MONASH University

Helix

DATA GOVERNANCE FRAMEWORK

DIANNE BROWN

NOVEMBER 2018

STEERING COMMITTEE SUBMISSION v1.2a

INTRODUCTION

To be a world leader in health data research, Monash University (Monash) has established Helix – a Health Data Platform – to support all health, epidemiological and translational research to:

- manage health data;
- collaborate over health data; and
- enable the application of advanced computing and analysis techniques and informatics for processing and analysing these often complex data sets.

This document articulates the Data Governance Framework of Helix and outlines how health data activities will be undertaken. Once endorsed by the Helix Steering Committee, it will not only underpin all policies, procedures and best practice recommendations but also provide the tools needed for researchers to manage their health data.

CONTENTS

1 CONCEPT OF A DATA GOVERNANCE FRAMEWORK

Monash undertakes a range of activities involving health data across a number of faculties and schools. Health data is considered sensitive personal data and requires high levels of governance covering a number of legal and ethical considerations.

Data Governance refers to the overall management of these activities and includes the articulation of the:

- Data Governing Body – the body entrusted to govern these data activities. Usually a steering committee, board or research committee.
- Data Governance Framework – a structured and well-defined description of these data activities upon which policy and procedures are built.

A number of strategic and operational benefits can flow from the development of a Data Governance Framework.

Figure 1: Strategic and Operational Benefits of Data Governance Framework

Firstly, the **strategic direction** of all data activities can be set through identifying where "value" lies along the Data Value Chain (see section 2). Monash will be able to assess where it has (or should have) a competitive advantage and if it has the **capabilities** needed to derive this value, given the current environment and expected long-term dynamics that will shape the future.

Resource allocation decisions can then be systematically made to ensure resources flow to where most value and opportunity lies.

Operationally the Data Governance Framework will be the foundation of the **policies** and **procedures** that will systematically identify best practice in all health data activities. From this **tools** and "how to" guides (both written guides and human guides) can be developed that support researchers as they undertake these activities and guard against repeating previous mistakes. This leveraging of accumulated corporate knowledge across Monash will be crucial in enabling it to become a world leader in health research.

Finally, and perhaps most importantly, the Data Governance Framework will provide a **common language** that can be used across the university and amongst external stakeholders to effectively communicate about these activities. Whether it is a strategic review or the building of an IT system, a common language will be used by everyone – from IT specialists through to lawyers, clinicians and researchers.

2 THE DATA VALUE CHAIN

The Data Governance Framework is underpinned by the **Data Value Chain**. The concept of a value chain was first introduced by Michael Porter in the 1980s and has been used as an analysis and decision making tool in management since. A value chain is not a flowchart describing a process or a sequential series of events. Rather it describes the necessary steps by which value is added, incrementally, to produce a final "product" or result.

This methodology when applied to health research (see Figure 2) identifies each step along the Data Value Chain that can be described as "adding value" from the initial purpose of the data collection/reporting to the final "product"; that is, the research output.

Figure 2: The Data Value Chain

The Data Governance Framework identifies seven Data Value Steps:

1) **Purpose of Data** – the articulation of the purpose of collecting and reporting data. This step requires the researchers to determine WHY they want to work with data and the "end product" of the value chain.

2) **Data Element Management** – management of Data Elements from their original selection and definition through to their variation during the lifetime of the research.

3) **Data Capture** – systems enabling the secure collection and capture of clinical and research data.

4) **Data Verification** – identification of data entry mistakes, omissions or data errors to ensure reliability of datasets.

5) **Data Repository Management** – secure holding and management of data (digital and analogue) through its lifecycle that adheres to FAIR principles – Findable, Accessible, Interoperable and Reusable.

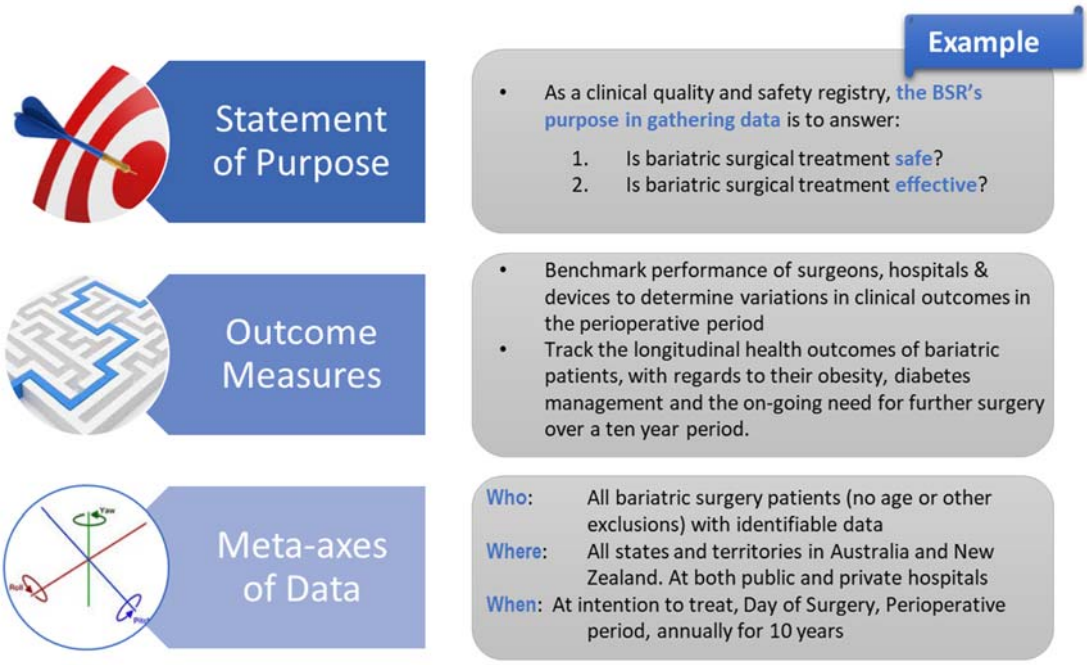
6) **Data Analysis** – systems and frameworks for the analysis of data to enable reporting that is replicable, robust and bio-statistically sound.

7) **Data Reporting** – provision of research outputs to agreed audiences, for specific purposes according to a schedule and in line with the purpose of the research.

2.1 PURPOSE OF DATA

This first step in the Value Chain requires the researchers to determine why they want to work with data and the "end product" of the research. By articulating the "Purpose of the Data" everyone is clear **WHY** data is being collected and reported, and it will provide the **lawful** basis for collection.

Figure 3: The Purpose of Data Articulated



In essence, a **Statement of Purpose** is the Research Question – that is, the explicit statement of the question(s) that are sought to be answered. These will usually be contained in one or more documents:

- research proposal;
- thesis outline;
- research protocol;
- business case; and/or
- articles of association/ terms of reference.

The **Outcome Measures** required to answer the Research Question will also need to be determined in this first step, as will the **Meta-Axis of Data**, which describes:

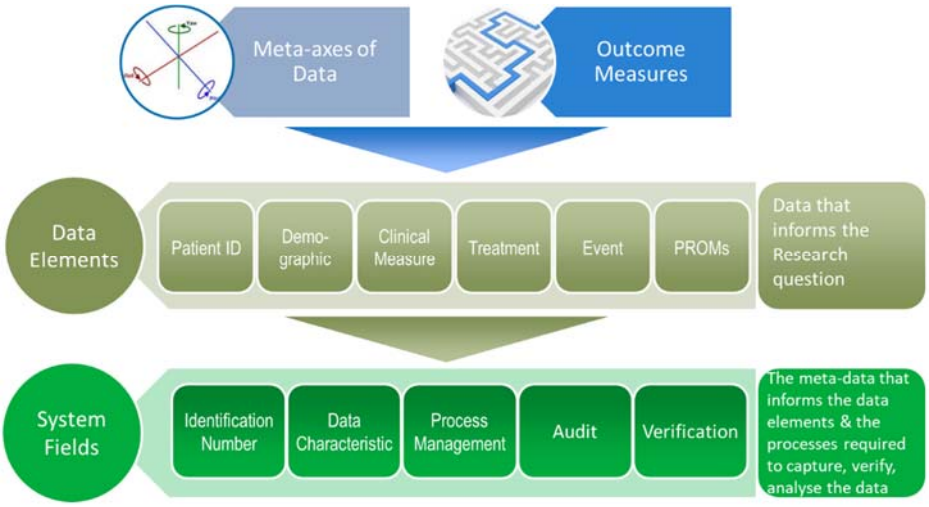
- upon **whom** the research will collect data – eg age group, gender/ sex, ethnicity, disease marker, other exclusion criteria;
- **where** the research will collect data – eg geography (global, country, state, city), private vs public health service provider, single vs multi-site; and
- **when** (or what time points) the research will collect data – eg at diagnosis, at intention to treat, at treatment, perioperatively, annually.

Finalisation of these three components may require a process (such as a Delphi process) to gain consensus amongst the various stakeholders, but it is vital this step is undertaken and completed prior to any data being collected.

2.2 DATA ELEMENT MANAGEMENT

This step in the Value Chain requires the Data Elements to be chosen and defined in the Data Dictionary. In addition a process to systematically change the Data Elements (if required) is also needed.

Figure 4: Data Elements and System Fields



Data Elements are the pieces of data that will be collected to inform the Research Question as either

- **Primary Data** – Researcher measures and collects the data directly. It would not be collected if not for the research; or
- **Secondary Data** – Researcher obtains data from another source. This data was originally collected for another purpose – usually as part of participant's medical record or administrative data related to their care.

Some Data Elements may also be **derived** as part of the research through calculation or modelling. Data Elements differ from System Fields which are the meta-data that informs the Data Elements and the processes required to capture, verify and analyse the data.

The Data Elements chosen must adhere to two principles:

(1) the Data Elements must be consistent with the stated purpose; and

(2) the more Data Elements collected, the higher the cost of collection and increased risk of data incompleteness.

Once chosen, these Data Elements must be included in a **Data Dictionary** and "defined". This involves a number of technical and clinical components, but at a minimum must include:

- Data Element Name – A name to be used by the researchers to describe this Data Element.
- Definition – A concise statement that expresses the essential nature of a Data Element and its differentiation from all other Data Elements. May be a simple explanation or technical reference and may include a Code-set.
- Justification/ Purpose for Collection – This is the purpose assigned by the researchers to collect this Data Element, and will be the lawful basis for collection of that specific Data Element.
- Prioritisation in Collection – An indicator of the obligation to collect this Data Element. It will either be mandatory (Priority 1), desirable (Priority 2) or optional (Priority 3).
- Representation Class – For all Data Elements there are 5 Representation Classes: Text, Value, Dates, Codes, True/False.
- Static vs Dynamic – Indicates whether the variable needs to be collected only once during the study (static) as it is not expected to change, or if it may need to be collected multiple times as there is an expectation it will change (dynamic).
- Collection Guide (including Permissible Values) – This is a detailed guide as to what should be collected and how it should be interpreted. It includes permissible values, formulas and any other relevant information to ensure that a consistent Data Element is collected/reported. It must be understandable to researchers, users of the system (including clinicians and their staff, hospital staff and research staff) as well as software/IT developers.
- Correspondence to Data Object – The relationship of the Data Element (one:one or one:many) to the *primary key* in the dataset. If a Data Element has *single* correspondence, there is only one value per field per primary key; if a field has *multiple* correspondences, there may be one or many (or no) values per field per primary key.
- Relationship to Other Data Elements – This describes in detail how this Data Element may be relied upon or used by other elements. This will generally specify Data Elements which may be derived from, or may

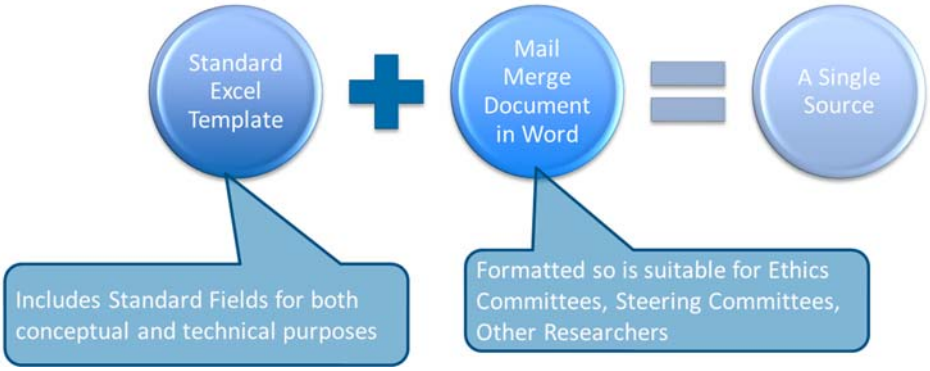
contribute to the derivation of another Data Element, or, alternatively, whose collection is conditional upon this Data Element being answered in a particular way..

- Source of Definition and Relationship to Other Datasets – Documents listed here have been used as references when designing the Data Element. Also listed are names of the organisations that developed the source document or provided advice on the Data Element. Also the Relationship to other Datasets is defined (including If METeOR has been used to source the definition of this Data Element).
- When Collection Commenced.

To select and define these Data Elements may again require a process to gain consensus amongst stakeholders such as a Delphi Process.

Once these Data Elements have been selected and defined there may be a requirement to further define technical elements and include in a more IT oriented **Data Dictionary** that focuses more on functionality and uses IT specific language and concepts. It is critical that there be only one source of truth for the definition of Data Elements. It is therefore recommended that if a more IT oriented Data Dictionary is required that a standard excel template be used and maintained that can accommodate a more technical approach but can then be outputted into a more suitable word document if it needs to be shared with other researchers or stakeholders (see Figure 5).

Figure 5: Data Dictionary that Uses Excel Template and Outputs to Word



Where research is carried out over a long timeframe (> 2 years), there is an expectation that some Data Elements may need to be varied.

> Example 1: All diabetes medications that were prescribed, other than insulin, were taken orally. So the options to describe the diabetes treatment included (i) Oral (mono) therapy; (ii) Oral (poly) therapy; (iii) Insulin. Injectable antihyperglycaemic drugs were then introduced onto the market and so “oral” therapy no longer properly described this type of diabetes treatment.
>
> Example 2: A study of retinal vascular imaging used one field in the database to collect a yes/no question about clinical abnormalities. However, the field was relabelled twice after it was set up, changing the meaning of the data collected at each point. The labels were:
>
> - ‘Are there any abnormalities?’
> - ‘Are there any reportable abnormalities?’
> - ‘Have abnormalities been reported?’

Ad hoc changing of Data Labels, Definitions, Field Options, etc, CAN impact significantly on the analysis and interpretation of the data. Therefore, a systematic **Data Element Variation** process is required to maintain the integrity of the data and ensure that when Data Analysis and Data Reporting stages are reached, there is a clear understanding of what is contained within that Data Element.

This Data Element Variation process will be used for additions, changes to definitions and for the retirement of Data Elements and must, at a minimum, include:

- Short Description of Current Data Element;
- Background of Proposed Change;
- Outline of the Proposed Change to the Data Element;
- Impact and Tracking of the Change; and
- Record of Approvals for Change.

2.3 DATA CAPTURE

This Value Chain step revolves around how the data that is gifted to the research about the patient, clinician or health service is gathered securely and efficiently. It includes all the systems enabling the secure collection and capture of clinical and research data whether manual or digital.

The Data Capture Value Chain step includes both the gathering of first hand data specifically for the research – Primary Data Capture – as well as the gathering of Secondary Data from other sources. Data Linkage can also be used in this Value Chain Step (see Section 3.1 for further explanation of Data Linkage specifics).

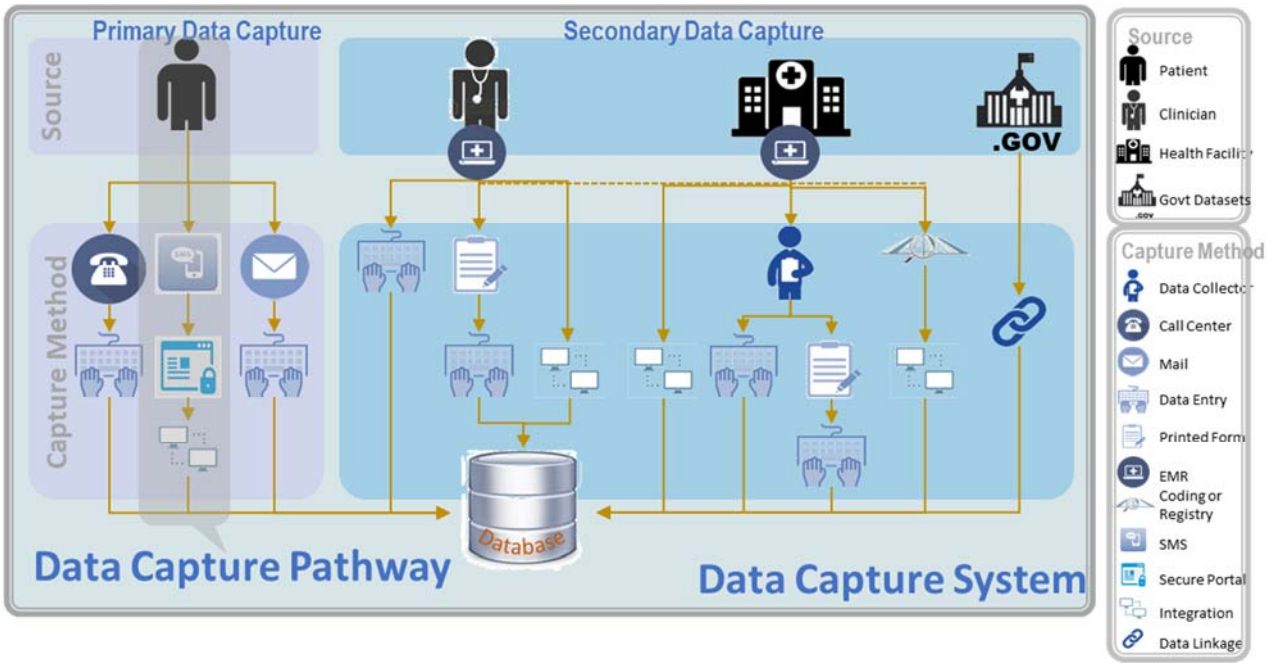
Importantly, this Value Chain Step requires Researchers to identify and map three things:

1) the individual **Data Capture Pathways** that will be used to gather the data – these include the digital and analogue capture methods that will be used;
2) the **Workflows** that are required to implement these Data Capture Pathways; and
3) the **Data Solution** that will use an IT system to corral the data into a systematic structure and may also assist in the capture process, workflows and Data Verification.

Regardless of whether the **Data Capture System** (Figure 6) involves a single Data Capture Pathway or multiple paths, it is essential that:

- all the processes are mapped, including the key Workflows underpinning the Pathway (this includes material such as Call Scripts and Call Protocols – how do you protect privacy when an answering machine is reached or you need to identify a participant – as well as SMS, Portal or Letter Content);
- the interplay is clear between all the Data Capture Pathways and Workflows;
- security of the data needs to be tested at all points along the Data Capture Pathways; and
- a change process is in place to ensure any variation in the process is systematic in order to maintain the integrity of the Data Capture System.

Figure 6: Data Capture System



The mapping of a Data Capture System and the underlying workflows will then be the basis for the selection of a Data Solution. Three factors dictate the sophistication (and therefore the cost) that is needed of the Data Solution:

1) The degree to which Roles need to be differentiated for Data Capture within the Data Solution.
2) The Workflows that are to be included within the Data Solution.
3) The Level of Validation built into the Data Solution (see next Section 2.4 Data Verification for more discussion).

The importance of each of these factors to the researcher will determine the complexity of the Data Solution and the technology that will be employed. If no differentiation of roles is required and there are no workflows to be controlled by the Data Solution with minimal internal system validation, then an EXCEL spreadsheet may suffice. Similarly, if a higher level of internal system validation is required, ACCESS may be a suitable Data Solution. If some role differentiation is required as well as more complex internal system validation, then REDCap may be suitable. Where complex role differentiation is required and complicated workflows need to be built into the Data Solution, then the more costly Custom Build may be needed.

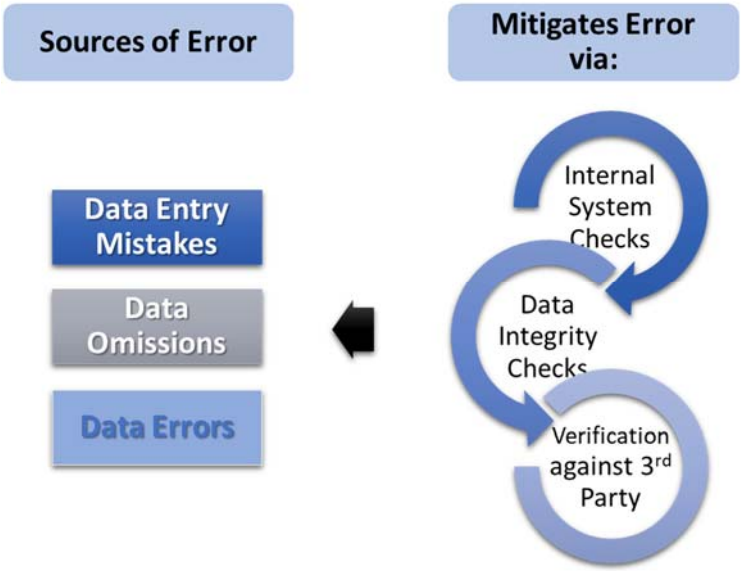
These trade-offs must be made and will be determined in large by the funds available. It should be remembered when choosing a Data Solution:

1. advice from Helix or RSU can help guide Researchers in making these trade-offs;
2. if customising a solution – whether ACCESS, REDCap or Custom Build – that just because functionality CAN be built, doesn't mean it SHOULD be built. It can be better to do some functionality offline/ manually, particularly if it complicates other parts of the Data Solution, by including it in the customisation; and
3. consideration must also be given to how the customisation is to be maintained into the future – have Style Guides been followed? Has it been fully documented? Is it scalable? Is the business logic sitting at the front-end (client-side) or in the database itself (server-side)?

2.4 DATA VERIFICATION

This Value Chain step of Data Verification involves the mechanisms that are used to provide confidence in the dataset. It includes those activities that identify Data Entry Mistakes, Omissions or Data Errors via Internal System Checks, Data Integrity Checks or Verification against Third Parties to ensure the reliability of the dataset.

Figure 7: Data Verification



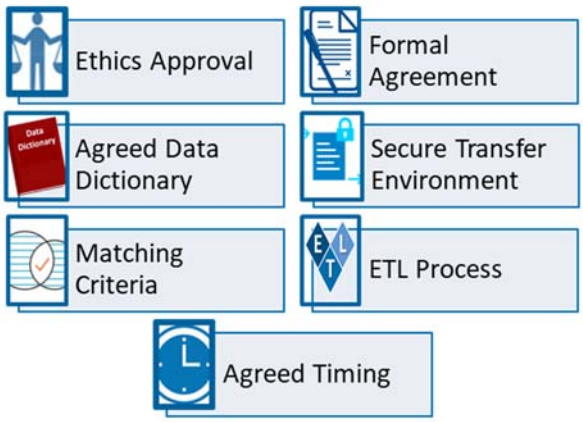
Data Entry Mistakes arise during the inputting of individual fields by humans or machines. Data Omissions is where whole records or individual fields are omitted from the dataset (either with intent or accidently). Data Errors are individual fields recorded with the incorrect value (either with intent or accidently).

To mitigate these errors there are three types of verifications that can be undertaken.

Firstly, **Internal System Checks** can be built into the Data Solution to guard against error at time of input. This will usually be to avoid Data Entry Mistakes. These checks do not allow data to be added that is not "valid" and are generally managed through the Permissible Values of each Data Element. For instance, the Data Solution will only allow a height between 1-3 meters. Or a clinicians name can be entered only from an administrator controlled list. Internal System Checks can also be used to ensure Data Omissions do not occur. A record can be flagged as incomplete or not allowed to be created if specific data is not provided.

Data Integrity Checks are run after data has been input and can be run according to a specified time either within the Data Solution or externally. In some instances, it may be better to run a monthly Data Integrity Check outside the Data Solution than build a very complex Internal System Check or an automated Data Integrity Check. This is a good option where the Business Rules governing the data's integrity are complex and will also be needed whenever free text options (eg, "other") are available. Data corrections that will need to be made after errors have been identified can be recorded either manually or may be automated. Regardless of the method, the correction must be systematically recorded to ensure integrity of the data at any subsequent audit.

Figure 8: Requirements for Data Linkage



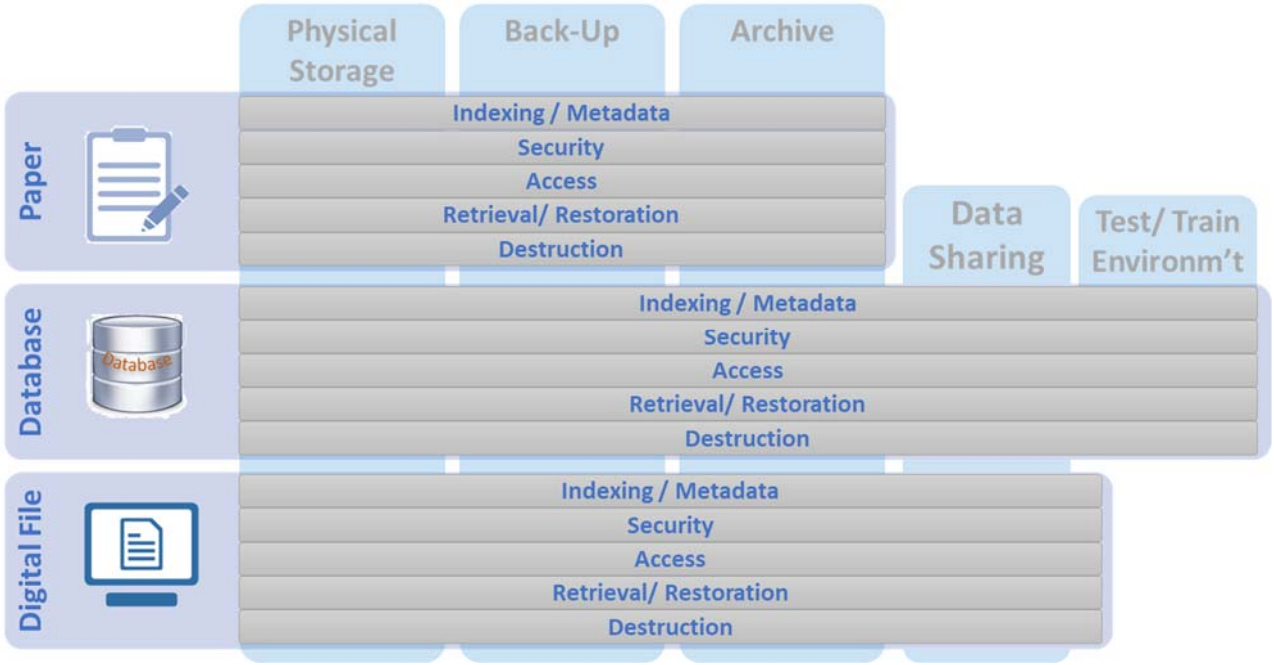
Finally, **Third Party Verification** may be required in some research. This type of Data Verification can avoid bias and confounding, particularly in self-reported data[1]. Data Linkage will be required whereby one dataset is compared to another. There are seven requirements for any Data Linkage (see Figure 8 and Section 3.1 for further discussion) but an additional requirement for Data Verification Linkage is the identification of which dataset is to be taken as "truth".

[1] *Frieden TR. Evidence for health decision making - beyond RCTs. NEJM 2017*

2.5 DATA REPOSITORY MANAGEMENT

This Value Chain Step outlines how data, obtained during Data Capture or Data Verification (whether paper, digital file and/or database), is to be held through its lifecycle, regardless of its format, so that it adheres to FAIR principles while meeting all data protection (privacy and security) requirements.

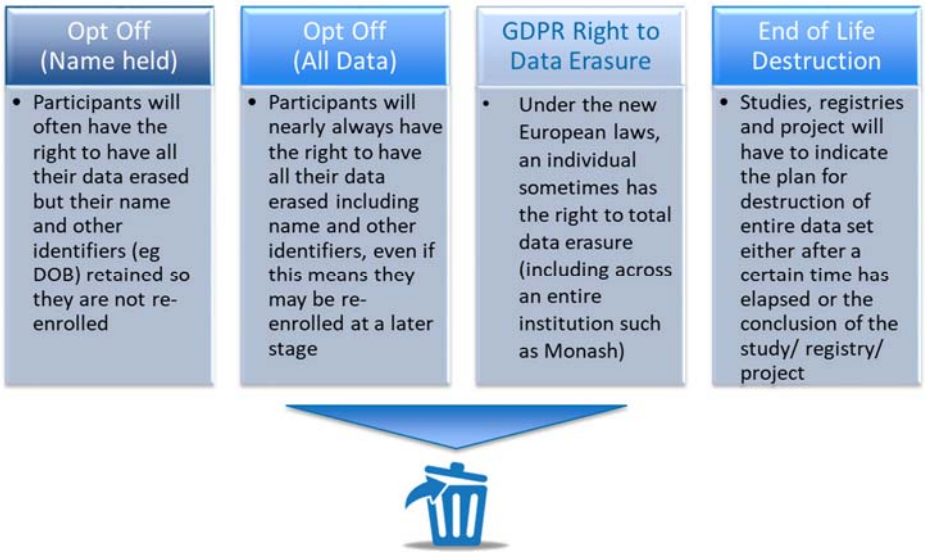
Figure 9: Data Repository Management Components



The lifecycle of the data includes the **Physical Storage**/ holding of the data, the **Backup** instances of the data, **Archiving** of the data as well as the data's existence in a **Sharing** environment or **Test/ Training** environment. How the data in each of these states is managed must be recorded with regards to:

- Indexing/ Metadata
- Data Protection (Security/ Privacy)
- Access
- Retrieval/ Restoration
- Destruction

Figure 10: Destruction of Health Data



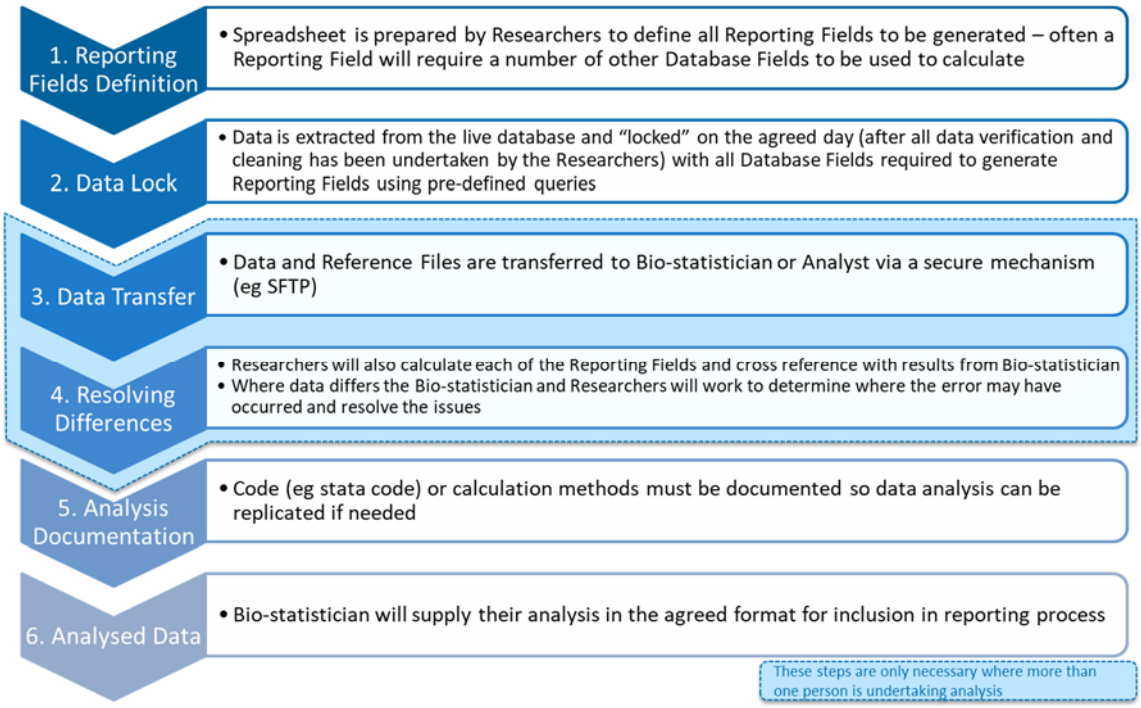
Destruction is particularly important in the context of health data given the privacy/ethical obligations involved. Systems are required to ensure all forms of data (paper, database, digital) in all their states (stored, back up archived, test and data sharing environments) can be destroyed securely and efficiently. If they cannot be destroyed (eg, if in a sharing environment), then it is essential this is communicated to the participants of the research.

A Data Repository Log that summarises this information is a useful resource (see Appendix A.1), to not only work through these issues and ensure compliance, but also to ensure that all other Research Documents (eg, protocol, ethics applications, data sharing agreements) are consistent with regards to how data is being held.

2.6 DATA ANALYSIS

This Data Value Chain Step outlines the systems and frameworks for Data Analysis so that it is replicable, accurate, consistent and valid. A number of activities are required to ensure this, including:

Figure 11: Data Analysis Requirements



It should be noted that steps 3. Data Transfer and 4. Resolving Differences are only required where more than one person is analysing the data. There may be some instances where only one researcher analyses data, but it is strongly recommended that all analysis be checked by more than one person.

Data Locking is an important but often over-looked step. It is particularly important where reporting is periodical and research is ongoing. It ensures that the analysis dataset is date/time stamped and when reported upon, the date/ time that it was extracted is made clear. Very few datasets are captured in real time, and participants often have the option to opt-out at any time and/ or may have periods where their consent is "pending" and needs to be excluded. This means that data extracted today may be different to data extracted tomorrow, even if it is referencing events in the same time period. Through this "locking" process researchers are able to have confidence that all the Data Verification has been undertaken on the dataset and in the level of data completeness.

The degree to which a dataset is complete will affect the validity of the analysis being performed. A record may appear to be "complete" as all fields have a valid entry, but if the system has not been designed correctly the designation of "complete" may be misleading.

> Example 1: A patient was phoned 5 times for their follow up; after the 5th call the patient is allocated to Lost to Follow Up (LTFU) and the follow up requires no further attention and drops off the worklist. In this case the designator of whether the follow up is due/current (FUVAL) is changed to 2 "complete", but the data fields are not "complete" as the patient hasn't been followed up. This can be further compounded where any of the data fields in that record include code sets or true/false options that default to a value other than "unknown".
>
> Example 2: A patient's treatment has been entered and all fields entered. Record is submitted and considered complete. In every field, however, the value "Unknown" has been selected.

Care must be taken, therefore, in the Data Element Management phase in the Data Element's definition, in the Data Capture phase in the design of the Data Solution, and in the Data Analysis phase to ensure that "completeness" reflects valid responses other than "Unknown".

2.7 DATA REPORTING

The Data Reporting Value Chain Step involves the final output of the research; it may take the form of a report, research paper, thesis or presentation, but will always align with the original purpose of the research.

Figure 12: Data Reporting Requirements

Within this Data Reporting step it should be very clear:

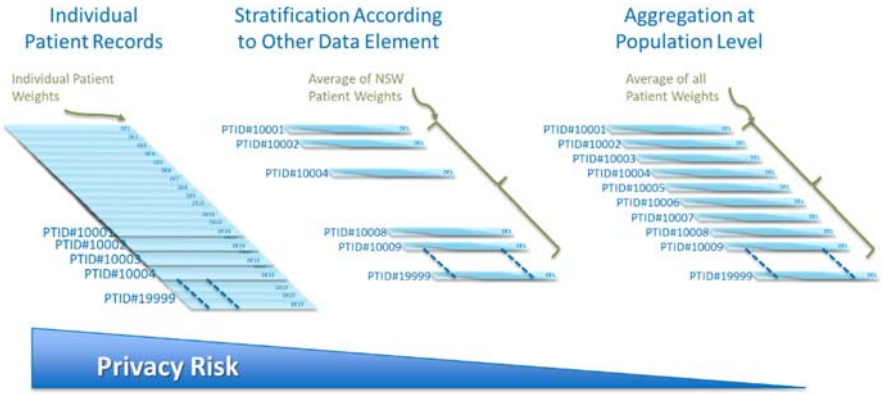
- to whom Reporting will be provided;
- when Reporting will be provided;
- what Reporting will be provided; and
- why Reporting will be provided.

Reporting in all its forms – a glossy report, online reporting tool, research paper, presentation –will be undertaken for one of three purposes:

- it is the actual research output;
- it is quid-pro quo, as the provision of reports is in exchange for data provision (eg individualised, confidential clinician and/ or site reports); or
- it is a contractual agreement in return for funding from governments, industry, research organisation, etc.

In providing reports it is also essential that all parties understand the degree of privacy and confidentiality that is attributable to the report. The level of aggregation and/or stratification of the data reported will be crucial in determining if these requirements are being met. Stratification – that is, segmenting the data into distinct groups – can increase privacy breach risk in and of itself. For example in Figure 13 if there is only one clinician in NSW, the aggregation of patient weight at the stratified level of State, may breach the clinician's privacy.

Figure 13: Aggregation and Stratification's Effect on Privacy



When deciding on the timing of reports it is important to understand the trade-off between accuracy and timeliness. Very few research projects will have the capacity for real-time capture and reporting of research data. Thus it is imperative to:

1. set expectations with regards to the timing of reporting. Finding the right time by which data is to be reported must take into account not only this need for accuracy but also how its timeliness will affect the research's usefulness, relevance and ability to impact change. If all other steps of the Data Value Chain have been properly undertaken – particularly the Data Verification step – then the time required to ensure accuracy should be reduced as the "Due Diligence" (comprehensive appraisal of the data's quality) is already complete when data is locked for analysis;
2. cite the defined period the data covers **AND** the date data is extracted – eg, *The data contained in this document was extracted from the Registry XYZ as at 28 February 2015, but pertains to procedures and follow-up that has occurred up to 31 December 2014. As the registry does not capture data in real time, there can be a lag between occurrence of an event and its capture.* The only exception to this is where data is "closed off[2]" and then reported, in this instance the dataset will always be "fixed" within the live data; and
3. ensure all participants whose data is included in a report are **consented as at the date of extraction** – for those research projects that use opt-out consent this may require people's records to be excluded until the 2/3 week opt-out period since they were sent their Explanatory Statement has expired.

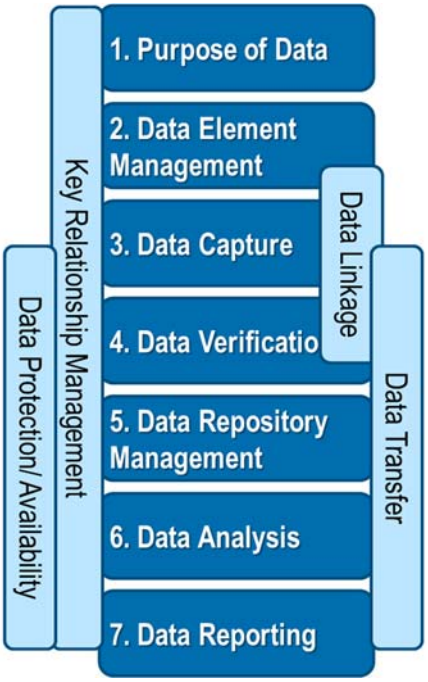
[2] Closed off" indicates that no data relevant to that period can be added after "close off" date – this is similar to the accounting notion of closing off accounts at the end of the financial period

3 OTHER KEY ACTIVITIES

In addition to the Data Value Chain steps the Framework also identifies key activities that cut across the Data Value Chain and their corresponding policies/ procedures. These include:

1) Data Linkage
2) Data Transfer
3) Data Sharing
4) Data Protection/ Availability
5) Key Relationship Management

Figure 14: Key Activities Across The Data Value Chain



3.1 DATA LINKAGE

Data Linkage involves the formal agreement between two parties to share their dataset for either Data Capture or Data Verification purposes. These will often be administrative datasets from the government such as Medicare, State Admissions and Episodic Data, PBS, AIWH data or ABS data. There may also be administrative or clinical data provided by hospitals in the form of ICD-10 data. Or it may simply be other research datasets that have been agreed to be linked.

If Data Linkage is for Data Capture purposes, it will either provide **depth** to the original dataset – that is, provide additional Data Elements to existing participants – or it may provide **length** to the original dataset – that is, provide additional participants.

If Data Linkage is for Data Verification purposes it will be used to verify the data captured by the research and/or fill any gaps that may be missing from the original data capture. In order to use Data Linkage for Data Verification purposes it is essential that the dataset of "truth" be established so that it is known which dataset will be corrected when differences arise.

All Data Linkage activities have a standard set of requirements (See Figure 8 in previous section, 2.4 Data Verification):

1. Ethics Approval: Both datasets will require ethics approval to link their data. This can be time consuming and should be considered in initial protocol development.
2. Formal Agreement: A Memorandum of Understanding (MOU) between the parties should be created to cover the roles and responsibilities of each. This MOU must comply with the legislative framework of both parties.
3. Agreed Data Dictionary: Data Dictionaries of both Datasets will need to be analysed and mapped to ensure that Data Elements used to match and/ or compare are consistent.
4. Secure Transfer Environment: A mechanism to transfer data between parties will be required.
5. Matching Criteria: Some Datasets may have a single identifier to match (eg, UR numbers), while others will require criteria and algorithms to properly match datasets.
6. ETL Process: An ETL process (Extract, Transform and Load) is required to ensure data can be matched, compared and/or appended or corrected systematically.
7. Agreed Timing: Agreed timing between the parties will need to be determined so that data remains relevant and useful.

The identification of **Criteria to Match** patients/records within the Australian health environment can be difficult as there is NO unique individual identifier, unlike other jurisdictions such as NZ that has a National Health Index (NHI). Matching within a site can confidently be done with UR Numbers, but each site or group of hospitals will have a different system. It should be noted that Medicare numbers **are not** a unique identifier for a person (see Meteor 270694), as a person's Medicare number can change (eg, when child leave parent's card, when a new family card is created, or following divorce).

Deterministic matching (that is, an exact match) can be used on just one field, such as Medicare Number (including reference), but may under-match for the reasons identified above. Deterministic matching on a combination of fields (eg, full name, date of birth and address) may also under-match as Australian addresses tend to have unstructured data and someone's legal name and recorded "full name" will often differ in Australia.

Probabilistic matching can be used where several field values are compared between two records and each field is assigned a weight that indicates how closely the two field values match. The sum of the individual field weights indicates the likelihood of a match between two records and researchers can then decide what level of confidence suits them.

Stepwise deterministic matching with appropriate decision rules, weightings, edit distances, etc, and using multiple fields can produce matching rates similar to probabilistic matching.

It should be noted that NO APPROACH IS PERFECT and whatever method is used, **manual review** is required for non matches and matches with low probability weightings.

3.2 DATA TRANSFER

Data Transfer involves the secure transmission of sensitive data between parties. Its purpose may be for reason of Data Capture, Data Verification, Data Repository Management or Data Analysis and in this way it is an activity that cuts across the Data Value Chain.

Data Transfer mechanisms can be assessed along three axes to determine the appropriate mechanism for the researchers:

1) Risk of Malicious Attack
2) Risk of Human Error Breach
3) Cost/ Benefit.

Figure 15: Assessment of Data Transfer Mechanisms

	Description	Risk of Malicious Attack	Risk of Human Error Breach	Cost/ Benefit
Custom Built Portals for Upload	Web based portal allowing upload/ download of files.	Low	Low	High cost to develop/maintain. Easiest option for end users as fully customisable.
SFTP	Secure File Transfer Protocol. (Differs from FTP in that both credentials and data are encrypted in transit.)	Low	Low	May require a change management approach with those sending the data – provision of user friendly training/"how to" guides can assist in this. Difficult to manage users (manual process).
Encrypted Email	Using a plugin to email that encrypts the message, the attachment, or both, before being sent to the mail server.	Low	Low	Medium cost. Need to manage private/public encryption key. Need to manage plugins on email clients.

	Description	Risk of Malicious Attack	Risk of Human Error Breach	Cost/ Benefit
Mail	Mail remains a well accepted transfer mechanism for health data, mainly in paper format.	**Low risk**, due to its distributed nature it is costly and difficult to maliciously attack on a large scale.	**Medium risk**, sending to the wrong address is always a risk	Provision of pre-addressed envelopes to those sending the data can mitigate some of the risk. Ensuring delivery is clear within Monash's postal system is also required.
Fax	Uses audio carrier to send an image over telephone line.	Low	Medium	Still widely adopted due to ease of use. Can be intercepted but very unlikely. Risk of wrong number. Some organisations only have email gateways.
Unencrypted Email	Email is NOT encrypted between servers. It is only encrypted with SSL if using a web portal to receive, and both the sender and recipient are on the same host.	High (compromised accounts or interception of data)	High (accidental forwarding)	**Health data should never be transferred using these services**
Publically Available Cloud-based File Sharing	Services such as Dropbox, Google Drive, Microsoft OneDrive, CloudStor.	Low (however need agreements)	Medium-accidental sharing	**Health data should never be transferred using these services without signed agreements and transborder flow review.**

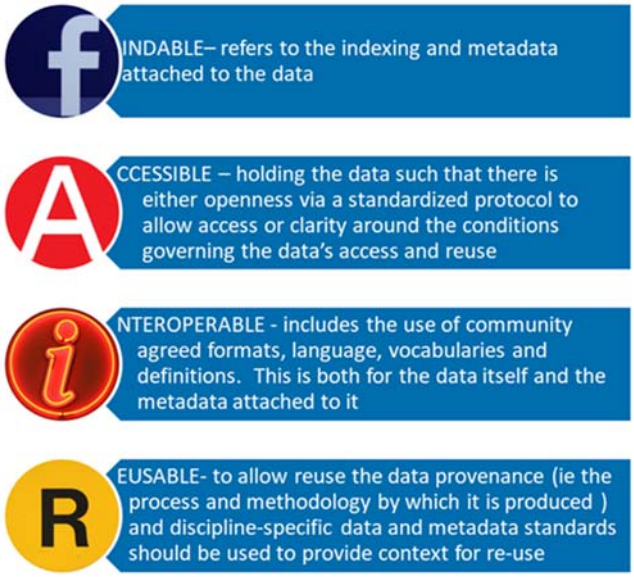
Data Transfer systems may fail due to human error or technical faults, so it is important to have a redundancy process/system in place in the case the first system fails. File transfer systems can be made resilient (multiple servers) to lessen the chance of this happening. If the approved data transfer system fails – there must be clear communication to stop users from going back to bad habits like sending data in an unencrypted email. Where data is transferred by another party via an insecure platform, researchers should ensure:

- the other party is informed immediately of the insecure nature of the transfer and that data should be deleted immediately;
- all copies of the data that are currently residing in an insecure location (eg, in email) are fully deleted from Monash's side, including Back Ups and "Trash" folders; and
- a secure option is re-offered (including training in its use) to other party.

3.3 DATA SHARING

Data Sharing describes the arrangements made with third parties to share data, rather than reports, to facilitate collaborative research or replicate results. It is essentially the "sub-contracting" of the Data Analysis and Data Reporting Value Chain Steps to other parties.

Figure 16: FAIR Data Principles



Data Sharing has become an integral part of all Research across the world. The FAIR data principles developed in 2015 and have been adopted by Monash and international bodies such as the NIH and European Commission to facilitate this Data Sharing.

These principles are applied to ensure that through the sharing of data, the dataset's maximum benefit is derived via its reuse. It is important, however, to keep top of mind the privacy and ethical requirements of the data that has been collected.

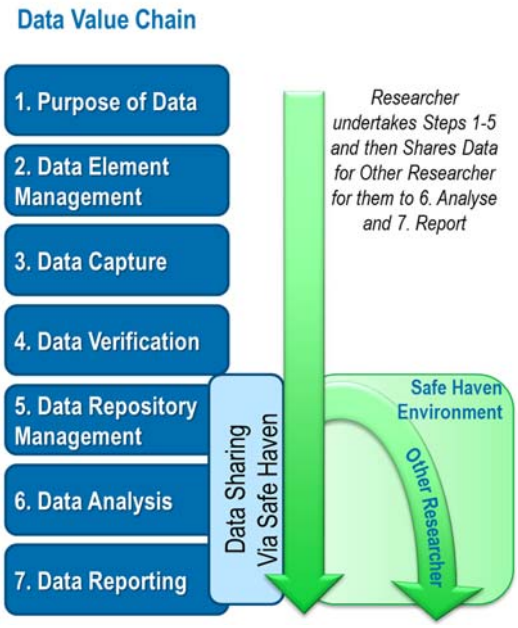
Like Data Linkage there are requirements to undertake Data Sharing including:

1. **Ethics Approval**: Permission to share data will need to be included in any Ethics Process to ensure participants' rights are maintained in how their data is used
2. **Data Access Policy**: A Data Access Policy will generally need to be provided in the Protocol that outlines when Data Sharing will be permitted and how it will be handled (including the governance body that will decide on whether data will be shared)
3. **Data Access Application/ Agreement**: A standardised Data Access Application or a formal Data Sharing Agreement will need to be in place before data is shared
4. **Secure Sharing/ Transfer Environment**: A mechanism to create a sharing environment (see Safe Haven discussion below) or to transfer to data to the other party will be required

The risks faced in "sub-contracting" parts of the Value Chain to third parties by Data Sharing can be mitigated via:

1) the use of **Safe Havens;** and/or
2) data undergoing **Anonymisation Processes**

Figure 17: Data Sharing Via Safe Haven



Safe Havens are a mechanism to protect the privacy/security of the data when it is shared. These are environments that are created to control particular aspects of the Data Value Chain when Data Sharing occurs.

Within a Safe Haven, a researcher can provide others access to their data but control the Data Repository Management as the data cannot be removed from the Safe Haven environment. Similarly the Data Analysis and to a limited extent the Data Reporting steps can also be controlled through the provision of approved analysis tools and the need to approve outputting of graphs/tables, etc from the environment.

These control mechanisms over these Data Value Chain steps is very important given the security and privacy requirements of health data.

The concept of "**Anonymised**" or "De-identified" data is fading. Omitting data that directly identifies the patient (eg medicare number, name, date of birth, address) is no longer sufficient to protect the participant's privacy. Conversely, if too much data is removed the usefulness of the dataset may be compromised (eg removing the day of the procedure and providing the month, may lose the granularity needed to measure days since surgery) or the ability to link datasets in the future may be lost.

The ability to re-identify data with publicly available information and large amounts of computing power have dispelled the notion that data can be in a state of "anonymised" or "de-identified". Instead, best practice anonymisation processes such as those outlined in ISO/IEC 20889 must be applied to the data to protect privacy when Data Sharing.

3.4 DATA PROTECTION/ AVAILABILITY

All “reasonable” steps must be taken to keep health data secure in all its formats – digital and analogue. Data Protection requires the privacy of the participant AND the security of the data to be maintained ACROSS the Data Value Chain – whether data is being captured, verified, held in a repository, analysed or reported.

Health information is considered “Sensitive” information – a subset of personal information – and as such all patient and clinician information will be handled in accordance with the *Commonwealth Privacy Act (1988)* including The *Privacy Amendment (Enhancing Privacy Protection) Act 2012* and other relevant state and territory laws and regulations relating to the collection, storage and dissemination of such information. There will be some instances where other nations’ laws may also apply (eg, in a global registry, the European Union’s GDPR – General Data Protection Regulation or US regulation such as HIPAA – may have jurisdiction).

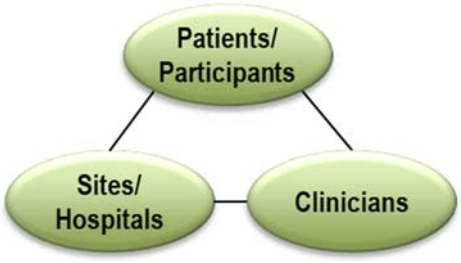
Four key aspects of Data Protection (security/ privacy) need to be considered:

1. Jurisdiction – depending on the type of data collected and the location of persons when collected, different jurisdictions may apply. Eg, an Australian citizen filling in a follow up while on a European holiday technically falls under GDPR, but a European citizen on holiday in Australia participating in a trauma registry does not.
2. Level required – depending on the type of data collected and the degree of anonymisation and/or aggregation that has occurred, different levels of privacy may apply.
3. Breach – notification requirements for privacy breaches differ between jurisdictions and Monash’s Office of General Counsel must make the assessment on whether a breach is notifiable or not.
4. Updates – what is considered “secure” today may not be considered secure tomorrow. Advances in hacking techniques requires vigilance in ensuring systems are updated regularly to maintain the required levels of security.

3.5 KEY RELATIONSHIP MANAGEMENT

The final key activity that cuts across the Data Value Chain is one that is peculiar to health research – that is, Key Relationship Management. This refers to the unique relationships within the health data environment between participants, sites and clinicians. These three parties are those **from whom data** is obtained (source) and/ or whom **data is about** (subject):

Figure 18: Health Data Key Relationships



These relationships between the three parties and with the researcher are managed through processes such as:

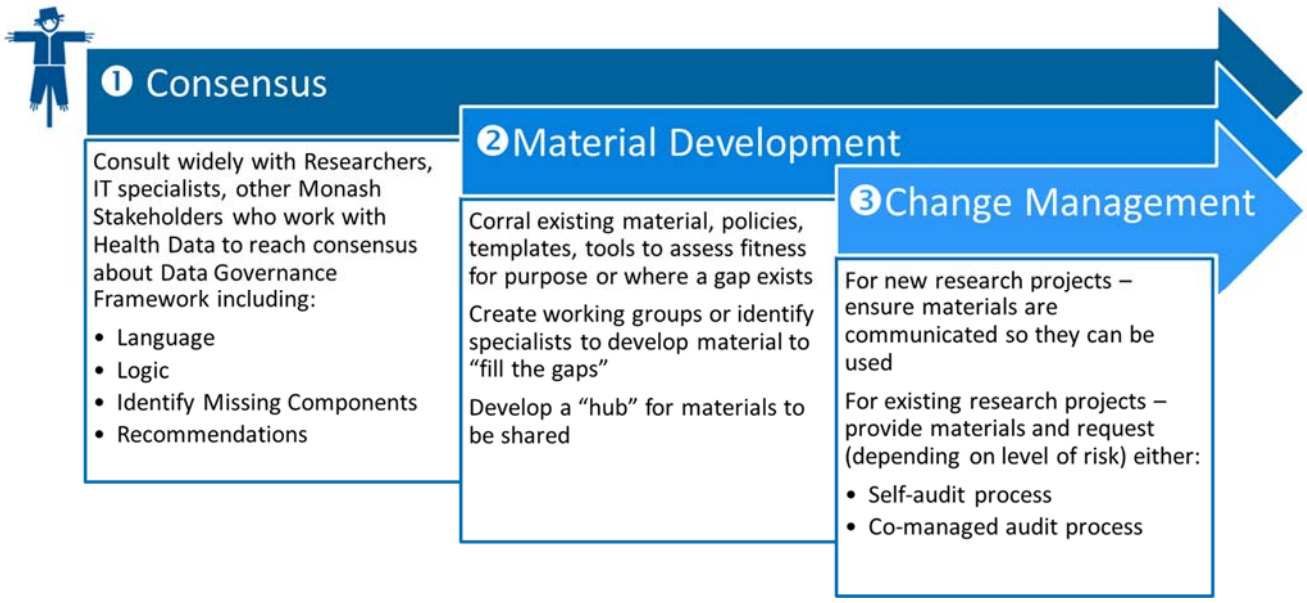
1) Ethics approvals to undertake research.
2) Consent processes during recruitment.
3) Recruitment of clinicians and/or sites to be involved in Data Capture and/ or recruitment of participants.
4) Maintenance of relationship with sites, clinicians and ethics committees.
5) Effective reporting back to patients, sites, clinicians and ethics committees.

These activities require both expertise and focus to ensure that relationships are maintained across all three groups.

4 IMPLEMENTATION OF DATA GOVERNANCE FRAMEWORK

This Data Governance Framework has been developed as a "strawman" – a basis for discussion. Extensive consultation has been undertaken with health researchers, IT specialists and other Monash stakeholders such as the Office of General Counsel and the Monash Research Office. Consensus has been reached around the language, logic and components of the framework.

This will always be a "work-in-progress" document and will need to be updated as the environment and circumstances change, however it will become an anchor from which material can be developed to guide all the policies and procedures of health data activities undertaken at Monash.



Material will now be developed to guide these activities. Prioritisation will be required to ensure those activities at highest risk are addressed first. A preliminary set of material has been identified including:

- GDPR related materials such as consent documentation and Data Sharing agreements;
- Data Repository Log;
- Data Transfer Guidelines;
- Guidelines for Technology Choice; and
- Best Practice Patient Matching Criteria.

Helix will work with other stakeholders across Monash and convene working groups to develop and endorse material. This will ensure any material already developed and working well will not be re-invented. Helix will also develop a "hub" where these materials can be shared and accessible to all Researchers and interested parties.

Changing practice will take time and once Helix has developed sufficient material, it will work with researchers and other vested stakeholders to identify what is needed and how the materials can assist.

A.1 Appendix – Data Repository Log

Data	Patient Case Record Form	Form of Data	
		⊙	Paper
		○	Database
		○	Digital File

		STORE	BACK-UP	ARCHIVE
Description		Filing Cabinet "Prior to conversion to electronic files by Registry staff, forms completed in hard copy will be stored securely in a locked filing cabinet, behind locked and swipe card-only accessible doors." Protocol v4.0	NA	To be determined when files will be sent to archive "Archived information will be stored indefinitely in a secure location." Protocol v4.0
Index/ Meta-Data		Indexed via PatientID	NA	Indexed via PatientID
Security		Behind card swipe door in locked filing cabinet. Keys are secured in locked box with single key secured overnight/ weekends in drawer	NA	Sub-contractor responsibility
Access		Registry Staff have access to key whenever they have access to building		Request via contractor (Business Hours)
Retrieval/ Restoration		Manual search by PatientID	NA	Request via contractor
Destroy	Opt off – leave name	All paper records on file are destroyed via Monash's secure document destruction service (SteriHealth) "Disposal of any information will be in accordance with the National Statement on Ethical Conduct in Research Involving Humans 2007 (2015) and New Zealand's Ethical Guidelines for Observational Studies: Observational Research, Audits and Related Activities (2012)." Protocol v4.0	NA	Request via contractor
	Opt off - full			
	Single Record Full Deletion			
	End of life Destruction	Currently protocol states it will be held indefinitely so not destroyed	NA	Currently protocol states it will be held indefinitely so not destroyed

Data		Patient Case Record Forms – Scanned Image	Form of Data	
			○	Paper
			○	Database
			◉	Digital File
		STORE	BACK-UP	ARCHIVE
Description		Paper Patient Teleforms are scanned into Digital files (each file is separate pdf) and stored on S:/Drive (see process below) "Other electronic files will be stored on the Monash University secure shared drive which is password protected." Protocol v4.0	S:/Drive is backed up by e-solutions nightly and held for 30 days	To be determined when files will be archived and how "Archived information will be stored indefinitely in a secure location." Protocol v4.0
Index/ Meta-Data		Indexed via PatientID which is in file name	Each of the 30 backups would mirror the s:/drive thus files would be indexed via PatientID which is in the file name (as per stored data)	To be determined
Security		Monash's secure s:/drive	eSolutions to outline	To be determined
Access		Available to staff with access to s:/drive 24 hrs a day and 7 days a week if have VPN access	eSolutions to outline	To be determined
Retrieval/ Restoration		Via access to s:/drive and search by PatientID – unable to search by patient name	eSolutions to outline	To be determined
Destroy	Opt off – leave name	Digital file(s) deleted from S:/drive "Disposal of any information will be in accordance with the National Statement on Ethical Conduct in Research Involving Humans 2007 (2015) and New Zealand's Ethical Guidelines for Observational Studies: Observational Research, Audits and Related Activities (2012)." Protocol v4.0	eSolutions to outline	To be determined
	Opt off - full			
	Single Record Full Deletion			
	End of life Destruction	To be determined	ESOLUTIONS to outline	To be determined

MONASH University

Further information

Monash University
Wellington Road
Clayton, Victoria 3800
Australia

E: Helix-DGF@monash.edu

monash.edu.au

CRICOS provider: Monash University 00008C