Authoritative information and statistics
to promote better health and wellbeing

Mr Peter Harris AO
Chairman
Productivity Commission
Level 12, 530 Collins Street
Locked Bag 2, Collins St East
Melbourne VIC 8003

Dear Mr Harris

Inquiry into data availability and use

The Australian Institute of Health and Welfare (AIHW) welcomes the opportunity to make a submission to the Productivity Commission's Inquiry into data availability and use. This submission is based on the extensive experience AIHW has gained from working closely with national health and welfare data, its stakeholders and using data linkage methods to explore fundamental policy questions.

The AIHW is a major national agency set up by the Australian Government in 1987. It is an independent corporate Commonwealth entity within the Health portfolio. Our mission is to provide authoritative information and statistics to promote better health and wellbeing. We provide timely, reliable and relevant information and statistics on hospitals and other health services, aged care, child care, services for people with disabilities, housing assistance, youth justice and other community services.

We collect data and manage national data collections in these areas, producing over 150 information products each year. Our work is frequently referenced by the media. We also provide information to other government bodies and cross-jurisdictional councils, to external researchers and also directly back to data providers. We report in formats that suit their purposes and allow them to place their service provision in a wider context.

We also develop, maintain and promote data standards to ensure that data collected are nationally consistent. In all these activities, we enable governments and the community to make better informed decisions to improve the health and wellbeing of Australians.

We trust that you find this information useful to your inquiry. Should the committee have any queries about the information I have provided or wish to seek additional data from the AIHW, please contact Mr Geoff Neideck, Head, Data Strategies and Information Technology Group,

Yours sincerely

Barry Sandison
Director (CEO)
Australian Institute of Health and Welfare
2 August 2016

Table of Contents

Introduction

Recent years have seen a rapid evolution in the ability to efficiently handle large amounts of data and produce meaningful insight through analysis of big data resources. There has also been an ongoing cultural shift during which individuals and organisations have become both more connected and more dependent on interoperable information systems. The effect of this shift has been more than just a change in the availability and quantity of data but it has also resulted in a reported shift in societal expectations for how this data should be used (see, for example, Cate et al (2013)). The Australian Government through its Public Data Policy Statement supports the use of data to *"stimulate innovation and enable economic outcomes"* and has committed to improve *"the use and reuse of public data"*.

Advances in technical opportunities to use data must continue to move in line with and be constrained by the ability to appropriately and ethically access available data resources. This is especially apparent in the health sector where large potential efficiency gains from the management of the health system and in public health outcomes due to wider usage of data must be balanced against the particular sensitivities of handling personal health information (see for example, Raghupathi and Raghupathi 2014). Health data is different to many other sources of data in this regard as many of the gains in research come from examining small clusters for which standard confidentialisation approaches could render the data ineffective. Some of the nuance associated with handling health data is currently acknowledged in general terms by *"permitted health situations"* in the Australian Privacy Principles and the waiver provisions for data used in medical research contained in section 95A of the *Privacy Act 1988*. In line with the technological advances for usage of data there has also been parallel development of enhanced tools to handle data safely, this was recognised by the European Public Health Association who said,

> *"For valid and efficient public health monitoring and research, it is sometimes necessary to use personal health data. Modern ICT technologies can ensure sufficient safeguards for personal health data."*

The Australian Institute of Health and Welfare (AIHW) is the national agency set up to provide information and statistics on Australia's health and welfare. It is a corporate Commonwealth entity accountable to the Australian Parliament, and is governed by a management Board. AIHW's work informs debate and decisions on policy and services. We have a strong commitment to providing independent and objective information and advice in all our work. The AIHW has a long history in providing data, information and reporting on health and welfare issues and as an accredited data integrating authority (see NSS 2015a) has an audited ability to uphold the highest standards in the appropriate handling of sensitive and personal data. This is of particular import in the handling of health data which can be considered to be of a particularly sensitive nature.

The AIHW has a strong relationship with key data suppliers and has established technological, administrative and personnel infrastructures to support data sharing and compilation activities. Examples of this include the construction of National Minimum DataSets (NMDSs) to effectively bring together data from multiple sources on a common basis and the establishment of AIHW's Data Integration Services Centre. Due to its role in reporting on health and welfare data and its experience in bringing together combined datasets source across multiple jurisdictions, AIHW will continue to play a key role in ensuring health and welfare data is made available for appropriate usage in order to realise the gains that can be achieved through effective use of these assets.

Where data is available, the majority of data sharing activities and usage arrangements have proven to be able to be successfully negotiated. In practice however, this can be a time consuming and resource intensive exercise, especially under current Commonwealth

integrating arrangements that result in the destruction of established data sharing infrastructure at the end of a project.

Experiences in data sharing and linkage activities, supported by recommendations from a review performed for the Cross Portfolio Reference Group on Data Integration for Statistical and Research Purposes (Linacre 2015) are now leading towards an evolved data systems approach where a number of key data resources are being either pre-enabled for sharing (for example, AIHW's linked hospitals to MBS trial) or linked into an ongoing resource (for example, the Australian Bureau of Statistics' Multi Agency Data Integration Project (MADIP)).

AIHW advocates a data systems approach as an efficient evolution of current arrangements and mirrors already established infrastructure available across all states and territories. Enhancing data availability and use, based on this approach, recognises that:

- A data systems approach to handling data is needed.

- To manage a data systems approach effectively, data assets need to be enabled for use across the system. This includes preparing the data to a set of standards and pre-processing aspects of the dataset.

- Clarification of guidance around privacy and other legislation is needed. This is of particular relevance for data collected from multiple sources and across jurisdictional borders.

- To maximise the value of investment in data resources and reduce the costs of research enduring arrangements be established so data assets can be retained and re-used.

There are a number of models that could be pursued to enhance data availability and use across the Commonwealth, however the availability of data systems in states and territories and our ability to enhance our current infrastructure to achieve this goal make further evolution of this approach particularly attractive. Under the data systems approach, data could be made more accessible through enhancing data sharing activities and this will raise a number of issues to be addressed as they are experienced across this system. Examples include the following.

- Multiple paths through the Australian Privacy Principles, state and territory legislation and individual Commonwealth agency legislation for data collection, access, use and release, meaning unequal opportunities for different stakeholders and data custodians in the data system.

- An expectation that investment in pre-enabling data to interact productively with an established data system will require ongoing support. Greater investment in metadata and national standards will greatly assist in preparing data to make sharing and linking easier and faster. Infrastructure, such as the open release of the Geocoded National Address File (G-NAF) and AIHW's METadata Online Repository (METeOR) provides a key role here.

- Securing consent for ongoing holding and usage of data specifically to support approved research and a data system which will be used to access data across a number of potentially unspecified uses. Inconsistencies in consent collection, wording and intent across Commonwealth agencies create a large impost on data sharing activities where consent is unclear. Broader, clearer and consistent definition of intended usage of personal information when seeking consent would enable data to be more widely used.

- Identification of high value datasets that should form part of the 'core' data system. This is especially important if some data should be pre-enabled with a higher degree of rigor.

- A data catalogue of enabled data resources. This would complement specification of a core combined, de-identified dataset, such as the MADIP. It is likely to be inefficient to hold all content data together in a single massive resource, so a key step in the data systems approach is to provide pointers (a master linkage key) to pre-link and enable many data resources to be brought together efficiently and effectively for high value approved projects.

- Social licence for sharing, holding and usage of shared data resources. Combined with this is broad governance and usage principles for a data systems approach. This would need to clearly state the basis for holding and use of the data and specify when, how and who can share data into and out of the data system.

- Unlocking public value through a tiered approach, enabling wide-spread access under appropriate protections and provisions with data tailored to be fit for purpose, for example for research and analysis.

Overall data has been able to become increasingly available for use under our current systems, provided some upfront investment in establishing a data system to enable data sharing and use. Using processes that have not been designed for effective re-use of linked data has proven very inefficient in practice and data systems are now being developed for efficient re-use of data linkage and sharing information. Tweaking and supporting current data linkage systems to continue to evolve in order to further enhance data sharing activities would seem to be a natural starting point for developing a comprehensive data system for increasing data availability and use.

The AIHW recognises the impact of recent, rapid technological changes in the ability to collect, transmit and analyse data and welcomes the opportunity to make a submission to the inquiry on data availability and use. AIHW's experience in bringing together data assets and making them available for research and reporting purposes, as well as our knowledge in the handling and linkage of sensitive and personal data, places us in the centre of development of an effective data system for Australia. This submission presents information on current data sharing and linkage arrangements and how this already available infrastructure can be evolved into a functional data system to enhance data availability and use. Additional supporting information on the governance processes and legislative arrangements which would enable AIHW to handle and use data within this system are presented in Appendix 1 and 2.

Increasing data availability through data sharing and the evolution of a data system

Data sharing activities can be classified into 2 broad types, those relying on an underlying transfer of identified data (personal information) and those that only require de-identified data (for example, open data). In practice there are many shades of grey between these 2 categories as high quality linkage using identifiers done using the separation principle (see National Statistical Service Guide for Data Integration Projects) means that the data merging step only relies on transfer of de-identified data. In these cases the pointers used to enable the data linkage have previously been established using the identified data.

In recent years, Australian governments have made substantial investments in building data linkage infrastructure (including establishing the Population Health Research Network– a major facilitator and proponent of data linkage) resulting in the establishment of robust data linkage models that ensure the safety, privacy and security of data used in integration projects. This infrastructure includes:

- state and territory linkage nodes – the Centre for Health Record Linkage (CHeReL, NSW/ACT), the Queensland Centre for Health Data Services (Health LinQ, Qld), SA/NT DataLink, the Tasmanian Data Linkage Unit (TDLU, Tas), Victorian Data Linkages (VDL, Vic), and the Western Australian Data Linkage Branch (WADLB, WA).

- the three current accredited Integrating Authorities (IAs) established under the Commonwealth Governance and Institutional Arrangements for Statistical Data Integration involving Commonwealth Data, are the AIHW, the Australian Institute of Family Studies and the Australian Bureau of Statistics.

Demand for data linkage services is increasing. The AIHW undertook 38 data linkage projects mostly on behalf of the medical research community during 2014–15 and we expect to complete over 60 projects during 2015–16. There has also been increasing levels of expressions of interest and inquiries into data linkage suggesting further growth would be expected. At the same time the sophistication of linkage activities has also been increasing. As enduring linked resources are developed within an evolving data system progressively more complex linkage activities can be expected.

The success of these linkage activities must be tempered by the amount of time it takes to establish each linkage with the biggest hold ups in the system not being the linkage of the data itself, but the data preparation, where data is not pre-enabled, pre-cleaned and pre-standardised and the approvals for release and use of the data. The majority of the linkages performed each year at AIHW are conducted on datasets for which standardised arrangements are already in place, for example the National Death Index and the Australian Cancer Database. The next step in this data systems approach to make the system more flexible and realise substantial speed gains going forward is to retain established infrastructure after each linkage, such as the established linkage keys as ongoing pointers between data sources, so that future work is substantially expedited. A trial to establish arrangements for maintain ongoing linkage keys and point to the state and territory master linkage keys as part of a national data system is currently underway between AIHW, the Commonwealth and New South Wales and Victorian health departments. This project is being conducted under the auspices of the National Health Information and Performance Principal Committee (NHIPPC). This trial will establish fully the arrangements through which an ongoing national data linkage can be maintained driving the next phase in data sharing efficiencies under our current processes.

Barriers (real and perceived) to data sharing and linkage

International (for example, New Zealand's Integrated Data Infrastructure) and Commonwealth (for example, the Australian Bureau of Statistics' (ABS') Multi-Agency Data Integration Project (MADIP)) initiatives have shown there are rapidly decreasing technological impediments to sharing data and bringing together a large-scale evidence base sourced from multiple data custodians. These examples show the technical ability to bring together data but have not addressed issues relating to fragmented data sources split across multiple jurisdictions.

Despite this increasingly accessible infrastructure and repeated recommendations to increase data sharing across the Commonwealth (for example, Management Advisory Committee Report on Connecting Government: Whole of Government Responses to Australia's Priority Challenges. 2004, Australian Government Information Interoperability Framework 2006, National Government Information Sharing Strategy 2009), the level of Commonwealth data sharing today is generally below the level required to appropriately harness the value of Commonwealth data holdings. This is recognised in the recent *Guidance on Data Sharing for Australian Government Entities* (PM&C 2016) which found that it is now increasingly,

> '...important that Australian Government entities streamline data sharing processes to enable greater efficiencies and improve decision making for policy and implementation processes.'

Another consideration in data sharing is the complexity of the relevant legislation. This, and the conservative way in which the legislation has been interpreted, have led to a perception that legislation is a key barrier to information sharing. This is acknowledged in the National Government Information Sharing Strategy (2009),

> 'The complexity of privacy laws often results in the default response to requests for information (that might be considered sensitive) as: "We cannot share our information because of privacy laws." This response is often given instead of determining (through the appropriate channels) whether the information can, in fact, be shared.'

This can be viewed in direct contrast to more recent advice from the OAIC (for example, Pilgrim 2016) and the Victorian Government CIO council's (2013) information management statement on sharing personal data between government agencies

> 'In principle, privacy should not be viewed as an inhibitor to the government's ability to improving service delivery based upon appropriate data sharing arrangements.'

Further considerations that can be perceived as barriers to data sharing include the size and complexity of the combined data resources and even though the technical impediments to handling shared data have been decreasing there may be costs associated with adapting to new and appropriately resourced infrastructure. These types of capacity constraints along with developing capabilities to handle and analyse large and complex shared data resources can present as a substantial barrier to entry for smaller entities wishing to effectively share and use data.

Overall, where data can be appropriately shared under legislation this has not been occurring regularly enough and AIHW's ability to access key datasets has been limited. Despite this the applicability of processes and methods for sharing data under data linkage arrangements have been clearly demonstrated and appropriate infrastructure is now in place that can be supported to enable a more effective data systems approach to improve data availability and use.

Case study: access to multi-jurisdictional data for linkage projects

The process involved in obtaining access to multi-jurisdictional data for linkage purposes can be complex and time-consuming. The AIHW Enhanced Mortality Database (EMD) project provides an example of this. The project involved four data sets, three of which are owned by the various jurisdictions. The AIHW has permission from the Australian Coordinating Registry for deaths data to use the NDI for approved projects, including data linkage. Each of the remaining jurisdictional data sets has a separate custodian located within a separate department in each jurisdiction, from whom permission to access the data must be sought.

Application to data custodians to access data for the original EMD project was made once approval was obtained from the AIHW Ethics Committee. It took a considerable time for access to be obtained for the jurisdictional data sets, as almost all of these also required ethics approval from each jurisdiction's own ethics committee, and in some cases also an Aboriginal health ethics committee, as well as obtaining formal support for 'technical feasibility' from the jurisdictional data linkage centres in some states. Many data custodians required follow-up information before approving access, including making presentations to registrars and other stakeholders regarding the data linkage protocol. Negotiations to obtain permission to use identified data for the linkage were extensive in some cases. Further, some approvals were for a relatively short period of time, and so the delay in obtaining all permissions and relevant data meant that negotiations for extension were occurring in some cases before the linkage was able to be started.

To extend the study a further 5 years using the same data sets and protocols, further data custodian clearance and new ethics approvals were required. These same clearances and new ethics approvals were required again in order to establish an ongoing data collection based on the EMD protocol. In terms of this ongoing collection, maximum time limits on approvals in some jurisdictions require obtaining renewed ethics approval every few years, even where the data items, linkage protocol and purpose of the collection do not change.

While the AIHW fully appreciates the need for clarity, robust methodology and protection of privacy and confidentiality in data linkage and other studies involving identified or identifiable data, any possible streamlining of the process for sharing data between agencies in different jurisdictions and at all levels of government, and in particular the processes for extending existing projects or collections where the data items and protocols have not changed, would greatly assist with undertaking projects aimed at improving the evidence base underpinning policy and planning.

Application of data linkage to improve data availability and use

People may come into contact with a range of services (for example, health, education and welfare) throughout the course of their life. As such, it is important to know the relationships between use of various services and outcomes for the people accessing these services. Effective use of these services would also require analysis of most common and effective pathways travelled through the services and also information about costs of provision against outcomes obtained. This can help to gauge the effect of welfare-related events during these interactions with the service delivery sector and assist in developing effective interventions. By using data linkage, different pathways can be analysed and critical points for intervention determined. Data linkage is a powerful tool for identifying multiple occurrences of individuals within a data set and for linking information across data sets.

Data linkage is a practicable method to address data availability and usage concerns as it make effective use of available data and reduces inefficient duplication of collection. Data linkage works as a powerful tool for identifying multiple appearances of individuals and for linking their information across data sets. This allows datasets, which collect different information about the same individual to be brought together in a secure and appropriately managed way. Where feasible and appropriate, data linkage can reduce data provider and respondent burden and maximise efficiency of data collection.

Privacy protection for ongoing sharing of data via linkage

The protection of the privacy of individuals is very important when linking data sets. There are several aspects to consider throughout a data linkage project to preserve privacy and confidentiality, including legislative and policy frameworks, governance arrangements, any necessary approvals by human research ethics committees, use of identifying and analysis variables, managing risks around dissemination of data, de-identification of personal information, when if/when it is no longer required and data archiving/destruction.

The privacy regime in Australia is complex, particularly where the data supplied involve multiple jurisdictions. The Australian Privacy Principles (APPs) in the *Privacy Act 1988* (the Privacy Act) establish requirements for Commonwealth agencies and organisations when dealing with personal information. Additionally, many Commonwealth agencies are subject to other legislation, policies and established procedures covering aspects of how data related to their functions may be collected, stored, used and/or disclosed.

For example, as well as complying with its obligations under the Privacy Act, the AIHW is also subject to requirements under the *Australian Institute of Health and Welfare Act 1987* (the AIHW Act). The confidentiality provisions in s. 29 of the AIHW Act apply to information obtained and/or held by the AIHW 'concerning another person'. The definition of 'person' under s. 29 of the AIHW Act is broader than that in the Privacy Act as it includes, for example, bodies or associations of persons whether incorporated or not, bodies politic and deceased persons.

To further complicate the data linkage landscape, provision of data by states and territories must be compliant with relevant state and territory privacy legislation and any other legislation relevant to the particular data involved.

The ability to link large data sets is problematical in circumstances where appropriate consent for this activity has not been sought from information subjects in respect of future acquisitions of data. Rules for the waiver of the need for consent do not allow for the granting of waivers 'for convenience', and Human Research Ethics Committees will require

agencies to demonstrate that it is not practicable to seek consent in respect of data not yet collected.

For each data linkage activity, each of the applicable legislative and policy considerations must be addressed as applicable to the particular dataset(s) and data custodian(s) involved. Nonetheless, they do not present insurmountable obstacles to sharing data. In fact, as well as presenting requirements and prohibitions, such legislative and policy regimes acknowledge the value of research and establish usable principles for when this can appropriately occur.

Improving data sharing using linkage—lessons learnt and directions for the future

The success of data sharing activities through data linkage has been limited by standard procedures which destroy linked data resources after a project has been completed. Data linkage arrangements and approvals have also been a limiting factor as these have generally been slow to establish. As has been previously highlighted arrangements challenging these limiting administrative constraints include the national master linkage key work at AIHW and the MADIP at the ABS. States and territories are also progressing enduring arrangements with methods such as state-specific master linkage keys and legislative changes to enable ongoing and regularised use of government data (recent examples of jurisdictions updating their privacy legislation to enhance data availability include NSW). These approaches recognise the positive aspects and the technical feasibility of using existing data linkage infrastructure to enable effective data sharing and are trialling approaches to regularise data sharing arrangements in a data systems approach.

The evolution of aspects of a regularised data systems approach to data linkage to enable data sharing and use have resulted in the development of a number of pieces of supporting infrastructure to manage metadata standards, data transport and consistency and arrangements within a data system. A comprehensive and fully realised national data system to support data availability and use through data sharing activities would likely adopt many elements of this infrastructure.

Enhancing data availability and use

Data sharing activities occupy a wide spectrum, fulfilling a variety of needs and occur on a wide variety of data types. The infrastructure needed will therefore vary depending on the type of data being shared and the purpose to which the shared data resource will be put. For example, in policy development an initial, but approximate result may be appropriate provided this can be produced in a very timely fashion, whereas in an epidemiological examination of cancer clusters, completeness and accuracy of the shared data may be paramount. Ideally a data system to enable data availability and use should be able to meet a variety of needs successfully in a flexible and cost effective manner.

Improving data sharing is a key pillar in making data more effectively available for use. Increased data sharing and visibility will reduce duplication of efforts, allow more contestable and transparent analysis and ultimately improve outcomes for entities and individuals who have contributed their data.

The benefits to improved data sharing will be high. This is clearly recognised in the Australian Government Public Data Policy Statement, which states our data is a key

> *'strategic national resource that holds considerable value for growing the economy, improving service delivery and transforming policy outcomes for the Nation'*

and the Australian Government has committed to the improved use of public data and to release non-sensitive data by default. Our opportunity is now to leverage our current infrastructure to design a data sharing system that will allow us to realise these gains in the most efficient and effective way.

The current data sharing system

It is widely recognised that the current data sharing system which relies on once-off linkages has proven to be slow and cumbersome to the point where it has not been effective at enhancing data sharing activities (see for example, Linacre 2015). It is AIHW's view the current system provides a good starting point and base infrastructure for a more effective system provided arrangements can be established for the maintenance of linked infrastructure and datasets. It is clear that there is little from a technical standpoint that is restricting current data sharing activities; however, as shown above, the key barriers are perceived to be complexity in data sharing legislation, conservative processes that lead to time blow outs in the approvals process, and the resources required to 'enable' data for linkage, including standardisation, understanding, cleansing and development, once approvals are in place.

In response to these 2 areas, there are few shortcuts to enabling high value data resources for sharing, while in comparison modern automated machine learning and semantic techniques for auto-categorisation and cleaning may provide rapid fit for purpose solutions for lower value datasets. For high value data, support can be provided from experts in this area and collaborative examples, such as in the preparation of NMDSs, are particularly effective. Development of a sharing enabled dataset will be an ongoing process so a data sharing system should be flexible enough to incorporate data sources in various stages of enablement. Furthermore where investment in enabling has not occurred this does not preclude a dataset from being shared, it will merely delay its usage, with the advantage the information about its preparation can be fed back to improve its accessibility into the future.

With regards to legislation, conservatism in data sharing and the complexity and inconsistencies in the legislation could be reviewed and consideration given to simplifying clauses related to data sharing activities. The New Zealand *Privacy Act 1993* provides a

useful example for how a minor difference such as explicit recognition of statistical activities in the legislation (Information privacy principle 3 – (4)(f)(ii)) could potentially make a difference to activities such as data sharing. Under the current environment, there is wide recognition of the benefits of data sharing activities so it would be hoped that simplification of the legislation and harmonisation of requirements and legislation across jurisdictions and agencies would reduce inherent caution in data sharing activities by empowering data custodians.

The current data sharing system has developed over a number of years and has provided successful examples demonstrating how data sharing exercises can be conducted. Any changes to the system may therefore need to recognise the built up trust inherent in the current system and consider ways in which a social license for enhanced future activities can be developed.

What could a data system to enhance data sharing look like?

The key step to enhancing data sharing arrangements is to first recognise that at some point personal identifier information must be exchanged to allow records about the same entity or individual in different datasets to be reconciled. If this step is not done or is not required then only de-identified and/or non-sensitive data can be exchanged which can be enacted relatively simply but will achieve a much lower value proposition from the data sharing activity. Furthermore a data sharing system that cannot handle identifiers and link data will restrict many data sharing activities that could be applied to produce high value datasets, such as the MADIP. To realise the maximum value and utility from our data resources a data sharing system must therefore consider how to, and whether it can, appropriately handle identifiers.

Identifier information can flow under a number of different secure methodologies, ranging from privacy preserving linkage methods, under which the identifiers are encrypted (for example, hashed) at source to the integrating authority model in which identifiers are securely transmitted to a third party organisation who has been accredited to be able to securely and appropriately handle personal data under the highest possible standards. Any of these linkage approaches can be used to produce linkage keys, or pointers, which can be placed back on the source data to pre-enable them for linkage and sharing activities. However it must be recognised that privacy preserving approaches may not be fit for all purposes as they will invariably miss a portion of the links that could be found if identifiers are shared. A secondary concern with missing a portion of links is that there is a strong propensity for this missingness to correlate with a specific characteristic of the population, for example married women may be more likely to be missed if maiden name is not available. This can be adjusted for under the integrating authority model where the characteristics of the linkage can be examined, but may never be apparent under a privacy preserving linkage approach.

Under a flexible data system the data sharing activity can adopt a fit for purpose method depending on the needs of the data activity and the purpose to which the data will be applied, however the data system should be designed so that it can facilitate all types of data sharing activities effectively. Along with pre-enabling data to increase speed of availability and use a data systems approach will improve security and probity of data across the data system (see for example, Trutwein et al 2006) this is because the use and transfer of identifiable information is minimised when brought together by a small number of central data linkage agencies. Under a working data system the identifiers flow less frequently, as they are not required to be transferred in every single linkage project but the sharing and availability of data will increase as de-identified data flows are used to bring together combined data resources.

Along with streamlining legislative arrangements for data sharing the 2 pieces of key infrastructure that would need to be developed to enable all effective data sharing activities would then be:

1. National master linkage key – this is the central pointer that is used to pre-enable or pre-link data on the source dataset ready for sharing. A national key would contain pointers to the existing master linkage keys in each state and territory unlocking their already existing data system for the management of state and territory data. Note that a trial is already in place, between NSW, Victoria, Health and AIHW which is developing this key based on MBS and hospitals data.

2. Curation of the enabled source datasets. This could be done either centrally, at source or under a hybrid approach where integrating authorities with experience in data sharing activities can advise and support on the best approach for a specific dataset.

Subsequent development of the data system would then rely on this infrastructure in the gathering of data into enduring linked datasets and any roll-out of resources, applications and software to support this data system.

The above data systems model to enhance data sharing is an evolution of the current Commonwealth data integration approach which leverages off already developed infrastructure as the most efficient way forward. This approach acknowledges the inefficiencies in the current system and recognises that these can be addressed through re-use of data and linkages and streamlining of approvals processes. Much of the work to demonstrate how the current system can evolve and the value that can be achieved is already in the trail phase, with projects such as ABS' MADIP demonstrating how an enduring data resource can be maintained and AIHW's national master linkage key trial demonstrating how to interact effectively with the jurisdictions' data systems. Provided a flexible approach is taken to the development of any data system, approvals processes are streamlined and data sources continue to be enabled in preparation for data sharing activities, this evolved system could be applied to meet our data sharing needs.

Discussion and learnings

Where data is available it may be brought together from a number of sources and subsequently transformed into a more valuable analytically-enabled information resource. The barriers to sharing data and creating these data resources have rarely proved insurmountable in practice, but administrative and cultural issues can create a high impost when attempting to make this happen. Strong examples highlight the potential gains from investing in establishing arrangements for an ongoing cross-agency data system, such as National Minimum DataSets (NMDSs), the Australian Bureau of Statistic's Multi Agency Data Integration Project (MADIP), Expanded Analytical Business Longitudinal Dataset (EABLD) , AIHW's National Death Index (NDI), Australian Cancer Database (ACD) and Specialist Homelessness Services Collection (SHSC). These collections demonstrate how an ongoing data systems approach and investment to pre-'enable' data in readiness for sharing has led to the creation of a number of valuable data assets. The majority of these datasets are all linkage enabled, ie ready to be brought together with other datasets for appropriate purposes in approved projects, and are successfully used for both research and reporting purposes as well as in linkage work.

Key learnings from instigating a data systems approach to sharing data are as follows.

- Investment is necessary to pre-'enable' data resources for sharing efficiently and on a regular basis, particularly with regards to establishing the metadata, common definition and standards. This investment does not have to be onerous; however, provided relevant information can be retained after linkage is performed using each dataset. Essentially this infrastructure preparation can be viewed as a data 'curation' function which can be supported either through a central agency (or a small number of agencies) or done collaboratively through shared resources and recording portals and with open supporting infrastructure and software. For example data can be:
  - geocoded (if possible) to become geospatially enabled,
  - pre-linked (linkage enabled) by incorporating a linkage key or standardising for use in a linkage engine
  - standardised (sharing enabled) so that the data is cleaned and uses common definitions where necessary. Raw source datasets are often not in standardised formats, without standardised metadata, and require substantial preparation for use as analytics resources.
- Agreement and collaboration with data custodians is central to producing an ongoing data resource. Each data custodian must be comfortable with the provisions under which data will be shared and the governance arrangements under which it will be collected, stored and used. High level directives and legislative changes could potentially have a big impact on the way in which these processes occur.
- To pre-enable data sharing activities it must be recognised that identifier data will be transferred at some point between contributing data custodians across the data system in order to link records and produce keys or pointers lining up records belonging to the same individual or entity. This can be done in a number of ways to balance the competing needs of privacy, efficiency and effectiveness. Privacy preserving linkage methods can be applied within this data system, however these will not be fit for all purposes as they do not achieve as high quality results as linkage methods which rely on the transfer of personal identifiers to a secure environment, such as in an accredited integrating authority.
- A flexible approach to data sharing will be critical to developing a useable data system. Different types of data and different uses of data will require different levels of

protection, handling and curation. This continuum could extend from open-by-default and non-sensitive data that is not required to be of high quality on one end. to highly sensitive unit record level data that requires high levels of curation and accuracy at the other. A light touch approach is possible here where each dataset is used differently, in a fit for purpose and resource appropriate way, however for this data system to be flexible enough for all purposes it should also have the capability to handle the most sensitive and highly curated datasets.

- Interoperability issues can be addressed with machine readable formats, data transmission checks and agreed data standards. AIHW has developed the VALIDATA system that enables custodians to develop their data submissions in a consistent way in collaboration with the source agency to address these issues.

- A systematic approach to metadata development allows external comprehension of the data and enables appropriate usage. Where possible this should be made compatible with commensurate data items in other datasets to enable comparability. AIHW has developed the METeOR system that enables collaboration in metadata development.

- Ongoing and regularised usage of data resources builds up key knowledge and skills in their use allowing additional value to be unlocked.

Adopting a full data systems approach to data sharing activities would build on existing infrastructure governance and data linkage experience but may initially require investment decisions by agencies in preparing data for linking and establishing arrangements within an overarching data sharing environment. When a data systems approach is applied it enables more effective and efficient use of high value resources, particularly when they are accessed regularly. In cases where this investment is not warranted the datasets can still be linked or brought together on a once-off basis but additional resources will be required in data preparation and understanding before appropriately reliable outputs can be produced. Efficient data sharing can from this perspective be viewed as a value proposition related to the fitness for purpose of the shared data resource as many questions can be assessed at an approximate level and still provide appropriate results. However where a dataset has been shared, or linked at least once, the intelligence gained through this process should be stored and fed back to prepare and pre-enable the source data for possible future sharing and analysis activities.

The effective re-use of data and the metadata that can be developed through sharing activities is paramount in a data systems approach as the majority of efficiency gains come from principles of:

- preparing data once and then using often, and

- recognising high value data sources and investing in their ongoing development so they are enabled for regularised usage across the data systems at an appropriate level, ie either an authorative and highly reliable source or a quickly accessible source of approximate information.

As high value data becomes more commonly used across a data system, data gaps will become more evident. Simultaneously there will be less requirement for isolated data sources that maintain duplicate information when this can instead by gathered from a central, authorative source.

AIHW has faced many barriers when bringing together datasets, as required, to appropriately monitor and report on health outcomes and, through this process, has found that few of these barriers prove insurmountable given sufficient time and resources. These activities have resulted in some highly valuable work (see for example Mathews et al 2013). Where there are gains from regularised access to high value and reliable datasets, a data systems approach, in which data is pre-enabled in expectation of future sharing activities,

can be employed to enhance ongoing data activities and realise efficiencies. Bringing together data appropriately unlocks value and insights both for program monitoring, research and analysis and ultimately for improving individual outcomes. As a general rule the greater the number of relevant data items that can be employed during analysis the more diverse the insights that can potentially be realised. By way of example, the Pathways In Aged Care (PIAC) project allows monitoring across the aged care systems and has enabled analysis of patient outcomes but further gains could be realised from matching this to expenditure data to monitor efficiencies in the system.

When sharing, using and reporting on health data the provisions of the Privacy Act may be addressed in a different way, often based on *"permitted health situations"* or through section 95A of the Privacy Act, as compared with other types of data sharing activities. While this is an appropriate recognition of the special needs of health data and indicates that we currently have a system that enables appropriate handling of health data, it also is reflective of a system in which a number of inconsistencies and inefficiencies can appear due to multiple paths leading towards the same goal. Where data is available, health data sharing activities have generally been able to occur under current legislation and practices but the challenge has been to increase the speed, flexibility and utility of these processes. There is an opportunity to now further invest in enabling a data system to continue to improve the practice of sharing data and usage across the Commonwealth. In summary, while a substantial re-working of processes for data sharing and usage is possible, an Australian data system has been working, albeit inefficiently and informally, and is now in a position to evolve into an effective data system. There remains much work to be done to realise this vision including.

- improved consistency of data sharing methods, processes and outputs across the Commonwealth

- updates, particularly with regards to harmonisation and streamlining of privacy legislation

- investment in 'enabling' of data – many barriers in bringing together (and usage) are up front from not being prepared

- social license for data sharing activities.

Responses to selected inquiry questions

Questions on high value public sector data

What public sector datasets should be considered high-value data to the: business sector; research sector; academics; or the broader community?

A good starting point to consider in compiling this list is the Australian Bureau of Statistics' 2013 report on Essential Statistical Assets for Australia. This report takes a top down approach, and additional data sources should be added when consideration of specific sectors are included. There has also been substantial further development of data since this time and currently existing linkage enabled data resources should now be considered of much higher value from a researcher and academic perspective than they would have been in 2013. Such datasets that could now be added include: Multi Agency Data Integration Project (MADIP), The Longitudinal Social Security Administrative Dataset (JASON), longitudinal census dataset, Expanded Analytical Business Longitudinal Dataset (EABLD) and MBS/PBS analytic datasets. .

What characteristics define high-value datasets?

Again the 2013 Essential Statistical Assets for Australia, contains a good discussion of the elements for what makes a high value dataset in terms of being able to compile a broad understanding of Australia's development. There are other aspects that make a data resource high value from a data availability and use perspective however including,

1. Uniqueness of content, both in terms of data items, scope/coverage and level of detail/disaggregation.

2. Previous investment in development and enabling for sharing and use (not a sufficient criteria in itself).

3. Utility of content

    a. demand for use

    b. value, or potential value, realised through use of data

    c. usability and usefulness

    d. fit for purpose.

Note that not all high value data is currently available in traditionally accessible formats and datasets. The growth of "data lakes" and other unstructured data archives indicates the potential for high value data to exist outside of established datasets.

What benefits would the community derive from increasing the availability and use of public sector data?

A large portion of the benefits that could be unlocked would currently be unknown as it is hoped that increasing availability of data will drive innovative use. In terms of the uses that we currently know about, increased availability of health data has a very clear value proposition as we already have examples of how research into health outcomes based on public sector data is already substantially improving health services. Some examples of these data linkage projects undertaken by AIHW are included in appendix 2: Examples of data linkages undertaken by the AIHW.

Questions on collection and release of public sector data

What are the main factors currently stopping government agencies from making their data available?

The main factors will depend on the agency and the level at which the data is to be released. When making highly detailed and non-open data available, appropriate data governance regimes, expertise in data management and safe data analysis environments are key factors to ameliorate concerns of possible breaches.

There needs to be substantial investment in enabling data so that agencies are ready to make their data available. Data cleansing and management skills and appropriate funding are the keys to breaking down these barriers.

In some cases, agencies are adopting standards of confidentialisation that exceed what would otherwise be required by compliance with relevant privacy laws. In many instances this can be justified in relation to the special characteristics of particular datasets where there exist high levels of risk. However, frequently, these standards are applied without any real attempt at balancing the levels of risk against the research benefits of releasing finer-grained data.

Other issues:

- the need to obtain multiple HREC approvals for individual projects, resulting in additional costs and time delays to researchers

- a number of recommendations of the Australian Law Reform Commission report *For Your Information: Australian Privacy Law and Practice* (ALRC 108), have yet to be implemented, including several relating, for example, to simplification and streamlining of the Privacy Act, and measures to facilitate health and medical research.

How could governments use their own data collections more efficiently and effectively?

See the main body of this submission for more detail. Data should be pre-enabled for sharing and be ready to be brought together across agencies regularly. There are multiple models for how the data could then be brought together, either on an ongoing basis or as needed (based on pre-enabled linkage keys). The main limitation currently is the effort to establish new arrangements to get data ready for effective use when the data is siloed and no formal arrangements have been established for data sharing.

Should the collection, sharing and release of public sector data be standardised? What would be the benefits and costs of standardising? What would standards that are 'fit for purpose' look like?

Yes, many aspects of collection, sharing and release of data should be standardised; however, to do this fully for all public sector data is neither cost effective nor necessary. The costs involved with standardising all collections to the highest level (for example, linkage enabled, geospatially enabled, complete and standardised metadata, sharing enabled for a data system, transfer enabled and cleaned, standardised data items) would be prohibitive. The benefits of standardising have been discussed in the main text of this submission, which are essentially to prepare the data to be made available quickly and sufficiently fit for the purpose that the data is to be used for. Where this preparation has not happened addressing concerns about the quality of the data and the implications for its use and the technical preparations (and edits/checking) throughout the process of sharing the data can be timing consuming.

Standardising data to enable the highest quality linkage and analysis should be reserved for the highest value datasets as this will incur the highest costs. On a fit-for-purpose basis, lower value datasets may not need this level of standardisation and could, for example, be auto-standardised using semantic and machine learning techniques. To make this workable it may be necessary to choose a small number of levels of governance and standardisation of

quality for a variety of categories of value for specific datasets. Fit-for-purpose standards for low value data may include deterministic, automated (for example, SLK-581 style) and privacy preserving linkage keys, auto categorisation of data items into standard metadata categories and limited quality/probity checking of the microdata.

What specific government initiatives (whether Australian Government, state, territory or local government, or overseas jurisdictions) have been particularly effective in improving data access and use?

There are several initiatives that provide examples, however it should be recognised that the advantages and disadvantages of these approaches may prove different when applied in another context. An example of this is New Zealand's Integrated Data Infrastructure (IDI) which provides some good lessons in the Australian context, particularly with respect to their supporting legislation, but may be less relevant to aspects of our federated system where many health and education data assets are held by the states and territories.

A relevant example for Australian health data is the Health and Social Care Data Integration initiative of NHS National Services Scotland and this follows a similar model to that being applied by all states and territories of Australia across the Population Health Research Network (PHRN). The PHRN has linkage nodes in all jurisdictions across Australia and this initiative has provided an effective data access mechanism across all states and territories. A national master linkage key is currently being developed to support, complement and enhance this infrastructure and extend it to interact with the Commonwealth.

Questions on data linkage

Which datasets, if linked or coordinated across public sector agencies, would be of high value to the community, and how would they be used?

Large scale datasets held in more than 1 agency would be of particular use if linked as this provides common ground for establishing pointers to pre-enable other datasets held in each agency for future sharing activities. The Medicare Benefits Schedule data is an example of a dataset which could fulfil this purpose, however, in practice the maintenance of compatible copies of the data across multiple agencies further highlights the need for a data system to manage data availability and use. Another option would be to create a standardised linkage resource through the accredited integrating authorities containing available pointer information for agencies to use to pre-enable their data resources.

Which rules, regulations or policies create unnecessary or excessive barriers to linking datasets?

There are a number of incompatibilities across agencies in terms of their legislation (or interpretation of legislation) that potentially limit data linkage activities.

How can Australia's government agencies improve their sharing and linking of public sector data? What lessons or examples from overseas should be considered?

As has been stated in the main text of this submission, much of the infrastructure to enable a working data system to increase data availability and use is already in place across Australia. Further evolution and adjustments to make data re-usable and data provision more efficient would build on the infrastructure already in place across all state and territories.

Questions on high value private sector data

What private sector datasets should be considered high-value data to: public policy; researchers and academics; other private sector entities; or the broader community?

Recognising the high sensitivity of some source of health data, private hospital services data, GP records and E-health records would be of particularly high value.

In each case cited, what characteristics define such datasets?

Where the private and public sector each provide competing or complimentary services access to both the private and public data would be required to get an holistic view of the health system, patient outcomes and pathways through the service sector.

What would be the public policy rationale for any associated government intervention?

It is impossible to fully monitor, evaluate and enhance service delivery effectively where competing and complimentary private sector services and the interactions between the sectors remain unknown.

What benefits would the community derive from increasing the availability and use of private sector data?

This very much depends on the data. The absence of private sector hospital services data and primary care data produces a massive gap in the effective analysis of the health system. This data is an essential statistical asset for Australia which if more widely available would enhance medical research and monitoring and evaluation of health services.

Questions on resource costs of access

How should the costs associated with making more public sector data widely available be funded?

The costs of preparing data for public access and the costs of infrastructure to enable access, particularly secure online access, can be significant. Where the data has potential benefits for many different categories of users then the principle of government funding of those costs should apply. However, an agency like the AIHW with a small appropriation has limited capacity to divert resources to invest in infrastructure and preparing large datasets for public use on a routine basis. If the data has benefit for only a limited number of users, particularly commercial users, then cost recovery principles could apply. On the other hand there is also an opportunity for strategic investment in capabilities that support all data users that would reduce potential duplication among entities performing similar functions. The open Geocoded National Address File (G-NAF) is an example of this which supports geospatially enabling data ready for sharing and other examples include AIHW's METadata Online Registry (METeOR) for collaborative production of metadata to sharing-enable dataset.

To what extent are data-related resources in agencies being directed towards dealing with data management and access issues versus data analysis and use?

Regularised use of data resources not only reduces the proportion of time spent on data management as compared with analysis and use but also builds up valuable expertise in appropriate usage that can unlock higher value analytical opportunities. Where there has not been regularised use and provision of data, the data management costs are determined by the level of previous investment in terms of pre-enabling, cleansing and curation of the data. The question of the level of resources directed towards these data management steps depends on the intended usage of the data and the level of fit-for-purpose curation that is required, although in general terms big data will require more substantial resources that single origin smaller datasets.

Is availability of skilled labour an issue in areas such as data science or other data-specific occupations? Is there a role for government in improving the skills base in this area?

The highest level of value from analysis is generally achieved through work conducted through multi-disciplinary teams containing complementary skills. In a broad sense, analysis teams should include a mix of:

- subject matter expert analysts
- data scientists or data management/IT experts
- mathematically oriented analysts and

- policy experts.

Although there is a shortage in skilled data scientists who understand both data management issues and mathematical analysis issues and are able to apply these tools effectively to big data assets, there is also a lack of subject matter analysts with deep experience in a particular area or field of data. While there are already initiatives in place, particularly in the tertiary education sector, to bring through more data scientists, developing partnerships with tertiary education agencies to encourage widespread understanding of data resources at an appropriate level to provide an ongoing supply of expert subject matter analyst is an area where providing access to government data assets could encourage skilled labour development in this area.

Questions on privacy protection

What are the benefits and costs of allowing an individual to request deletion of personal information about them? In what circumstances and for what types of information should this apply?

From a statistical and data analytics perspective, there is a serious risk that allowing individuals to delete their data will bias the dataset and produce misleading findings as the people withdrawing their information will often have different characteristics to the remaining population. This can be considered as related to the issue of withdrawal or with-holding of consent, which can be of particular concern in the collection of reliable statistics about minority groups. While it may be possible to compensate for any unrepresentativeness in the analysis, this will not be possible in all circumstances and more reliable analysis may result from situations where consent to use the data is waived with the dataset being de-identified in a secure, trusted and regulated environment before it is made available to researchers.

Questions on other restrictions

Having regard to current legislation and practice, are further protocols or other measures required to facilitate the disclosure and use of data about individuals while protecting privacy interests? What form should any such protocols or other measures take?

See main text of submission and appendices for further information. In summary the current legislation and practice seems to appropriately meet these requirements and additional protocols would not seem to be warranted. Clearer and consistent guidelines on the application of privacy principles are needed however. It would also be more helpful in this area to streamline and harmonise the legislation in order to make it more accessible.

Questions on data security

Are security measures for public sector data too prescriptive? Do they need to be more flexible to adapt to changing circumstances and technologies?

Although this could be reviewed, this has not proven to be a major impediment to integrating authorities participating in data availability and use activities.

Appendix 1: What does the AIHW do?

The Australian Institute of Health and Welfare (AIHW) is Australia's national agency for information and statistics on Australia's health and welfare. Statistics and data developed by the AIHW are used extensively to inform discussion and policy decisions on health, community services and housing assistance. Under Australia's constitution, health and welfare services are primarily delivered by the states and territories, who are also mainly responsible for the collection of statistics on these services. A fundamental aim of the Institute is to promote consistency among national, state and territory statistics, in order to produce comprehensive national data of the highest standard.

The AIHW is committed to providing high-quality national data and analysis across the health, housing and community services sectors, presented in meaningful and relevant ways and delivered in a timely manner. Accurate statistical information, comprehensive data development and high-quality analyses support an increased understanding of health and welfare issues. This evidence base is critical to good policy-making and effective service delivery, which have a direct impact on the lives of Australians.

We are custodians of several major national health and welfare data collections, and maintain close engagement with our data providers to ensure the quality and integrity of our work. We aim to communicate our data, information and analytical products as widely as possible in accessible formats to key stakeholders and the broader public.

Provided that a project fits with our mission, values and legislation, AIHW can carry out virtually any task related to health and welfare data. This can include surveys and other data collection activities through to developing national data standards and data sets, linking data, analysing it, and writing objective and independent reports that are valued and respected by governments and the community. We have 320 expert statistical, administrative and communications staff in Canberra, backed by numerous data holdings, unmatched knowledge of the health and welfare sectors in Australia, robust governance arrangements, strong privacy and confidentiality controls, state-of-the-art business processes and infrastructure, and access to a network of government, research, education and private sector collaborators from around Australia.

In the interests of independence and transparency, nearly all of AIHW's work is released to the public—while also protecting individual privacy.

AIHW's products and services

AIHW publishes over 180 outputs each year, ranging from comprehensive national reports to technical documents and guides, to innovative web-based products. As required by the Parliament, flagship reports, *Australia's health* and *Australia's welfare*, are published in alternate years.

We have an expanding range of web sub-sites, including for mental health services, specialist homelessness services and Indigenous health and welfare. We are increasingly using interactive data visualisation services on these sites.

Several interactive health and welfare data sets are publicly available on our website, and we have produced 3 apps available for free download from Apple's App Store—*Ozhealth, Ozwelfare* and *Indigenous info.*

Robust and reliable data and evidence

AIHW plays a central role in developing performance indicators for national performance reporting in health, ageing, early childhood, education, housing and homelessness, disability care and closing the gap on Indigenous disadvantage.

In recent years AIHW developed the Specialist Homelessness Services data collection, including an online data collection portal and client management tool.

We are also continuing to work with the Department of Social Services to build and develop the National Aged Care Data Clearinghouse for policymakers, researchers and consumers. In a similar vein, for many years we have operated the National Cancer Statistics Clearinghouse, in collaboration with the Australasian Association of Cancer Registries.

We have also been custodians of the National Hospital Statistics databases for over 25 years.

Summary of AIHW capabilities and experience

• *Expertise* — we are the leading health and welfare statistics agency in Australia, with an enviable track record of over 2,500 health, welfare and housing reports over the last 25 years.

• *Data linkage experience* — we carry out 50–80 linkage projects a year of varying size, complexity and scope. We were the first organisation, for example, to undertake data linkage between hospitals and residential aged care, and between child protection, juvenile justice and homelessness services.

• *Extensive networks* — the AIHW has a strongly collaborative approach to its work, with strong relationships with the Commonwealth, and state and territory governments, universities and other expert groups, national committees, and with the private sector.

• *Access to data* — because of our reputation and our networks we have an enviable record in being able to arrange access to a wide variety of useful government and non-government datasets for data linkage purposes. As both a national 'node' — or linkage centre — within the Population Health Research Network, and as an Integrating Authority, we are uniquely placed to enable safe and secure linkage to data from state and territory databases as well as to Commonwealth data. And we also hold some of the most useful large health, welfare and housing databases ourselves. This can still take time but we work with data custodians to enable safe use of their data for important research.

• *Trust* — our extensive and effective data governance framework helps ensure that we deliver on our mission of providing authoritative information and statistics to Australians while also complying with our legal, regulatory and governance obligations in acquiring, handling and releasing data.

• *Ethics* — we work with an experienced, effective and independent AIHW Ethics Committee with efficient online application and approval processes (all data linkage projects need ethical approval).

• *Value-for-money* — our rates are comparable to our private and public sector competitors, and we have been carrying out projects under contract since we were established in 1987.

Governance of data sharing, collection and release activities

The AIHW is an Australian Government statutory agency established under the *Australian Institute of Health and Welfare Act 1987*. The Act contains very strong confidentiality protections for all data held, and requires the AIHW to publish two key biennial reports in alternate years: *Australia's health* and *Australia's welfare*. Numerous other reports are produced each year, all of which are available free of charge on the AIHW website.

Accountability and independence

As an independent statutory entity AIHW is accountable to the Australian Parliament, and is governed by a Board comprising independent and government members.

The Board is subject to the general oversight of the Minister for Health, but cannot be directed by the Minister unless he/she undertakes a consultative process with the states and territories, as set out in the AIHW Act.

Governance, privacy and confidentiality

In support of data sharing infrastructure and initiatives, data governance arrangements for data sharing are becoming more streamlined. Alongside recent guidance on data sharing activities for Australian Government Entities published by the Department of Prime Minister and Cabinet in April 2016 a simplified committee structure has been established to discuss data issues across Commonwealth agencies. AIHW is currently a member of the Deputy Secretaries Data Group, DSDG, (reporting to the Secretaries Data Group) and also has representation in the Data Champions network. Groups such as the DSDG are available to consider cross-agency data issues that will impact on the availability and use of data, such as in the health, welfare and education spheres.

AIHW manages data professionally, with due respect for its sensitivity, and with privacy and confidentiality assured through legislation, rigorous data policies and procedures, and the scrutiny of a legally-constituted and independent Ethics Committee.

The AIHW is bound by 2 sets of confidentiality and privacy requirements — those contained in the Privacy Act and those contained in section 29 of the AIHW Act.

The Privacy Act establishes obligations on all private and public sector organisations for collecting, using or disclosing personal information.

Section 29 of the AIHW Act places very strong protections on the release of information concerning individuals and organisations. The AIHW cannot be forced to divulge such information, even by a court of law. The AIHW Act also recognises that AIHW must comply with written terms and conditions imposed by data providers.

The public has the right to expect that AIHW will manage data professionally, with due respect for its sensitivity, and with privacy and confidentiality assured.

As an information agency, we rely on strong data governance to retain the trust of our many data providers, data recipients and other stakeholders.

The AIHW's Data Governance Framework is a 24-page document that explains general data governance concepts, and details the AIHW's robust data governance arrangements. A copy of the full Framework is available from the AIHW.

AIHW's Data Governance Framework identifies and provides an overview of the AIHW's robust data governance arrangements, including:
- a description of key concepts in data and data governance;
- the legal, regulatory and governance environment in which AIHW operates;
- core data governance structures and roles;
- an overview of AIHW data-related policies, procedures and guidelines;
- systems and tools supporting data governance; and
- compliance regimes.

These data governance arrangements apply to data: collected and/or enhanced by the AIHW; collected on the AIHW's behalf (for example under collaborative or sub-contractual agreements); and data obtained from all external sources.

As an information agency, AIHW relies upon strong data governance to perform its functions effectively and maintain a trusted reputation amongst its many data providers, data recipients and stakeholders.

This Framework recognises that a combination of supporting legislation, roles, policies, practices and supporting tools and technologies is required to deliver effective data governance arrangements at AIHW.

In summary the data governance arrangements at AIHW help us to achieve our mission of providing authoritative information and statistics about Australians while complying with our legal, regulatory and governance obligations in acquiring, handling and releasing data.

Legislation supporting AIHW activities in data sharing and use

The AIHW Act

The *Australian Institute of Health and Welfare Act 1987* (the AIHW Act) establishes the AIHW as a Commonwealth corporate entity and sets out the Institute's functions, powers and governance arrangements. The main functions of the AIHW are to collect, analyse and disseminate health- and welfare-related information and statistics.

The AIHW Act enables the release of information for public benefit while protecting the identity of individuals and ensuring that data providers can be confident that the AIHW will adhere to data supply terms and conditions.

Section 29 of the AIHW Act contains a confidentiality provision which prohibits the release of information 'concerning a person' held by the AIHW unless one of three specific exceptions applies. These are:

• express written permission to release the information to a specific party has been given by the relevant data provider(s)

• release has been approved by the AIHW Ethics Committee, and such release is consistent with the terms set by the data provider when supplying data to the Institute

• release is in the form of publications containing de-identified statistics, information and conclusions.

The confidentiality requirements in section 29 protect a broader range of information than the Privacy Act, including information about deceased persons.

Privacy Act

The *Privacy Act 1988* creates obligations on Commonwealth agencies and private sector organisations in relation to collecting, using or disclosing personal information. The AIHW is therefore bound by two sets of confidentiality and privacy requirements—those contained in the Privacy Act and those contained in section 29 of the AIHW Act.

Freedom of Information Act

Under the *Freedom of Information Act 1982* (FoI Act), 'documents' held by agencies (which includes information in databases) are subject to access by members of the public unless a legislative exception applies. Section 32 of the FoI Act provides that data protected by the confidentiality provisions of section 29 of the AIHW Act are exempted from release under freedom of information.

Decisions on whether AIHW documents can be released under the FoI Act are made by the AIHW's Freedom of Information Officer in the first instance.

Roles and organisational structures to support AIHW's data governance activities

Board

The AIHW is governed by a Board established under the AIHW Act. The Board has 15 members who have knowledge or experience relevant to the work of the AIHW and who come from a wide range of Commonwealth, State and other organisations. The Board has approved the Data Governance Framework.

Ethics Committee

The AIHW Act establishes an AIHW Ethics Committee whose main functions are to form an opinion on ethical grounds of the acceptability of any activities being conducted by the AIHW, or the release of identifiable data for research purposes. It may impose any conditions it thinks appropriate in performing its functions.

Significantly, the Ethics Committee may authorise:

• the release of personal information for medical research that would otherwise be a breach of a privacy principle in the Privacy Act; and/or

• the release of health-related or welfare-related information as provided for in section 29 of the AIHW Act.

Further information about the AIHW Ethics Committee and its processes, including a list of approved projects is available at www.aihw.gov.au/ethics/.

Data Governance Committee

The Data Governance Committee, comprising 7 senior staff of the Institute, makes recommendations on data governance and data-related matters to the Executive Committee.

Data custodians

A data custodian is a staff member with delegation from the Director to exercise overall responsibility for a specific data collection, in accordance with legislation, policies, guidelines and any specific conditions for use applicable to that data collection. Subject to these requirements, a data custodian has the power to release data to other bodies or persons.

Prescribed security roles

Effective security risk management is required of *all* AIHW personnel and contractors. More specifically, the Australian Government's Protective Security Policy Framework (PSPF) prescribes key security roles for agencies. In AIHW, the Head of the Business and Governance Group is the Security Executive, responsible for agency protective security policy and practices, while the AIHW Facilities Manager is the Agency Security Adviser, responsible for day to day protective security functions. The Unit Head, ICT Operations, is the AIHW's Information Technology Security Adviser responsible for co-ordinating information technology security.

Data linkage

The AIHW is accredited to integrate Commonwealth data for high-risk research projects. The AIHW performs data linkage through its Data Integration Services Centre (DISC), a physically secure area within AIHW that can be accessed solely by authorised, specialist staff. Within the DISC, all data integration projects are conducted on a separate secure network and best practice data protection methods are employed.

Policies, guidelines and procedures

The AIHW's internal data policies, guidelines and procedures are designed to ensure that all staff, and especially those with delegated authority to make data-related decisions, have clear sources of information to guide them in performing their roles effectively and appropriately. These documents cover the acquisition, management, linkage, security, confidentiality, quality, publication and retention of data.

Systems and tools supporting data sharing activities

ICT systems

The AIHW's ICT systems support secure and auditable data governance processes. In particular, access to data is separately restricted at the network, data server and database level, requiring individual authorisation.

AIHW data catalogue

The AIHW's Data Catalogue is the official listing of all AIHW data collections. It:

• identifies past and present data custodians for each AIHW data collection; and

• describes each AIHW data collection, including its scope, format, period of coverage, sub-collections, availability for research, links to relevant publications, whether the collection contains identifiable data, and related datasets (if any).

A limited public version of the *Data catalogue* is available on the AIHW website.

METeOR

METeOR, which is managed by the AIHW, is Australia's registry of national metadata standards for the health, community services and housing assistance sectors. Through METeOR users can find, view and download over 2,600 data standards.

These standards provide a structured description of the characteristics of data. METeOR allows free online access to a wide range of nationally endorsed data definitions, and tools for creating new definitions based on existing already-endorsed components.

Validata™

The AIHW works with its data providers to maximise the currency and quality of data collections. The AIHW's online data receipt and validation product, Validata™, has been designed to improve the quality and timeliness of data supplied by jurisdictions and non-government organisations. It enables data providers to quickly check and validate the quality of their data submissions against a set of validation rules.

Compliance and monitoring of appropriate data usage

The AIHW Ethics Committee requires regular monitoring of the progress of projects it has approved and audits of particularly sensitive data collections. The audits occur as part of the internal audit program, with outcomes reported to the Board through the Audit and Finance Committee.

Data collections held by AIHW may also be subject to audit by data providers (for example, under conditions specified in data supply agreements) and by statutory office holders such as the Auditor-General and the Privacy Commissioner.

The AIHW Act and Australian Public Service Code of Conduct require staff to be diligent in preventing breaches of information security. A breach of the confidentiality requirements of

section 29 of the AIHW Act constitutes an offence punishable by law, including imprisonment.

Appendix 2: Data linkage at the AIHW

The AIHW is an accredited Commonwealth Integrating Authority with a strong record and capacity for data linkage. The operation of the Data Integration Services Centre provides infrastructure and supports governance for the AIHW to deliver an accessible service facilitating cost effective health research. The centre maintains AIHW's standing as an accredited Integrating Authority by ensuring compliance with the newer and more stringent security and governance arrangements required of agencies wishing to undertake data linkage involving Commonwealth data where the project is classified as high risk. The centre also manages AIHW's membership of the Population Health Research Network.

The AIHW undertakes a wide range of data linkage work for researchers, both internal and external to the AIHW. This work supports a variety of innovative analyses about the health and welfare of Australians, such as longitudinal analysis and movements between health and welfare services. This work thereby acts to increase the value of individual data sets.

Data security for linked data

To secure accreditation as one of only three Commonwealth Integrating Authorities AIHW has met extremely stringent criteria covering project governance, capability, data management, security, and the protection of privacy and confidentiality.

A few of the many measures involved include: use of the 'separation principle' to ensure privacy — where personal information is stored in separate files from transactional information; a physically secure area with access available only to data integration staff; a separate secure network; all access logged; integrated data confidentialised; and data encrypted as part of the archiving process.

In carrying out its data linkage activities the AIHW abides by the National Statistical Service (NSS) *High level principles for data integration involving Commonwealth data for statistical and research purposes* and *Best practice guidelines*.

Consistency of data enhances/enables linkage and sharing activities

The consistency with which linkage variables are reported across the data sets affects how accurately linkage processes can match records for the same person across collections. Inconsistencies can result from simple reporting or processing errors (for example, in date of birth and sex). However, some variables can be expected to change over time. Three areas that have been identified as likely to impact on data linkage between education and training collections include:

• change of residential address

• consistency of reported name

• stability of person identifiers.

Examples of data linkages undertaken by the AIHW

At the AIHW, data linkage projects come in all sizes, grades of complexity and scope, on a very wide variety of topics related to the health and wellbeing of Australians.

A 'routine' health data linkage project might involve matching people who received treatment 'X' with subsequent hospital attendances or how long they lived after treatment.

At the other end of the spectrum it could be possible to link various health, welfare, education and income and support payments datasets to show how to get better outcomes and value from health and welfare programs and expenditure.

Examples of projects under way or completed at AIHW include:

• *Pathways in Aged Care (PIAC)* — this linked dataset covers aged care assessments and use of 7 different Commonwealth aged care service programs from 2002 to 2011, as well as deaths. The study showed that even where people are eligible for residential aged care they prefer to remain in their homes for as long as they can.

• *Cancer risk in people exposed to computed tomography scans* — Medical Benefits Schedule data on 680,000 CT scans were linked to the Australian Cancer Database to examine if there was increased incidence of cancer after CT scan exposure. The study showed exposure to CT scans in childhood increased the incidence of cancer.

• *Impact of a population-based HPV vaccination program on cervical abnormalities* — this collaborative study between the Australian Institute of Health and Welfare (AIHW) and the Victorian Cytology Service (VCS) linked the National HPV Vaccination Program Register with Victoria's Pap test Register. The study was the first in the world to show a population-based HPV vaccination program resulting in a fall in cervical abnormalities within 5 years of implementation.

• *Dementia care in hospitals* — this study, commissioned by Alzheimer's Australia, used linked data to investigate the experience of people with dementia in the New South Wales hospital system. It showed that people with dementia stay in hospital longer than other patients and have higher costs of care, and that the condition is often not recorded in patient records.

• *Diabetes Care Project* — this 3-year pilot study was conducted by McKinsey and Company, with AIHW being contracted to coordinate the data linkage components across a range of health data sources. The study evaluated whether new models of care deliver better-quality outcomes for people with diabetes.

• *Homelessness, income support and employment pathways* — we are undertaking a project funded by the Department of Social Services to link Specialist Homelessness Services client data with income support and employment services data from the Commonwealth Departments of Human Services and Employment. The study will to identify the service delivery patterns and pathways used by clients.

• *Child Protection data and educational achievement* — this project involves linking Child Protection data with educational achievement data as measured by NAPLAN (National Assessment Program — Literacy and Numeracy) testing.

• *Child immunisation and health* — two projects we are undertaking in 2014–15 involve linking child immunisation and health outcomes data.

References

AIHW 2014. Australia's health 2014. Australia's health 2014. Canberra: AIHW. http://www.aihw.gov.au/australias-health/2014/

Australian Government Information Interoperability Framework 2006.

Australian Bureau of Statistics 2013, *Essential Statistical Assets for Australia* Cat no. 1395.0.

Cate, Fred H.; Cullen, Peter; and Mayer-Schonberger, Viktor, "Data Protection Principles for the 21st Century" (2013). *Books by Maurer Faculty*. Book 23.

Department of the Prime Minister and Cabinet 2016. Guidance on Data Sharing for Australian Government Entities. https://www.dpmc.gov.au/resource-centre/public-data/guidance-data-sharing-australian-government-entities

Department of the Prime Minister and Cabinet 2015. Australian Government Public Data Policy Statement. https://www.dpmc.gov.au/sites/default/files/publications/aust_govt_public_data_policy_statement_1.pdf

Flowerdew, R. Manley, D. Sabel, C. 2008 *Neighbourhood effects on health: Does it matter where you draw the boundaries?* Social Science and Medicine 66 1241–1255.

Holman, C.D., Bass, A.J., Rosman, D.L., Smith, M.B., Semmens, J.B., Glasson, E.J., Brook, E.L., Trutwein, B., Rouse, I.L., Watson, C.R., de Klerk, N.H., Stanley, F.J. 2008, 'A decade of data linkage in Western Australia: strategic design, applications and benefits of the WA data linkage system', Australian Health Review, 32, 4, pp. 766–777.

Linacre, S. 2015 *Interim report on review of Commonwealth arrangements for data integration for statistical and research purposes: Feb 2015.* Background paper for the cross portfolio reference group on data integration for statistical and research purposes.

Management Advisory Committee Report on Connecting Government: Whole of Government Responses to Australia's Priority Challenges. 2004.

Mathews JD, Forsythe AV, Brady Z, et al. *Cancer risk in 680 000 people exposed to computed tomography scans in childhood or adolescence: data linkage study of 11 million Australians.* BMJ 2013;346:f2360.

National Government Information Sharing Strategy 2009.

National Statistical Service 2015a. *Accredited Integrating Authorities.* http://www.nss.gov.au/nss/home.nsf/pages/Data%20Integration:%20Accredited%20Integrating%20Authorities

National Statistical Service 2015b. *Statistical Data Integration involving Commonwealth Data* http://www.nss.gov.au/nss/home.NSF/pages/Data+Integration+Landing%20Page?OpenDocument

National Statistical Service 2015c. *The separation principle.* http://statistical-data-integration.govspace.gov.au/topics/applying-the-separation-principle/

Parenteau, M. Sawada, M. 2011. *The modifiable areal unit problem (MAUP) in the relationship between exposure to NO_2 and respiratory health.* International Journal of Health Geographics 10:58.

Pilgrim T 2016. Privacy, Data and De-identification. https://www.oaic.gov.au/media-and-speeches/speeches/privacy-data-de-identification

Raghupathi and Raghupathi *Big data analytics in healthcare: promise and potential Health Information Science and Systems* 2014, 2:3.

Trutwein, B., Holman, C.D.J., Rosman, D. 2006, 'Health Data Linkage Conserves Privacy in a Research-Rich Environment', Annals of Epidemiology, 16, 4, pp. 279–280.