breach of APP 11.

Selection of technical measures to de-identify personal information

The Department of Health applied technical de-identification measures that had been used, without incident, previously for releases of PBS data to approved researchers. The PBS de-identification approach was subject to a report by ABS, which found that there was a minimal disclosure risk.

The Department of Health sought to rely on the conclusions drawn from previous reviews of the 10% PBS dataset in assessing the risks of re-identification in the MBS/PBS dataset.

However, the earlier PBS data disclosures occurred in different circumstances to the MBS/PBS release, most relevantly:

- The PBS data was released only to researchers who had completed an application, detailed the purpose and methodology of their research proposals, and obtained approval by a committee of representatives from the Department of Health and DHS.
- The PBS releases were each for a single use of the dataset, and a new application must be made for each and every other use.
- Specific conditions were imposed around storage, security and destruction of the data.
- The level of exposure of the PBS dataset was significantly lower than that of the MBS/PBS dataset. In 2016, the PBS dataset was accessed by less than 20 approved clients, compared to the 1500 downloads of the MBS/PBS dataset that occurred in the one month that it was available online.

Given the different context, the Commissioner considers that it was inappropriate for the Department of Health to rely on the ABS review in assessing the risk of disclosure.

Other than the earlier ABS report, the Department of Health did not seek external expert advice on its encryption or de-identification method.

Given the circumstances of the publication, the Commissioner considers that the Department of Health should have conducted a further targeted risk assessment in relation to the re-identification of patient and provider data, supported by appropriate processes. The Commissioner considers that the Department of Health breached APP 1 and APP 11 by failing to conduct such a risk assessment.

Conclusion

The Commissioner considers that the Department of Health breached APP 1 and APP 11 in the course of publishing the dataset.