

Deanonymizing the Medicare Benefits Schedule (MBS) and Pharmaceutical Benefits Schedule (PBS) dataset

Chris Culnane s47F ,
Ben Rubinstein s47F ,
and Vanessa Teague s47F

University of Melbourne

September 8, 2016

Abstract

This paper describes a problem in the encryption of Service provider ID numbers in the Medicare Benefits Schedule (MBS) and Pharmaceutical Benefits Schedule (PBS) dataset recently published at <http://data.gov.au/dataset/mbs-sample-10pct-1984-gz>. We show here that the anonymization is easily reversed, at least for providers. Using only publicly available information, we have decrypted all the service provider ID numbers. We did not decrypt Medicare Patient ID numbers, but online documentation suggests that they may be encrypted in the same way.

The data set does not include names or addresses, but the implications of exposing individual IDs are still serious.

1 Introduction

We have had some early discussions with the Australian Bureau of Statistics about a potential project relating to privacy-preserving secure data linking. In addition to trying to design improved techniques, the questions included investigating the security of current methods.

We find that the method used in the MBS 10% dataset to encrypt provider IDs is not secure. Documentation suggests that the same method is used to encrypt Medicare patient ID numbers, but we have not decrypted them and are not certain about their security.

1.1 Background on the data set

The MBS 10% sample was constructed by choosing a random 10% of Medicare patient ID numbers (PINs). For each selected PIN, (almost) all medicare claiming activities were included in the data. Most of the data are not encrypted, including Medicare item codes for tests and doctors' visits, pharmaceutical benefits (PBS) item codes, and other information such as pricing and location. The codes describe *what* medical activity occurred, for example a particular kind of test, a particular prescription, or a visit to a GP or specialist. The ID numbers describing *who* received or provided the service are encrypted.

The data set does not include names or addresses of patients or doctors.

The website states that "A suite of confidentiality measures including encryption, perturbation and exclusion of rare events has been applied to safeguard personal health information and ensure that patients and providers cannot be re-identified." Strictly speaking, the method of obscuring patient and provider ID numbers is not encryption, though we will use that term to refer to it for consistency.

In summary, the data set contains medicare data treated in three different ways:

1. some data, such as Medicare and PBS item codes, are recorded in the clear,
2. some data, such as the date of the item, are randomly perturbed,
3. individual patient and provider IDs are encrypted.

1.2 Background on deanonymizing data

Generally there are two main ways of identifying individuals in this kind of data set:

linkage attacks use the unencrypted data to identify people by linking the record with other known information;

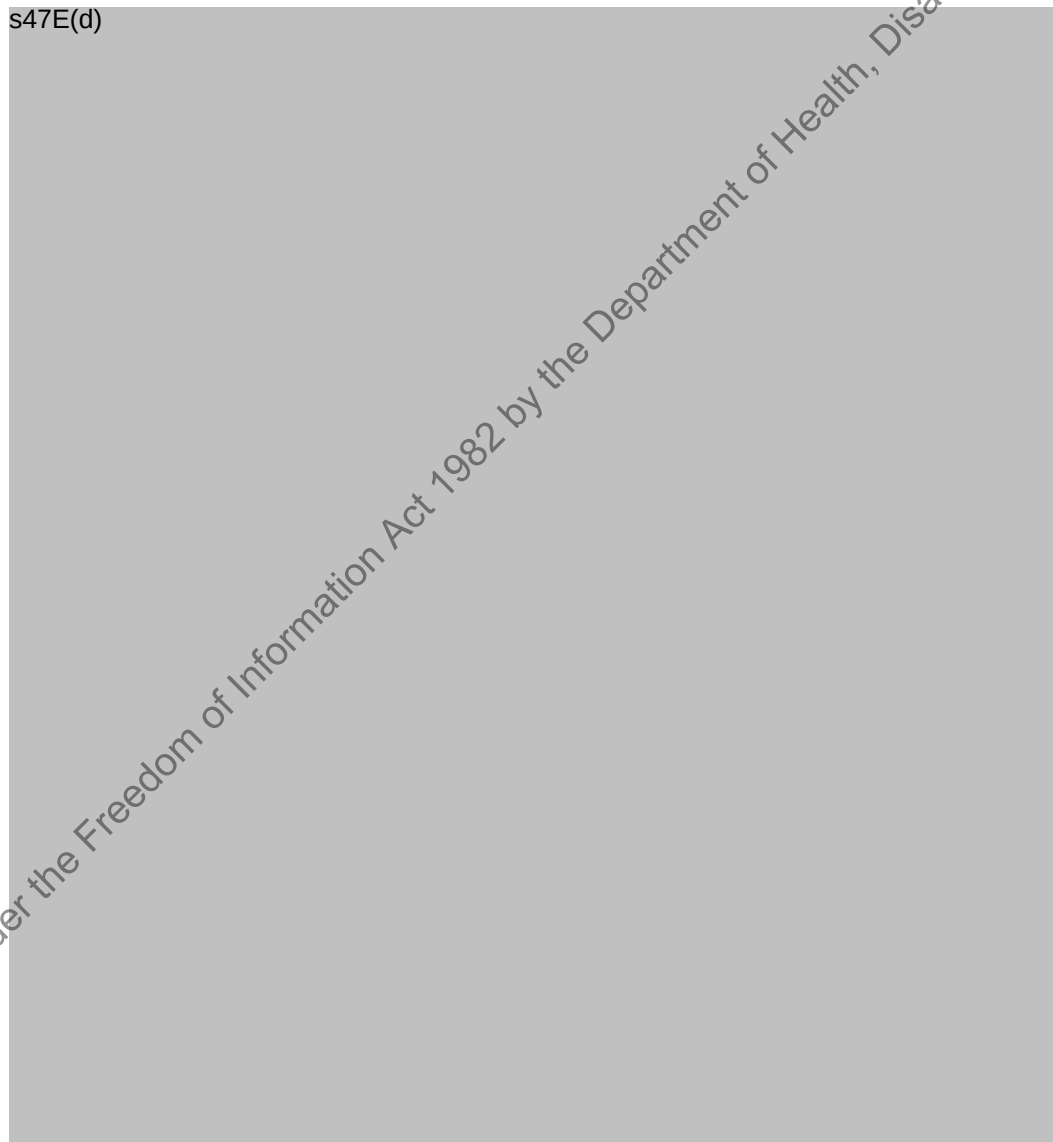
cryptographic attacks work by reversing the encryption algorithm to recover encrypted data.

This disclosure concentrates on successful cryptographic attacks. Linkage attacks have been demonstrated on other datasets [Swe02, NS08, NSR11] but we have not attempted them here.

The next section explains why the encryption of provider IDs is not secure.

2 How the encryption of individual IDs can be reversed

s47E(d)



Released under the Freedom of Information Act 1982 by the Department of Health, Disability and Ageing

s47E(d)


3 Conclusion

Publishing this sort of data brings great benefits for research but also great risks to privacy. The technical details matter: it's a specialist task to understand whether a particular algorithm securely encrypts data or not.

Explaining the anonymization algorithm publicly at data.gov.au was the right thing to do. Keeping the details of the algorithm secret would not have made it secure. "Security through obscurity doesn't work" is widely accepted among cryptographers. Omitting some of the algorithmic details didn't keep the data secure—if we can reverse-engineer the details in a few days, then there is a risk that others could do so too. However, keeping it secret would have made it more difficult for us to identify the problem, without making the encryption scheme any more inherently secure.

The best approach would be to publish the anonymization and encryption algorithms long before the data set, so that errors or vulnerabilities could be identified and corrected before the data were released.

The Australian government is making important decisions about what personal data to publish and how it should be anonymized, encrypted, or linked. Making good decisions requires accurate technical information about the security of the system and the secrecy of the data. We should have an open public discussion about what to publish and link, and how to keep it secure.

References

- [Aus16] Australian Government Department of Health. Mbs 10% sample technical information, 2016. <http://data.gov.au/dataset/a8e3c0bc-44ac-4e9a-8b3c-b779438ddb10/resource/a33103be-17d8-4857-b621-334a813f4b99/download/>

mbs-10-sample-metadata-part-5-of-5---technical-information.pdf.

- [NS08] Arvind Narayanan and Vitaly Shmatikov. Robust de-anonymization of large sparse datasets. In *2008 IEEE Symposium on Security and Privacy (sp 2008)*, pages 111–125. IEEE, 2008.
- [NSR11] Arvind Narayanan, Elaine Shi, and Benjamin IP Rubinstein. Link prediction by de-anonymization: How we won the kaggle social network challenge. In *Neural Networks (IJCNN), The 2011 International Joint Conference on*, pages 1825–1834. IEEE, 2011.
- [Swe02] Latanya Sweeney. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(05):557–570, 2002.

Quantitative analysis of re-identification risk from MBS PBS data

Released under the Freedom of Information Act 1982 by the Department of Health, Disability and Ageing

Confidential

Document administration

Approvals

Version	Date	Approver	Role
V3	11/09/2016	s47F	Director, Engineering & Design

Version history

Version	Date	Description	Author
V3	11/09/2016	Initial version for consideration by PMC and other stakeholders	Privacy team, Networks Group (Data61)

Released under the Freedom of Information Act 1982 by the Department of Health, Disability and Ageing

Executive Summary

We quantify the risks of re-identification attacks of the released MBS dataset. The attacks consist in identifying one or several known records in the release dataset and use them as identifiers or auxiliary information to de-anonymise other records associated with a selected target or the entire population in the dataset.

We first identify three categories of attributes based on the level of accessibility and availability for re-identification. These are categorized as commonly available, moderately available and rarely available. Based on the categorization we investigate (1) the impact of individual attributes on re-identification, (2) the impact of the combination of attributes on re-identification, and (3) the prevalence of re-identification risk amongst individuals in the sampled MBS data. We quantify the risks of re-identification risk by analyzing the uniqueness of records in the dataset which captures the likelihood of linking the released attributes to a background knowledge an "attacker" might acquire. Our analysis is based on the year 2014 of the MBS 10% dataset. However, we note that our findings are statistically significant and applicable to the entire released dataset.

Our results show that (1) **the re-identification risks in the MBS dataset vary from very low (less than 0.1%) for attributes classified as commonly available** (e.g. Gender, obfuscated year of birth, knowledge of one day of visit to the practitioner) **to significantly higher risks (more than 10%) for selected moderately available attributes** (e.g. knowledge of the Field of specialty of the practitioner visited) **and severe risks for all attributes from the category of rarely available attributes** (2) **The combination of attributes, however, does increase the risk of re-identification significantly even for the commonly available category of attributes.** Furthermore, the combination of two or more moderately available attributes raises re-identification risks to a notable level (3) **The impact of the high-identification risks is very limited to a small population in the dataset**, with 100% of individuals encountering less than 0.1% re-identification risks when a commonly available attribute is used. Notably the risk does not spread much when using rarely available attributes such as the Fee Charged, with **only 1.2 % of the individuals** with over 10% of re-identification risks.

Re-identification risks in the 10% MBS and PBS sample data

Re-identification means to identify an individual in an anonymized dataset by using public information and linking it to individual rows in the anonymized dataset. In general, this approach can link multiple rows in the anonymized dataset to the target individual. The risk of re-identification is higher if there is a smaller number of rows linked to the target individual. The risk of re-identification is the highest if there is a unique row in the anonymized dataset linked to the target individual. This means that if the anonymized dataset contains unique rows, the likelihood of re-identification is the highest.

s47E(d)

Confidential

s47E(d)

Released under the Freedom of Information Act 1982 by the Department of Health, Disability and Ageing

Confidential

s47E(d)

Released under the Freedom of Information Act 1982 by the Department of Health, Disability and Ageing

Confidential

s47E(d)

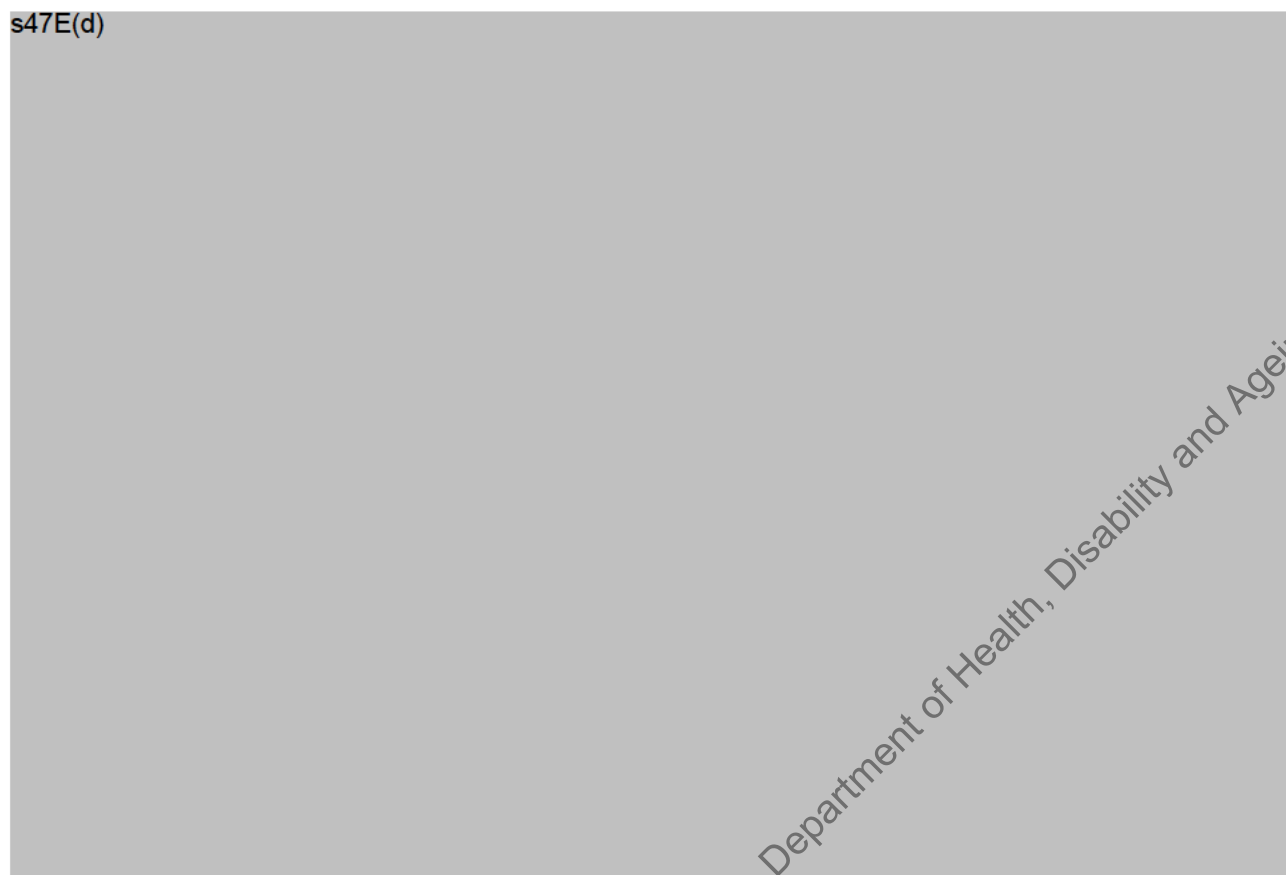


Figure 1: Re-identification risk against individual attributes.

2. Re-identification risk for combinations of attributes

s47E(d)



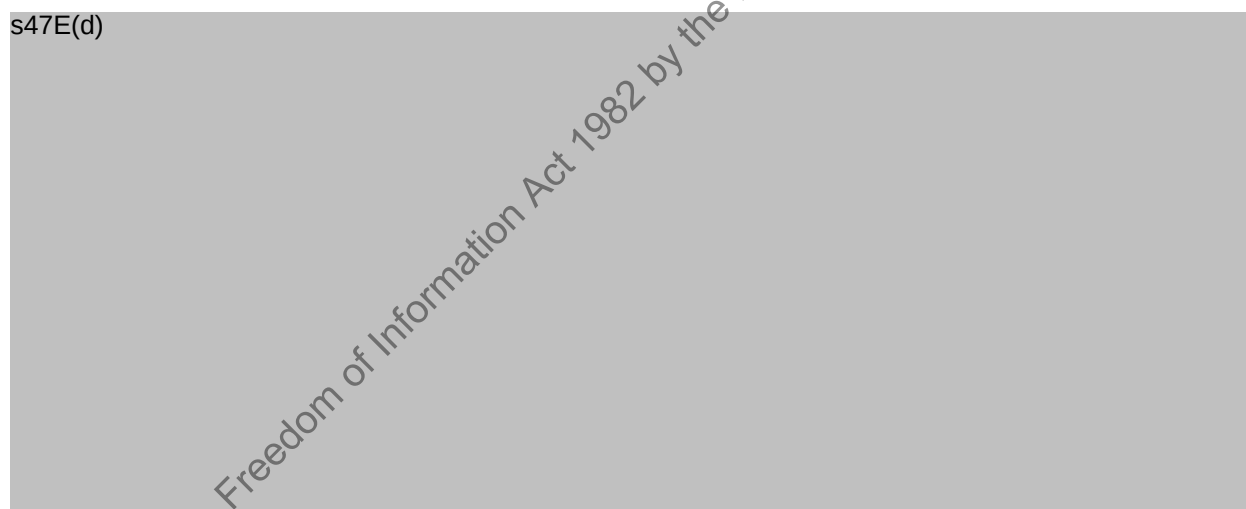
s47E(d)



Figure 2: Re-identification risk against number of combined attributes.

3. Prevalence of the risks among individuals

s47E(d)



s47E(d)

Released under the Freedom of Information Act 1982 by the Department of Health, Disability and Ageing

Confidential

s47E(d)

Released under the Freedom of Information Act 1982 by the Department of Health, Disability and Ageing

Confidential

s47E(d)

Appendix B

The risk of re-identification given an attribute A with value a is calculated as follows. First we calculate the frequency of each value a_i of attribute A as follows:

$$f_i = \text{count}(A = a_i)$$

then the re-identification risk is calculated as:

$$RI = \frac{\sum_{i=1}^{|A|} \frac{1}{f_i}}{N}$$

Confidential

where N is the total number of unique individuals in the dataset and $|A|$ represents the support of A (number of attribute values existing in the dataset). A value of RI close to 1.0 means high risk of re-identification. A value closer to 0 means low risk of re-identification.

For a combination of attributes, the procedure is the same, except that now we count each possible tuple of values.

Released under the Freedom of Information Act 1982 by the Department of Health, Disability and Ageing

Deanonymizing Provider IDs in the Medicare Benefits Schedule (MBS) 10% Dataset

Chris Culnane Benjamin Rubinstein Vanessa Teague

s47F

Department of Computing and Information Systems
University of Melbourne

September 17, 2016

1 Introduction

This article describes a problem in the encryption of Service provider ID numbers in the Medicare Benefits Schedule (MBS) dataset recently published at <http://data.gov.au>. Using only publicly available information, we found that it was possible to decrypt the service provider ID numbers. We did not decrypt Patient ID numbers.

The dataset does not include names or addresses, but the implications of exposing individual IDs are still serious.

Once notified, the Department of Health immediately removed the dataset from the website, opened lines of communication with us to further understand the issue, and conducted their own careful investigation of the problem.

1.1 Background on the dataset

The MBS 10% sample was constructed by choosing a random 10% of patient identifier numbers (PINs). For each selected PIN, (almost) all Medicare claiming activities were included in the data. A separate dataset lists Pharmaceutical Benefits Schedule (PBS) claims for the same patients. Most of the data are not encrypted, including Medicare item codes for tests and doctors' visits, pharmaceutical benefits (PBS) item codes, and other information such as pricing and location. The codes describe *what* medical activity occurred, for example a particular kind of test, a particular prescription, or a visit to a GP or specialist. The ID numbers describing *who* received or provided the service are encrypted.

The dataset does not include names or addresses of patients or doctors—the MBS dataset includes an encrypted PIN and provider ID for each activity. Note that the patient's PIN is not the number written on their Medicare card, and that the PBS dataset does not include the prescribing doctor's ID. The website states that “A suite of confidentiality measures including encryption, perturbation and exclusion of rare events has been applied to safeguard personal health information and ensure that patients and providers cannot be re-identified.” Strictly speaking, the method of obscuring patient and provider ID numbers is not encryption, though we will use that term to refer to it for consistency.

In summary, the data were treated in four different ways:

1. some rare events were removed;
2. some data, such as Medicare and PBS item codes, are recorded without encryption;
3. some data, such as the date of the item, are randomly perturbed; and
4. individual patient and provider IDs are encrypted.

1.2 Background on deanonymizing data

There are two main ways of trying to identify individuals in this kind of dataset:

Linkage attacks use the unencrypted data to identify people by linking the record with other known information; and

Cryptographic attacks reverse the encryption algorithm to recover encrypted data.

This article concentrates on successful cryptographic attacks. Linkage attacks have been demonstrated on other datasets [Swe02, NS08, NSR11, DMHVB13] but we have not attempted them here.

The next section explains why the encryption of provider IDs is not secure.

2 How the encryption of individual IDs can be reversed

s47E(d)

3 Conclusion

Publishing data can bring great benefits to research but also great risks to privacy. The technical details matter: it's a technically challenging task to understand whether a particular algorithm securely encrypts data or not. Datasets containing sensitive information about individuals clearly deserve more caution, and may not always be suitable for open public release.

The Australian government's open data program provides numerous benefits, allowing better decisions to be made based on evidence, careful analysis, and widespread access to accurate information. Decisions about data publication itself should follow the same philosophy.

The best approach would be to publish technical details about privacy protection long before the dataset, so that errors or vulnerabilities in the proposed scheme could be identified and corrected before the data were released. The privacy protections themselves should be detailed in public in advance. They can then be subject to empirical testing, scientific analysis, and open public review, before they are used on real data. Then we can make sound, evidence-based decisions about how to benefit from open data without sacrificing individual privacy.

References

- [Aus16] Australian Government Department of Health. MBS 10% sample technical information, June 2016. <http://data.gov.au/dataset/a8e3c0bc-44ac-4e9a-8b3c-b779438ddb10/resource/a33103be-17d8-4857-b621-334a813f4b99/download/mbs-10-sample-metadata-part-5-of-5---technical-information.pdf>.
- [DMHVB13] Yves-Alexandre De Montjoye, César A Hidalgo, Michel Verleysen, and Vincent D Blondel. Unique in the crowd: The privacy bounds of human mobility. *Scientific reports*, 3, 2013. <http://www.nature.com/articles/srep01376>.
- [NS08] Arvind Narayanan and Vitaly Shmatikov. Robust de-anonymization of large sparse datasets. In *2008 IEEE Symposium on Security and Privacy (SP 2008)*, pages 111–125. IEEE, 2008.
- [NSR11] Arvind Narayanan, Elaine Shi, and Benjamin IP Rubinstein. Link prediction by de-anonymization: How we won the Kaggle social network challenge. In *The 2011 International Joint Conference on Neural Networks (IJCNN)*, pages 1825–1834. IEEE, 2011.
- [Swe02] Latanya Sweeney. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(05):557–570, 2002.



Department of Health

Independent Review of 10% Sample MBS and PBS Data Decryption

October 2016



Australian Government
Department of Health

protiviti[®]
Risk & Business Consulting.
Internal Audit.



Review Sign-off

Timeline

Milestone	Date Completed
Planning	12 September 2016
Agreed Review Plan	12 September 2016
Fieldwork Complete	23 September 2016
Exit meeting	23 September 2016
Draft Review Report to Management	28 September 2016
Management Comments	28 October 2016
Final Review Report	2 November 2016
Report Tabled at Audit Committee	

Sign-off

I acknowledge the findings and recommendations enclosed in this report and undertake to have the agreed actions completed within the identified time frame

.....
October 2016

.....
October 2016



Table of Contents

1. Executive Summary	4
2. Background, Scope and Approach.....	6
3. Chronology.....	9
4. Risk Analysis	13
5. Positive Practices	15
5.1. Consistency with the Government's espoused principles for data access.....	15
6. Findings and Recommendations	16
6.1. Encryption method applied to the MBS data in the dataset had not been tested for its purpose	16
6.2. Data governance arrangements at Health are in a state of change, reducing clarity of requirements and expectations in data management and release	18
6.3. Formalities for approval of release of the dataset were not appropriately robust	20
6.4. Project and risk management disciplines associated with the development and release of the dataset would have benefited from increased rigour.....	21
7. Observations.....	23
7.1. Consistency of standards and testing in central data release forums.....	23
Appendix 1: Personnel Consulted	24
Appendix 2: Documents Reviewed.....	25



1. Executive Summary

1.1. Background

On 1 August 2016, a sample of de-identified and encrypted Medicare Benefits Scheme (MBS) and Pharmaceutical Benefits Scheme (PBS) patient level claims data was publicly released onto the website www.data.gov.au. This followed over a year of work to assemble, test and confirm the dataset for release. A suite of confidentiality measures had been applied to the data to protect the identity of patients and providers.

On 8 September 2016, the Australian Bureau of Statistics (ABS) was advised by researchers from the University of Melbourne that they had been able to decrypt the Service Provider Number from the 2014 MBS dataset contained within the broader dataset. After being notified, the Department of Health confirmed that the sample data had in fact been decrypted.

1.2. Overall Assessment

This independent review found that the Department's initiatives to publicly release this data were consistent with the principles of the Government regarding enabling broader access to information for research, policy-making and decision-making.

Scope Item 1: Understand and Define Risk Exposure to the Department from the Incident

The likelihood of public re-identification of the data in the dataset is possible given for some components of the dataset it has been proven to be possible, and this will likely become public. There are varying levels of likelihood depending on attributes of the dataset.

The primary consequence of this public re-identification is reputational, and the reputational consequence is likely to be high, given the sensitivity of the data and the level of public expectation. There is also some risk of legislative breach being an outcome if there is public reidentification; however this is likely to be contested.

Scope Item 2: Define root cause of the unexpected disclosure of sensitive information to the Department

This independent review has found that there were four primary contributing factors that led to the unexpected re-identification of sensitive information from the publicly released linkable 10% MBS and PBS Sample dataset:

- The encryption methods utilised for the Service Provider Number had not been tested to confirm fitness for purpose given the unique nature of data and the more public release of the data.
- Data governance arrangements at Health were in a state of change, to a more controlled model of management, reducing clarity at the time of requirements and expectations in data management and release.
- Formalities for approval of release of the dataset were not appropriately robust.
- Project and risk management disciplines associated with the development and release of the dataset would have benefited from increased rigour.



1.3. Summary of Findings and Recommendations

Report Section	Finding	Recommendation	Risk Rating
6.1	Encryption method applied to the MBS data in the dataset had not been tested for its purpose	Data release policies for the Department should include a requirement for specific testing of confidentialisation measures for all public releases of unit-level or patient level data to confirm that they are appropriate to meet the Department's risk appetite for data release. Such testing should be undertaken by qualified personnel with relevant skills including (as required) cryptography and data analysis.	High
6.2	Data governance arrangements at Health are in a state of change, reducing clarity of requirements and expectations in data management and release	The Department should continue and accelerate its implementation of centralisation of certain aspects of data governance in order to increase consistency of approach and decision making for data access. Further, the central governance body should develop a comprehensive set of guidance for line areas that hold data to ensure consistent approaches.	High
6.3	Formalities for approval of release of the dataset were not appropriately robust	Data release policies for the Department should include a requirement for all public releases of unit-level or patient level data to be reviewed and approved by the responsible Deputy Secretary.	Medium
6.4	Project and risk management disciplines associated with the development and release of the dataset would have benefited from increased rigour	Data release policies for the Department should include a requirement for all public releases of unit-level or patient level data to apply Departmental project management disciplines and risk assessment and management disciplines consistent with those applied for data integration projects.	Medium

1.4. Overall Management Comments

.

1.5. Limitation

This report is intended solely for the use by the Department of Health, and should not be distributed to any third party without the consent of Protiviti, which will not be unreasonably withheld.



2. Background, Scope and Approach

2.1. Background

The Department of Health is the custodian of extensive data holdings relating to health care and health care costs in Australia. This is an important resource to inform decision making by Governments, providers, pharmaceutical companies and other relevant entities.

The *Public Sector Data Management* report issued by the Department of the Prime Minister and Cabinet in July 2015, noted the following:

“Currently, the Commonwealth’s capacity to fully derive value from public sector data is constrained by competing priorities and the lack of an overarching strategy.

- There is no clear mandate for the Commonwealth to use and release public sector data.
- There are barriers (perceived and real) to sharing data within the Commonwealth and with jurisdictions to improve policy and service delivery.
- The APS lacks sufficient incentives, skills and organisational arrangements to capitalise on its data.
- The Commonwealth does not have a strong culture of publishing data to foster economic opportunities.”

As a consequence, a series of recommendations were made, which the Government has committed to implement. In total, 15 recommendations were made, which can be grouped as follows:

1. Build confidence and momentum (six recommendations) to be implemented over 1-6 months.
2. Systematise the use and release of public sector data (nine recommendations) to be implemented over 1-18 months.

On 1 August 2016, a sample of de-identified and encrypted Medicare Benefits Scheme (MBS) and Pharmaceutical Benefits Scheme (PBS) patient level claims data was publicly released onto the website www.data.gov.au. This followed over a year of work to assemble, test and confirm the dataset for release. A suite of confidentiality measures had been applied to the data to protect the identity of patients and providers.

The dataset comprised linkable MBS and PBS claims data for a random selection of 10% of the Australian patients, including MBS data from 1984 and PBS data from 2003 up to 2014. The data release represents approximately 3 billion lines of data relating to approximately 2.5 million Australians. The intention of the Department is to update this dataset annually as additional data becomes available.

The release of the data is expected to provide researchers with the opportunity to undertake research without having to bear the costs and delays associated with requests to government for unit record linked data. It would facilitate longitudinal analysis of patient progression through the Australian health system and provide the basis for cohort analysis.

The release of the dataset was an extension of a long-standing arrangement that had been established by the Department of Human Services (DHS) whereby a 10% sample of de-identified PBS (only) claims data had been released to a selection of approved research organisations on a contracted basis, with certain restrictions on how the data could be used by those approved research organisations. This arrangement had been in place since 2005. The structure and methodology of the sampling and confidentialisation of the data for the linkable MBS and PBS claims data that was publicly released was based on the approach that had been taken for the PBS (only) data.



On 8 September 2016, the Australian Bureau of Statistics (ABS) was advised by researchers from the University of Melbourne that they had been able to decrypt the Service Provider Number from the 2014 MBS dataset contained within the broader dataset. After being notified, the Department of Health confirmed that the sample data had in fact been decrypted. The Department immediately requested the suspension of access to the dataset on www.data.gov.au.

It is understood that approximately 1,500 downloads of all or part of the dataset had been made prior to the suspension of access.

2.2. Objective

Following the identification of this issue, the Department of Health commissioned an independent review to assess and document the circumstances associated with the recent event of decryption and re-identification of certain MBS and PBS data and assess and document the controls that were undertaken to manage the risk of privacy breach or unexpected disclosure of the sensitive information with a view to:

1. Understand and define exposure to the Department;
2. Define root cause of the unexpected disclosure of sensitive information to the Department; and
3. Identify recommendations for the future to ensure the risk is managed consistent with the Department's risk appetite.

2.3. Scope

All aspects of the selection, assembly, management (including encryption and perturbation) and subsequent re-identification of data associated with the reported incident are included within the scope of the independent review.

Scope Limitations

The independent review does not:

- assess response strategies once the incident was identified;
- assess the Department's data management and publication strategy, but rather has assessed the processes and decision making in the case of the data release in question (consistent with that strategy);
- test the encryption and perturbation mechanisms;
- provide legal opinion in connection with compliance with legislation, although it did consider legal advice sought by the Department;
- make any judgement on the accuracy or otherwise of the content of the dataset.

The assessments made during this independent review have been provided in good faith and in the belief that such statements and opinions are not false or misleading. Due to the limited duration of the independent review, Protiviti has relied on information that was provided by Health and other stakeholders. (A full list of officers consulted in the independent review is included in Appendix 2 to this report.) Protiviti does not express an opinion as to whether the information supplied is accurate and no warranty of accuracy or reliability will be given. Furthermore, we have not implied and it should not be construed that we have verified the information provided to us, or that our enquiries could reveal any matter that a more extensive examination might disclose.



2.4. Methodology

The approach to the independent review is defined below.

- Through consultation with data owners, stakeholders including DHS and ABS, and personnel who identified the ability to decrypt the data, and consideration of existing risk assessment documentation associated with the released data, document the risk associated with the potential decryption of the released data (considering both likelihood and consequence of decryption and identification);
- Through consultation with the University of Melbourne, identify how the decryption was identified and implemented, with a view to assisting in identification of root cause;
- Identify all aspects of the management of the project to select, assemble and, manage (including decryption and perturbation) the data, with a focus on identification of key controls to manage the risk. Key controls may relate to (but not be limited by): due diligence regarding release project, selection of encryption and perturbation techniques, testing of encryption and perturbation, information authorisation, quality assurance and risk assurance processes, and decision making and authorisation for release;
- Identify the root causes of the unexpected decryption of information;
- Through consideration of root cause and controls, identify:
 - Whether any weakness in control contributed to the occurrence of the event, and whether this is outside the Department's documented risk appetite.
 - What (if any) recommendations should be implemented in the future to manage the risk to a level consistent with the Department's risk appetite.
- Make comment on records retained to support the overall project and management of risk associated with the data release.

2.5. Review Team

Team Member	Position
s47F	Managing Director



3. Chronology

The following table presents a chronology of relevant events leading up to the identification of decryption of the data

Date	Event
1990's	DHS records indicate that the method of scrambling/encryption of data used for the 10% PBS Sample dataset was potentially being used by the then Health Insurance Commission and was later adopted for use in the 10% PBS Sample dataset. (Based on advice from DHS.)
Mid 2005	DHS was approached by a research organisation requesting a PBS sample data. From this the 10% PBS sample dataset was created and the research organisation contracted on a fee-for-service basis to provide quarterly 10% sample datasets. This arrangement has continued to operate ever since, with new approved research organisations being included in the arrangement.
16 July 2010	Minister for Finance makes <i>Declaration of Open Government</i> .
May 2011	Office of the Australian Information Commissioner releases <i>Principles on open public sector information</i> .
19 July 2011	<i>Enterprise Data Management – High Level Principles</i> released by the Chief Information and Knowledge Officer of the Department of Health.
March 2012	DHS, in consultation with Health, proposed a review of the existing 10% PBS sample dataset provision arrangements (at that time, to six private sector organisations), including review of the contract and the arrangements. It was agreed to utilise the assistance of an outposted officer from ABS in Health. Review was part of a broader workplan which included a "review of the 10% PBS dataset in its current form", including "is there an ability to identify consumers, providers or pharmacies within a three month period or over a greater length of time". The workplan also included broader work in development of guidelines and review of fee-for-service contract." The intent was the work was to be completed within one year.
25 May 2012	Terms of Reference of Department of Health Data Governance Council is approved.
14 June 2012	<i>Enterprise Data Governance Framework</i> re-released (prior version April 2012), which provides guidance on data governance and management. (The framework includes limited detailed instruction regarding data release procedures.)
30 August 2012	Meeting between representatives of Health and ABS to progress ABS review. It was agreed at this meeting that the project would be split into two phases.
27 November 2012	ABS (Methodology and Data Management Division) issued quote for conduct of the review
23 May 2013	Memorandum of Understanding between ABS, Health and DHS in agreeing the <i>Review of the 10% PBS Sample File</i> , including definition of scope which includes two main components (summary only): (1) Is there an ability to identify consumers, providers or pharmacies in the current file?; and (2) Is the 10% PBS sample file fit for purpose?.
28 August 2013	Australian Bureau of Statistics issues <i>Review of the 10% PBS Sample File – Phase 1 Report</i> following their review of the previously released 10% PBS Sample dataset.
11 November 2013	Health accepts findings of <i>Review of the 10% PBS Sample File – Phase 1 Report</i> and confirms requirement to continue to <i>Review of the 10% PBS Sample File – Phase 2 Report</i> .
29 July 2014	Australian Bureau of Statistics issues <i>Review of the 10% PBS Sample File – Phase 2</i>



Date	Event
	<i>Report</i> following their review of the previously released 10% PBS Sample dataset.
29 January 2015	Meeting between Health (Paul Storey, Angela Mikalauskas, Martin Butler) and ABS representatives to discuss and scope a potential review of the disclosure risks associated with the 10% MBS Sample dataset (based on similar review done for 10% PBS Sample dataset). This review was never commissioned by the Department of Health.
July 2015	Discussion Paper <i>Data – getting it, governing it, making it available, strategic analytics and research</i> issued by Paul Madden, Special Adviser proposing new arrangements for data governance and management at Health. These new arrangements involve a centralisation of certain aspects of data governance.
24 July 2015	LSB issues advice (LEX 23633) to Therapeutic Goods Administration <i>Linking MBS and PBS data</i> . Advice responds to the following questions: (1) Can you advise how the TGA can link, work with and store MBS and PBS data for the Project?; (2) Is a Public Interest Certificate required?; (3) can TGA staff or their contractors link the data – is the Secretary's approval required for this activity?; (4) For how long can the linked file be stored on TGA systems?; and (5) How should the separate MBS and PBS files be stored?
25 August 2015	Issue of latest version of Department's <i>Data Access and Release Policy</i> . The policy defines principles for access and release of "identifiable" and "de-identified" data but does not make statement on what constitutes de-identified.
14 October 2015	Data Governance Council (DGC) meeting at which the proposed release of the 10% MBS/PBS Sample dataset. Minutes of the meeting note: • "Data will be put through a confidentialisation process that the department is confident will make re-identification of the owner of the data a practical impossibility." • "Health is confident the re-identification is near impossible as the information held is of an administrative nature and doesn't include diagnostic data." • "The DGC agreed to the public release of linkable MBS, PBS and public hospital 10% sample data set, noting the release proposed will be presented to the Executive for final approval."
28 October 2015	<i>Bilateral Head Agreement 2015-18</i> between Secretary of DHS and Secretary of Health signed. This is an umbrella agreement on broad ranging working arrangements.
3 November 2015	Advice from LSB to MBD: "We have reviewed the methods being proposed to ensure no individual is able to be identified as set out in the Data Governance Council agenda paper. Based on this review and your statement that you are satisfied that once those actions are taken no individual will be able to be re-identified, we are of the opinion that the linkable 10% sample of MBS and PBS data will be de-identified and therefore not information 'with respect to the affairs' of a person. Therefore the secrecy provisions in s130 of the Health Insurance Act 1972 and s135A of the National Health Act 1953 will not apply to prohibit disclosure. The information will also not be 'personal information' for the purposes of s6 of the Privacy Act 1988."
November 2015	Draft ¹ Information Brief to Deputy Secretary advising of planned MBS data releases, including reference to 10% linkable sample. Notes that "careful attention to confidentiality requirements and other data sensitivities".
1 December 2015	Australian Government Solicitor issues advice (15194245) <i>De-identified MBS and PBS data</i> to the Department of Health (LSB). Advice responds to the following questions: (1) In what circumstances will a court find that a 'de-identified' MBS and PBS data is

¹ The original and final Information Brief has not been able to be obtained by this independent review.



Date	Event				
	not 'information relating to the affairs of [a] person' such that the prohibitions in s135(1) of the National Health Act 1953 and s130(1) of the Health Insurance Act 1973 will not apply?; (2) In what circumstances, will a court find that this data is not information relating to an individual and therefore not subject to the guidelines made under s135AA of the National Health Act 1953?; (3) If a court finds that the MBS and PBS data is information relation to the affairs of [a] person ... what sanctions could apply to the officer in handling the data in breach of the prohibitions in s135A(1) of the National Health Act 1953 and s130(1) of the Health Insurance Act 1973"; and (4) Is 'information relating to the affairs of [a] person' the same as 'personal information' for the purposes of the Privacy Act 1988?				
3 December 2015	Department of Prime Minister and Cabinet issue the <i>Public Sector Data Management Report</i> advocating better utilisation of data. Report recommendations include: <table> <tr> <td>1-6 months: Build confidence and momentum</td><td>1-18 months: Systematise the use and release of public sector data</td></tr> <tr> <td> 1. Secretary of PM&C signals data is a priority 2. Commission several high value projects 3. Build external partnerships 4. Publish readily available non-sensitive datasets 5. Build data and analytics capability 6. PM&C coordinates projects and progresses data policies and governance </td><td> 7. Implement a data policy framework 8. Build and maintain trust 9. Establish a trusted-access model for sharing integrated data 10. Create and publish a searchable whole-of-government data catalogue 11. Develop a Commonwealth Government high-value dataset framework 12. Publish data management standards 13. Establish a consistent and transparent approach to user charging. 14. Create a legislative environment that supports data use while maintaining privacy. 15. Promote innovation in public administration. </td></tr> </table>	1-6 months: Build confidence and momentum	1-18 months: Systematise the use and release of public sector data	1. Secretary of PM&C signals data is a priority 2. Commission several high value projects 3. Build external partnerships 4. Publish readily available non-sensitive datasets 5. Build data and analytics capability 6. PM&C coordinates projects and progresses data policies and governance	7. Implement a data policy framework 8. Build and maintain trust 9. Establish a trusted-access model for sharing integrated data 10. Create and publish a searchable whole-of-government data catalogue 11. Develop a Commonwealth Government high-value dataset framework 12. Publish data management standards 13. Establish a consistent and transparent approach to user charging. 14. Create a legislative environment that supports data use while maintaining privacy. 15. Promote innovation in public administration.
1-6 months: Build confidence and momentum	1-18 months: Systematise the use and release of public sector data				
1. Secretary of PM&C signals data is a priority 2. Commission several high value projects 3. Build external partnerships 4. Publish readily available non-sensitive datasets 5. Build data and analytics capability 6. PM&C coordinates projects and progresses data policies and governance	7. Implement a data policy framework 8. Build and maintain trust 9. Establish a trusted-access model for sharing integrated data 10. Create and publish a searchable whole-of-government data catalogue 11. Develop a Commonwealth Government high-value dataset framework 12. Publish data management standards 13. Establish a consistent and transparent approach to user charging. 14. Create a legislative environment that supports data use while maintaining privacy. 15. Promote innovation in public administration.				
7 December 2015	Prime Minister issues <i>Australian Government Public Data Policy Statement</i> . (Statement notes that Australian Government entities will "uphold the highest standards of security and privacy for the individual".				
15 April 2016	Deputy Secretary acknowledges information brief noting pre-launch release of subset of 10% MBS/PBS Sample dataset at PharmHack 2016. Brief notes: "The design of the datasets has been rigorously tested both within the Department and vetted by the Australian Bureau of Statistics. The datasets have also undergone extensive internal testing. A suite of confidentiality measures including encryption, perturbation and exclusion of rare events has been applied to safeguard personal health information and ensure that patients and providers cannot be identified."				
16-17 April 2016	Pre-launch release of subset of 10% MBS/PBS Sample dataset at PharmHack 2016.				
May 2016	Draft High Level Principles for the Release of Health Information and Data released.				
May 2016	Draft Protocol for the External Release of Health Information and Data released.				
25 May 2016	Half-day workshop between Health and DHS to establish a "Protocol for the External				



Date	Event
	Release of Health Information and Data". (Draft Protocol at this time was draft with considerable gaps.).
2 June 2016	Feedback received on a subset of the 10% MBS/PBS Sample dataset, which was circulated to some contracted recipients of the previous 10% PBS Sample dataset. Comments focussed on usability, not on security.
July 2016	Information brief released to Minister Ley, Assistant Minister Wyatt and Assistant Minister Gillespie informing them of the release of linkable de-identified 10% sample of Medicare and PBS claims data (MB16-001178). (Cleared by FAS and AS in MBD and PBD on 26 July 2016). Brief notes: "A suite of confidentiality measures including encryption, perturbation and exclusion of rare events has been applied to safeguard personal health information and ensure that patients and providers cannot be re-identified."
July 2016	Department of Prime Minister and Cabinet issues an Implementation Report arising from the Public Sector Data Governance Report advocating better utilisation of data.
1 August 2016	Linkable de-identified 10% sample of MBS and PBS data, including description of dataset, explanation of confidentiality measures and data adjustments and high level definition of encryption is published on www.data.gov.au .
8 September 2016	Draft Terms of Reference of proposed Data Governance and Analytics Committee is issued for comment.
8 September 2016	University of Melbourne representatives draft and deliver a report to the ABS describing the problem in the encryption of the Service Provider ID from the MBS 10% Sample dataset.
8 September 2016	ABS representative provides a copy of the University of Melbourne report to Health.

4. Risk Analysis

The data set includes two types of personal de-identified data relating to Australian citizens who have received medical services or pharmaceuticals (patients), and medical service providers who have received medical benefits under the MBS.

In the context of this review, the primary risk relates to the broader (public) access to information that can identify either of those types of personal information.

Re-identification can occur broadly through one of two means: (a) decryption of patient or service provider identifiers that have been encrypted using a variety of means; or (b) identifying one or several known records and using them as identifiers, or other known information outside the dataset, to re-identify other records relating to an individual or the entire population.

Likelihood of Decryption of Service Provider Identifier

The event that triggered this independent review indicates that the re-identification of service provider identifiers is at least possible. While it requires some knowledge of encryption, there is sufficient capability in society to suggest that re-identification is **possible**. Further, there are tools and information that can assist people to link a service provider identifier to personal or practice information about the service provider.

This possibility is reinforced by the fact that there are approximately 1,500 downloads of the dataset (in part or in full) that are now in circulation and uncontrolled.

Likelihood of Decryption of Patient Identifier

The Department has sought the assistance of an Australian Government agency with deep expertise in cryptography to see if they can re-identify the patient data from the dataset. At this time of this independent review, that agency has been unable to re-identify patient information from the encrypted identifier. While not absolute, this suggests that the likelihood of reidentification of patients through decryption is **low**.

Likelihood of Re-identification Through Data Linkage with Other Known or External Data

The Department has sought the assistance of Data 61 to quantify the likelihood of re-identification from the dataset. The results of this analysis identified that:

- (a) Re-identification likelihood varies from very low (<0.1%) for attributes classified as commonly available, to significantly higher risks (>10%) (which can be classed as **possible** using the Department's risk assessment methodology for selected moderately available data attributes, and then to severe risk for rarely available attributes.
- (b) The combination of attributes does increase the risk of re-identification; and
- (c) The impact of high-identification risks is limited to a small population in the dataset (1.2% of individuals).

Consequence of Reidentification

In the context of the specific event that prompted this independent review, the re-identification of service providers from the data has a **moderate** consequence to the Department. The data set provides the opportunity, for identified service providers, to observe billing practices (including billing in excess of MBS rates) and to observe services provided. For some service providers that may provide visibility of provision of sensitive services which some service providers will not wish to be publicly available.



Alternatively, should it transpire that patients are re-identified from the dataset, this is likely to have a higher consequence to the Department as the patient confidentiality is the very reason that structures are in place to protect health record data.

From a legal perspective, reidentification of service provider or patient data raises a risk of breach of any of the Privacy Act, the National Health Act and the Health Insurance Act. The organisation has received legal advice that if data was released in a form that the Department genuinely believed it had been de-identified, and then is subsequently re-identified, this would not represent a breach. However, this independent review noted that there is some contention on this matter, as it is untested. It is likely that this legal principle may be challenged in the event of publication of re-identification of data. It is further noted that the Department's Legal Services Branch is aware of this risk and is preparing accordingly.

Ultimately the main immediate consequence of re-identification of this data is reputational damage for the Department. The Department is tasked to hold this data and is expected to do so carefully. The very re-identification of the data, and some of the circumstances surrounding the event, are likely to attract considerable criticism.

The independent review has been made aware that other jurisdictions are (or are planning to) release similar data for similar purposes. This may mitigate some of the consequence of the risk.

Released under the Freedom of Information Act 1982 by the Department of Health, Disability and Ageing



5. Positive Practices

5.1. Consistency with the Government's espoused principles for data access

The Department's initiatives to publicly release this data are consistent with the principles of the Government regarding enabling broader access to information for research, policy-making and decision-making. These principles are noted in:

- Minister for Finance's *Declaration of Open Government* (2010)
- Office of the Australian Information Commissioner *Principles on open public sector information* (May 2011)
- Department of the Prime Minister and Cabinet's *Public Sector Data Management Report* (2015); and
- Australian Government *Public Sector Data Policy Statement* (2015).

Further, feedback from researchers and representatives of the Department has noted that the information is useful and valuable to researchers to inform Government and marketplace activity in the provision of healthcare.

Released under the Freedom of Information Act 1982 by the Department of Health, Disability and Ageing



6. Findings and Recommendations

s47E(d)

Released under the Freedom of Information Act 1982 by the Department of Health, Disability and Ageing



s47E(d)

Released under the Freedom of Information Act 1982 by the Department of Health, Disability and Ageing



s47E(d)

6.2. Data governance arrangements at Health are in a state of change, reducing clarity of requirements and expectations in data management and release

Finding

At the time of the incident, the Department was transitioning to more controlled governance arrangements for data custody and release which are expected to drive improvements in control and consistency in decision making on data management and release. These were not in place to inform the management arrangements for the 10% MBS and PBS Sample data initiative.

Discussion

Over the past four years, there has been considerable change in the data management arrangements at Department of Health.

In June 2012, the Department issued its first *Enterprise Data Governance Framework*, as a component of the then Department of Health National Alignment Project. This framework was designed to be consistent with the globally recognised Data Management Body of Knowledge. It is a set of data governance principles, responsibilities and structures designed to ensure data is managed effectively and securely.

In August 2015, the Department issued its latest *Data Access and Release Policy* which defined a policy statement on data release and certain guidelines, including guidelines relating to “low-risk de-identified, confidentialised or non re-identifiable data”. The guidelines note a focus on public accessibility of data, and that public accessibility and release should “maintain the security of data holdings and individual and organisational policy”.

However there is no further guidance on Department standards on how to maintain security and confidentiality.

Since mid-2015, and the release of a discussion paper entitled *Data – getting it, governing it, making it available, strategic analysis and research*, there has been consideration of changes in structure regarding the responsibility for data management in the Department. These considerations remain ongoing at the time of this independent review.



This environment requires data custodians to make judgements on how and when to take action to comply with the *Enterprise Data Governance Framework* and related policies, in the absence of clarity on limits and tolerances or expectations on how to apply the policies. The data custodians for the MBS and PBS data in effect exercised their judgements in that context in the release of the 10% MBS and PBS Sample dataset.

The following represents a description of the anticipated arrangements (which were not in place at the time of the release of the 10% MBS and PBS Sample dataset:

- New, more centralised arrangements are being implemented for the management of access to the department's data. This is to enable the greatest possible use of the data whilst ensuring that privacy and confidentiality of data are preserved, and that consistent decisions are made with respect to data access requests.
- The new arrangements mean that all data requests will be centrally coordinated via a "single front door". This will make it easier for researchers and others to know how to apply for access, as well as allow us to monitor progress and ensure requests are assessed in an appropriate timeframe relative to the complexity of the request.
- The arrangements will also ensure that all data requests are assessed using standardised criteria, and that a record is kept of who has received health data and the uses to which they have put it.
- In addition, a new senior oversight committee has been formed, chaired at deputy level, which advances the work already completed by the department in implementing a data governance framework and risk management framework to the department's data management, including the management of data requests.
- These arrangements are part of a broader strategy to continuously improve the management of the department's data and analytics capability. The new "hub and spoke" model for data and analysis was implemented at the beginning of September, following Executive agreement in August.

It is understood that as part of the new arrangements, consistent central approaches to certain elements of data management and access are being considered. One of these is the "five safes" framework for access to sensitive data.

Risk Exposure

The absence of consistent data management disciplines likely contributed to the de-cryption by the University of Melbourne. Further, the absence of such disciplines indicates poor risk management which may be noted when the event is analysed by stakeholders.

Recommendation 2	
Risk Rating	High
The Department should continue and accelerate its implementation of centralisation of certain aspects of data governance in order to increase consistency of approach and decision making for data access. Further, the central governance body should develop a comprehensive set of guidance for line areas that hold data to ensure consistent approaches.	
Management Comments	
Action Officer	Estimated Completion Date



6.3. Formalities for approval of release of the dataset were not appropriately robust

Finding

No Executive or Deputy Secretary approval was obtained prior to the release of the 10% MBS and PBS Sample dataset to the public. While this may not have been required in a formal delegation context, it is not unreasonable to suggest it would have been appropriate given the unprecedented nature of the data release and the commitments made that such approval would be sought.

Discussion

The Secretary of the Department of Health issues privacy guidelines for MBS and PBS data, consistent with the *Privacy Guidelines for the Medicare Benefits and Pharmaceutical Benefits Programs* issued by the Privacy Commissioner under s135AA of the *National Health Act 1953*.

The Secretary's privacy guidelines (and delegations) do not specifically cater for the use and release of de-identified linkable data, which therefore provides some lack of clarity on where approval to release the data should be sought.

The proposal to release to 10% MBS and PBS Sample data was considered by the Department's Data Governance Council in October 2015. The minutes of that meeting record the following:

"The DGC agreed to the public release of linkable MBS, PBS and public hospital 10% sample data set, noting the release proposed will be presented to the Executive for final approval."

Following that consideration, this independent review has noted the following briefs regarding the release of the dataset:

- November 2015: Draft Information Brief to Deputy Secretary advising of planned MBS data releases, including reference to 10% linkable sample.
- 15 April 2016: Information brief noting pre-launch release of subset of 10% MBS/PBS Sample dataset at PharmHack 2016.
- July 2016: Information brief released to Minister Ley, Assistant Minister Wyatt and Assistant Minister Gillespie informing them of the release of linkable de-identified 10% sample of Medicare and PBS claims data (MB16-001178). (Cleared by FAS and AS in PBD on 26 July 2016).

Each of the April 2016 and July 2016 Briefs note the following: "A suite of confidentiality measures including encryption, perturbation and exclusion of rare events has been applied to safeguard personal health information and ensure that patients and providers cannot be re-identified."

The independent review has not identified any documented approval from the Deputy Secretary or the Departmental Executive to release the dataset. Given the guidance from the Data Governance Council and the fact that this data release was unprecedented, it is reasonable to expect that such an approval would have been obtained prior to release.

Risk Exposure

While Executive review and sign-off may or may not have identified the weakness in encryption, the failure to obtain such review and sign-off represents a process failure which will likely be criticised by stakeholders when this event is analysed.



Recommendation 3	
Risk Rating	Medium
Data release policies for the Department should include a requirement for all public releases of unit-level or patient level data to be reviewed and approved by the responsible Deputy Secretary.	
Management Comments	
Action Officer	Estimated Completion Date
	The centralised data governance and release framework will be submitted to the Data Governance and Analytics Committee and then to Executive for consideration for endorsement in December and implementation from January 2017.

6.4. Project and risk management disciplines associated with the development and release of the dataset would have benefited from increased rigour

Finding

The 10% MBS and PBS Sample dataset development and release was not conducted as a formal project with associated disciplines regarding business cases, approvals, budget and schedule management and formal testing and clearance processes.

Further, although there was clearly a deep consideration of the risks associated with the development and release of the dataset, no formal risk assessment was prepared, reviewed and approved by a project owner or senior executive.

Discussion

The independent review was advised that the 10% MBS and PBS Sample dataset development and release was not run as a formal project, but rather was "fitted in with other obligations". This means that many of the disciplines associated with good project management (business cases, approvals, budget and schedule management and formal testing and clearance processes) were not applied.

This is not to suggest it was not well-accepted and known that the initiative was underway at Senior Executive and broader Departmental contexts. There was reasonable visibility of the initiative and relevant staff were keeping their Branch and Divisional leaders informed.

However, the absence of business case and certain gate approvals likely missed an opportunity for consideration and guidance on issues such as confidentialising of the data in the dataset. It is impossible to know if such disciplines would have resulted in a different course of action; however the opportunity was not available.

In the context of data integration processes, the Department has a comprehensive risk assessment method and model which is applied to ensure risks are identified and mitigations approved. As this initiative was not a data integration process (it is a data linkage process), such a risk assessment was not undertaken.

It is clear that the officers responsible were highly attuned to the risks to be managed and were using their judgement and best endeavours to manage the risks accordingly. However,



the absence of a documented and rigorous assessment missed an opportunity for assurance of completeness of the identification of risks, and the confirmation that approaches being taken to manage risks were appropriate and within the risk appetite of the Department.

Risk Exposure

The absence of effective project and risk management disciplines likely contributed to the de-cryption by the University of Melbourne. Further, the absence of such disciplines indicates poor risk management which may be noted when the event is analysed by stakeholders.

Recommendation 4	
Risk Rating	Medium
Data release policies for the Department should include a requirement for all public releases of unit-level or patient level data to apply Departmental project management disciplines and risk assessment and management disciplines consistent with those applied for data integration projects.	
Management Comments	
Agree The new centralised data governance and release framework will include a formal risk assessment framework which will be applied to all releases of data external to the department. This framework will categorise data releases into low, medium and high risk releases, with standardised assessment, risk mitigation and approval processes to deal with each. This framework will be consistent with that used for data linkage projects, as well as any new whole-of-government processes introduced to deal with unit record data releases. The framework will be designed such that the release approval process will be auditable, reportable and consistently applied across the Department. Project management disciplines consistent with those applied to data integration projects will be applied to data releases generally.	
Action Officer	Estimated Completion Date
	The centralised data governance and release framework will be submitted to the Data Governance and Analytics Committee and then to Executive for consideration for endorsement in December and implementation from January 2017.



7. Observations

The independent review identified the following area for improvement in knowledge management. This does not relate to a specific control weakness and therefore is not presented as a finding. However, it is suggested that this represents an opportunity for more effective and efficient practices.

7.1. Consistency of standards and testing in central data release forums

It is understood by the Department that there is limited central guidance on de-identification, or data structures, from the administrators of www.data.gov.au. In order to achieve a centralised consistency around the nature, structure and confidentiality of data being loaded on that central release site, there is an opportunity to impose certain standards and/or undertaken certain pre-loading checks to ensure only data that meets the Government's data release standards are released.

Released under the Freedom of Information Act 1982 by the Department of Health, Disability and Ageing



Appendix 1: Personnel Consulted

The following personnel were consulted as part of this independent review. We appreciate their assistance and candour throughout:

Name	Title/Role
Mr Andrew Stuart	Deputy Secretary Health Benefits
Mr Paul Madden	Special Adviser Strategic Health Systems and Information Management
Ms Maria Jolly	First Assistant Secretary Medical Benefits
Ms Penny Shakespeare	First Assistant Secretary Pharmaceutical Benefits
Ms Alanna Foster	First Assistant Secretary Research, Data and Evaluation
Mr Matt Yannopoulos	First Assistant Secretary Portfolio Investment
Ms April Purry	Assistant Secretary Legal Services
Mr Jaye Smith	Assistant Secretary Primary Care and Diagnostics
Mr Nick Henderson	Assistant Secretary Pharmaceutical Policy
Mr Ian Crettenden	Assistant Secretary Health Analytics
s47F, s47E(c)	Director Information and Knowledge Management
	Director PBS Information Management
	Director MBS Analytics
	Director Information Management and Relationships
	Assistant Director PBS Information Management
Ms Malisa Golightly	Deputy Secretary, Department of Human Services
Mr Mark Whithnell	First Assistant Secretary, Department of Human Services
Ms Joanne Mason	Assistant Secretary, Department of Human Services
s47F	General Manager, Australian Bureau of Statistics
	Director, Australian Bureau of Statistics
	University of Melbourne
	University of Melbourne



Appendix 2: Documents Reviewed

The following documents were considered as part of this independent review:

Documentation Title
Question Time Brief "10% MBS/PBS Sample Decryption", 12 September 2016
File 2013/017801, Pharmaceutical Benefits & Access: Agreements – Review of 10 percent PBS Sample
File 2015/000477, Pharmaceutical Benefits & Access: Agreements – Review of 10 percent PBS Sample – Part 2
Talking Points for Minister's Office "Public Release of Linkable 10% sample of Medicare Benefits Scheme (Medicare) and Pharmaceutical Benefits Scheme (PBS) Data, clearance by FAS PBD
Information Brief, "Release of linkable de-identified 10% sample of Medicare and PBS claims data", various drafts, including comments from PBD Executive.
Various emails relating to release of the data, from PBD staff.
Protocol for the External Release of Health Information and Data (Draft v 0.9), May 2016
High Level Principles for the Release of Health Information and Data (various drafts), May 2016
Information Management and Release Protocol, November 2012 (extract)
Bilateral Head Agreement 2015-2018 between The Secretary of the Department of Human Services and the Secretary of the Department of Health (signed, October 2015)
Department of Health Data Access and Release Policy (August 2015)
Frequently asked questions: further information about the Data Access and Release Policy
Australian Government Public Data Policy Statement (December 2015)
Minute to Deputy Secretary Andrew Stuart, "Pre-launch release of five years of linkable, line-by-line Medicare (MBS) and Pharmaceutical Benefits (PBS) data (April 2016) (final and various drafts)
Administrator of the National Health Funding Pool, "Data Linkage: Governance Framework Considerations" (October 2015) (including request for comment)
Undated document "MBS, PBS Open Source Data Release 2016"
Email exchange February 2016 "MI5073 – Linking MBS & PBS data to CHeReL dataset"
Email February 2015 "Notes from ABS/Health meeting on the 10% PBS and MBS samples"
Email exchange January 2016 "10% MBS/PBS Sample"
Minute to Deputy Secretary Andrew Stuart "For Information: MBS Data Release Plans" (November 2015) (draft)
Minute to Deputy Secretary Andrew Stuart "Approval to Link MBS and PBS Data to Analyse Whether Drugs Used to Treat Multiple Sclerosis are Related to the Incidence of Melanoma" (3 November 2015)
Email exchange November 2015 "MBS data release – update"
Email exchange November 2015 "Response from legals re release of linkable 10% sample of MBS, PBS"



Documentation Title
Email exchange October 2015 "AGS request for advice re de-identified data"
Data Governance Council Meeting 23 Minutes 14 October 2015
Data Governance Council Meeting 23 Paper Agenda Item 3: Public Release of Linkable MBS, PBS and Public Hospital 10% Sample Data Sets
Data Governance Council Meeting 23 Paper Agenda Item 3i: Public Release of Linkable 10% sample of MBS, PBS and Hospital Admissions Data
Request for comment on paper: Public Release of Linkable 10% sample of MBS, PBS and Hospital Admissions Data
Email exchange September 2015 "For comment: draft papers for DGC – release of 10% MBS, PBS and hospital data"
Email exchange June 2015 "ABS report Review of the 10% PBS Sample File".
Minute from David Cullen, Strategic Policy Unit "Providing the Strategic Policy Unit with Access to PBS Data", 1 April 2015
Minute from Paul Creech, Pharmaceutical Policy Branch "Providing the Strategic Policy Unit with Access to PBS Data", 20 February 2015
Minute from David Cullen, Strategic Policy Unit "Providing the Strategic Policy Unit with Access to PBS Data", 2 February 2015
Email 6 February 2015 "rules for creating 10% sample"
"Review of the 10% PBS Sample File – Phase 1 Report", Australian Bureau of Statistics, 28 August 2013
"Review of the 10% PBS Sample File – Phase 2 Report", Australian Bureau of Statistics, 29 July 2014
Paper "Key events in preparation of 10% sample for release"
Australian Government Solicitor advice "De-identified MBS and PBS data" (Ref: 15194245), 1 December 2015
Australian Government Solicitor "Executive Summary: De-identified MBS and PBS data" (Draft 7 December 2015)
Email exchange March 2016 "Analytics Strategy – some question to assist the development of the strategy"
Title: Linkable De-identified 10% Sample of Medicare Benefits Scheme (Medicare) and Pharmaceutical Benefits Scheme (PBS) (undated)
"Deanonymizing the Medicare Benefits Schedule (MBS) and Pharmaceutical Benefits Schedule (PBS) dataset", University of Melbourne, 8 September 2016
Email exchange March 2016 "Request for PBS data for PharmHack"
List of PBS 10% Sample Licensees (including addresses)
Letter from Paul Storey "PBS/MBS 10% Data Sample" dated 11 May 2016 (advised that sent to PBS 10% Sample Licensees (with data release)
Minute to Deputy Secretary Andrew Stuart (draft) "To seek your in-principal agreement to publicly release de-identified, linkable 10% sample of the MBS and PBS via data.gov.au" (October 2015)
"Confidentiality practices in DHS", undated paper.
Legal Services Branch Advice 24 July 2015 "Linking MBS and PBS data".



Documentation Title
Deputy Secretaries Meeting Discussion Paper "Data – getting it, governing it, making it available, strategic analytics and research", July 2015
Download from www.data.gov.au (dated 14 September 2016) "Linkable de-identified 10% sample of MBS and PBS".
"Public Sector Data Management Report", Department of Prime Minister and Cabinet, December 2015
"Public Sector Data Management Implementation Report", Department of Prime Minister and Cabinet, July 2016
Data Governance Council Terms of Reference, version 1.0
"Enterprise Data Management High Level Principles", version 1.0
Database Alignment Project Enterprise Data Governance Framework, Version 1.01
Commonwealth Health Data Integration Risk Assessment Tool
Data Governance and Analytics Committee Terms of Reference
Five Safes: designing data access for research, University of the West of England, 2016
Data Access and Release Plan – Pharmaceutical Benefits Division v0.3, 31 July 2015
Pharmaceutical Benefits Division 10% MBS and PBS Sample Data Release Chronology
"How to confidentialise data: the basic principles:", National Statistical Service
"Principles on open public sector information", Office of the Australian Information Commissioner, May 2011
"Declaration of Open Government", Minister for Finance, July 2010
Quantitative analysis of re-identification risk from MBS and PBS data, Data 61, September 2016
MBS/PBS re-identification scenarios, Data 61, September 2016

DATA DEIDENTIFICATION: DOES IT WORK?

A REPORT ON THE FEASIBILITY OF DEIDENTIFYING THE MBS/PBS DATASET AND THE
IMPLICATIONS FOR FUTURE RELEASES OF AUSTRALIAN GOVERNMENT DATA.

Chris Culnane, Benjamin Rubinstein and Vanessa Teague¹,
Department of Computing and Information Systems
The University of Melbourne, 5 Dec 2016

s47F

@unimelb.edu.au

We examine the deidentification of the MBS/PBS dataset and the likelihood that improved methods might allow a secure version to be re-released. Our findings replicate those of similar studies of other deidentified datasets:

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

Although our specific example is the linked MBS/PBS 10% sample, we do not intend criticism of the people who published it or the methods they used. The methods were bound to fail, because they were trying to achieve two inconsistent aims: protection of individual privacy and publication of detailed individual records. We demonstrate why that is probably not possible for 30 years of medical billing data. We expect similar results to apply to other rich datasets in the government's care, including census data, tax records, mental health records, penal data and centrelink data.

We support the program of making more data more easily available to facilitate research, innovation and sound public policy. However, there is an important technical and procedural problem to solve: there is no good solution for publishing sensitive unit-level records that protects privacy without substantially degrading the usefulness of the data. Rather than committing to full open release of deidentified data, we suggest an open process of research and discussion of more secure alternatives. Australians can think of innovative ways to solve this problem if we can understand why existing solutions do not work.

1

All of the sensitive database queries were conducted by V Teague.

There are two ways to identify individuals in a deidentified dataset: decryption and linking. The first section of this report asks how hard it was to recognise and correct the weakness in the encryption of supplier IDs, which we notified the government about in September. We then examine the difficulty of linking attacks, which involve combining other sources of information with the published dataset to reidentify individuals.

How hard is this reidentification?

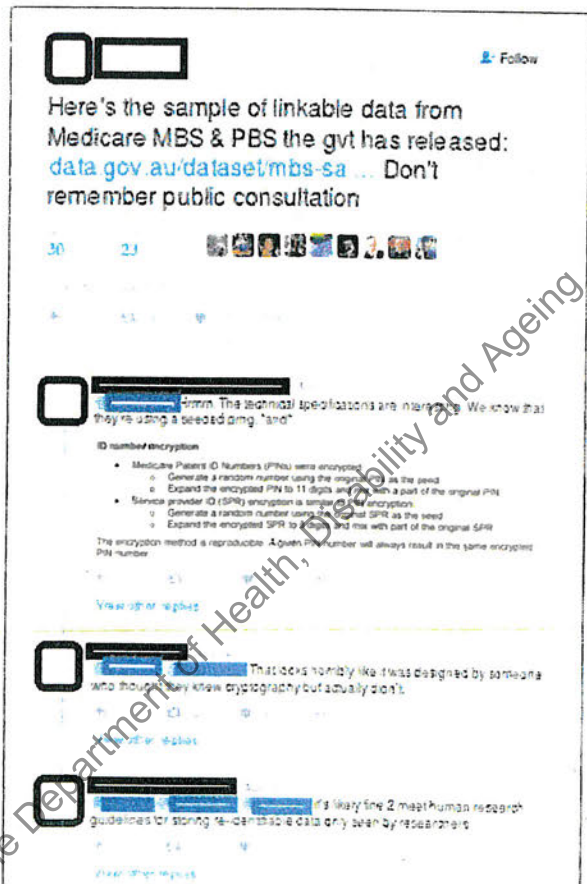
Unfortunately, it is straightforward for anyone with technical skills about the level of an undergraduate computing degree.

DECRYPTING SUPPLIER IDS

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.

We learned later that the inappropriateness of that encryption scheme had been discussed extensively on Twitter weeks beforehand. (See picture.) So we were not the first to discover this problem - it was obvious to many other people.

This 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.



INDIVIDUAL LINKAGE ATTACKS

"Succinctly put, 'Deidentified' data isn't, and the culprit is auxiliary information."

--- Cynthia Dwork

Computer scientists have used linkage attacks to reidentify deidentified data from various sources including telephone metadata (Srivatsa & Hicks, 2012), social network connections (Narayanan, Shi, & Rubinstein, 2011), health data (Sweeney, 2002) and online ratings (Narayanan & Shmatikov, 2008), and found high rates of uniqueness in mobility data (De Montjoye, 2013) and credit card transactions (de Montjoye, Radaelli, Singh, & Pentland, 2015). Linkage attacks work by identifying a "digital fingerprint" in the data, meaning a combination of features that uniquely identifies a person. If two datasets have related records, one person's digital fingerprint should be the same in both. This allows linking of a person's data from the two different datasets – if one dataset has names then the other dataset can be reidentified. This is not necessarily sophisticated: reidentification based on simply linking with online information has also been reported (Barbaro & Zeller Jr, 2010) (Duhigg, 2002).

EXAMPLE: THE NETFLIX MOVIE RATINGS DATASET

Netflix, a US DVD rental and streaming service, published a deidentified dataset of users' movie ratings. The IMDb movie knowledgebase website listed users' names along with their ratings. Many users rated a similar (or identical) list of movies on both sites. Different people watch different movies: a person's collection of movies served as a fingerprint, uniquely identifying many users. Narayan and Shmatikov (2008) demonstrated that the Netflix dataset could be reidentified by matching it with IMDb movie ratings and the attached name. One user reidentified in this way had rated sensitive movies privately on Netflix, but not on IMDb. The reidentification technique still worked even given perturbations in the data and some errors in the linking assumptions, because a person's film choices are often very different from everyone else's.

We attempt to apply these techniques to the MBS/PBS dataset.

BACKGROUND ON THE MBS/PBS 10% SAMPLE DATASET

The MBS/PBS dataset contains billing information, including PBS (prescription) and MBS (medical) records for 10% of Australians born in each year. Each patient has an encrypted ID number and a year of birth and gender. Each record attaches a medical event to a patient, including a code identifying the service or prescription,

the state the supplier and patient were in, a date, the price paid by the patient and reimbursed by Medicare and, in the case of MBS records, an encrypted supplier ID. Some rare events were removed before publication, and all the dates were perturbed randomly by up to two weeks in an effort to protect privacy. Some MBS/PBS item codes are generic (code 00023 occurs millions of times, indicating a visit to the GP); others are highly specific and sensitive, such as prescriptions that are only for HIV patients.

How easy is it to identify individuals in the dataset by linking the unencrypted parts of the record with known information about the person?

We start by investigating whether the sort of health, demographic, or movement data typically available in a person's public profile is sufficient to identify them uniquely if they are in the released dataset. This is the minimum information that could be available to a malicious attacker.

Many people include on their public profiles their year of birth and gender. This typically puts them in a crowd of more than ten thousand in the MBS/PBS 10% sample. When new information is added, however, the set of possible matches shrinks rapidly. Each new item of known information reduces the set of possible matches within the dataset. When the set of possible matches is small enough to inspect manually, the person's privacy is seriously at risk, if that person is in the dataset.

SEARCHING FOR MYSELF (V TEAGUE)

There are nearly 3 million people in the dataset, but only 17,310 women share my year of birth. Two of my children were born in Australia, one in 2006 and the other in 2011. Specifying these years of birth gives 59 possible matches, 25 in Victoria. Knowing their exact birthdates, and incorporating the two-week perturbation of dates, leaves zero records.

This shows that I am not in the dataset, but if I was I would be easily isolated based on no more information than a typical personal Wikipedia page. This is consistent with other privacy analyses in the literature. A person doesn't need to have a single rare event in order to be identifiable - often a collection of ordinary facts suffices.

Of course, reidentification can be mistaken, but these sorts of results give us some idea of how much information is necessary to identify a person with confidence. If

we found that, for about 90% of people, two exact dates of childbirth returned zero matches, then we could infer a fairly low probability of mistaken reidentification based on that much information.

SEARCHING FOR OTHERS

s47E(d)



Released under the Freedom of Information Act 1982 by the Department of Health, Disability and Ageing

s47E(d)



s47E(d)

HOW COULD PRIVACY OF THESE RECORDS BE PROTECTED?

s47E(d)

The next obvious suggestion is to remove unusual individuals completely, or at least to remove unusual events such as giving birth very late or very early in life. However, this is highly problematic. A person's "anonymity set" is the number of records consistent with all known facts about that person. Mothers in the table above have an anonymity set of 1 if their year of birth and year of giving birth are known. What size of anonymity set is acceptable? For large states like NSW and Victoria, there are quite a few women giving birth in their late 40s. Should we demand the same level of anonymity for smaller states such as SA/NT and WA? How much extra information should be necessary to isolate a particular patient? For example, information about at least one other child is likely to produce a unique match in this set. Is that acceptable?

These important questions do not seem to have good answers. Removing these individuals or events would reduce the accuracy of the data and bias or even prevent some possible studies. Leaving them in is a serious risk to privacy, even with a large perturbation of dates.

Removing unusual individuals is a false option anyway, because everyone is unusual given enough information about them. Indeed, for childbirth as for many other facts about people, a small collection of ordinary events is usually enough to isolate that person. In the next section we discuss identification based on knowledge of 2 or 3 dates of childbirth among women who are not particularly old or young.

UNIQUE CHILDBIRTH PATTERNS

We read Wikipedia pages and online news stories about famous Australians who are also mothers, recording their years of birth and the birthdates of their children if we could find them. V Teague then entered that information as queries in the

database. The first few, including Prof Gillian Triggs, former Senator Natasha Stott-Despoja, Olympic athlete Cathy Freeman and deputy opposition leader Tanya Plibersek, returned zero matches. Hence we can be fairly confident that those people are not in the sample.³ This sounds comforting, but really isn't, because it suggests that if they were there they would probably be unique and therefore easily identifiable.

Some other queries returned unique matches, though we do not have enough corroborating information to be certain of the identification. These are listed in a separate Appendix.

We examined the sample data to understand how much information about a woman's childbirth history was needed to identify her. An individual in the dataset matches someone else if they were born in the same year and if every time they have a child, the dates are within 4 weeks. (This is the furthest that two events truly on the same day could be shifted apart by the 14-day perturbation.) We asked how much information was necessary to limit one person to a very small number of (other) matches. The results are shown in the dark blue bars of Figure 3. We have already seen that some very old or very young mothers are uniquely identifiable. So are women who have unusually many children. We found that 102,593 women in the dataset have at most 6 possible matches (including themselves). More than 55,000 are unique.⁴

We then asked whether a larger perturbation of dates would improve privacy, and if so by how much. The results are shown in the same figure, for 28 days (red), 60 days (yellow) and 90 days (green). Like other similar studies, we found that larger perturbations improved privacy only a little. Doubling the perturbation to 28 days still leaves 30,000 unique individuals. Even with a 90-day perturbation, more than 9,000 mothers are unique. In the context of pregnancy and childbirth, a 3 month perturbation of dates would seriously affect the accuracy of many scientific studies.

³ Many childbirths occur in a public hospital setting that is never billed to medicare and therefore not in the dataset. Hence it is possible that some of those people are in the dataset but their childbirths are not.

⁴ Again, these uniqueness results apply only to women who gave birth in a setting that was billed to medicare. This probably correlates highly with financial resources and hence is relatively easy to guess.

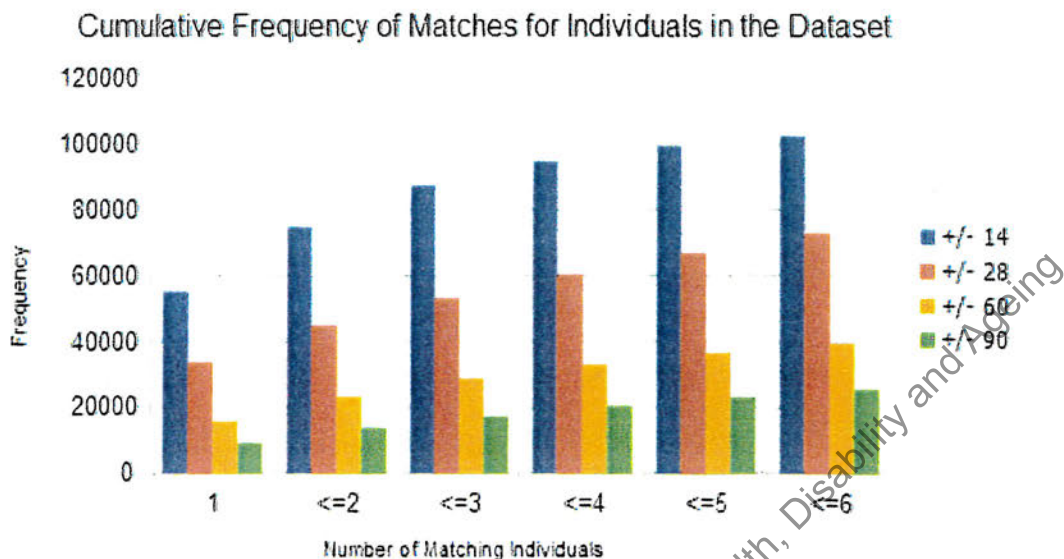


FIGURE 3: MATCHES BY CHILDBIRTH DATES

OTHER HEALTH-RELATED FINGERPRINTS

Many other health-related characteristics could be enough to identify a person. Many other types of people make a great deal of information about their health available on the Internet. V Teague conducted a brief study of professional footballers, noting their injuries and surgeries over the years from online information and translating them into database queries. The results were very similar to those for childbirths. Some people's distinctive histories are clearly absent. For one AFL team captain, a unique record matches the publicly available information about his medical history, year of birth, and interstate movements. This is also listed in the Appendix. There was not enough information to validate this identification with certainty, but generally the level of publicly available detail about elite athletes' health is enough to identify some of them in the dataset.

SINGLE EVENTS

Sometimes a news story gives very precise dates for someone's surgery or hospital admission. In two instances a news story about a relatively common operation on a (former or current) member of a state parliament matches a unique record. These are listed in the Appendix. There is a chance that these reidentifications are wrong because we do not have much extra information, but if those people are in the dataset then their records are obvious.

LARGE-SCALE LINKAGE WITH OTHER DATASETS

So far we have used linking information from one-off sources. However, some entities have access to large databases, with names, that could be a source of linking information for many people at once.

Australian privacy law refers to what can "reasonably" be reidentified, but this depends on what other data is available for linking. Latanya Sweeney purchased the electoral roll database of Cambridge (USA) and used it to reidentify the state governor's medical record (Sweeney, 2002). The Australian electoral roll is not available for purchase, but is widely distributed. What other datasets is it reasonable to assume an attacker might be able to access?

BILLING DATA AND HEALTH INSURERS

One example is the billing data itself: the fees associated with MBS records. Private health insurers have access to much of their own customers' Medicare claim information. A simple way to compute the links is to add up the numbers and compare the totals.

The MBS/PBS dataset lists, for each transaction, the amount paid by the patient (or private insurer) and the amount paid by the government. We computed, for each patient and each year, the government's total payments and the patient's total payments, for MBS and PBS separately, for all the transactions in that year. This produces four numbers each year, which constitute a fingerprint. Of the 2.1 million patients with records in 2014, nearly 900,000 received a unique total of paid MBS benefits, while over 900,000 paid a unique contribution. As we add more years, almost everyone is unique.

Most health insurance companies would be missing some of this data, for example if the item is bulk billed or paid by the patient. The same approach would work on whatever data they did have. If the rates of uniqueness were a little smaller on the restricted dataset, totals for different years of data could be used together.

This demonstrates that a private health insurer (for example) could efficiently track the medical records of past customers through the decades of data, or derive extra information they didn't know about from current customers. This would be a clear breach of privacy that would possibly never be reported, even though the data could lead to detrimental decisions for the individual in the future.

The proposed amendments to the Privacy Act criminalizing reidentification would not stop this. Rather than explicitly reidentifying the government dataset, the health insurance company could link it to a de-identified version of their own database, derive some conclusions, delete the MBS/PBS data, and then link their conclusions back to their named clients using the intersecting data.

PHARMACY RECORDS AND DATA BROKERS

In 2015, data broking firm IMS reportedly purchased “anonymous” records of patients’ prescription histories from pharmacists (Dunlevy, 2015). Depending on the effectiveness of the anonymization, it might be possible for these to be deanonymized or linked with the published PBS data.

PAYMENT DATA AND BANKS OR CREDIT CARD COMPANIES

Credit card records alone can be used to identify an individual. A recent paper on the uniqueness of credit card billing data (de Montjoye, Radaelli, Singh, & Pentland, 2015) found that only 4 transactions were needed to make 90% of people unique, even when prices were removed. Perturbing dates by 15 days improved privacy a little, but 80% of people were still unique based on 10 transactions. The exact numbers might not translate across to the Australian setting, but the implication would: if someone paid for even a handful of doctor’s visits or PBS prescriptions with their credit card, then their bank or credit card company would probably have enough information to reidentify their MBS/PBS record.

LINKING WITH OTHER GOVERNMENT DATASETS

The more individual records the government puts online, the more data can be used in a linkage attack. For example, raw census data includes dates of birth for the entire household.⁵ We are not aware of any public ABS microdata that reports this information, but if there was one it could be linked with dates of childbirth in the MBS/PBS dataset. Based on our analysis above, this would provide a unique fingerprint for many women and hence many households. Again some perturbation of dates would make this a little harder, but any reasonable perturbation would still allow some rate of successful linkages. The guarantee that “no personally identifiable data will be released from the census” could

⁵ So does data held by a school or public library.

therefore depend on what other data is already published and how easily it can be reidentified and linked to census data.

MALICIOUS ATTACKERS

The information available to well-intentioned researchers represents the absolute minimum information available to someone with criminal intent. The same approach could be used to link information extracted by social engineering, derived from social media by deception, or purchased from leaked or illegally exfiltrated datasets of personal information such as the recent Red Cross leak, which included names and addresses as well as some medical information.

Discoveries about data are often more powerful in combination than alone. The decryption of supplier IDs might not immediately reveal very much about patients, but it could be combined with other data to make a linkage attack much easier. Even without decryption, reidentification of one individual could make others easier, for example if one elite athlete's record reveals which surgeon treats other elite athletes.

The threat model here is very strong: a dataset released today might still be scrutinized in the presence of extra information decades into the future.

SAMPLING, UNIQUENESS AND CONFIDENCE

Several letters in recent issues of *Science* have been devoted to an argument over the meaning of uniqueness results in a sample (Sanchez, Martinez, & Domingo-Ferrer, 2016). One side of the argument emphasizes that

- uniqueness within the sample does not necessarily imply uniqueness within the whole population: a unique match might actually be a coincidental resemblance to someone who is not in the sample, so reidentification cannot be made with confidence, and
- uniqueness does not imply successful reidentification, if you don't know enough points about someone to make them unique.

Although these are both true, they are not as comforting as they sound.

- If you know the facts that make a person unique, and that person is in the sample, then uniqueness does imply reidentification. If those facts are easily available on the Internet, then uniqueness implies easy reidentification.

- There might be other ways to learn that a particular fingerprint is unique in the whole population. If that fingerprint appears in the sample, then the match is right.
- There might be more information available about a person than is required to make them unique. For example, if 3 facts produce a unique match, and another 2 known facts also match the retrieved record, this would increase confidence.
- Matches of infrequent types could be made with more confidence than matches of more common types. The frequency of occurrences of particular profiles in the sample could be used to estimate the frequency in the whole population.
- Confidence may not be necessary anyway. The identification is accurate if the person is in the dataset.

EXAMPLE: POPULATION UNIQUENESS BASED ON DHS STATISTICS

The department of human services releases statistics about the rates of MBS billing codes, for the whole population (not just 10%).⁶ These are aggregated into patient age ranges of ten years, and reported for each state and each month. Some codes are very uncommon, and are either absent or charged only once in particular months for particular ranges of patient age. Overall, 27% are unique.

For example, former Prime Minister Kevin Rudd had an unusual surgery (an aortic valve replacement) in Brisbane in August 2011. According to the DHS data, this surgery was unique in that month and age range. His record is not in the dataset, but if it was this could form evidence of a correct reidentification.⁷

This sort of data can also be used to reduce confidence in false identifications. For example, Senator Cory Bernardi was diagnosed with tuberculosis in February or March of 1996. The rate of this disease in Australia is so low that, when V Teague found a man of the right age in the dataset who had had this test, she was fairly confident that the match was right. However, a careful examination of the DHS

⁶ <https://data.gov.au/dataset/medicare-benefits-schedule-mbs-group>

⁷ This is complicated by the two-week perturbation. Although there were no other patients in August or September, there was one in July whose dates could have been perturbed to form a false positive. There would need to be some evidence against this possibility, e.g. if that person was also in the dataset. If not, there would be a 50% chance of accurate reidentification without other information.

data shows that, although in most age ranges and months there are zero or one tuberculosis tests in South Australia, in those particular dates and age ranges there are 4.⁸ This turned out to be a false match, and that fact was made more obvious by population-level uniqueness data.

EXAMPLE: ESTIMATES OF CONFIDENCE FOR REIDENTIFICATIONS

Our reidentification results are a concrete example.

s47E(d)

- Two reidentifications are by uniqueness based on age, gender, state and dates of surgery, but no other supporting information.

Some of these could be false matches, but if the person is in the dataset, then their record is easily identified.

A separate appendix contains the list of encrypted patient IDs with names. We will never release the list publicly.

If we had analyzed uniqueness and attempted reidentification over the whole population rather than a 10% sample, this would be a list of perhaps 30 or 40 very confident reidentifications. Instead we have a list of seven, of which 3 or 4 are likely to be right. For the 90% of people who are not in the sample, this is a huge gain. For the 10% who are, the incomplete confidence has little benefit.

⁸ This may not be a coincidence: tuberculosis is highly infectious and when one person tests positive their friends are often tested too. Perhaps the 3 other South Australian men of the same age were friends of Sen Bernardi.

s47E(d)

DISCUSSION: WHERE TO FROM HERE?

Taking advantage of the benefits of big data without seriously compromising privacy is one of the most difficult engineering challenges of our time. It makes no sense for the government to insist on one solution to this problem – the open publication of deidentified data – despite conclusive evidence that this solution does not work.

The MBS/PBS dataset release was motivated by convincing arguments about the utility of that data for medical research that saves lives. We strongly support this sort of research, and the general aim of informing public policy and inspiring innovation with scientific analysis of data. The question is how to engineer that without destroying privacy.

Our reidentification of some people in the MBS/PBS dataset is not an isolated case, but a replication of numerous other results in which similar techniques have been shown to work on other datasets. Reidentification will only become easier as more information is released. The combination of attributes that could form fingerprints is difficult to predict and safeguard against, which is why privacy criteria like k-anonymity (Sweeney, 2002) are inherently flawed (Machanavajjhala, Kifer, Gehrke, & Venkatasubramanian, 2007) (Li, Li, & Venkatasubramanian, 2007). It is very unlikely that even the most well-informed and well-intentioned set of guidelines on deidentification can guarantee privacy protections appropriate for sensitive data such as the MBS/PBS sample while retaining the usefulness of the data.

A well-intentioned official might carefully follow a deidentification process, but some individuals might still be “reasonably identifiable”. If compliance with deidentification guidelines is wrongly assumed to imply proper mathematical protections of privacy, this represents a serious loophole in privacy law, not just for government but also for private companies.

Of course it is possible to remove information from records until nothing meaningful can be derived about individual people. The question is whether this approach solves any of the problems that open data is intended to solve, or whether the manipulation necessary to protect privacy destroys the research value of the data. If scientists still need to apply for the unperturbed or unaggregated version, then we haven’t really solved the problem of data access.

SOME SPECIFIC QUESTIONS AND SUGGESTIONS

There are many alternatives to full open release of deidentified data. It is better to specify a good process for encouraging research and fact-based decisions about privacy protection and data sharing, than to commit in advance to one particular solution that probably doesn't work.

Different solutions might solve different problems. It is important to ask *exactly what problem are we trying to solve?*

Open public release of data about government is a great idea. Public transport data, expenditure, and other non-sensitive data should be publicly released.

Data about people should be much more carefully considered. It is important to ask why the data is being released, who needs to see it, what they are allowed to do with it, and what the implications of a breach would be. Here is a short list of possible solutions. All of them still include some risk of information leakage, and some inconvenience to researchers. Different approaches address different problems and have different risks - evaluating the tradeoffs would require input from the scientists who use this data.

- Basic deidentification, such as removing names, could be combined with other methods of securing the data.
- Access to data could be controlled either physically or electronically. Researchers could sign up and have their access controlled by a license with penalties for misuse or leakage.
- The Productivity Commission's *draft report on Data Availability and Use* recommends giving deidentified data to "trusted users." These could be required to meet standards for data security.
- Researchers could apply for only those parts of a dataset that they actually need to do their particular research.
- Narayan and Shmatikov (2010) suggest "*An interactive, query-based approach is generally superior from the privacy perspective to the 'release-and-forget' approach.*"
- Aggregate data rather than individual records could be released publicly. Frameworks such as Differential Privacy (Dwork, 2008) could be used to assess what aggregates are privacy-preserving.
- In some cases cryptography can be used to make inferences from hidden information, for example to compute totals without decrypting individual

records, or to link records from different datasets without decrypting individual IDs. However, this doesn't guarantee that the results of the linking or tallying will protect privacy.

- Portals such as data.gov.au could list all the available datasets, including sets not published openly on the website. This could include instructions and requirements for access to data that was available under stricter conditions.

Australia can learn from overseas examinations of the same issues. The European Union recently released guidelines for the release of public datasets (<https://www.europeandataportal.eu/en/content/how-address-privacy-concerns-when-opening-data>). Step 1 is "Understand the data. Consider potential use cases, the value of the data and potential risks." A deidentification algorithm might be suitable for a scientist working under a legal responsibility to keep the data private, but not suitable for public release of the dataset.

A US presidential commission on cybersecurity received a number of submissions on privacy and data sharing. An MIT submission (Pentland, Shrier, Hardjono, & Wladawsky-Berger, 2016) emphasised "Open Algorithms" and "Permissible Use." Open algorithms means that details about the methods and processes should be available for public scrutiny; permissible use emphasises the consent or expectations of the people whose data is shared.

Privacy protection is a subfield of computer science with a peer-reviewed literature that can underpin good decisions about the release of sensitive data. Privacy should be designed into these processes by people who understand this science.

Publishing details about the deidentification techniques for the MBS/PBS data was the right thing to do because it allowed the problems to be detected. This approach should be retained and extended: before data is released, details about the process should be published so that they can be examined by privacy experts and the public. Then the government can make good decisions based on full information from free and open research and public comment.

CONCLUSION

Recent government policy relies on secure deidentification of detailed personal data, but secure deidentification of rich data is probably not possible without substantially degrading the data (Ohm, 2010). This report shows that some MBS/PBS records can be reidentified, thus adding another example to a long list of unsuccessful attempts at deidentification of sensitive datasets.

These failures are not simply a result of choosing a bad method of deidentification. They reflect the inherent statistical fact that a small number of ordinary points of information is often enough to identify a person. Perturbing the data or decreasing its precision can improve privacy gradually, at considerable cost to accuracy. Removing rare individuals or rare events is a false hope because everyone is unique if enough information about them is known.

The proposed amendment to the Privacy Act to criminalise reidentification will not solve these problems. It will make them harder to detect, understand and avoid.

There are exciting new ideas for provably privacy-preserving computation on sensitive data. Google has experimented with differential privacy (Abadi, et al., 2016), using a statistical technique for computing valid aggregate statistics with a high degree of uncertainty about individual data. Microsoft researchers have used homomorphic encryption to compute on private medical data without ever learning the individual records (Bos, Lauter, & Naehrig, 2014). We even know of some Australians who routinely use this kind of maths.

The Australian government holds vast quantities of information about individual Australians. It is not really "government data". It is data about people, entrusted to the government's care. Data about government should be published openly and freely - not so for sensitive data about people. That should be published only when a clear, public, explanation of the encryption and anonymization methods has received enough peer review and public scrutiny to convince everyone that personal information will remain private. For some datasets, including the MBS/PBS record-level data, this is probably not possible.

Making more data available more widely could have many benefits, but the approach needs to be re-engineered with a better understanding of the risks. Australia really can become a leader in the data sciences by encouraging free and open research in privacy-preserving technologies for data sharing. Understanding which ideas don't work is a first step in innovation towards techniques that do.

BIBLIOGRAPHY

- Abadi, M., Chu, A., Goodfellow, I., McMahan, B., Mironov, I., Talwar, K., et al. (2016). Deep Learning with Differential Privacy. *23rd ACM Conference on Computer and Communications Security (ACM CCS)*.
- Barbaro, M., & Zeller Jr, T. (2010, April 9). A Face Is Exposed for AOL Searcher No. 4417749. *The New York Times*.
- Bos, J., Lauter, K., & Naehrig, M. (2014). Private predictive analysis on encrypted medical data. *Journal of Biomedical Informatics*, 50, 234-243.
- De Montjoye, Y.-A. e. (2013). Unique in the crowd: The privacy bounds of human mobility. *Scientific reports* 3.
- de Montjoye, Y.-A., Radaelli, L., Singh, V. K., & Pentland, A. ". (2015). Unique in the shopping mall: On the reidentifiability of credit card metadata. *Science*, 347 (6221), 536-539.
- Duhigg, C. (2002, Feb 16). How Companies Learn Your Secrets. *The New York Times Magazine*.
- Dunlevy, S. (2015). Some chemists are selling patients' prescription information so drug companies can increase their profits. *news.com.au*.
- Dwork, C. (2008). Differential privacy: A survey of results. *International Conference on Theory and Applications of Models of Computation*.
- Erich, Y., & Narayanan, A. (2014). Routes for breaching and protecting genetic privacy. *Nature Reviews Genetics*, 15 (6), 409-421.
- Li, M., Li, T., & Venkatasubramanian, S. (2007). t-closeness: Privacy beyond k-anonymity and l-diversity. *2007 IEEE 23rd International Conference on Data Engineering*, 106-115.
- Machanavajjhala, A., Kifer, D., Gehrke, J., & Venkatasubramanian, M. (2007). l-diversity: Privacy beyond k-anonymity. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 1 (1).
- Narayanan, A., & Shmatikov, V. (2010). Myths and fallacies of personally identifiable information. *Communications of the ACM* 53.6, 24-26.

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

Narayanan, A., Shi, E., & Rubinstein, B. I. (2011). Link prediction by de-anonymization: How we won the Kaggle social network challenge. *The 2011 International Joint Conference on Neural Networks (IJCNN)* (pp. 1825-1834). IEEE.

Ohm, P. (2010). Broken promises of privacy: Responding to the surprising failure of anonymization. *UCLA Law Review*.

Pentland, A., Shrier, D., Hardjono, T., & Wladawsky-Berger, J. (2016). *Towards an Internet of trusted data: a new framework for identity and data sharing - input to the commission on enhancing national cybersecurity*. MIT Connection Lab.

Sanchez, D., Martinez, S., & Domingo-Ferrer, J. (2016). Comment on "unique in the shopping mall: on the re-identifiability of credit card metadata". *Science*, 351 (6279).

Srivatsa, M., & Hicks, M. (2012). De-anonymizing mobility traces: Using social network as a side-channel. *Proceedings of the 2012 ACM Conference on Computer and Communications Security* (pp. 628-637). ACM.

Sweeney, L. (2002). k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10 (05), 557-570.

Released under the Freedom of Information Act 1982 by the Department of Health, Disability and Ageing

s47E(d)

Released under the Freedom of Information Act 1982 by the Department of Health, Disability and Ageing