Horizon Scanning Series
The Future of Precision Medicine in Australia

Data

This input paper was prepared by Adrian Turner with contributions from Cheryl George, Bill Simpson-Young, Dr Stephen Hardy, Dr Chelle Nic Raghnaill and Jane Polak Scowcroft (Data61).

Suggested Citation

Turner, A, George, C, Simpson-Young, B, Hardy, S, Raghnaill, C, Scowcroft, J (2017). The Future of Precision Medicine in Australia: Data. Input paper for the Horizon Scanning Project "The Future of Precision Medicine in Australia" on behalf of the Australian Council of Learned Academies, www.acola.org.au.

Data

This paper was prepared by Adrian Turner with contributions from Cheryl George, Bill Simpson-Young, Dr Stephen Hardy, Dr Chelle Nic Raghnaill and Jane Polak Scowcroft of Data61.

1. Introduction

Advances in data-driven research and technology development in areas including processing, informatics and artificial intelligence hold the potential to unlock new opportunities for the development of personalised health solutions, targeted to the specific care – or genetic – needs of the individual.

Precision medicine aims to create targeted therapies for individuals based on unique factors such as their genetic makeup, environment and lifestyle. Developments in our ability to rapidly collect, analyse and safely share data between individuals and organisations - without compromising individual privacy - will undoubtedly support the development of such target therapies, by enabling health professionals access to cross-system and population patient outcomes information. For policy makers, access to a broader base of information will also drive new insight into ways to improve patient outcomes, prevent and treat disease and anticipate future health service needs.

The impact of advances in data processing power is most evident in areas including genomics, medical imaging, and point-of-care diagnostics. For example, where the first successful mapping of a human genome – the "Human Genome Project" completed in 2003 - took 13 years and cost around US$3 billion to complete, a similar result is now possible in under 48 hours, for less than US$1,000[1].

In parallel, developments in materials science and sensor technologies are also driving a new wave of customized health, providing clinicians and patients with access to personalised data and diagnostics on the spot, at the point of care. In 2014, health and wellness monitoring applications accounted for 66.3% of biosensor revenue globally, demonstrating the rise in the use of sensors as a data generation and diagnostics tool in healthcare[2].

In order to realise this potential, advances need to be made in the areas of data integrity and standards, data sharing and interoperability, data security and privacy, and in skills development to support a new approach to healthcare. This chapter explores how some of those aspects offer new opportunity for the precision medicine and health sector, and some of the emerging technologies or approaches being explored.

2. Data integrity and standards

A key factor impacting the usability of health related data is inconsistency. This is both in terms of the way information is collected and recorded - ranging from verbal and paper-based records to sensor networks - and the resulting range in its level of quality or reliability.

In order for a dataset to hold value, in terms of its potential to contribute to new health insights, we need to be able to trust in the integrity of the data, meaning we need to be sure that appropriate quality controls and processes – such as ethics and consent – are in place when it was collected and that the methodology of collection is well documented and available to users of the data.

[1] Nature 2014, *Technology: The $1,000 Genome*, online https://www.nature.com/news/technology-the-1-000-genome-1.14901, Hayden E C, accessed 9 August 2017.

[2] Commonwealth Scientific and Industrial Research Organisation 2017, *Medical Technology and Pharmaceuticals: A Roadmap for unlocking future growth opportunities for Australia*, https://www.csiro.au/en/Do-business/Futures/Reports/Medical-Technologies-and-Pharmaceuticals-Roadmap, pg 16

2.1. Data Integrity

Setting in place accepted and common standards to ensure the integrity of health data will act to accelerate the potential for data sharing and linkage, in turn offering opportunity for the development of new treatments, technologies and predictive systems, targeting individual and system-wide needs[3].

The use of common metadata registries, such as those conforming with ISO11179, will facilitate the accurate capture and management of descriptive and structural health metadata (including assumptions and methodologies used in data capture) will aid more precise data combination and linkage, reuse of data and its governance.

2.2. Data Standards

Organisations within Australia's health and medical sector currently adhere to an evolving variety of data management standards and requirements, including those set by governments – in the form of commonwealth and state-based legislation – and domain-related common standards.

Key governing bodies, including Australian Medical Association (AMA) and National Health and Medical Research Council (NHMRC), are active in setting best practice standards and the Office of the Australian Information Commissioner (OAIC)'s health and digital health guidelines and fact sheets provide ready guidance on the application and implications of the Privacy Act 1988 to health information, and initiatives[4].

These standards also align with best practice standards being issued by domain experts, such as the Global Alliance for Genomics and Health's *Framework for Responsible Sharing of Genomic and Health*-Related Data and International Cancer Genome Consortium's (ICGC) global policies for good research practice[5].

Achieving greater unification in standards across Australia, particularly in relation to privacy will greatly support organisations in ensuring compliance, and enable the development of cross-jurisdictional platforms and shared approaches.

One way to achieve this, is to simplify the action of interpreting and mapping the applicability of regulation to an activity. For example, Data61 and the Department of Industry are developing **Regulation-as-a-Platform**. This offers strong potential to simplify this process, by enabling practitioners to readily review the applicability of regulations and standards in place, to their activity. This is achieved by converting legislation into a form of machine readable logic, which means that it can be reviewed and interpreted rapidly, enabling a common view of all applicable rules and standards, potentially via a single platform.

This approach could also be utilised to provide feedback on compliance states, including when regulations are altered or updated.

2.3. Systems Interoperability

Finding ways to make health information systems more interoperable, that is enabling seamless digital records across all care settings, based on open standards will support access to more data, and provide policy makers, professional and individuals with new health insights.

For Governments, or policy makers, being able to access information stemming from a variety or interoperable sources will support use of data for insight into public health measures, medical outcomes and care delivery outcomes and trends.

---

[3] Commonwealth Scientific and Industrial Research Organisation 2017, *Medical Technology and Pharmaceuticals: A Roadmap for unlocking future growth opportunities for Australia*, https://www.csiro.au/en/Do-business/Futures/Reports/Medical-Technologies-and-Pharmaceuticals-Roadmap, pg 16

[4] OAIC 2017, Health and Digital Health Fact Sheets, online https://www.oaic.gov.au/individuals/privacy-fact-sheets/health-and-digital-health/
(accessed 11 August 2017)

[5] NHMRC 2017, *NHMRC Statement on Data Sharing*, online (accessed 11 August 2017), https://www.nhmrc.gov.au/grants-funding/policy/nhmrc-statement-data-sharing

For the medical practitioner, being able to seamlessly interface and/or share information with other organisation within the health and medical support chain, or with individual patients would alleviate a number of current challenges, including delay, duplication of activity and information gaps. And for individual patients, being able to receive and share data on their own health progress and outcomes would enable a more personalized level of care.

In all of these cases, greater systems interoperability, with strong data privacy and system security at all points, sets the foundation for enabling data to be shared and accessed.

3. Data Sharing

The Productivity Commission's *Inquiry Report into Data Availability and Use*, released in March 2017, highlighted Australia's health and medical sectors as a prominent example of where the opportunity presented by data is not being fully realised, due to impediments (including legal, technical, cross-jurisdiction issues, etc) and mistrust[6].

In an effort to address this, the Inquiry recommended establishing a new National Data Custodian (NDC) body, responsible for identifying and designating particular data sets as being of national interest, with a view to mandating their release.

The Inquiry also suggested the establishment of Accredited Release Authorities, tasked with ensuring datasets marked for release in particular domains – such as health – are made available with the right level of sharing restriction (open or secure) and standard in place (metadata, API).

These developments offer strong potential to drive a significant shift in the availability of health datasets, however before they can be realised, systems to enable the secure sharing of health data are needed. These will have different settings depending on the perceived sensitivity of the data, level of private or confidential content they contain, and need to be able to manage a variety of scenarios, from sharing between health providers and patients, to sharing across operators in the health system.

When sharing is realised, the potential for new insight into health trends and outcomes, and opportunity for precision, or personalised treatment and monitoring will be staggering, while still preserving patients' privacy.

3.1. Approaches to Data Sharing

The process of sharing and publishing health data involves many aspects. Figure 1 (below) visualises the core steps currently involved in bringing data together - which includes separate processes of data extraction, joining (through linking or integration), protection, access and use – in order to utilise data to achieve outcomes (such as innovation, improved care policy, or cost savings).

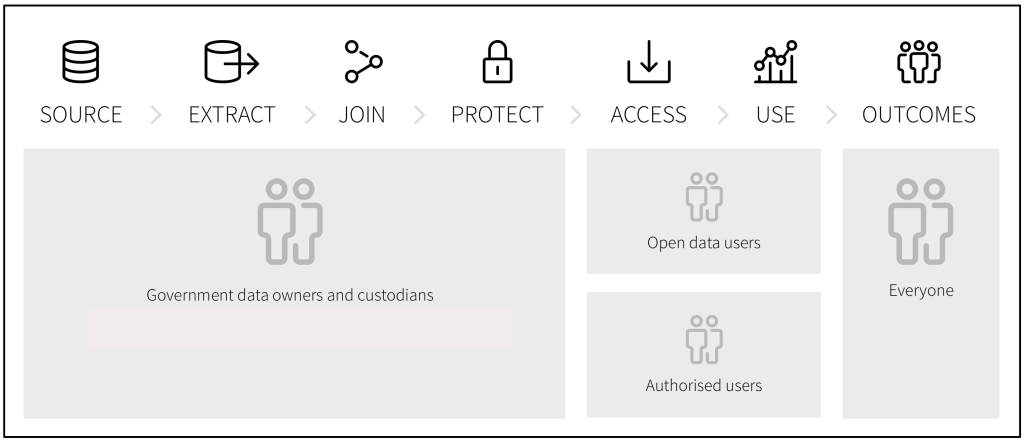

Figure 1: Steps in data publishing

[6] Productivity Commission 2017, *Inquiry Report into Data Availability and Use,* online, http://www.pc.gov.au/inquiries/completed/data-access/report/data-access-overview.pdf (accessed 9 August 2017)

Data sharing or integration involves bringing together data held within separate organisations or sources, in order to provide a unified view and/or access to the combined pool. There are three generalised approaches currently taken to data sharing and integration[7]:

- **Point-to-point:** where data is exchanged ad-hoc and periodically between organisations, with little consistency in standards, format or quality across datasets. This is the dominant data exchange pattern for most Australian government and health organisations at present.
- **Centralised:** where data is collected and/or aggregated by a central intermediary that has responsibility for transforming and/or analysing it, providing a uniform interface for users of the data. Examples include the Australian Bureau of Statistics.
- **Federated:** where data is exchanged in a co-ordinated way between agencies, using agreed standards and/or shared platforms able to process and transform data (e.g. API gateways, data linkage, on-the-fly virtual dataset generation, etc.). Under this model, source data remains with its original custodian, with no persistent data stored on shared platforms, or with an intermediary. This is an emerging model in use globally, and by Australian Government agencies. Examples include ATO's Standard Business Reporting (SBR)[8] platform and Australian Government's NationalMap[9] federated spatial visualisation platform.

Data61 is currently working with the Department of Prime Minister and Cabinet on a project to improve the searchability, quality, indexing and discoverability of available datasets. The software developed in the project - known as **MAGDA (Making Australian Government Data Available)** – supports better ways for locating and accessing data from across the country and this data can be used together with personal data for more targeted health analytics.

3.2. Data-driven insights

The opportunities for machine learning to support rapid and personalised predictions will continue to grow as more data becomes available. These technologies will enable better diagnoses, suggest new treatment options, and ultimately lead to an improved understanding, at the individual and population level, of the factors that contribute to disease and have impacts on prevention.

The ability to draw on a growing field of data sources will allow machine learning to generate insights that would not be apparent from a single data set in isolation. This data can come from anywhere, not just from formal sources such as private patient information such as genetic profiling, test data and patient records, but also from more informal data sources, such as fitness trackers and other personal devices.

Data-driven insights will enable practitioners to more effectively utilise their expertise by supporting their decision-making, helping them achieve better patient outcomes. Machine learning algorithms will bring the most relevant knowledge from the vast corpus of medical research to the practitioner's fingertips, and will be able to place a patient's individual responses in the context of the broader patient population in real-time.

Ultimately, machine learning technologies will not just process data but actively collect it to improve their performance. They will understand when their predictions are uncertain and what additional data will reduce that uncertainty. They will not only suggest tests for a particular patient, but also understand what other tests will improve their own predictive performance in the future, so that they can continue to learn and improve patient outcomes.

---

[7] NICTA 2015, *Enabling Business to Government Digital Interaction: A Report to the Australian Government*, prepared for the Australian Taxation office, June 2015

[8] http://www.sbr.gov.au

[9] https://nationalmap.gov.au

3.3. Building on our research strengths

Data-driven research work in precision health in Australia is already well underway.

In March 2017, the Garvan Institute of Medical Research and the Centre for Pattern Recognition and Data Analytics (PRaDA) at Deakin University launched a new investigative analysis into patterns in human genomic data. Utilising genomic data from around 14000 full genomes, the *Garvan-Deakin Program in Advanced Genomic Investigation (PAGI)* is a step in exploring the use of AI and machine learning to gain insight from mapping an individual's genomic information with large clinical datasets[10].

A further example of the potential for precision medicine from data-driven analysis is a branch in the area of pharmacogenomics, which examines how an individual's genetic makeup affects a person's response to particular drugs. Technology will soon make it possible for a practitioner to compare the outcome information gathered across a large group of individual patients, with an individual's own genes, to determine the most effective disease treatment pathway.

4. Data Security and privacy

Ensuring the privacy and confidentiality of individual's health information remains intact, at all stages of engagement with the healthcare system, must remain a central consideration in the design of any health data sharing or analytics undertaking.

To date, concern regarding the risks associated with the holding or sharing of private or sensitive data, including risks of data breach or loss, have proven barriers to organisations exploring data sharing. Similarly, for individuals, the fear of misuse of personal data, or access being provided to third parties without consent means that most people are wary of their data being collected or shared.

The key to changing our willingness to share data, is to build trust. That includes trust in the systems and architectures which support the sharing of our data, and also trust in the existence of suitable and robust standards to ensure professional practice, in the handling and use of our data.

4.1. Building trustworthy systems

A number of methods have developed to quantify confidentiality risks that arise from sharing and releasing of private and confidential data. The most general and mathematically formal method of disclosure risk assessment is based on probabilities of re-identification, that is, the level of probability that a third party could learn information about an individual from released data and a set of assumptions about the third party's behaviour. A number of different methods are currently used to alter or perturb data before release. These include:

- **Suppression / Masking** – where sensitive values are masked or removed before release;
- **Aggregation -** where data is expressed in summary form and so reduces disclosure risks;
- **Data swapping –** where data values for selected records are swapped, to discourage users from matching, since matches may be based on incorrect data;
- **Perturbation or noise -** protecting numerical data by adding random data or noise to datasets;
- **Synthetic data -** replacing original data values with values simulated from probability distributions. These distributions reproduce as many of the relationships in the original data as possible.

Each of these approaches has its limitations, and all methods in some way limit the level of insight that could be derived from the raw, detailed data being analysed.

[10] Deakin University 2017, *Garvan and Deakin University join forces to accelerate precision medicine through machine learning, Woodhams E S, online http://www.deakin.edu.au/about-deakin/media-releases/articles/garvan-and-deakin-university-join-forces-to-accelerate-precision-medicine-through-machine-learning, accessed 9 August, 2017.*

4.2. Preserving Privacy and Confidentiality

Techniques are emerging which hold promise to enable confidential data to be accessed for insight, without putting its content at risk. For example, new federated, privacy-preserving analytics methods developed by Data61 – known as **Confidential Computing** – enable encrypted queries and responses to be sent and received by third party datasets, without requiring raw data to be shared or exposed.

The technique combines three underlying technologies - distributed machine learning, homomorphic encryption and secure multiparty computing - to provide a platform to draw insight across organisations, or from individuals, without requiring direct access to raw data.

As an example of the potential applications for this capability to precision medicine, there is potential that confidential computing could enable an individual - holding all of their own health information on their own computer or smartphone (such as their genome or personal health record) - to securely map their data against a third party database or health provider's system, and receive personalized results back, without their private data ever being disclosed to another party.

5. Skills

Australia already has a depth in capability in medical and health related research, including in fields relevant to precision medicine, particularly in genomics.

Examples include the Australian Genome Referencing Facility, with the Melbourne branch based at the Walter and Eliza Hall Institute, and the Kinghorn Centre for Clinical Genomics at the Garvan Institute. In 2015 the National Health and Medical Research Council also allocated $25 million to establishing the *Australian Genomic Health Alliance*, a national network of 47 partner organisations including research institutes, hospitals and universities[11].

For Australia to take best advantage of its medical research strengths and emerging opportunities for data-driven technologies, measures will need to be taken to ensure availability of the required workforce data skills to support this work.

6. Data Architecture and Infrastructure

The Australian Government seems to be moving towards adopting a more federated approach to data sharing and management, enabling coordination and accessibility, but where control of the raw data continues to reside with its custodian organisation.

These models are attractive as they ensure the most current, available data is used at all times, and that data always resides with the group best aware of its acquisition and context, confidentiality, privacy and limitations. As this trend continues, and more organisations – both within and outside of Government – adopt federated models, the potential for cross-organisation, privacy preserving data sharing and analytics improves.

Supported through the National Innovation and Science Agenda (NISA), Australian Government agencies and CSIRO Data61 are currently collaborating on a number of projects to test techniques for allowing trusted access to high-value, Government datasets, whilst preserving the data's confidentiality and integrity.

Known as the **Platforms for Open Data** (PFOD) initiative and initiated in 2016, the projects underway apply a mix of existing and new techniques to provide cross organisation sharing, in a federated, but secure environment.

For example, one project in the initiative allows data platforms to interactively access aggregated data which is confidentialised on-the-fly from sensitive unit record datasets. The prototype API is being developed

[11] Australian Financial Review 2017, *Australia will lead world in medical technology, says Bill Ferris,* Redrup Y, published 16 March 2017, online http://www.afr.com/technology/australia-will-lead-world-in-medical-technology-says-bill-ferris-20170314-guy969 (accessed 10 August)

initially to provide secure access to the Multi-Agency Data Integration Project (MADIP) dataset but can be used with many other sensitive data sets that are otherwise difficult to access.

Another exciting new area also offering potential for application in a precision medicine setting is **blockchain** technologies. Blockchain has been widely celebrated as a potential mechanism for generating and supporting distributed trust on the internet, where trust is difficult to establish, by providing a common register and platform to support information audit and rapid consensus[12].

These technologies offer strong potential to enable data to be shared and agreement to be reached across operators in the health system, generating a common audit trail of interactions, approvals and reviews.

7. Data ecology

A further and significant barrier to achieving new insight to support precision medicine is the difficulty in understanding what data is being generated or is available across the health system. This gap in knowledge not only means that health data collection is duplicated across the system, but also that valuable data is not being accessed or utilised to support decision making or to support cross-population insights.

Using data federation techniques to develop sharing platforms, to keep a running catalogue of generated and available data, along with the use of a combination of machine learning and AI to make datasets more discoverable and searchable, holds promise to resolve this.

This ecology could be mapped to a particular disease domain, both to inform programmatic activities as well as to identify where new data sources/proxies are needed.

Such a resource could also potentially support the observation of health outcomes data in parallel to other, non health focused datasets, to explore correlation. For example, the intersections between health outcomes and exogenous factors such as conflict, migration, natural disasters or climate change, etc., could point to the impact external factors may have on health[13].

[12] Commonwealth Scientific and Industrial Research Organisation 2017, *Distributed Ledgers: Scenarios for the Australian economy over the coming decades,* Hanson RT, Reeson A, Staples M.

[13] World Bank 2017, *Big Data in Action workshop: 2017 World Government Summit Report*, online: http://documents.worldbank.org/curated/en/750841491276077629/pdf/114010-WGSBigDataWorkshopv.pdf, accessed 10 August 2017.