Australian Research Data Commons

Elicitation of contexts for discovering clinical trials and related health data: An Interview Study

ARDC user study of data discovery project report (2)

Ying-Hsang Liu (0000-0001-6504-4598), Wissen Insight
Mingfang Wu (0000-0003-1206-3431), ARDC
Megan Power (0000-0001-7345-559X), Monash University
Adrian Burton (000-0002-8099-7538), ARDC

Version 1.0
14 April 2023

NCRIS National Research Infrastructure for Australia
An Australian Government Initiative
ARDC is enabled by NCRIS

Executive Summary

In this project, we investigated the range of contexts of data discovery faced by clinical trials researchers, to better understand crucial factors that influence data discoverability and lead to data reuse. Our findings connect ways researchers both initiate personal data journeys and follow highly managed processes of data selection criteria. The project demonstrates the need for a holistic view in understanding data discovery approaches by clinical trials researchers.

Our findings suggest the significance of prioritising user information needs for selecting and using datasets. As using a dataset involves a complex process of conducting secondary data analysis within a research project, it is important for data repositories to focus on meeting the objectives of system design at this level, which go beyond the scope of traditional information retrieval systems such as online databases and search engines. Data repositories need to consider ways to specify their objectives of system design for the data discovery systems and services.

Project aim: This project was conducted with the Australian Research Data Commons (ARDC) and forms part of larger activities through the ARDC data discovery project and the Health Studies Australian National Data Asset (HeSANDA) initiative. The purpose of the HeSANDA initiative is to enable discovery of health data assets via a federated data catalogue. For this project, we aim to contribute a more detailed understanding of the ways clinical trials researchers discover, access and manage data of value to their work within the context of open clinical trials and related health data.

The project focuses on identifying requirements for improving the data discovery experience for users. Our research questions are:

1. How do researchers approach data discovery?
2. How do researchers search for data?
3. What data attributes matter to researchers' data search?
4. What criteria do researchers apply for assessing relevance and usability of datasets?
5. What are the contexts of data reuse by researchers?

Project methodology: The study design aimed to elicit narratives on the contexts of data discovery through in-depth interviews of clinical/health guideline developers, researchers conducting secondary research studies and clinical trial designers. A pre-interview survey captured participant profiles (i.e., research areas, career stage, job roles and sources of data in recent research projects) to inform the interview phase. The research and interview design loosely followed the data lifecycle to gain a holistic understanding of ways researchers' approach searching for, accessing, collecting and handling data as the research process unfolds. However, the interviews also aimed to capture the initiating points of data discovery in 'personal data journeys.' Participants were recruited directly via the ARDC. Groups were identified who had either participated in projects in relation to the HeSANDA Program or were known by the project lead to be engaged in clinical trials related activity. This mixed-method approach provided rich material for a more detailed characterisation and thematic analysis based on the user pre-interview responses and the interview transcripts.

Participant profile: A total of 17 participants were recruited for the study based on their experience in discovering other party's data for their own research, or discovering data within any data repository/catalogue or via any other data discovery tool. The participants held varied roles and represented a range of career stages. While the majority of participants were engaged with collaborative clinical/health research projects and have extensive experiences discovering and reusing data, we also interviewed clinical trial designers and clinical/health guideline developers who brought multiple perspectives to the study. We also interviewed researchers who we noted engaged more intensively in data discovery to support the tasks of systematic reviews, meta-analysis and guideline development.

Summary from the pre-interview survey: Some findings of relevance from across the participant responses which informed the interviews and supported analysis of the interview transcripts included:

- An overall pattern of high reliance on both published literature and word of mouth via colleagues for data discovery, particularly at the starting point for a project. This was emphasised more notably by Early Career Researchers (ECRs) and Middle Career Researchers (MCRs).
- Engagement in sourcing information from conferences and seminars from both within and outside the participants' main field of interest.
- An indication of role differentiation with doctoral and ECRs committing more time to their nominated projects and MCRs and (Late Career Researchers (LCRs) conducting a broader range of roles in relation to clinical trials research.
- A pattern of data activities which aligns with expectations for the data lifecycle for early-mid and mid-end stage projects.
- An apparent limited engagement with the data lifecycle during maturity phases of data archiving or data publishing activities.

Summary from the interview analysis: The full interview transcripts of all 17 participants provided rich insights for exploring the research questions. Here is a summary of findings:

- Participants commonly identified important international clinical trials data sources, including registries, such as ClinicalTrials.gov and the WHO International Clinical

Trials Registry Platform (ICTRP), online search databases, such as PubMed, MEDLINE and Embase, and preprint servers.

- Participants who specified more advanced search strategies for systematic reviews, worked in consultation with content and search experts as well as experts in the related medical field.
- Participants described enormous demands in negotiating the data licensing of clinical trials data for meta-analysis. Increasingly, complex data sharing and reuse agreements were required for further research collaboration.
- Data quality was identified as the most critical data attribute by researchers in their data discovery efforts. In most cases quality was associated with the data integrity (e.g. screening for fraudulent data or poorly developed trials), reliability of both the source and the methodology (established longitudinal studies), the granularity level of data (e.g., individual participant data) and the double-blind randomised controlled trial studies published in peer-reviewed journal articles.
- A relatively higher reliance on data from outside both the home institution and Australian government sources. Australia is behind many European countries in aggregating primary or clinical datasets and providing access to both aggregated or individual patient datasets. As a result, many Australian researchers source datasets from European countries (e.g. UK, Finland).

Specific challenges were described in both the discovery and access phases by a number of researchers using current search and access practices. These included: cost of data curation and sharing of clinical trials data; data linkage to medical records requiring individual negotiations with data custodians; complex processes for making data sharing and reuse agreements; lack of common data attributes, such as data dictionary and metadata standards for the purposes of making sense of data and data harmonisation; improvement of organisation of trials in registries linked to publications; and recognition of social licence for medical data and data governance are necessary conditions for data reuse.

The following is the summary along each activities of the data lifecycle:

Identifying data source and discovering data

- As part of domain expertise, most participants have good knowledge of their data sources. Social networks and word of mouth are the most used sources of identifying additional relevant data sources. Peer-to-peer engagement as well as senior researchers are important guides in the process of data discovery.
- A clinical trials research operations team includes clinicians, clinical research associates, data managers and statisticians. Some team members have a designated role of identifying data sources and datasets, and some negotiating with data custodians on data use agreement/licence.
- The role of identifying datasets is usually performed by subject librarians, who are trained to search by using sophisticated query construction operators and utilising standard controlled vocabularies, such as MeSH (Medical Subject Headings) terms.

- Data aggregators (clinical trial registries, government data, data repositories, search engines and paid services and data products) are consulted, depending on the research scope and available resources.
- In some teams, online search databases, such as; PubMed, MEDLINE and Embase, are routinely assessed for up to date clinical trials data published in peer-reviewed journal papers. Systematic searches within these databases are part of the meta-analysis and clinical guidelines development.
- Clinical Trial Registries, hospital records and personal health data, such as Pharmaceutical Benefits Scheme (PBS) data, are often consulted for the design of clinical trials. Researchers often conduct secondary data research, such as scoping and (Cochrane) systematic reviews, individual patient data (IPD) meta-analysis and network meta-analysis.
- Participants noted that Australia’s Clinic Trial Registries has information about trials, however, not all registered trials have associated datasets, and in many cases, there is a lack of information for those registered but not completed or dropped trials. Access to key datasets is arranged by individual negotiations, with data custodians, through complex data sharing and reuse agreements.
- Barriers to accessing data include (but are not limited to): finding contact information regarding data custodians or data managers; materials transfer and negotiation of data sharing agreements between institutions and across countries; data restrictions of confidentiality, such as data from clinical trials conducted by pharmaceutical companies and data from national medical records; and challenges with data linkage.
- Important data attributes in the clinical trial included: measurement, level of granularity, quantity and coverage; familiarity with data capture technologies; and the relative completeness and quality of metadata (i.e. data dictionary, standards, reliability and re-usability of metadata).
- In addition to providing details of data quality and integrity checks, notably using Cochrane risk of bias checklists, participants also describe additional processes to identify whether: the study received ethics approval; has been retracted from a journal; or has any indicator of potentially fraudulent data..

Data processing

- Participants describe data harmonisation from multiple sources of datasets involving a series of operations to conform to data structure, organisation and granularity, data integration and data manipulation.
- Data linkage of health records through a third party with de-identified data was noted as particularly complex and difficult to organise.
- Participants highlighted their documentation of data, including their development of metadata schemata and creation of data dictionaries. In most cases international standards were co-created or were in development through a community of specialists.
- Participants described working with others at the commencement of trials where possible to ensure data capture conformed to comparable standards where possible..

- The role of researchers' tacit knowledge was noted by the researchers as a less tangible but important aspect of approaching data processing in their particular context.

Data analysis

- Participants commonly used quantitative data analysis techniques and tools, such as statistical software packages (SPSS, SAS, R) and other analysis tools (Python, R markdown script, toolbox for MRI data analysis).
- Tasks performed include data modelling, data extraction, and model validation through a list of identified data analysis techniques, such as individual patient data (IPD) meta-analysis, network meta-analysis, and systematic review.
- In collaborative research, data analysis is usually performed with a common set of research questions and outcome measures. However, some challenges were noted with data transformation to a common scale for synthesis and specification of common data elements or data dictionaries.
- New technologies and data specificity were described which created ongoing considerations for interoperability, data formats and potential to integrate data. Magnetic Resonance Imaging (MRI) data was a common example where comparability of datasets was challenging due to upgrades in technology and software.

Data reuse

- Participants noted particular challenges in clinical trials data reuse included: the cost of data curation and storage; complex processes of arranging materials and data transfer agreements, and lack of data sharing culture.
- Data quality is critical for data reuse in terms of organisational data expertise, provision of data in suitable formats and data coverage.
- Data needs in view of the specificity of research objectives affect data selection criteria and data reuse.
- A meta-analysis team invests enormous resources in negotiating the data licensing of clinical trials data to be included in the analysis, or developing further research collaborations.
- Social licence for medical data and data governance are necessary conditions for data reuse.

Data sharing and publishing

- Participants emphasised sharing data through interpersonal interactions (social networks and word of mouth), data aggregators (data repositories and data archives), clinical trials registries and research literature in scholarly communication.
- While participants described the principle of publishing their data for public good, noting public investment in health research, they also noted the need to satisfy publishers´ requirements and meet institutional policy.
- Participants described the importance of bringing clinical trials researchers into their meta-analysis as collaborators. In this way they could agree on shared research questions and common outcome measures, as well as incentivise collaborators to engage as co-authors in published outputs.

- Arranging materials and data transfer agreements and ethics approval for the specific purpose of data sharing was noted by a number of participants as challenging.

Implications/Recommendations: The implications for the design and evaluation of data repositories and suggestions for service provisions of data repositories are presented as follows:

Data publishing and associate datasets to the registered clinical trial record:

1. Promoting datasets: Organising training workshops and research conferences for promoting benchmark datasets, their access and use methods, would be useful for reaching out to clinical trials researchers - noting they are receptive to pursuing interdisciplinary and collaborative research projects.
2. Publishing data papers: A published data paper is seen as valuable in linking the dataset with information about steps by which a dataset is collected and processed. Also, clinical trials researchers trust a dataset more if it has gone through an editorial review process.
3. Providing consistent guides (or standards) about data documentation and data dictionary: While clinical trials researchers will seek to identify any potential information on an identified topic, they also create specified search strategies. This indicates data repositories need to provide detailed documentation of data, high-quality metadata and data dictionary, both to enhance interoperability of metadata across repositories and help researchers to discover and integrate datasets from multiple repositories.
4. Enhancing provenance and common licence information: Data aggregators need to establish common licences for reducing negotiation time, including terms of data sharing, data use and reuse. Providing provenance information, such as data processing details and contact information of the data custodians and data managers is also of high value for clinical trial researchers.

Data discovery system supporting both novice and expert users in health studies:

5. Enrich and link metadata with sources external to a data repository/catalogue: Clinical trials researchers often create their own data repositories by scanning known sources for the most current clinical trials activities. Data repositories which reliably source research data from clinical trials registries, data papers, supplementary materials attached with journal papers, links to data processing codes published in other platforms are likely to be more valued by clinical trials researchers. Building a database which tracks data from journal papers, data papers and clinical trials registries would also be highly valuable.
6. Enhance user interface design by providing customised data formats: Data repositories can expand the sources of data as suggested by researchers, and build search interfaces that consider the researchers' needs of both data discovery and data processing. In most cases, the provision of raw data with specific structure, organisation and granularity, with minimum requirements of data dictionary is needed for clinical trials data reuse in meta-analysis and clinical guidelines development.

7. Enhance user interface design by understanding search behaviour, and supporting data searchers from novice to expert: Data repositories need to prioritise the system objectives of meeting user information needs when they are engaged with selecting and using a dataset. For example, enrich index and search interface with data variables, such as clinical metrics of outcome measures and methods.
8. Make metadata available together with datasets: As noted earlier, clinical trials researchers depend on adequate metadata when they analyse a dataset for assessing the fitness of the dataset and use conditions for their research project. In many cases, the data needs to be traced back to its source before it is fully utilised in clinical trials research.
9. Data paper and dataset metadata record linkage: Metadata records describing the reported data in data papers should be linked to the dataset in repositories for data discovery and reuse.

This project has been approved by the Monash University Ethics Committee with the application No. 30889.

Table of Contents

Acknowledgement

We acknowledge the first nations peoples of Australia, custodians of the land and sea upon which this research was conducted.

We extend our thanks to all the interview participants from across Australia who so generously shared their experiences and knowledge. This report represents a summary of the rich contributions which we hope to share in fuller detail in further publications.

To our partners, HeSANDA node network, our thanks for supporting and encouraging participants to take part in this project.

Wissen Insight (via Dr. Ying-Hsang Liu, ABN 69 892 903 501) and Monkey Gully Research (Dr. Megan Power) ABN 92 531 631 874) provided consultation services to this project.

1. Introduction

With the expansion of the open data movement in recent years, the issues of data discovery and data reuse have received more attention from various stakeholders, including (but not limited to); funding agencies, research organisations, researchers, data service providers, data practitioners and end-users. In response to growing global demand, many countries have invested in managing and cataloguing data, promoting data sharing and data citation, setting up infrastructures for streamlining data access and process. However, fewer resources have been invested to investigate the range of contexts of data discovery faced by researchers, which is a crucial factor influencing data discoverability and leading to data reuse.

As the primary agency to support shared research data resources in Australia, this project by the Australian Research Data Commons (ARDC) aims to fill this gap. The project focuses on characterising the contexts of data discovery both within and across disciplines to inform the design and evaluation of data discovery services. Our research questions are:

1. How do researchers approach data discovery?
2. How do researchers search for data?
3. What data attributes matter to researchers' data search?
4. What criteria do researchers apply for assessing relevance and usability of datasets?
5. What are the contexts of data reuse by researchers?

We find that there are various barriers to data access in the health research domain in Australia. In general, there is a lack of data sharing culture in the health research domain, partly due to the strict data sharing and reuse protocols of participant data implemented by the research ethics committees of research institutions, as well as the bureaucracies of accessing the medical records in different states. Specifically, data linkage of medical records requires individual negotiations with data custodians and there are complex processes for making arrangements of data sharing and reuse agreement across institutions.

2. Methods

This study has been designed to elicit the contexts of data discovery by adopting a mixed-method approach within the post-positivist research paradigm (Williamson, 2018). Specifically, we used the methods of survey and in-depth interview to understand the broader contexts of data discovery within people's information-seeking processes. A pre-interview survey was administered to learn more about the participant's background, including their research areas/topics, stage of career, job roles and their data sources in recent projects. This information was also used in the follow-up semi-structured in-depth interview.

A critical incident technique was used to elicit the contexts of data discovery by asking probing questions (Davenport, 2010; Flanagan 1954). The interview protocol was developed and organised by referencing stages of a data lifecycle.

2.1 Interview instruments

After a literature review, we piloted an initial interview protocol with three researchers and subsequently designed a semi-structured interview protocol which more closely followed each step of the data lifecycle (after Yong et al, 2021). These steps were:

- Collecting/Discovering/Accessing: Gathering data from surveys, censuses, voting or health records, business operations, web-based collections, and other relevant, accessible sources.
- Processing: Removing irrelevant or inaccurate information, reformatting contents to be interpretable by an analytic software, and otherwise validating the data collection.
- Analysing: Assessing the data collection with a goal of extracting insights about the issue they are studying.
- Using: Acting on the insights derived. These actions can affect data collected for future operations.

For the purpose of this study about data discovery context from researchers, we have added the following more granular stage/activity to the data life cycle to explore aspects of the research cycle:

- Data discovery: A researcher may engage with a data discovery system for discovering and identifying datasets that were collected by others but could be used for their research projects. In this report, we will reference data collected by others as secondary data.

The interview protocol structured by the data lifecycle provides us with a holistic view of how the research data has been managed from the perspectives of both data practitioners and end-users within research projects. Specifically, since user data discovery is an iterative process that involves the data attributes (i.e. the characteristics of data that users attend to when accessing data), it helps to characterise the information-seeking activities by the information objects sought at different stages of a research project. In the context of a collaborative research project that reuses data collected by other researchers or sources (i.e. performing secondary data analysis, such as systematic reviews and meta-analysis), it helps to characterise the context of data reuse by identifying the data attributes, which is one of the main research questions in this study.

For each step of the data lifecycle, we asked a few questions in relation to our research questions as discussed in the introduction. The full interview protocol is described in Appendix I. During an interview, we applied Critical Incident Technique (CIT) (Flanagan, 1954). CIT is a research method in which the research participant is asked to recall and describe a time when a behaviour, action, or occurrence impacted (either positively or negatively) a specified outcome[1]. We started an interview by asking participants to talk about

[1] Critical incident technique: https://en.wikipedia.org/wiki/Critical_incident_technique

a research project that required discovering and reusing data not generated by themselves. Interviewers gently guide participants step by step along the data lifecycle. The protocol is used by interviewers more as a guideline than a question-answering tool. Only when certain aspects were not covered by participants, interviewers then prompted a question, for example, if a participant jumped into and talked about a data source straightaway, an interviewer would ask: Could you tell us how you discovered that data source?

2.2 Pre-Interview Survey

We designed a pre-interview survey and invited participants to fill the survey ahead of their interview. The purpose of the survey is twofold: 1) to help researchers/interviewers to understand a participant's work context and to tailor interview questions accordingly, 2) to prepare participants/interviewees for the interview. The survey collected baseline information about each participant's profile, their research field of focus and provided an overview of their data discovery processes related to informing the interview process.

The pre-interview questionnaire, as attached in Appendix I, has two parts: the first part was to get participants' consent in participating and recording the interview, the second part asked: 1) a participant's background (employ organisation, research topic, job role), 2) information about a recent research project that includes a short description of the project, participant's role in the project, the funding sources of the project, the stage of the project and any related publication to the project, and 3) activities related to data (e.g. data scoping, data sourcing, data gathering, data transformation, data publishing, etc), types of data and sources to get data.
The pre-interview survey also provided a valuable prompt for starting a conversation with participants, setting up participants' expectation of an interview and for interviewers to be familiar with terminologies and acronyms from a research topic.

The survey was developed in GoogleForms and made available to participants upon confirmation of an interview date. A project overview and consent information was provided to participants via email at this point. The consent form for the project was incorporated into the survey form to simplify this process for the participants. All participants completed the consent form but one participant in the interview elected not to complete the pre-interview survey.

2.3 Recruitment of Participants

Participants were recruited directly via the ARDC. Groups were identified who had either participated in projects in relation to the HeSANDA Program or were known by the project lead to be engaged in clinical trials related activity. The participants were invited to participate in the study via direct invitation sent to the HeSANDA node network. In some cases the researchers for this project contributed to the recruitment process by directly contacting groups who were known or recommended by others. The interviews were all conducted via Zoom conference software with at least two members of the research team conducting the interview. Members of the HeSANDA team were observers for two interviews after gaining consent from the interviewees.

2.4 Data Analysis

Appendix I provide a summary of the pre-interview survey. We used the nVivo qualitative data analysis tool to annotate transcripts and summarise findings. The interview notes were also used to triangulate the research findings in this study.

3. Findings and Recommendations

This section summarises the findings from the analysis of the pre-interview survey and the annotated interview transcripts. We have included some quotes from the interviews to illustrate the Australian context as identified by the interviewees.

3.1 Summary from the pre-interview survey

The pre-interview survey aimed to support the interview process through an understanding of participants' profiles, key research areas and projects of focus. The survey also provided an indication of the participants' data discovery and secondary data use processes in relation to a nominated project. Each of these was explored in more detail through the interview.

Some findings of relevance from across the participant group which informed the analysis of the interviews included:

- An overall pattern of high reliance on both published literature and word of mouth via colleagues for data discovery, particularly at the starting point for a project. This was emphasised more notably by Early Career Researchers (ECRs) and Middle Career Researchers (MCRs).
- Engagement in sourcing information from conferences and seminars from both within and outside the participants' main field of interest.
- A relatively higher reliance on data from outside both the home institution and Australian government sources.
- An indication of role differentiation with doctoral and ECRs committing more time to their nominated projects and MCRs and (Late Career Researchers (LCRs) conducting a broader range of roles in relation to clinical trials research.
- A pattern of data activities which aligns with expectations for the data lifecycle for early-mid and mid-end stage projects.
- An apparent limited engagement with the data lifecycle during maturity phases of data archiving or data publishing activities.

3.2 Findings from the analysis of the interview transcripts

> "… the only other data that we deal with is registry data. So reporting for the registry … the team would get data exported from the IT managers or the system. We get all of that information as well as from …clinicaltrials.gov, because …we generally look at what's happening in Australia with a lot of these projects. So with those two registries we can cover about 95% of trials in Australia." (P2)

Interview participants include clinical trial designers, clinical/health guideline developers and secondary studies researchers. The participants hold varied job roles at different stages of their career. They are all engaged with collaborative research projects and have extensive experiences discovering data for the research purposes. Their data needs are triggered by specific work tasks within the workflows. Data attributes associated with data quality have been identified as a prominent requirement by researchers for data discovery and reuse.

Data search behaviour is characterised by searching well-regarded online databases, including PubMed, MEDLINE and Embase and discipline specific data repositories, as well as querying with advanced search strategies. Specific challenges include the cost of data curation and sharing in clinical trials; data linkage of medical records requiring individual negotiations with data custodians; complex processes for making data sharing and reuse agreement; lack of common data attributes, such as data dictionary and metadata standards for the purposes of making sense of data and data harmonisation; improvement of organisation of trials in registries linked to publications; and social licence for medical data and data governance are necessary conditions for data reuse.

Overall, the interview results have provided a detailed understanding of user data discovery from the perspectives of research data management, supplemented by work task analysis. In this section, we summarise the overall findings of work task analysis by the participant groups in Table 3.1 and provide some implications or recommendations for enhancing user experiences of data discovery.

Table 3.1: Work task analysis by clinical trial designers, clinical/health guideline developers and secondary studies researchers

Tasks	Clinical trial designers	Clinical/health guideline developers	Secondary studies researchers
Literature review	Scoping review Systematic review Meta-analysis	Scoping review Systematic review Narrative review	Systematic review Meta-analysis
Data discovery	Data sources: trial registries, PubMed, Embase for cohort observational registry data Search strategies development: Source of registry data Meta-analysis conducted before a trial	Data sources: trial registries, PubMed, Embase, trials listed in guidelines, GIN (Guidelines International Network) Search strategies development: Definition of topic scope and topic mapping Identification of gaps: Guidelines and recommendations used and adapted in Australia; Consultation with steering committee; ensure nothing relevant is missing Evidence summary: Sub-committees to develop (Search strategies development and evidence summary could be contracted out, with mutually agreed research questions, format, data searching and synthesis; Trial ID useful for tracking information)	**Meta-analysis** Data sources: trial registries, PubMed, Embase, preferably to the granularity of individual patient data (IPD), reach out to data custodians Search strategies development: Database searching; consultation with clinicians; search interviews Evidence review: Quality assessment, integrity checks, data extraction, data transformation, making sense of data Data analysis and synthesis: Cochraine review, RevMan for forest plot, evidence summary **Secondary studies** Access to source data requires ethics approval, negotiations with data custodians, complex data sharing agreements Data (pre)-processing is complex for some types of data, e.g., MRI data Data linkage: Be able to access multiple datasets, include data harmonisation;

Tasks	Clinical trial designers	Clinical/health guideline developers	Secondary studies researchers
			de-identified data linked by a third party; Re-identify data for data commons
Trial design	Collaboration by supervision Informed by research literature		
Trial study	Clinical research coordinator: Recruitment of participants Clinical research operations team (including data managers and statisticians): Measurement, trial analysis		
Guideline development	Work with guidelines group, reaching out to clinicians and doctors	Guideline process: steering committee and sub-committees (registries as important source of information of review process; including context experts, clinicians and clinician researchers) Make recommendations External review and stakeholders consultation	
User engagement	Patient outcome change	Final guidelines: Communication tool with stories (adapting to training context), living guidelines, risk factors algorithms	

3.2.1 How do researchers approach data discovery?

> “*... I guess from the guideline point of view, we're only looking at well - I say published trials, so we would pick these up from pubMed or from the preprint servers, and those are the two main sources that we use to identify trials.*” (P17)

Researchers’ data discovery journey is embedded in their information-seeking environments and specific work tasks. This is reflected in the use of multiple data sources for seeking data. The data sources include interpersonal interactions (social networks, conferences and collaborative research networks), data aggregators (government data, data repositories, search engines and online search databases), research literature in formal journal publications or grey literature from clinical trials registries. Access to clinical trials data and data linkage was frequently mentioned by the participants as the main data source.

However, there are also differences in the data sourcing requirements between the groups interviewed. Clinical trial designers perform meta-analysis before a new trial, while clinical/health guideline developers do a systematic or narrative review for specific topics. Secondary studies researchers seek clinical trials data from published articles, clinical trials registries and direct contact with data custodians for meta-analysis.

In the case of clinical trials researchers, finding eligible studies in the research literature for inclusion in meta-analyses and systematic reviews typically also involved content experts, health librarians and professional search services. Search interviews with content experts form part of the search strategies development. Evidence reviews and synthesis involves data integrity checks, quality assessment and data extraction. A common challenge across the study participants is the enormous resources invested for accessing the raw clinical trials data. As such, we recommend:

Recommendation: Data repositories can expand the sources of data as suggested by researchers, and build search interfaces that consider the researchers’ needs of both data discovery and data processing, such as provision of raw data with specific structure, organisation and granularity, with minimum requirements of data dictionary for clinical trials data reuse in meta-analysis and clinical guidelines development.

> “*... So you then start getting this knitted … study with multiple references that relate to that study and you use that information to then start extracting data and then to assess whether … you think [the trial’s] reliable or not in terms of the information that's being sent. So, yeah, we use all those multiple sources to make all those judgment calls.*” (P11)

Regarding data aggregation, our findings suggest that the clinical trials data from published research literature and clinical trials registries are important sources. However, there are some barriers for data access. For example, not all registered trials have associated datasets, and, in many cases, there is a lack of information for trials which are registered but subsequently discontinued or where data is not complete or where the originator cannot be traced, particularly in the case of dropped trials.

Access to key datasets is arranged by individual negotiations, with data custodians, through complex data sharing, materials transfer and data reuse agreements. Approaching data discovery by data aggregators is linked to the data processing since researchers need provenance information about the methods of data collection. This is also required for clinical trials researchers to make sense of data to support their tasks of data extraction, data harmonisation and evidence review and synthesis. Other important considerations include the documentation of data, metadata and data dictionary associated with data, as well as engaging with data custodians and data managers for data access or further research collaborations. That is, getting access to researchers' tacit knowledge about data processing, provenance and licensing information are very important considerations of reusing data.

Recommendation: Enhance provenance and licence information: Data aggregators need to establish licence of data use, including terms of data sharing, data use and reuse use and provide provenance information such as data processing details, and contact information of the data custodians and data managers.

Since clinical research literature in formal journal publications, as well as grey literature, such as research reports, are important data sources, researchers have found the data by using data repositories, domain specific databases, clinical trial registries and preprint servers. For example, the data source can be the reported number or statistical table in the published paper, links via the paper to a local repository or other data repositories when researchers perform meta-analysis. Publishing data as data papers (i.e. peer-reviewed descriptions about dataset) in high profile venues has demonstrated the usefulness of data for the research community, as well as the recognition of data curation efforts by researchers and data practitioners. Since data papers are assigned DOI (digital object identifier) and can be accessed and cited within the peer-review system, it has been considered a measure of research impact for stakeholder interests, including (but not limited to) governments, research funding agencies, universities and research institutions, researchers and data practitioners.

Recommendation: Enrich and link metadata with sources external to a data repository/catalogue: Data repositories can source the research data from clinical trials registries, data papers, supplementary materials attached with journal papers, links to trial ID and grant information published in other platforms and build a database of keeping track of data from traditional journal papers, and data papers and clinical trials registries.

3.2.2 How do researchers search for data?

> "*... So we did systematic searches of big databases like Medline and EMBASE, but we also searched trial registries - so clinicaltrials.gov and WHO's ICTR (WHO International Clinical Trials Registry). So because ongoing or planned trials are also able to be included in our meta-analysis.*" (P15)

Researchers' search behaviour of data discovery is characterised by searching well-regarded online databases, including PubMed, MEDLINE and Embase, discipline specific data repositories, as well as querying with advanced search strategies. In line with the findings of social scientists' literature research is integral to dataset search (Krämer et al., 2021), our findings suggest that as part of the research processes, the phases of data discovery and data processing are intertwined. For example, our findings reveal that important considerations of performing data processing include the documentation of data, metadata and data dictionaries, data integrity checks, data extraction and data harmonisation were also described as distinct processing stages, often requiring direct engagement with data custodians and data managers.

Since health research is typically a highly collaborative effort, some secondary studies researchers invite data custodians for further research collaborations. As such, researchers contact data custodians to help them make sense of data and contribute to data harmonisation by filling out data format templates or preparing data in prescribed formats. Reciprocal values can be seen for new collaborators in opportunities to co-publish, while the initiating researchers benefit from greater assurance of data provenance and quality.

Recommendation: Data repositories need to provide detailed documentation of data collection and processing processes, high-quality metadata and data dictionary, as well as approachable contact information for engaging researchers with data custodians and data managers.

> "*It tends to be that the research question will tend to dictate what sort of search ..., it's about teasing apart what aspect of a particular question they're interested in. You know, are they interested in the effects of drugs in terms of pain, or are they interested in some kind of qualitative assessment about you know patients' experiences of being in trials, or what aspect is it that they're really interested in. ...when I've got a clearer idea about what it is that they want, and then we'll often go through, kind of PICO framework*" (P17)

In this study most researchers have specific data needs for their work tasks. For instance, for clinical trial designers, they search for cohort observational registry data to perform a meta-analysis before a trial. Clinical/health guideline developers search for existing guidelines used and adapted in Australia and trials listed in the guidelines to identify the gaps and ensure that nothing relevant is missing. In some cases, search strategies development and evidence summary are contracted out, with mutually agreed research questions, format, data searching and synthesis. Secondary studies researchers who conduct meta-analysis try to find trials data, preferably to the granularity of individual patient data (IPD).

Other secondary analysis studies have used MRI data, medical records and Pharmaceutical Benefits Scheme (PBS) data, The latter which typically includes trials linked to commercialisation of new therapeutics, requires ethics approval and negotiations with data custodians, with complex materials transfer, data sharing and reuse agreements.

Some analyses require data linkage with multiple datasets, which is achieved by de-identified data linked through a third party. Re-identifying data for data commons is proposed as a possible solution to eliminating the barriers to data reuse. Some participants expressed frustration with the lack of data sharing culture in health studies in Australia, particularly when compared to open data sources internationally. In line with the findings of a recent report of digital health data in Australia, we also suggest “immediate, short-term solutions [are needed] to deliver tangible support to harmonise data and information governance” (Frean et al., 2023, p.18).

The study highlights clinical trials researchers who have established advanced search strategies in order to query specialised online databases and trial registries by well-trained information professionals. In addition to the searcher’s familiarity with the field, content experts are consulted for the development of search strategies by using the PICO framework for structuring the clinical questions in search interviews. In formulating the queries, searchers look up the database indexes and do simple searches to obtain a few relevant articles. Participants are also routinely performing search tasks and tracking up-to-date information, such as logging double-blind randomised controlled trials via API and Python scripts, to create local databases and trial registries for future reference.

Importantly, a Google-style search interface will not suit the needs of professional searchers for clinical trials related data access given the specificity of developing advanced search strategies and monitoring most up-to-date published trials. Importantly, such search strategies are also documented and published as part of the research processes.

Since clinical/health researchers are embedded in a collaborative research environment, professional networks and interpersonal interactions are also important for data discovery.

Table 3.2 presents a list of data sources organised by types of data sources.

Recommendation: Data repositories can expand the sources of data as suggested by researchers, and build search interfaces that consider the researchers’ needs of both data discovery and data processing, such as provision of raw data with specific structure, organisation and granularity.

Table 3.2 Types of data sources consulted by the participant groups

Types of Data Sources	List of Data Sources
Clinical trial registries	● ClinicalTrials.gov ● WHO International Clinical Trials Registry Platform (ICTRP) ● Australian New Zealand Clinical Trials Registry (ANZCTR) ● ANZDATA (Australia and New Zealand Dialysis and Transplant Registry) ● EU Clinical Trials Register ● Cancer registries ● Trauma registries ● Concussion registries
Online databases	● PubMed ● MEDLINE ● Embase ● ERT trials clinical reports
Data repositories	● DNA repositories ● TB (Tuberculosis) GAP (Genomic Analysis Portal) data ● UK Biobank ● GenBank ● NeuroVault
Professional organisations	● American Society of Clinical Oncology (ASCO) hub ● Guidelines International Network (GIN) ● Australian & New Zealand Neonatal Network (ANZNN) annual reports ● PROSPERO (International Prospective Register of Systematic Reviews) ● NICE (National Institute for Health and Care Excellence) ● Cochrane ● NMA (National Medical Association) ● ANZGOG (Australia New Zealand Gynaecological Oncology Group) ● International Cancer Genome Consortium ● Sax Institute (45 and Up Study)

Government data	● WA Health, Government of Western Australia ● NSW Health ● Australian Bureau of Statistics ● Therapeutic Goods Administration (TGA) trial reports by pharmaceutical companies ● Medicare for electronic medical records ● Pharmaceutical Benefits Scheme (PBS) data ● Primary Health Networks ● Electronic Medical Record (EMR) of acute care ● Medicaid data (USA) ● Danish Health Authority ● Department of Veterans Affairs
Search engines and archives	● Google ● Preprint servers
Professional services	● External contractor (professional search services for information searching and synthesis) ● Allscripts electronic health records system
Social networks	● Affiliated research team and collaborators ● Clinical trialists

3.2.3 What data attributes matter to researchers' data search?

> "*... The unique aspect of that (Australian National) dataset was it was huge. So at the time, it's like over hundreds and thousands of people so … the opportunity was massive because you could actually investigate this question at a large scale as opposed to, you know, prospectively collecting this data, which obviously will take a lot of time and effort.*" (P9)

In this study, we identify the data attributes by the broad categories of: 1) Data needs; 2) Data quality and integrity; 3) Metadata and documentation; 4) Data access. Specifically, data attributes can also be characterised in terms of measurement, level of granularity, quantity and coverage.

Data needs

Clinical trials researchers try to find data with common outcome measures for meta-analysis. Since data discovery is linked to specific research objectives, the important research considerations are whether the data is fit for purpose. Participants described a need for data that is fit for purpose in terms of their research scope, the study population and the clinical treatments or other interventions being studied.

Data quality and integrity

In conducting meta-analysis and developing clinical/health guidelines, double-blind randomised controlled trials are important sources of information. However, participants also described looking for any examples of clinical trials globally which met their specifications or which could inform their study. In most cases data quality was benchmarked in relation to whether the data was published or directly linked to publication standards in peer-reviewed journals. However participants also noted additional processes linked to assessing data integrity , to assure that data collection followed standard (ethical) procedures and to determine whether data was suitably accurate or potentially fraudulent.

Metadata and documentation

Availability of detailed data descriptions, metadata and data dictionaries were also important for participants' ability to make sense of the data. Accessing the raw data, preferably at the level of individual participant data (IPD) level is preferable because it can enhance the power of statistical analysis. Researchers also emphasise the access issues of individually negotiating data licensing agreements for data sharing and reuse, for example, dealing with multiple data custodians and data linkages of de-identified data.

Since researchers use their domain expertise to assess the quality of data, they follow the protocols by conducting integrity and statistical checks to detect the anomaly and possible fraud of data.

Regarding metadata and documentation of data, high-quality metadata consists of the following attributes: data dictionary, reliability and re-usability of metadata, data access conditions and data licence, documentation about data collecting process and whether the standard classification systems, such as the International Classification of Diseases (ICD) are followed. Since a data dictionary serves the function of filtering out the variables that researchers are not interested in or help researchers find the specific variables of interest in data discovery, it can be used to design search systems to support user search behaviour of querying with variables in data repositories[2]. The reliability and re-usability of metadata are concerned with the feasibility of data and associated metadata for data reuse and data linkages. To enhance the reusability of data, meta-analysis researchers have provided suggested coding forms, data dictionaries or co-development of common outcome measures when they try to obtain clinical trials data in collaborative research efforts.

Access

Finally, whether the data custodians or data collectors can be contacted is important for learning more about the data for data reuse, which may not be detailed in the metadata and documentation of data. This information is useful for further data processing and analysis, such as outcome harmonisation and evidence review in systematic reviews and meta-analyses.

Recommendation: Data repositories need to source high-quality primary data for data reuse by enabling querying with variables, providing user guides with metadata in compliance with community standards and updated contact information of data providers.

3.2.4 What criteria do researchers apply for assessing relevance and usability of datasets?

> "*So once we determine that something is eligible for inclusion, we will contact [the research team] and try and get them to share their data .... It's probably what we're going through now with a lot of these projects and looking at … standard risk of bias checks which changes for aggregate data and individual participant data. But we also are starting to do a lot of integrity checks. … there's a lot of information coming out about statistical checks you can perform to ensure that data has not been fabricated by them and come to ensure randomness of allocations and things like that that improve the reliability of clinical trials.*" (P2)

As discussed in the previous section, researchers take data attributes into account in their data search. Depending on the search contexts, researchers have applied these data attributes as selection criteria for assessing the relevance and usability of datasets.

Comparing the data attributes identified in the study and the information most important to properly use or select a dataset in the European Open Science Cloud (EOSC) data quality attributes (Lacagnina et al., 2023), we map the shared selection criteria to the data discovery contexts:

[2] https://www.icpsr.umich.edu/web/pages/ICPSR/ssvd/ is a good example of search/compare variables in social science research.

Table 3.3: Relationship between the selection criteria and data discovery contexts

Selection Criteria	Data Discovery Contexts
1. User guide (including a description of size, structure, abstract, typical usage, production methodology, dictionary)	● Making sense of data ● Proposal of sourcing data from repositories
2. Scientifically accurate (e.g. validated against reference, plausible)	● Quality checks in meta-analysis
3. Licence of use, including terms of use	● Access to individual participant data from journal publications ● Access to data from trial collaborators ● Data aggregation and linkages ● Data publishing (data availability statement for reused data) ● Data sharing across institutions
4. Version	● Data analysis and documentation (MRI data) ● Data publishing (sequencing data) ● Data collection (recruiting patients based on the ICD coding)
5. Data dictionary	● Proposal of sourcing data from repositories and data custodians (e.g. obtaining individual patient data with x number of variables) ● Data collection in meta-analysis (reaching out to potential collaborators for individual participant data) ● Data analysis (e.g. common data elements in meta-analysis) ● Data curation by data managers ● Data discovery (to download data or not)

6. Clarity about how to cite the dataset and availability of its product locator, like URL or DOI	● Data publishing (data availability statement for reused data)
7. Archiving policy	● Data sharing protocols (strict protocol and share within research team) ● Access to data from repositories (e.g. TB GAP from NIH) ● Data processing via virtual desktop (e.g. MRI data)
8. Compliance of metadata with community standards	● Data processing (e.g. data cleaning task) ● Data collection (sourcing of MEDLINE records for systematic reviews) ● Data collection in meta-analysis (reaching out to potential collaborators for individual participant data) ● Access to data from trial collaborators ● Data sharing across repositories
9. Provenance and traceability information	● Data collection in meta-analysis (reaching out to potential collaborators for individual participant data) ● Trials listed in guideline documents
10. Information about the data provider and point of contact	● Data collection in meta-analysis (reaching out to potential collaborators for individual participant data) ● Access to the original data via data custodians ● Proposal of sourcing data from repositories and data custodians (e.g. clarifying certain drugs in PBS data)
11. Format according to community standards	● Data analysis (e.g., coding to prescribed data format in meta-analysis) ● Data analysis (e.g., MRI data from different brands of machines) ● Data collection in trials (e.g., records by nurses and data managers) ● Guidelines in summary format for communication
12. Data are complete in space and time, adequate resolution (if applicable)	● Selection of eligible studies in meta-analysis ● Pre-processing of MRI image files

13. Up-to-date/currency, timeliness/novelty	● Selection of eligible studies in meta-analysis ● Systematic reviews in clinical/health guidelines development
14. Details on strengths, limitations, known issues, including availability of uncertainty information	● Making sense of data
15. Technically correct/having passed sanity checks (e.g., no unexpected gaps in the time series, consistency between data and metadata)	● Data checks in meta-analysis (e.g., statistical checks for data integrity and randomness of allocations)
16. Evidence of regard to ethical conduct, protection and confidentiality of data	● Human subjects ethics approval ● Clinical trials registries ● Linkage of de-identified data by a third-party
17. Evidence of data reuse	● Trials listed in guideline documents or living clinical guidelines ● Citations in journal publications

Our research findings suggest that the line between using and selecting a dataset is not clear, given the complexities of research contexts. Other factors at play in the way participants select and use datasets include the research environment, career stage, job roles and the nature of research projects (e.g., scoping, multidisciplinary/interdisciplinary and collaborative). Using and selecting a database for data discovery is also an iterative search process coupled with consulting multiple resources, as shown in the data sources (See Table 3.3). The identified selection criteria and the search context can be further investigated to determine the relative importance of these selection criteria in different stages of a research project and phases of a data lifecycle.

Recommendation: Data repositories need to prioritise the system objectives of meeting user information needs at the level of using and selecting a dataset by the identified data attributes. Using a dataset involves complex processes of performing secondary data analysis in a research project.

3.2.5 What are the contexts of data reuse by researchers?

> "*... Context and correct coding, so comprehensive code books are invaluable. Sometimes examples of … protocols, … facilitator information sheets that have been provided … all of these non-published documents that are really important to [our] understanding … could be the make or break between including or excluding an outcome.*" (P2)

In our previous study (Liu, Wu, Power, & Burton, 2022), we identify several dimensions of contexts that enable the data reuse, including 1) Anticipation of data reuse; 2) Expertise of managing data; 3) Provision of data; 4) Data coverage; 5) Data interpretation; 6) Data needs, and 7) Model validation. These dimensions of contexts of data reuse are applicable to the health domain in this study. However, we discuss the specific implications from the perspectives of the three main groups of health/clinical researchers, namely clinical trial designers, health/clinical guideline developers and secondary studies researchers.

Our findings suggest that the reuse of clinical trials data is anticipated by clinical trial designers since they have reached out to clinicians and clinical researchers in their work tasks. Health/clinical guideline developers have extensively drawn on the clinical trials data and published articles for systematic reviews in the guideline development process. For secondary studies research, such as systematic reviews and meta-analysis, the clinical trials data at the granularity level of individual participant data is highly sought-after to increase the statistical power of the analysis. Overall, these findings show that curated clinical trials data is crucial for health/clinical research for all the three groups of researchers, but there are usually limited resources left for data curation when a clinical trials study is finished.

As revealed in the study, data reuse is a social process. Researchers use multiple data sources for collecting and discovering data through their social networks, data aggregators and research literature. In many cases the research is conducted as, or they are part of an international consortium of collaborative research projects.

Our findings reveal that data needs, in view of the specificity of research objectives, affect data selection criteria and data reuse. Following the research objectives, the study scope is concerned with defining the boundaries of topics. The data selection criteria are reflective of the PICO framework, which is commonly used for structuring clinical questions: **P**atient or problem; **I**ntervention or exposure; **C**omparison or control and **O**utcome(s). The research findings show that the ICD (International Classification of Diseases) code is crucial for identifying the research problem in a study, and common outcome measures are important for data harmonisation in systematic reviews and meta-analysis.

The categories of data interpretation, data needs and model validation are concerned with the researchers' interaction with data, relevant to the aforementioned data quality metrics of 1) scientifically accurate (e.g. validated against reference, plausible); 2) details on strengths, limitations, known issues, including availability of uncertainty information, and 3) technically correct/having passed sanity checks (e.g., no unexpected gaps in the time series, consistency between data and metadata). In the context of health domain research, data checks in meta-analysis also involve statistical checks for data integrity and randomness of allocations. In larger research groups, a specialist evidence review team performs qualitative assessment by following the standard procedures such as those outlined in the Cochrane review.

The anticipation of data reuse is relevant to the archiving policy of data repositories and how the data can be curated to meet the needs of researchers. Within a collaborative research environment, there are data sharing protocols in place within the research team. However, there are barriers to materials transfer and data sharing between institutions and across countries. The issues of data restrictions of confidentiality and data linkages of de-identified data are raised by the study participants. Overall, our findings suggest that data repositories have facilitated the use of existing datasets for secondary data analysis in some disciplines in the health domain, such as NeuroVault for neuroimaging studies and Tuberculosis (TB) Portals data for genomic analysis. Data reuse can be facilitated by eliminating the barriers to accessing the raw clinical trials data and fostering data sharing culture within the health research community.

Recommendation: Data repositories need to re-imagine their roles within the changing scholarly communication system by integrating or collaborating with the stakeholders in the changing scholarly communication landscape.

References

Carter, P., Laurie, G. T., & Dixon-Woods, M. (2015). The social licence for research: Why care.data ran into trouble. *Journal of Medical Ethics*, *41*(5), 404–409. https://doi.org/10.1136/medethics-2014-102374

Davenport, E. (2010). Confessional methods and everyday life information seeking. *Annual Review of Information Science and Technology*, *44*, 533–562. https://doi.org/10.1002/aris.2010.1440440119

Flanagan, J. C. (1954). The critical incident technique. *Psychological Bulletin*, *51*(4): 327–359. https://doi.org/10.1037/h0061470

Lacagnina, C., David, R., Nikiforova, A., Kuusniemi, M. E., Cappiello, C., Biehlmaier, O., Wright, L., Schubert, C., Bertino, A., Thiemann, H., & Dennis, R. (2023). Towards a data quality framework for EOSC. https://doi.org/10.5281/zenodo.7515816

Liu, Y.-H., Wu, M., Power, M., & Burton, A. (2022). *Elicitation of data discovery contexts: An interview study (1.0)*. Zenodo. https://doi.org/10.5281/zenodo.7179526

Frean, I., Belgard, M., Zeps, N., Boyd, J., Shaw, T., Cavedon, L., & Gray, L. P. (2023). Digital transformation of healthcare in Australia constrained: A call to action for a national data governance framework. Accessed on March 7, 2023 at: https://digitalhealthcrc.com/wp-content/uploads/2023/02/DHCRC-Call-to-Action-for-a-National-Data-Governance-Framework_Feb-2023_FINALDESIGNED.pdf

Williamson, K. (2018). Research concepts. In K. Williamson & G. (Eds.), *Research Methods: Information, Systems, and Contexts* (2nd ed., pp. 3–25). Chandos Publishing. https://doi.org/10.1016/B978-0-08-102220-7.00001-7

Yong, A., Zahuranec, A. and Verhulst, S. (2021). A layered approach to documenting how the third wave of open data can provide societal value. Accessed on March 7, 2023 at: https://opendatapolicylab.org/articles/the-onion-model-a-layered-approach-to-documenting-how-the-third-wave-of-open-data-can-provide-societal-value/index.html

APPENDIX I - Analysis of the Pre-Interview Survey

The pre-interview survey is available at:
https://drive.google.com/file/d/1aeG_V-Vi3ZcwrYeq7IkbY7pErYH0WSdx/view?usp=share_link

1. Participant profiles

The total number of participants recruited was seventeen (17). Of these, sixteen (16) provided responses to the pre-interview survey and these are included in this analysis. The participants' main university affiliations were based in: New South Wales (7), Victoria (2) Queensland (1), Tasmania (1) and Western Australia (3). Participants represented seven (7) universities. Five (5) participants were drawn from a particular team and this provided an in-depth view of the data handling roles in the complex and specialist environment of clinical trials reviews. Note that a number of participants worked in specialist units and research centres and/or were affiliated with more than one university.

Participants were also drawn from a range of specialist areas of clinical health and health sciences. An Australia New Zealand Field of Research (FoRs) coding was applied based on a participant’s nominated project of interest to look at this breadth of representation. The following fields were identified; Community Child Health (2), Endocrinology (1), General practice (1), Gerontology (1), Immunology (1), Nephrology (1), Neurology (3), Oncology (2), Paediatrics (3), Primary Care (1) and Public Health (1). These were high level allocations based on the researcher’s interpretation of the field which was most discussed by the participant.

Participant roles were of interest to the study in gaining an overview of who engaged with data searching activities. In particular, participants were asked to indicate the ways they interacted with clinical trials data. The three options provided in the survey were those suggested during the recruitment of participants and were expected to be reaffirmed. These were: RO1 a clinical or health guideline developer; RO2 a research analyst using clinical trials data; RO3 a clinical trials designer or other. Participants were also asked to indicate their career stage through a suggested list and to indicate the period of time that they had been working at that particular career level.

The recruitment was successful in engaging participants from across the spectrum of research experience from current doctoral candidates (doctoral), to early career researchers (ECRs), mid-career researchers (MCRs) and late career researchers (LCRs). The participants also included two senior professional staff (LCPs) with research backgrounds who are noted here for their specific roles in specialist programme and information management. However, note that even doctoral candidates interviewed were not 'new' to the field of clinical research and had either long periods of experience working in the field of clinical trials research and/or were also qualified medical graduates with some research specialism in their respective field of interest.

As we would expect participants at different career levels to either have a growing level of expertise, or perhaps a greater breadth of expertise overtime, we compared responses on role types to the career phase of participants. As can be noted from Table 1 below, for the participants contributing to this study it was more typical for the early career group; doctoral researchers (2) and ECRs (5) to be analysts. However, two (2) ECRs were also engaged in clinical health guideline development. The later career group: Mid (3) or late career researchers (3) or late career professionals (2) were more typically engaged in clinical health guideline development, while retaining a level of research analysis. Participations who were grouped as later career professions were less definitive about their roles in relation to this question.

Table 1: Participant Career Stage and Roles related to clinical trials research

			Career Stage					
RO1 *Clinic/Health guideline developer*	*RO2* *Research Analyst*	*RO3* *Clinical trial designer*	**Doctoral**	**ECR**	**MCR**	**LCR**	**LCP**	**Grand Total**
			1					**1**
	YES		2	2	1			**5**
	YES	YES		1				**1**
YES				1		1	1	**3**
YES	YES			1	1	1	1	**4**
YES	YES	YES			1	1		**2**
			3	**5**	**3**	**3**	**2**	**16**

2. Funding profile for participant nominated projects

Participants were asked to indicate their time commitment for a project of focus. Most participants (12) committed either one day per week (7) or two days per week (5) to the nominated project, with four (4) committing three or four days to the project. This data was compared to participants' identified funding sources to see whether potentially larger projects allowed for a higher time commitment.

Federal funding, such as via NHMRC, was the predominant funding source with 7 projects identified as being wholly funded through Federal sources and a further four receiving a mix of Federal and other funding. The receipt of Federal funding did support 3 participants with higher time commitments to their projects. Three (3) participants indicated that they were either self-initiating projects which they worked on within their usual work hours or were undertaking work, such as Cochrane reviews, which were part of their professional activities and were often conducted outside their nominal work hours.

We also looked at whether career stages of participants affected either the type of funding sources and ability to commit more time to their projects. Six (6) participants who were beneficiaries of Federal funding were undertaking doctoral studies or were early career researchers (ECR), compared with four (4) more senior researchers. Those in earlier career stages were generally devoting more time to their nominated project than more senior researchers or professional staff. This suggests a research workforce effect where funding can be directed to support researchers in early careers to focus on particular projects while more senior researchers have a wider portfolio of projects being managed.

Table 2: Funding profile by career stage and work commitment

CAREER STAGE	*F1 Australian Federal gov*	*F2 Australian State gov.*	*F3 Internal Organisation*	*F4 International*	*F5 NGO*	*F6 Commercial*	*F7 Philanthropic*	*F8 Self-initiation*	*F9 Other*
Doctoral								YES	
			YES			YES			
	YES			YES					
ECR	YES								
		YES	YES		YES				
MCR			YES						
	YES								
			YES						
LCR									
	YES								
LCP									YES
	YES	YES					YES		

3. Project stage and starting points for sources of information

Most participants (11) provided information on a current project which was either at its midpoint (6) or in the writing up phase (5). The stages for other projects ranged from scoping (1) to closed (2), with one (1) project described as ongoing and one (1) project self-described as being in an analysis stage.

Participants were asked to select any information sources they accessed as starting points for their project from a list provided. As projects were at different stages we hoped participants could reflect on sources of importance in relation to their projects overall for discussion at the interview. The profile shown above indicates a diverse range of sources drawn on by participants.

Participants were provided choices for information and data from either within their own field of interest or focus and more widely. The most significant source of information selected from within their field of interest was via literature searches (13), asking colleagues (11) and via conferences and seminars (6).

However, participants also looked to colleagues and conferences outside their field (4) as well as literature searches outside their direct field of focus (3). In relation to the locations participants were most likely to look for sources, slightly more looked outside their immediate organisation (5) than within their organisation (4). Perhaps surprisingly there was limited engagement with data sources from Australian government organisations, either Federal government (2) or State governments (1). Unusually, even participants working on projects which received State funding did not include the State government as a source of information or data.

Participants were asked to select any information sources they accessed as starting points for their project from a list provided. As projects were at different stages we hoped participants could reflect on sources of importance in relation to their projects overall for discussion at the interview. The profile shown above indicates a diverse range of sources drawn on by participants.

Participants were provided choices for information and data from either within their own field of interest or focus and more widely. The most significant source of information selected from within their field of interest was via literature searches (13), asking colleagues (11) and via conferences and seminars (6).

However, participants also looked to colleagues and conferences outside their field (4) as well as literature searches outside their direct field of focus (3). In relation to the locations participants were most likely to look for sources, slightly more looked outside their immediate organisation (5) than within their organisation (4). Perhaps surprisingly there was limited engagement with data sources from Australian government organisations, either Federal government (2) or State governments (1). Unusually, even participants working on projects which received State funding did not include the State government as a source of information or data.

Table 3: Sources of information and data by career and project stage

											Project Stage				
CAREER STAGE	*S1 Asking colleagues in FoI*	*S2 Via conferences/ seminars in FoI*	*S3 Colleagues or Conf outside my FoI*	*S4 Data sources within my institution /org*	*S5 Data sources outside my institution/org*	*S6 Lit searches within FoF*	*S8 Internet search*	*S9 Fed Gov sources*	*S10 State Gov sources*	***S11 Other***	**Scoping stage - Midpoint**	**Writing up-Analysis**	**Ongoing**	**Closed**	**Grand Total**
Doctoral		YES				YES					1				**1**
	YES					YES					1				**1**
	YES		YES	YES	YES	YES						1			**1**
ECR	YES					YES					1				**1**
	YES				YES	YES		YES	YES			1			**1**
	YES		YES			YES					1				**1**
	YES	YES				YES				YES		1			**1**
	YES	YES	YES			YES						1			**1**
MCR	YES					YES						1			**1**
	YES	YES				YES					1				**1**
	YES	YES	YES	YES		YES	YES				1				**1**
LCR						YES								1	**1**
					YES									1	**1**
										YES	1				**1**
LCP					YES	YES		YES					1		**1**
	YES	YES			YES							1			**1**
											7	**6**	**1**	**2**	**16**

Table 4: Data activities by project stage

										Project Stage						
DA1 Data scoping	*DA2 Data search*	*DA3 Data access*	*DA5 Data cleaning*	*DA6 Data transformation*	*DA7 Data curation*	*DA8 Data archiving*	*DA9 Data sharing*	*DA10 Data publishing*	*DA11 Data other*	Scoping stage	Midpoint	Analysis	Ongoing	Writing up	Closed	**Grand Total**
									YES						2	**2**
			YES	YES										1		**1**
		YES	YES		YES		YES					1				**1**
				YES	YES						1			1		**2**
	YES												1			**1**
	YES		YES	YES							1			1		**2**
	YES	YES	YES	YES	YES		YES							1		**1**
YES		YES	YES	YES	YES		YES	YES						1		**1**
YES	YES									1						**1**
YES	YES	YES	YES	YES			YES	YES			1					**1**
YES	YES	YES	YES	YES	YES						2					**2**
YES	YES	YES	YES	YES	YES		YES				1					**1**
										1	**6**	1	1	**5**	2	**16**

**Within 3 months of the survey response*

4. Project stage and data lifecycle phases

Participants were asked to indicate the types of data activities (DAs) they had undertaken in the last three months at the point of the survey. This question was to help guide the interview to focus on data activities that were most common as well as to understand under-developed activities.

Project stages identified by participants were predominantly either in their midpoint (6) or writing up and analysis stage (6). These are further illustrated in Figure 1 below to demonstrate the changing emphasis across the data lifecycle. This indicates that participants are using a wide range of data handling skills, with data sharing and publication coming more into focus as projects mature. For the purpose of this analysis, an individual researcher working at the scoping phase is included with the midpoint group.

An important indication from this data is that no participants indicated that they had undertaken any data archiving activities (see DA8) and only one participant indicated they had considered data publishing. This could have direct relevance for the potential success of any repository that depends on effective archiving practices. One respondent, whose project was ongoing, similarly only recorded data searching as a data lifecycle activity.

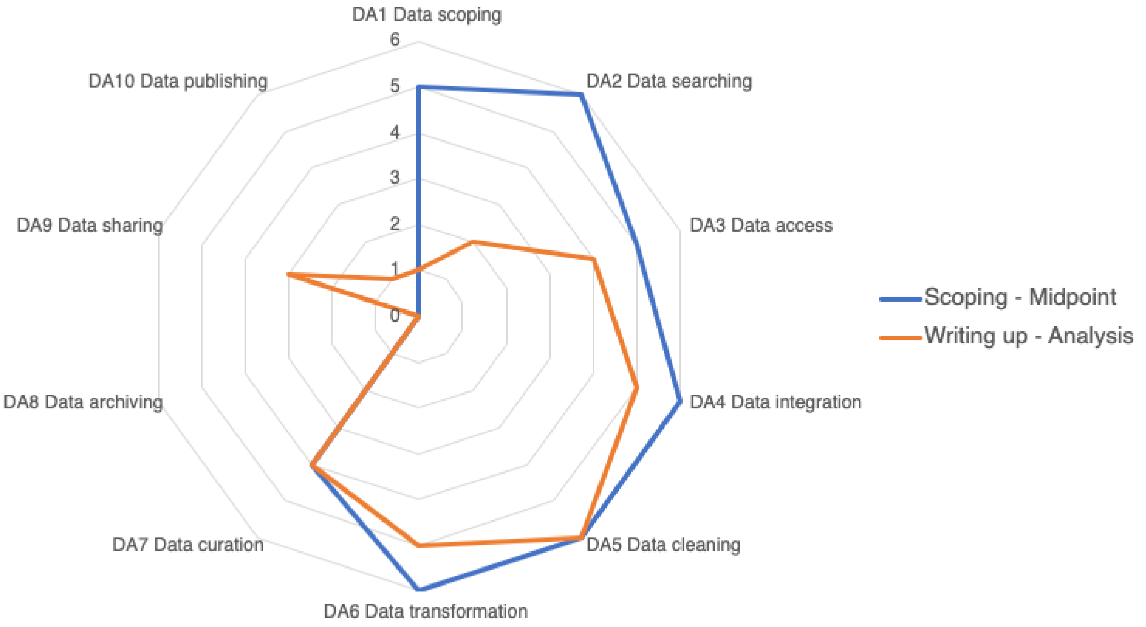

Figure 1 Data lifecycle activities for two key project phases (n=11)

5. Data types selected by Field of Research (FoR)

Participants were asked to indicate the types of *secondary* data sourced in relation to their field of study / nominated project. Participants could select any types of data from a list developed by the researchers based on initial consultation through the HeSANDA project.

We looked at this data in relation to the field of research most closely linked to the nominated project. As shown in Figure 2 below, since this study focussed on participants involved in clinical trials research, this particular data type was the most prominent option selected (8 examples).

The output also indicates that participants for this project were wholly focussed on clinical trials data in relation to paediatrics, oncology, primary care and public health. Those engaged in general practice, gerontology, immunology and neurology indicated they were seeking a range of data types. In one case the researcher was still scoping a project and had not yet identified data types.

For those indicating other types of data not listed, two noted that they were using their own data rather than selecting secondary data and one noted they were using data from research cohorts from both Australia and internationally. These points were explored with the participants in more detail in the interviews.

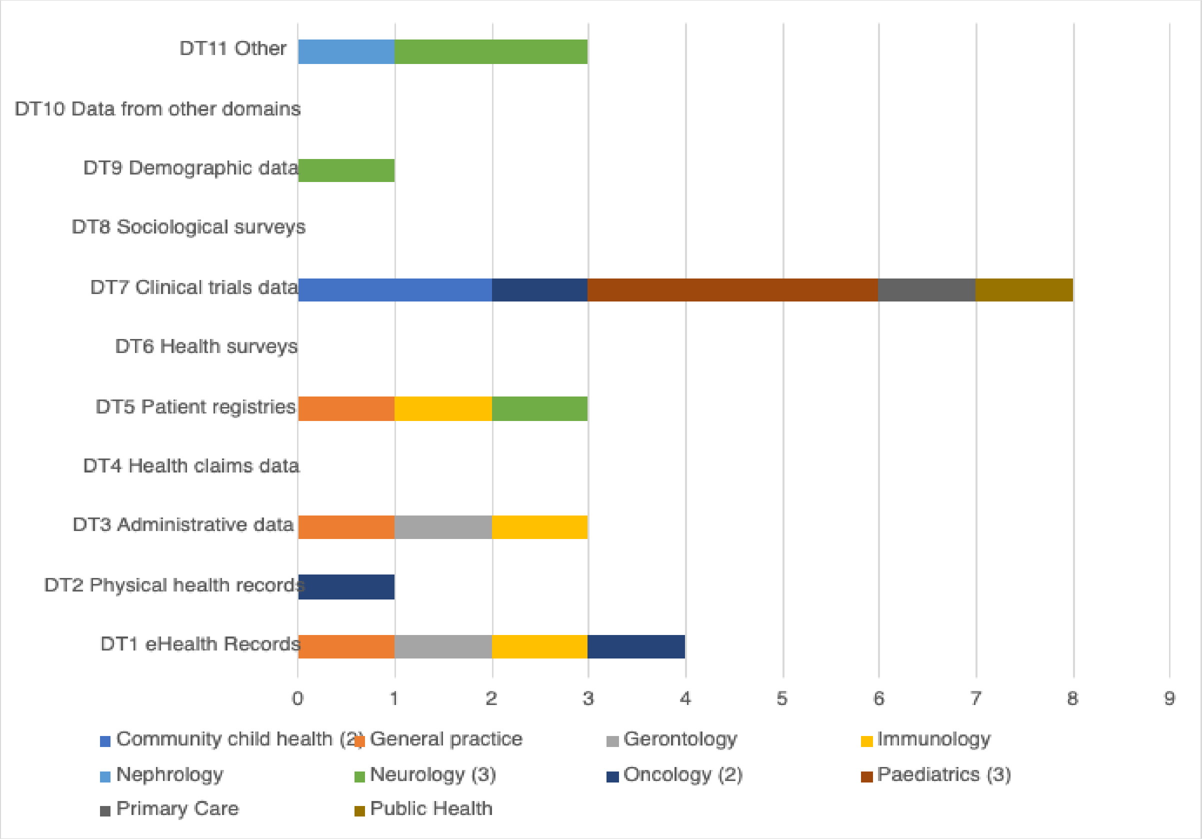

Figure 2: Data types of interest by main field of research

6. Summary

The pre-interview survey aimed to support the interview process through an understanding of participants’ profiles, key research areas and projects of focus. The survey also provided an indication of the participants’ data discovery and secondary data use processes in relation to a nominated project. Each of these was explored in more detail through the interview.

Some findings of relevance from across the participant group which informed the analysis of the interviews included:

- An overall pattern of high reliance on both published literature and word of mouth via colleagues were important at the starting point for a project, potentially more so for early to mid-career researchers
- Engagement in sourcing information from conferences and seminars from both within and outside the participants’ main field of interest
- A relatively higher reliance on data from outside both the home institution and Australian government sources
- An indication of role differentiation with doctoral and ECRs committing more time to their nominated projects and MCR and LCRs conducting a broader range of roles in relation to clinical trials research.
- A pattern of data activities which aligns with expectations for the data lifecycle for early-mid and mid-end stage projects
- An apparent limited engagement with data lifecycle maturity through data archiving or data publishing activities
- An apparent focus on clinical trials review vs dependence on broader types of data in different fields of research

The interviews provided an opportunity to gain more detail on participants’ survey responses and understand these in context. For example, the highly globalised nature of some fields of research means that primary sources of data are often only available from international repositories. The indications from the survey that researchers actively seek quality publications with reliable data attached is not mirrored by their engagement with data archiving or data publication was of particular interest. This was further explored in the interviews and the full report discusses this finding and the potential implications for establishment of an Australian clinical trials research catalogue.