Data Sharing and Release Legislation Issues Paper Consultation Response

Dr Chris Culnane, A/Prof Benjamin Rubinstein, and Dr Vanessa Teague

July 2018

The opinions expressed in this submission are the authors' own and do not reflect the views of The University of Melbourne.

1 Introduction

Our research team demonstrated the easy re-identifiability of suppliers and patients in the MBS-PBS 10% sample. We explain here the implications for better protection of personal data.

The scientific fact that should underpin all decisions in this area is that a person's detailed individual record cannot be securely de-identified while still retaining most of its scientific value.

The MBS-PBS 10% longitudinal sample includes 30 years of billing records for each of the 2.9 million Australians who were present. Many patient records can be easily and confidently re-identified, using just a few points of information about the person such as a few dates of surgery or childbirth. Re-identification subsequently reveals detailed information about the person's medical history and prescriptions. While the OAIC has confirmed our finding that doctors' IDs could be recovered, the OAIC's report finds that patients could not be "reasonably identified" according to the Privacy Act—contradicting our demonstration to the OAIC and Department of Health.

Rather than indicating a "risk averse culture" as the issues paper suggests, the MBS-PBS 10% incident indicates inadequate technical protections of privacy, insufficient understanding of the risks of sharing data, and inadequate legal protection. The OAIC's interpretation of "reasonably identifiable" signals that very easily and confidently identifiable information can already be shared or sold without the individual's consent. If Australia's privacy regulation framework permits the public release of such easily identifiable sensitive information on so many individuals—irrespective of specific laws prohibiting such data linkage[1]—then Australians' personal data requires expanded protections, not a weakening of existing legislation.

Australian Privacy Principle 6 specifies

> *6.1 If an APP entity holds personal information about an individual that was collected for a particular purpose (the primary purpose), the entity must not use or disclose the information for another purpose (the secondary purpose) unless:*
> *(a) the individual has consented to the use or disclosure of the information, . . .*

There then follow a series of specific exceptions.

The *Data Sharing and Release* proposal reverses this position, by assuming that personal information held by government may be shared or published unless the responsible authority decides not to. Remarkably, "consent" does not appear once in the issues paper. We see no evidence that Australians support this significant change. By contrast, recent news on *My Health Record* suggests that our concerns over secondary uses of identifiable data are widely shared.

[1] https://www.aph.gov.au/Parliamentary_Business/Committees/Senate/Health/Health/Sixth%20Interim%20Report/c04, 4.5.

2 Responses to Questions

Key Principles of the Data Sharing and Release Bill

1. Are these the correct factors to take into account and to guide the legislative development?

No, not all factors are taken into account in the proposal, for guiding important legislative development. Privacy and data protection are given inadequate consideration by the proposal in its present form. That "consent" does not make one appearance in the proposal, while being a central tenet to privacy best practice, is indicative of significant misalignment. Without acknowledging and protecting the fundamental right to privacy of individuals, there will be little chance of establishing trust in the use of public data.

2. What else should the Government take into consideration when designing the legislation?

Neither data releases nor legislation exist in isolation. The government should consider improving privacy legislation, *e.g.* by adopting provisions from the GDPR. Anything else is likely to be counter productive by weakening already inadequate protections and rights of citizens and consumers.

Scope of the Data Sharing and Release Legislation

3. Should the scope be broader or narrower?

The entity that collected the data is not the prime concern, it is the type and nature of the data that is more important. As such, unit-record level data about individuals should never be shared without individual informed consent.

4. Are there entities that should be included or excluded from scope? How would this be justified?

Any entity that collects sensitive personal data, for example, medical data, should be excluded from the default sharing and release arrangements. Such data should only be shared when those for whom the data relates have provided consent. Such consent should be revocable at any time, for example, by adopting a dynamic consent approach.

Data about government, but not about individual citizens, should be open by default. Parliament applying the principles of the *Data Sharing and Release Bill* to itself, prior to sharing any further citizen data, would help build social license to begin to explore data sharing more broadly.

5. Should any specific categories of data be specifically out of scope? How would this be justified?

Any unit-record-level data about people, including medical records and also other sensitive information derived from Centrelink, education, tax, or other data, should be excluded and require individual informed and dynamic consent.

There are other categories of data where groups, rather than individuals, should be asked for their consent. Ask indigenous data sovereignty experts.

6. Should exemptions, for example for national security and law enforcement, occur at the organisational level or for specific data categories?

A significant risk for national security and law enforcement exceptions is potential for abuse by government in avoiding publishing data that could hold them to account. Attempts to open Attorney General George Brandis's diary were frustrated on the grounds of exemptions to the Freedom of Information act, which were ultimately decided to be spurious.[2] Ministerial diaries are rarely a legitimate matter of national security, and nor is most data about the ordinary workings of government. Exemptions should be evaluated by an independent, not politicized, government entity. The general public do not have the luxury to classify their own information as a matter of national security to prevent sharing, neither should ministers who seek to build social license with the public.

7. Are there instances where existing secrecy provisions should prevail?

To override existing secrecy provisions risks serious erosion of public trust, and would likely have a long term negative impact on trust in data use and the role of government more broadly. The current public discussion around *My Health Record* data does not indicate that it would be popular to override the (already limited) protection in that act.

We highlight that the proposal seeks to honour commercial agreements—stating that "Existing contractual obligations, including around purchased data sets, will continue to apply"—yet proposes breaking the legislative secrecy obligations that apply to the public.

Each of the existing protections was written for some reason, and data was collected under that protection. Overriding that protection means breaking the commitment under which the data was collected.

Streamlining Data Sharing and Release

8. Do you agree with the stated purposes for sharing data?

It is far from clear that the data should be shared for all the stated purposes.

In particular, the details of "2. support the efficient delivery of government services or government operations" are too broad. In particular, "administering or enforcing compliance requirements" could be used to justify action against individuals. Whilst it may be desirable to facilitate strong enforcement, it may have a detrimental impact on society as a whole if it causes disengagement from essential services for those in fear of enforcement action.

10. What further detail could be included in the purpose test?

The nature of the data and the necessity to have consent from the individual or relevant group should be included explicitly. Currently the purposes are fundamentally consequentialist, prioritising the perceived greater good instead of respecting minimal rights of the individual. Furthermore not all data sharing is beneficial. Data sharing may have significant risks for individual (or group) harm if data is exposed. Furthermore, there may be a significant public cost from distrust in data privacy—people may stop seeking treatment for stigmatised illnesses, or may stop giving accurate information in surveys such as the Census.

It is also important to distinguish genuine scientific research in the public interest from commercial research for financial gain.

[2] https://www.theguardian.com/australia-news/2016/jan/20/george-brandis-challenges-ruling-process-request-release-diary

11. Should data be shared for other purposes? If so, what are those purposes?

Consider adding a purpose consistent with the open government declaration to make government "...more transparent, responsive, accountable, and effective."[3]

12. Should there be scope to share data for broader, system-wide purposes?

Only data that isn't unit-record-level data about individual people.

13. Should the purpose test allow the sharing of data to administer or enforce compliance requirements?

As noted above, absolutely not.

Data Safeguards

14. Is the Five-Safes framework the appropriate mechanism to ensure data is safeguarded?

The Five-Safes framework itself provides no protection, it is solely a decision-making framework to plan and evaluate a prospective release[4]. It is clearly better to think about a release before it happens, but just thinking about it does not guarantee a safe release. The biggest problem with the Five-Safes framework is visible in its name: the framework optimistically biases towards determining a release to be safe, even when it is far from safe. There is no such thing as a "safe person" or a "safe environment," only "lower risk" and "higher risk" ones.

If the Five-Safes framework was in fact called the 'Five-Risks framework', with a corresponding pivot as follows, it would be more effective. It would be better to force users to evaluate risk, as opposed to safety, and would encourage evaluation of vulnerabilities instead of protections. Evaluating such vulnerabilities is a skilled task, and not something that can be achieved through following a framework document. It requires a deep understanding of the underlying mathematics, as well as wide understanding of privacy-preserving techniques. Such skill sets are rare, and it is far from clear that sufficient capacity exists within Australia to undertake a large number of data releases. Government would more effectively promote data sharing by investing in and building capability in the area.

Five Safes also encourages compartmentalised thinking, which is evident in both this document and related government advice. Data sets cannot be treated as isolated instances, data sets that appear to be *less sensitive*, or even not sensitive, may enable linking of more sensitive data.

15. Are there any additional safeguards that should be applied?

Data should only be publicly released when protected by established privacy-preserving techniques, which does not include traditional de-identification techniques, including those described in the flawed Data61/OAIC *De-identification Decision Making Framework*. Where individual data is to be released publicly it should be protected by differential privacy; where individual data is to be transmitted, stored or processed on untrusted platforms, it should be encrypted.

[3] https://www.opengovpartnership.org/open-government-declaration

[4] where 'release' is used as a term to refer to release or access to data

16. Are there any instances when the Five-Safes could not be applied?

Since the *Five-Safes* framework does not provide any actual protection, but merely provides a framework to facilitate the thought process, it is difficult to think of any instances where it could not be applied. What is more concerning is the lack of appreciation of what Five-Safes actually provides: it is the techniques applied to achieve the notional "safety" that are critical, not the framework itself.

17. Is the Five-Safes appropriate when data is shared and used for the specific purposes in the purpose test above?

As noted above, the Five-Safes framework is only worthwhile if used as a carefully-considered "five risks" framework.

19. How would you envisage Five-Safes principles be applied over the life-cycle of data to ensure data safeguards are continually met?

Data about individuals should not be released as unit records, instead access should be granted. It must always be possible to revoke and recall data—this is possible in a controlled research enviornment, but not possible after open release. Consider for example the Department of Health's decision to take down the MBS-PBS 10% sample. This is laudable, but does not remove it from the storage of the 1500 people who reportedly downloaded it.

20. Under what circumstances should trusted users be able to access sensitive data?

Only in secure environments, which must be both physically and digitally secure. Such facilities should be air-gapped and offline. The security measures should be regularly audited, with audit reports made public. Access to such facilities should require some form of clearance, which expires annually. All actions a user has undertaken within the environment should be recorded and auditable, forming a part of the renewal process.

Public Sector Data Sharing Arrangements

21. Would this arrangement overcome existing barriers to data sharing and release?

Where such barriers do exist, they are often there for a good reason: protecting the privacy of personal data.

Furthermore, existing barriers do not seem to be very strong. The Productivity Commission did not understand the MBS-PBS release when it "...found that a risk averse culture and complex and inconsistent requirements among Australian Government agencies was a barrier to data sharing and release." There was nothing risk-averse about the release of 1 billion rows of sensitive MBS/PBS data, in which many doctors and patients can be easily and confidently identified. This exposes the affected individuals to substantial risks of discrimination, denial of credit or insurance, extotion, *etc.*

The OAIC report on the MBS-PBS release indicates that the PBS dataset was already being shared before its public release. "The Department of Health applied technical de-identification measures that had been used, without incident, previously for releases of PBS data to approved

researchers."[5] Our research showed that the PBS dataset was identifiable on its own. It is obviously highly sensitive, so even assuming no further public releases of this kind of data, data sharing already occurs. There should be an obligation to publish outcomes of this past sharing to allow for a public assessment of its benefits.

22. Would streamlined and template agreements improve the process?

Security, privacy protection, and evaluations of safety are inherently bespoke. Future challenges will not be in the writing of agreements, but will be in the evaluation of risks. If those agreements are a shortcut to that process, then the entire undertaking will be undermined.

23. Do you agree that data sharing agreements should be made public by default?

All data sharing agreements must be made public in full. In light of the lack of quality data about what sharing has already taken place, in advance of further consideration of the *DS&R Bill* a thorough evaluation of existing data sharing should be conducted, the results of which must be made public. Such an evaluation should require all recipients of public data to declare the data sets they hold, where they were sourced from, the size (number of rows) in the data set, and a full data dictionary (list of fields in the data set). There would be no privacy implications of releasing such information, but it would provide an accurate and informed base line from which to determine future directions.

There is no security through obscurity. Predicated on cases of data sharing being performed with appropriate justification and privacy protections in place, publication of agreements will serve to build public trust in government and social license in data sharing. Indeed best-practice privacy-enhancing technologies such as cryptographic protocols and differential privacy make guarantees under threat models permitting full disclosure of the protection technologies used.

24. What level of detail should be published?

Full detail should be published, because transparency is essential to establishing trust. It is public data and as such the public have a right to oversight of its use. Full transparency is also a suitable certificate that appropriate protections are in place: as above best practice protections are made no weaker by open-sourcing protective algorithms.

25. What else should a data sharing agreement contain?

There should be appropriate redress clauses for individuals who may be harmed as a result of their data being shared through the agreement.

26. What other transparency mechanisms could be mandated?

Inline with comment regarding making data sharing agreements public, all recipients of public data should be required to maintain a publicly available list of data sets they have received, and the properties of that data set as previously described. Independent audits of data protection mechanisms and facilities should be annually published.

[5] https://www.oaic.gov.au/privacy-law/commissioner-initiated-investigation-reports/publication-of-mbs-pbs-data

Roles and responsibilities within the system

27. How long should accreditation as an ADA or Trusted user last?

We do not believe that it is possible to establish "Safe" people, but if an organisation was going to try, the accreditation would need to be at least as stringent as that of Australian Government Security Vetting. Such accreditation should be reviewed annually, based on an audit of the queries and actions they have taken within the data environment, with a more thorough review every 3 years.

ADA's should require yearly audits, which must be independent audits, not self declared. Existing approaches of evaluating organisations based on what they believe their security posture to be are unsuitable. For example, a number of organisations will claim to apply the Separation Principle, yet will simultaneously have a single system administrator with global access to everything, which is completely at odds with the principle.

28. What could the criteria for accreditation be?

The criteria should include organisational accreditation, to ensure the infrastructure is adequately supported and maintained. Individuals must be at an accredited organsation, which must support and guarantee ongoing support of the data environment the data will be deployed in. Individual accreditation should include a review of academic qualifications and past publications, as well as a declaration and evaluation of any conflicts of interest—particularly in regards to funding and consultation. It should also require demonstration of a thorough knowledge of privacy risks and cyber security best practice.

29. Should there be review rights for accreditation?

Accreditation should always be open to review, with appropriate protections provided for whistle blowers to ensure that deviations from accepted behaviour can be raised and identified early.

30. Should fees be payable to become accredited?

It appears reasonable to have a base fee for accreditation, with a grant scheme to sponsor researchers who can demonstrate a particular public interest need for access, but who do not have funding to pay for accreditation.

National Data Commissioner

32. Are these the right functions for the National Data Commissioner?

Setting up the National Data Commission to "...champion greater data sharing and release..." positions it to undermine privacy protections of individuals. It is of utmost importance that the individual (or appropriate group) be represented in the process, and that privacy protection be given principle focus. It is folly to view data about individuals as government data: such data is often highly sensitive data about individuals who have a right to determine its use.

Promoting best practice is a laudable aim. However, depending on current de-identification advice such as the OAIC/Data61 guidelines is highly risky and will likely lead to further major privacy breaches. It is unclear that best practice has yet been established, or that government agencies presently have capacity to establish it alone. A culture for consultation with technical

privacy experts, as well as legal and human rights experts, is a pre-condition (but not a guarantee) for the Commission to achieve its remit.

37. What aspects should be taken into consideration when considering consequences for non-compliance with the DS&R Bill?

Individual redress: if someone's data has been misused that individual will have incurred some degree of harm, which must be redressed either by the party in non-compliance, or the releasing agency which failed to adequately protect that individual's data.

38. Should the consequences differ depending on the type of data involved or the type of misuse, e.g. harsher penalties for intentional misuse?

Intentional or unintentional misuse is not the key issue, it is the scope for harm caused to individuals. If someone accidentally leaves a sensitive data set on a train, it was clearly not intentional, but the scope for harm remains.

39. Should penalties be strict liabilities?

Yes.

The suggestion that there should be immunity for good faith indicates a failure to understand the serious consequences for an individual if their data is exposed. There may be very serious harms including discrimination, exclusion from credit, or even exposure to family violence. It is inadequate to say that there shouldn't be a penalty as long as these harms were caused "in good faith." No such immunity should exist. A lack of severe consequences will breed complacency in the handling of data.

The suggestion that strict liability would discourage openness about breaches is peculiar in the light of the existence of the *Notifiable Data Breaches* legislation. The idea that it might disincentivise data sharing is not an appropriate concern. If an agency is concerned a release might result in a privacy breach for which they would be held strictly liable, it is a good indicator that it should not be making that release. The purpose of the liability is to impose a threshold in confidence required before releasing data. If that threshold cannot be reached either the data is too sensitive, or the protection techniques inadequate.

40. What would be an appropriate penalty for intentional misuse of data?

For an individual, two years in jail and a lifetime ban from access to all data in the future. The organisation that supported that individual should also receive a 2 year ban from access to data, and have all existing data sets revoked. It is essential to have strict penalties to incentivise protection and oversight.

41. How would responsibility for misuse of data be shared across the data system?

In incidents of misuse there are two streams of responsibility that must be assigned. Firstly, the individual/organisation that perpetrated the misuse should be held responsible for their actions. Additionally, the agencies and entities that signed off on the initial release must also be held responsible, since misuse would demonstrate a failure in either process or judgment. The responsibility of the perpetrator and the facilitator are distinct, and as such, not shareable.

It is important to distinguish between those who are doing harm and those who are explaining a failure in existing protections. The proposal to outlaw re-identificaiton of public datasets would simply have prevented analysis of weak privacy protections, not corrected the mistakes.[6]

42. To what extent should there be a complaints mechanism and how should it work?

Without a complaints mechanism it will be difficult to evaluate and detect breaches, in all but the self-reported cases. Individuals must have a right to hold to account those using their data.

43. Should a complaints mechanism provide for complaints by the public?

Absolutely, it is the public whose data is being used, it is the public who ultimately hold the risk of a breach of data about them, and it is the public that will suffer the harm in the incidence of another breach.

Recent controversy over My Health Records suggests that ordinary members of the public were interested and concerned about secondary uses of their data, to an extent not predicted by those who administered the system. Respecting these sorts of concerns would go a long way towards building public trust.

[6] See submissions to the Parliamentary inquiry at https://www.aph.gov.au/Parliamentary_Business/Committees/Senate/Legal_and_Constitutional_Affairs/PrivacyReidentification/Submissions