Guide to Applying for Linked Data

Combined Protocol and Application Form

SECTION 1 Project Details

- Please cite your CHeReL project ID number. This is generated when you submit an Expression of Interest via CheckApp.

- If you have applied for data from other organisations or jurisdictions, you will receive other reference numbers. These should be included to assist with identifying your project when we discuss your project with these other parties.

- For multi-jurisdictional projects involving data from all other states other than ACT and the Commonwealth (Australian Institute of Health and Welfare; AIHW), please consult the Population Health Research Network (PHRN).

- Research studies utilising linked data must be approved by a ***Specialist Data Linkage*** Human Research Ethics Committee (HREC). Studies using linked unit record data comprising personal and health information from only NSW data collections are required to be reviewed by the NSW Population & Health Services Research Ethics Committee (PHSREC).

- Multi-jurisdictional data linkage studies that include NSW Health data may be reviewed by any of the specialist NMA data linkage HRECs. Please note that not all jurisdictions are party to the NMA. Please refer to the NMA Data Linkage Guide, which lists the specialist linkage HRECs, and can be found at https://www.clinicaltrialsandresearch.vic.gov.au/national-mutual-acceptance.

- If you have received HREC approval for the clinical portion of your project (e.g. for the collection of primary data), additional Data Linkage HREC approval must be sought for the data linkage component.

Check - Project Details

☐ I have submitted an EOI on CheckApp and have a CHeReL Project ID

☐ You have provided a data linkage HREC reference number

SECTION 2 Version Control (for ethics to review)

- A new submission constitutes version 1. When this protocol is updated as part of any project amendment this section should be updated.

SECTION 3 Investigators and Participating Institutions

- Ensure **all** investigators are listed in this section. You can add more tables rows to fit more researchers.

- Email addresses should be institutional addresses (i.e. no Gmail, Hotmail, etc).

- The principal investigator cannot be a student.

- It is a requirement of the NSW Ministry of Health under our Authority to Disclose Data that a Confidentiality Undertaking is signed by the person with the most senior position where unit record data will be accessed (this does not need to be the principal investigator). A Confidentiality Undertaking will therefore need to be signed by a representative at each site accessing linked data. Signatories do not necessarily need to be accessing data themselves but take responsibility for those accessing data at their respective site.

- All information should be the same as listed in the draft HREA (NOTE: only up to 10 investigators can be listed in the draft HREA/REGIS).

- If the contact person is also a researcher on this project, please include their details at Section 1 and this Section 3.

- All people to have access to unit record data are to be listed in the first table and in the *Investigator / Researcher Accessing Linked Unit-Record Data* table.

- Data and role separation rules apply to prevent the re-identification of data. Anyone who has access to identifiers in any primary dataset may not have access to linked data.

- Requests for overseas user access to unit record data will be assessed on a case-by-case-- basis. Note that circumstances for approval is rare. Very strong justification is required for the addition of overseas researchers.

Check - Investigators and Participating Institution

☐ All researchers are on my project are listed here

☐ I have indicated which researchers have access to linked data and the physical location and organisation (site) from which researchers are accessing data from is listed

☐ The researchers are the same on my HREA

☐ All researchers accessing data are specified in the *Investigator / Researcher Accessing Linked Unit-Record Data* table and the table is complete

☐ No researcher who has access to identified data (primary identifying information) will have access to linked data.

SECTION 5 Background / Rationale

- This application should focus on the data linkage component of your study.
- Background should cover what is known in literature, what gaps are sought to be addressed through the broader study and data linkage, and any possible research translation that may arise from this research.
- If you have primary data, describe the process for how primary / clinical data was collected (e.g. “a clinical trial was conducted…”) as background context/what has been achieved so far. Reference the ethical approvals that underpin the collection of this primary data. These approvals should be submitted for noting.
- If AH&MRC approval is being sought, please ensure that this is mentioned here.

Check - Background / Rationale

☐ A review of literature, research context, knowledge gaps and research goals are provided.

☐ The collection of primary data is described (if relevant)

☐ The ethical/legal basis under which primary data have been collected is described and attached (if relevant)

SECTION 6 Aims and Objectives

- Provide a statement of primary and secondary aims/objectives, key research questions, and/or a clearly defined hypothesis to be addressed through data linkage.

Check – Aims and Objectives

☐ A hypothesis is stated, and/or list of question/aims is provided

SECTION 7 Methods

Study Design

- Data linkage can be used in retrospective cohort studies, longitudinal studies, case control studies, to supplement randomised control trials and other interventional studies.
- CHeReL can identify a cohort for linkage comprising of cases and controls or can identify controls against a primary research cohort. For these projects, researchers may be required to complete a Case Control Specification Form.
  Case Control forms should not be submitted for HREC approval, these are for technical use by CHeReL only.

Cohort/Study Population

- A study population refers to the group identified by specific criteria to answer the research question. This may be people with a specific condition who were eligible for a specific treatment.

- The study population may be broader than the population/cohort that you wish to receive data for.

- For cross jurisdictional linkages, the study population should be described as persons from all involved states. Note that the request for data for NSW in linkage in Section 12 relates to the cohort for CHeReL linkage only.

Data Collection

- If you have collected data for a specific study, this data held by your research team is considered primary data. If primary data are being used in this linkage project, the data should be described as well as its method of collection within the "Primary Data" section. All relevant information about this dataset should be provided, i.e. is it a survey, variable list, consent form/participant information sheet, etc.

- Reference the legal and/or ethical approvals that underpin the collection of any primary data. These approvals should be submitted for noting.

- Secondary data refers to datasets that were not originally collected for the purpose of this study. These include administrative datasets. If you are looking to source and link secondary datasets, please list these all within the "Secondary Data" section.

- All datasets in your project must be listed including data from other jurisdictions that will be linked by other data linkage units.

- Variable lists for all datasets including those from other jurisdictions need to be supplied to CHeReL.

Consent

- If you plan to obtain consent for your data linkage for your cohort, please discuss the specific wording to be used for consent with PHSREC or other data linkage HREC.

- For CHeReL, patient information and consent forms should explicitly state:
    - that the data you collect is intended to be linked;
    - that the data you collect (including both personal identifiers and content data) will be handled by CHeReL (and other data linkage units);
    - where the data you collect are intended to be stored and analysed;
    - where the linked data are intended to be stored and analysed; and
    - that linked data will be de-identified for analysis and all efforts to prevent re-identification will be made.

- Refer to the National Statement for Ethical Conduct in Human Research Chapter 2.3.10 to structure your request for a Waiver of Consent. Strong justification should be provided.

Check - Methods

☐ The type of study being conducted is described

☐ I have completed and supplied CHeReL with a Case Control Specification Form (if relevant)

☐ The study population is described

☐ The cohort for linkage is described (if this is a subset of the study population)

☐ The primary dataset(s) is described (if relevant)

☐ All NSW datasets are listed

☐ All external datasets are listed

☐ A variable list is supplied for **all** datasets listed including the external datasets.

☐ Explicit consent for data linkage has been obtained, or a waiver of consent is justified

SECTION 8 Data Governance

Data Flow

- The data flow should outline:
    - How the cohort (for whom data will be linked and supplied) is defined
    - What data will be linked and supplied – specify if this includes identifiers only, or both identifiers and content data.
        - Note that CHeReL requires both identifiers and content data be supplied to CHeReL for linkage and release. If content data are not able or permitted to be provided to CHeReL, justification should be included here.
    - movement of data within and between sites including the methods of transfer used and by whom.
    - All data sources should be listed including those from other jurisdictions if applicable. (as the researcher may receive the data from each of the data custodians from whom data are requested, from CHeReL and/or other data linkage units).
    - where the data will be held
    - which sites or entities require access to analyse data
    - data separation practices where applicable.

- This section should be laid out in steps

  E.g.

  1. *The [role] at [location] will provide [what data e.g. identifiers only, or identifiers and content] to CHeReL for the [primary data] – cases - via Secure Transfer Protocol*
  2. *The [role] at [location] will provide [what data e.g. identifiers only, or identifiers and content] to CHeReL for the [external data] via Secure Transfer Protocol*
  3. *The [role] at [location] will provide [what data e.g. identifiers only, or identifiers and content] to CHeReL for the [external data] via Secure Transfer Protocol*
  4. *The [role] at [location] will provide [what data e.g. identifiers only, or identifiers and content] to CHeReL for the [external data] via Secure Transfer Protocol*
  5. *The [role] at [location] will provide [what data e.g. identifiers only, or identifiers and content] to CHeReL for the [external data] via Secure Transfer Protocol*
  6. *CHeReL will derive controls matched based on [criteria] to cases.*

7. *CHeReL will link [primary identifiers] to MLK datasets (e.g. APDC, EDDC, NAP, PDC etc.) with external datasets and primary datasets and assign a CHeReL PPN for cases and controls.*
8. *CHeReL will upload content data + CHeReL PPN for cases and controls for [MLK datasets, any external datasets CHeReL obtains content data for] to [secure data environment e.g. SURE].*
9. *Mapping files will be set back to [roles] with CHeReL PPN to data custodians for [external datasets CHeReL did not obtain content data for].*
10. *[role] at [site] will append CHeReL PPN to [external] content data and upload to [Secure Data Environment].*

- This should correspond with the flow diagram at Section 8, Question 1B.
- Flow examples are provided at Appendix B and Appendix C.

Data Separation

- When using administrative data, it is necessary to adhere to the Separation Principle to protect the identities of individuals and organisations in datasets.
  No one working with the data can view both the linking (identifying) information (such as name, address, date of birth) together with the merged analysis (content) data (such as clinical information, benefit details etc) in a linked dataset.
  Whilst ethics approval extends to the ability to link the data for use in research, it does not confer approval for researchers to have access to a person's highly sensitive, identifiable linked information.

  You may wish to emphasise the separation principle by stipulating at a minimum within your protocol something along the lines of:

  - Personal information about individuals is required to obtain satisfactory record linkage however the processes of record linkage and data analysis will be completely separated ensuring the identifying information is not available with personal/clinical (or otherwise identified) information and cannot be used for research purposes.
  - Researchers who will have access to linked data will only be supplied with linkage keys which can be used to link datasets. Identifying data will only be used for the purposes of data linkage by CHeReL.
  - Data Custodians with access to personal (identified) information will not have access to Linked Data.
  - Your data flow should articulate how the data separation is upheld including outlining the separation of roles throughout the movement, linkage and use of data.

Data Flow Diagram

- You may use your own data flow diagrams or use the templates that are available on CHeReL's website.
  - You may add and remove boxes, links, etc. as required.

- Please add as much detail as possible. Ensure the cohort aligns with the cohort for linkage description, and all datasets including external datasets are listed.

- Please show all data sources and their movement in at least one flow diagram – this includes data from other jurisdictions.

- For complex/large multi-jurisdictional linkage projects, you may prepare a separate diagram for the movement of all data at the jurisdictional level and a more detailed diagram which focuses on the movement and use of NSW data specifically.

Data Storage, Access, and Security

- Describe the Secure Data Environment where linked data will be stored.
  Include information on the platform, technology and who is responsible for operating the environment.
  Specify if the environment has any independent security assessment, e.g. PSAF, ISO27001, or other.
  Outline whether the environment relies on services outside of Australia (e.g. Cloud services for machine learning, artificial intelligence, or advanced analytics).

- Describe how data are securely transferred to and between environments (methods and procedures used) and what governance controls are in place.

- Outline plans, policies and processes for managing data / confidentiality breaches.

- Describe who is permitted to have access to the linked data and the controls in place to protect the data and ensure access is only available to those who are permitted to access linked data.
  Specify who is responsible for approving and provisioning/revoking access to the environment.

- Describe what data are permitted to be moved or copied from the Secure Data Environment to another location and the controls in place to ensure the outputs of the analysis are appropriate and non-identifying.

Use and Disclosure

- List the anticipated project outputs and describe what formats (i.e. publications, reports, presentations, etc) used to disseminate findings.

- Results should be presented in ways that do not risk a participant, or a small community of participants, being identified if information specific to them is disclosed to or inferred by a reader.

- Individual "cases" must not be described.

- Simple statistical descriptions (e.g. percentages or means) must be based on groups of at least five participants. It must not be possible to read or estimate individual subject's values from other results (e.g. OR or RR) or graphical presentations of data. Cell values of less than 5 should be presented as '<5'.

- Small populace (<1,000 population), geographically small (e.g. SA1) or sensitive (Indigenous) communities should not be identified, either directly or indirectly. Statistics relating to individual small or sensitive communities should not be presented unless they are one of at least five such communities being described.

Data Retention

- Specify the place and method of storage.

- Specify for how long data will be stored following completion of the project.

Data Disposal

- Specify when data will be deleted.
- Specify the method of deletion, and by whom it will be deleted.

Check – Data Governance

☐ I have provided a detailed data flow/methodology that outlines in sequential steps the who-what-how-where-(why) associated with the movement of data in my project.

☐ A flow diagram is provided in the form, or separately, which outlines all data movement including data external to NSW.

☐ I have described where data will be stored, the environment's security and the framework in place to manage access and breaches.

☐ I have specified what outputs are expected from this project, and how I will mitigate identification of individuals or communities.

☐ Data storage methods and duration is described.

☐ Data destruction methods are described.

SECTION 9 Analysis Plan

- Specify the primary and secondary study outcome measures and include information on exposure/s, covariates, and other factors.
- Describe how these are defined based using the data and how these reflect the aims.
- Describe how the linked data variables will be used to address the aims and stated outcomes of interest.
- Outline the statistical methods and tools that will be used for analysing the data.
- It may be useful to pre-define the desired results and the minimum requirements to achieve these results (power, significance, etc.).

Check – Analysis Plan

☐ Planned outcomes are described in detail

☐ A statistical analysis plan is described in detail

SECTION 10 Project Funding / Support

- Details of funding is required, including from where funding is sought. This is used in requisite reporting which monitors the use of research funding across data linkage.

- CHeReL will not on-provide any specific financial information relating to your project to other parties.

- A project is only technically feasible if it has sufficient funding to cover the costs of linkage and data storage for the duration of the project. Please consult the pricing matrix available on our website for the costs of NSW and ACT data linkage. Note this does not cover third party costs associated with data linkage in other jurisdictions or data storage within a Secure Data Environment.

Check – Project Funding/Support

☐ Funding amount and source is specified

☐ Funding is sufficient for CHeReL linkage, linkage with other data linkage units and data storage within a Secure Data Environment

SECTION 12 Data Linkage

- During which calendar year do you require your first data linkage? – specify the month and year when you wish to receive the data for analysis. For instance, this may be next year, or may be in some time following primary data collection.
  Note that specification of a date is not a guarantee that linked data will be delivered by this date.

- If you require subsequently linkages or updates to your linkage, CHeReL can retain PPNs for a small unchanged external cohort for future supply of additional records.
  We are unable to retain PPNs for cohorts derived from MLK data. Updates to MLK-data-derived cohorts must be entirely re-linked.

- The cohort definition (Section 12A) and the request for linked data (Section 12B or 12C) are two separate categories. The cohort definition yields no data. This just defines the parameters that CHeReL will use to identify the cohort. The linked data request is where you specify the data that you wish to have provided for the cohort.

- There may be overlap in the data that defines the cohort and the data that are requested for the cohort, or there may be no overlap (See examples 1 and 2 below).

SECTION 12 (A)

- *Cohort for Linkage* is the cohort for which data will be linked and supplied. It may be all or a subset of the study population. If there is more than one cohort, please outline them all here.

- CHeReL requires an approximate cohort size (number). For cohorts that are to be derived from administrative data, this should be estimated based on existing literature or available health statistics.

- If your project will involve Commonwealth Data Linkage by AIHW, the cohort size that is derived by CHeReL must fall within ±10% of AIHW's approved cohort size for AIHW to accept cohort identifiers from CHeReL.

- All criteria that are to be used to define your cohort should be listed. This may be all participants in an RCT or may be derived from MLK-data. Examples of criteria may include age, sex, specific admission codes (diagnoses or intervention codes), date ranges for interventions, etc.

- Linked data that is requested for analysis should **not** be specified here. Note that if your cohort is defined by a specific admission, this is used to select your cohort. However, you will not receive details of admissions unless you request the relevant dataset variables at section *12B.1/C.1 – Data Requested*.

SECTION 12 (B)

12B.1 – Data Requested

- Information about the datasets commonly linked by CHeReL including the currency of data is at: https://www.cherel.org.au/datasets.

- Within the Admitted Patient Data Collection (APDC), all records are created at separation. This means that a record will only exist after the episode has concluded (death, discharge, transfer etc). If an admission is ongoing, the record will not yet exist.
  Admission date is useful if a particular leading event is being considered (e.g. bushfire) for which you might be trying to see its impact on admissions soon after, or if you want to ensure everyone in your cohort was admitted after a certain intervention.
  If you are interested in counting admissions within a date range, Admission Date is a variable that can be supplied even where the records are being selected for by separation.
  E.g. If you are interested in obtaining APDC data for the date range
  1 January 2019 – 31 December 2020:
  - By admissions, you will receive records for all episodes that commenced during that date range, and that have also already concluded. Note that if a person was admitted in late December 2020 and is still admitted, you will not receive any record for this person.
  - By separations, you will receive records for all episodes that concluded during that date range, irrespective of when they commenced (a separation may of have occurred in Feb 2019, they might have commenced in late 2018 and you will receive this record).

12B.2 – External Datasets

- Ensure that all datasets listed here are also listed at the Secondary Datasets section at Section 7 – Methods.

- CHeReL requires personal identifiers to undertake probabilistic linkage. For external datasets to be linked by CHeReL, the following identifiers are recommended to facilitate linkage:
  - Surname
  - Given name(s) including middle name(s) if available
  - Statistical Linkage Key (if name is not available)
  - Sex
  - Date of birth

    - Street address
    - Suburb/locality
    - Postcode
    - Event date (admission, test, notification, diagnosis etc. this is a date by which the data can be filtered, and is relevant to the record)
    - A unique record number

  In addition, the following are recommended, where available:
    - Person ID / patient ID / Client ID
    - Facility ID / Hospital code (for medical records)
    - Medical Record Number (for medical records)

- CHeReL should receive both identifiers and content for linkage. If content data are not able to be provided to CHeReL, reasoning (e.g. the legal and ethical basis for why this data cannot be handled by CHeReL) is required.

- Please provide contact details of the data custodian - CHeReL will use these details to seek data custodian approvals.

- Please provide contact details for the data manager/supplier (if different to the data custodian) – CHeReL will use these details to source data.

- Where CHeReL is only provided with identifiers, note that CHeReL can only release mapping tables, linkage keys, and/or the StudyID back to the data custodian. Identifiers and StudyID must be removed from study content data prior to upload to the data storage environment with linked data. See Appendix C.

- It is prudent to speak with data custodians of external datasets to determine feasibility of data use in data linkage projects prior to applying to CHeReL.

Economic Evaluations

- If you are interested in economic analysis, please consult the activity based management team concerning the availability and suitability of data.

SECTION 12 (C - Family Linkage)

- Only complete this section if you are undertaking a family linkage project

- Please provide describe in as much detail as possible how you would like your linkage outputs to look.

- Note that:
    - Other parent can only be identified through concordance in birth records, so it is not possible to tell whether other parents are biological parents.
    - Mother-child relationship is reliable if we have PDC records, however ‘Other Parent’ information is sometimes missing or poor quality
    - RBDM Births registrations also includes adoptions so some people will have multiple mothers and other parents.
    - Siblings may be difficult to identify if there is no PDC record to tie a child to their biological mother.
    - Mother and Other Parent records in RBDM Births are only available from 2000 onwards.

    - Twins / siblings are recognised in PDC and RBDM birth records if the "mother" half of their records link to the same person. Twins / multi-births are distinguished further by a plurality flag which indicates that the record is part of a multi-birth. These flags exist in both the PDC and BDM Births collections.

- Please be advised that due to the nature of family linkage, the following may occur:
    - Mother and baby linked into the same person ID (false link)
    - Missed links between RBDM Birth and PDC records (both in the mother and the baby)
    - Missed links between records belonging to the same mother, i.e. we miss the sibling relationship. Often caused by the mother changing her name (married/re-married) and address at the same time.
    - Other false links (either between different babies or mothers)

12C.1 – Data Requested

- Please specify for which cohort you require data for from each dataset. E.g. you may only wish to receive Births Data for the children and siblings, and not parents.

Check – Data Linkage, Requested Data, Family Linkage

☐ I have specified when data are preferred to be received

☐ Selection criteria is defined for the cohort for whom linked data will be provided

☐ I have specified that PPNs are to be retained (if relevant)

☐ All requested datasets are listed with a requested date range

☐ A variable list is supplied for all requested datasets

For Economic Evaluations (if relevant):

☐ I have consulted and received the support of the Activity Based Management Data Custodian

For External Datasets (if relevant):

☐ A variable list is supplied for all external datasets

☐ Free Text variables are identified and flagged for review

☐ I have indicated whether CHeReL will be supplied with identifiers only or both identifiers and content data

☐ Identifiers are all listed

☐ I have spoken with, and obtained the support of, Data Custodians of external datasets

☐ Contact details are provided for Data Custodians and Data Manager

For Family Linkage (if relevant):

☐ I have described the desired family relationship

☐ I have indicated for which family members I wish to receive data (per dataset)

Variable Lists

- Available MLK dataset variables are subject to change. Please ensure that you are using the most current version of the variable lists on our website.

- CHeReL cannot comment on the specific variables requested and if these are suitable for your research. Data Custodians and ethics reviewers will consider and provide advice.

- The blank variable list template on CHeReL's website may be used for primary and other external datasets.

- Indicate if a variable is being requested by marking it with a "Y" on the variable list.

- Please provide strong justifications for all variables selected. Justifications should specify why the selected variable is required, and which aim/research outcome/measure it is needed to answer. Data Custodians will review justifications and ensure that only the variables that are required to answer the research aims have been selected.

- If Indigenous status has been requested and is being used as a covariate only, ensure that this is stipulated in the protocol – with justification on how this variable will be used. Evidence of AH&MRC HREC approval may be necessary to receive this variable.

- Sensitive variables such as full DOB, postcode, suburb, latitude/longitude, free-text variables etc are rarely approved due to their identifying nature and strong justification is required for approval and release.

- Ensure variables requested across datasets are consistent (e.g. if requesting 'month and year' format for DOB from NSW APDC, request this same variable from NSW EDDC and not full DOB).

- NSW Data custodians will review all dataset variables involved in the study, including for data that is being supplied and linked by another jurisdictional data linkage unit. This is so NSW data custodians can best consider the risk to their own datasets when linked to and stored alongside other data.

- Variables from external datasets, such as those from other jurisdictions, may be submitted to CHeReL in the format that is submitted to the relevant data custodian or DLU.

- CHeReL may progress with a draft version of the AIHW TA and variables for review by NSW data custodians. However, AIHW data variables that CHeReL approves need to be the same as those that are eventually approved by AIHW. If there are discrepancies, this will require an amendment either with PHSREC and/or AIHW HREC, which may significantly delay the delivery of your linked data.

- CHeReL can only progress your application once we have received a copy of all data variables that NSW data will be linked with, including those from other jurisdictions.
  This is because linkage increases the risk of data being re-identifiable as combinations of variables may be unique and can identify individuals. Review of all data variables involved in

linkage is to ensure data custodians are comfortable that the risk of re-identification of data through linkage is minimal.

Free Text Variables

- The use of free text analytical data is deemed high risk and is therefore subject to additional scrutiny.

- It is CHeReL policy that CHeReL does not handle or release free-text fields.

- If your study requires the use of free-text variables, CHeReL requires a formal statement from the data custodian that specifies that these variables have been de-identified, and a confirmation that these will be released by the custodian or representative (such as the data manager) for research directly to a Secure Data Environment without handling by CHeReL. This can be in the form of an email or other document, which will be included in the suite of documents that is presented to the executive who signs off on the release of NSW health data for your project.

- For the statement on the use of free-text data:
    - List all free text variables in your written statement
    - For each free text variable:
    - Confirmation that this variable has been de-identified
    - The methodology for de-identification[1]
    - Confirmation the variable may be used for research
    - Confirmation that the custodian will supply free text data to the researcher within the designated research environment directly without handling by CHeReL.
    - A guarantee that all free text information has been reviewed and is free of identified information.

- If NSW health data custodians are not satisfied that these data are adequately de-identified, they will not sign off on the release of their data alongside these variables and it will be necessary to exclude any variables deemed too risky from the data linkage study before approval will be granted.

[1] CHeReL do not mandate methodology, as there are many available options researchers may employ, e.g. Rule Based method could, for example, be used to find names, addresses, or email addresses in data records. Machine learning models and other applications may be employed. Conversely, manual approaches may have been utilised by the research team to check, delete identifying information (names, age-related variables, spouse names, geographic information) from the free text responses, pre-summarising and categorising free text information etc. There is some literature available that may be helpful: https://pubmed.ncbi.nlm.nih.gov/35546422/.

Variable List Check

☐ I have selected all requested variables with "Y"

☐ A justification is given against each variable requested with "Y".

☐ I have sought AH&MRC approval for the use of variables relating to indigenous status.

☐ All project and dataset details at the top of the variable list are complete (e.g. name of collection, name of project, CHeReL ID, etc).

☐ Variable lists do not contain identifiers to be used for linkage, only content data

☐ I have provided all variables including those to be linked by other jurisdictions

☐ Formal written advice is provided from the data custodians of external datasets regarding obtainability and supply of free text variables (if relevant)

NSW Privacy Form

- You must respond **Yes** to ***Question 1***, as linked data research utilises personal health information that can be re-identified.

- You must respond **No** to ***Question 2***, as data linkage requires the use of identifiers.

- You must respond with "**The project involves linkage of data**" at ***Question 3***.

- Please list all the datasets you intend to use at ***Question 5*** and describe how they were collected (if relevant) and who is the custodian.

- You must respond **Yes** to ***Question 6.***

- Please respond **No** to ***Question 8***. While it may be the case for any primary data that data are collected for the primary purpose of research, with reference to administrative data research is not the primary purpose for data collection.

- At ***Question 13***, ensure the information provided here aligns with the information provided in the Data Governance Section of your combined application and protocol form.

HREA

- Where a data linkage project is being reviewed/has been approved by another specialist data linkage HREC under NMA, please provide your approved or draft HREA to CHeReL.

- Where a project is to be reviewed by NSW PHSREC, CHeReL needs to cite and review the draft HREA from NSW PHSREC.
  Please download and provide CHeReL with a draft version of your HREA in PDF form by doing the following:
    - After creating your project in REGIS, navigate to the Project page – which has Applications, Contacts, Details etc.
    - Click on the link to your Application (the Identifier will begin with something like 2024/ETHXXXXX). This will open all the sections of the HREA.
    - Select the Preview button at the top right, and then open the Zipped file, you will see a PDF Draft version of your HREA, which you can then save.

    - ➢ Please ensure that all researchers listed in the HREA are also listed on your combined protocol and application form.
      Note only up to 10 Investigators can be entered in the draft PHSREC HREA/REGIS – if there are additional investigators, these need to be listed in the CHeReL Protocol.

    - ➢ Ensure that your application has been made out to the correct ethics committee (Cancer Institute NSW; NSW Population and Health Service Research Ethics Committee).

    - ➢ See Notes specific to submitting to PHSREC via REGIS at https://www.cancer.nsw.gov.au/research-and-data/nsw-population-health-services-research-ethics-com/how-to-apply.

    - ➢ Please select 'Data linkage research' at question 1.17

    - ➢ Please ensure you specify that the data being used in the project is described as '**Re-identifiable (coded) information**' at section 3 Data and Privacy.

    - ➢ If a project is submitted in error, please contact the PHSREC team on cinsw-ethics@health.nsw.gov.au with the Application Identifier (2024/ETHxxxxx) and ask the team to release it.

Appendices

Appendix A - Cohort Selection and Data Linkage Request Examples

Example 1 – you are interested in outcomes of those who have had cardiac arrest

Cohort Definition:

"Adults aged 35 years and older during Jan 2020- Dec 2022 who had a cardiac arrest and were hospitalised"

Dataset	Date Range	Parameters
APDC (Admitted Data)	Jan 2020 to Dec 2022	Cardiac arrest diagnosis codes (supplied)
EDDC (Emergency Data)	Jan 2020 to Dec 2022	Cardiac arrest diagnosis codes (supplied)

← Note that the date range is limited to limit the size of the cohort.
Diagnosis codes are limited so that only people who have cardiac arrest are included.

CHeReL uses the above parameters to isolate a cohort.

The Linked Data Request:

Dataset	Date Range	Data requested
APDC (Admitted Data)	Jan 2020 to Latest Available	All diagnosis codes, cause of admission, treatments etc.
EDDC (Emergency Data)	Jan 2020 to Latest Available	All Dx
RBDM Death Registrations	Jan 2020 to Latest Available	
Cause of Death Unit Record File	Jan 2020 to Latest Available	
MBS (from Cwth)	Jan 2020 to Latest Available	MBS code and costs
PBS (from Cwth)	Jan 2020 to Latest Available	PBS code and costs

← From the same range of the incidence event, to latest available tracks the outcomes of individuals.
Note records are requested for all diagnosis types to see what other things the cohort has subsequently come into hospital for. This helps account for co-morbidities and track outcome measures.

← To determine if and how death occurs

← To identify any treatments to track outcomes.

← To identify if drugs are used to track outcomes.

CHeReL uses the above information to supply you with linked data for the above cohort.

Example 2 – you are interested in determining downstream arthritic impact (outcomes and economic) of fracture following motor accidents. You wish to compare a case group (those who have been involved accident and develop arthritis) with a control group (all others who have an accident and do not go on to develop arthritis).

Cohort Definition:
“Adults aged 18 and over who were admitted to NSW hospitals with any fractures following traffic accidents during January 2000 – Dec 2000” – specific fracture codes are supplied.

Dataset	Date Range	Parameters
APDC (Admitted Data)	Jan 2000 to Dec 2000	fracture diagnosis codes (supplied)

← Note that the date range is limited to limit the size of the cohort. This cohort is based on historical dates.
Diagnosis codes are limited so only people who have factures are included

CHeReL uses the above parameters to isolate a cohort.

The Linked Data Request:

Dataset	Date Range	Data requested
APDC (Admitted Data)	Jan 2010 to Latest Available	All diagnosis codes, cause of admission, treatments etc.
EDDC (Emergency Data)	Jan 2010 to Latest Available	All Dx
Non-Admitted Patient (NAP) Data	Jan 2010 to Latest Available	All clinic types
DNR (Cost Data)	Jan 2010 to Latest Available	Cost data across admitted, emergency and non-admitted clinics to determine costs associated with patient treatment.
MBS (from Cwth)	Jan 2010 to Latest Available	MBS code and costs
PBS (from Cwth)	Jan 2010 to Latest Available	PBS code and costs

← Note that the date range is a modern timeframe and does not capture the original admission.

Records are requested for all admission types to see what other things the cohort has subsequently come into hospital for. This helps identify if someone came in for arthritis treatment or other – **to identify cases vs controls.**

← Cost data for economic evaluation

← To identify if arthritis treatment is used. Used to distinguish cases and controls, and estimate costs for both cases and controls.

← To identify if arthritis drugs are used. Used to distinguish cases and controls, and estimate costs for both cases and controls.

CHeReL uses the above information to supply you with linked data for the above cohort.

Example 3 – you have enrolled adult participants into a clinical trial for a specific renal function procedure “Kidney Trial” during 2020 and collected primary data through surveys at baseline and follow-up between 2020 and 2022. You also wish to identify other adults admitted to hospital for treatment of kidney disease who did not receive your clinical trial intervention.

Cohort Definition:
“All adults who are enrolled in the ‘Kidney Trial’ in 2020 (cases) and all adults who were not participants who were enrolled in the Kidney Trial but who were otherwise admitted to hospital for the treatment of kidney disease in 2020 (controls).”

Dataset	Date Range	Parameters
“Kidney Trial” Dataset	Jan 2020 to Dec 2020	All enrolled participants
APDC (admitted data)	Jan 2020 to Dec 2020	Adults aged 18 or older admitted to Hospital for treatment of kidney disease in 2020 (codes supplied)

← Note that the date range is limited the enrolment period of the trial.

← Note that the date range captures an incidence event like when the cohort’s intervention took place

CHeReL uses the above parameters to isolate a cohort (cases and controls).

The Linked Data Request:

Dataset	Date Range	Data requested
“Kidney Trial” Dataset	Jan 2020 to Dec 2022	
APDC (Admitted Data)	Jan 2015 to Latest Available	Only records relating to nephrological, urological, and gynaecological admissions (codes supplied)
EDDC (Emergency Data)	Jan 2015 to Latest Available	Only nephrological, urological, and gynaecological presentations (codes supplied)
Non-Admitted Patient (NAP) Data	Jan 2015 to Latest Available	Clinics related to the treatment of kidney disease (Tier 2 Clinic codes supplied)
RBDM Death Registrations	Jan 2020 to latest Available	
CODURF	Jan 2020 to latest Available	

← Note the date range covers the total date range for the data collection.

← Note that the date range is a pre-dates the recruitment of the cohort (2020). This provides lookback data for the cohort.

Records are requested for only admission types that may be related to kidney disease to determine associated morbidity only.

← to determine mortality of the cohort

CHeReL uses the above information to supply you with linked data for the above cohort.

Appendix B - Cross-jurisdictional data flow (with no AIHW involvement)

Data Flow Diagram

The recommended data flow diagram template is available here:

External cohort(s), linked with external data linkage facilities (e.g. cross jurisdictional linkages)

Data Flow

1. The Research Team's Data Custodian / Data Manager (who will not handle or analyse linked data) prepares primary dataset(s) / cohort (study) data. A **StudyID** (arbitrary project specific person ID) is appended to both the study personal identifiers and to the study content data.

Study ID	Identifier1	Identifer2	Identifier3	...

Study ID	StudyContent1	StudyContent2	StudyContent3	...

2. Research Team's Data Custodian/Manager transfers personal identifiers and **StudyID** to all Data Linkage Units (DLUs)

Files sent to other DLUs from Data Custodian/Manager:

Study ID	Identifier1	Identifer2	Identifier3	...

3. CHeReL Data Sourcing Manager will contact the Data Custodian/Data Manager to obtain the personal identifiers, study content data (*including MBS and PBS from Services Australia where applicable*) and **StudyID** to CHeReL.

Files sent to CHeReL from Data Custodian/Manager:

Study ID	Identifier1	Identifer2	Identifier3	...

Study ID	StudyContent1	StudyContent2	StudyContent3	...

4. CHeReL generates a new **Universal_PPN** and creates mapping file of **Universal_PPNs** to **StudyID** (This is a direct replacement of **StudyID**, no deduplication or processing of records occurs) and CHeReL sends mapping file of **Universal_PPNs** to **StudyID** to other DLUs.

Mapping File to DLUs:

Study ID	Universal_PPN

5. CHeReL will identify the cohort within NSW administrative datasets using the personal identifiers, links to their NSW/ACT administrative data collections.

6. CHeReL removes all identifying information and **StudyID** and appends **Universal_PPNs**.

Study ID	Identifier1	Identifer2	Identifier3	...	NSWContent1	NSWContent2	NSWContent3	...	Universal_PPN

Study ID	StudyContent1	StudyContent2	StudyContent3	...	Universal_PPN

7. Other DLUs identify the cohort within their administrative datasets using the personal identifiers and link their jurisdictional data.

8. DLUs remove identifying information and **StudyID** and append their jurisdictional specific **DLU_PPNs**.

Study ID	Identifier1	Identifer2	Identifier3	...	DLUContent1	DLUContent2	DLUContent3	...	DLU_PPN

9. DLU/jurisdictional data custodians upload jurisdictional content data and **DLU _PPNs** to [Secure Data Environment].

DLU(s) uploads Files:

DLUContent1	DLUContent2	DLUContent3	...	DLU_PPN

10. CHeReL uploads NSW/ACT content data, study content data and **Universal_PPNs** to [Secure Data Environment].

CHeReL uploads Files:

StudyContent1	StudyContent2	StudyContent3	...	Universal_PPN

NSWContent1	NSWContent2	NSWContent3	...	Universal_PPN

11. Each DLU uploads mapping file of their **DLU-PPNs** to **Universal_PPNs** to [Secure Data Environment].

Mapping File uploaded by DLUs:

Universal_PPN	DLU_PPN

12. The research analysts replace **DLU-PPNs** with **Universal_PPNs** on all datasets within [Secure Data Environment].

DLUContent1	DLUContent2	DLUContent3	...	DLU_PPN

+

Universal_PPN	DLU_PPN

DLUContent1	DLUContent2	DLUContent3	...	DLU_PPN	Universal_PPN

DLUContent1	DLUContent2	DLUContent3	...	Universal_PPN

13. The research analysts merge datasets using **Universal_PPNs** within [Secure Data Environment]

Dataset 1	Universal_PPN

+

Dataset 2	Universal_PPN

+

Dataset 3	Universal_PPN

Dataset 1	Dataset 2	Dataset 3	Universal_PPN

Appendix C - Data Flow for CHeReL Linkage of External Cohorts where CHeReL cannot receive content data and must return linkage keys

Data Flow Diagram

The recommended data flow diagram template is available here:

External cohort(s) where CHeReL receives identifiers only

Data Flow

1. The Research Team's Data Manager, who will not be handle or analyse linked data, will prepare the [primary dataset(s) / cohort (study) data]. Identifiers will be separated from the study data and a StudyID is appended to both the Identifiers and to the content data by the Data Manager.

2. CHeReL Data Sourcing Manager will contact the Data Manager to obtain the study data Identifiers via secure file transfer.

File sent to CHeReL:

*Study*ID	Identifier 1	Identifier 2	Identifier 3	…

3. CHeReL will identify the cohort within NSW administrative datasets using the identifiers, link the identifiers of cohort to those within all requested health and external datasets

4. CHeReL will remove identifying information and append a project specific **NSW_PPN**

*Study*ID	Identifier 1	Identifier 2	Identifier 3	…	NSW PPN

5. CHeReL links de-identified content data of requested datasets and appends the project specific **NSW_PPN**

NSW PPN	NSW Dataset Content 1	NSW Dataset Content 2	NSW Dataset Content 3

6. CHeReL creates a mapping file between the StudyID and project specific **NSW_PPN** and sends this to the Research Team's Data Manager

Mapping File

*Study*ID	NSW PPN

7. The Research Team's Data Manager uses the StudyID and **NSW_PPN** mapping file to add the **NSW_PPN** to [primary dataset(s) / cohort data] content data

StudyID	NSW PPN

+

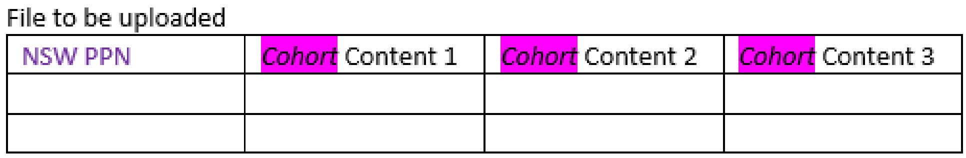
StudyID	*Cohort* Content 1	*Cohort* Content 2	*Cohort* Content 3

=

StudyID	NSW PPN	*Cohort* Content 1	*Cohort* Content 2	*Cohort* Content 3

8. Research Team's Data Manager removes the StudyID from the [primary dataset(s) / cohort data]'s content data and uploads the [primary dataset(s) / cohort data]'s content data into [Secure Data Environment]

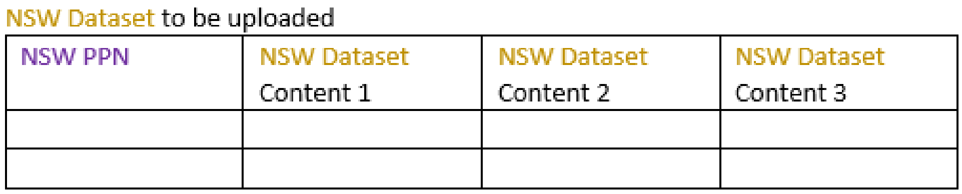
StudyID	NSW PPN	*Cohort* Content 1	*Cohort* Content 2	*Cohort* Content 3



File to be uploaded

NSW PPN	*Cohort* Content 1	*Cohort* Content 2	*Cohort* Content 3

9. CHeReL Uploads Linked Datasets to [Secure Data Environment]



NSW Dataset to be uploaded

NSW PPN	NSW Dataset Content 1	NSW Dataset Content 2	NSW Dataset Content 3

10. Approved analysts (who do not have access to the identified [primary dataset(s) / cohort data]) access [Secure Data Environment] to analyse data.