Foundation for Alcohol Research & Education

# Protocol and training resource:

## Researching people with complex needs using data linkage

Daren Fisher

Melissa Clarence

Dr Leanne Dowse

Professor Eileen Baldry

Ruth McCausland

UNIVERSITY OF NEW SOUTH WALES

APRIL 2011

# Protocol and Training Resource

## Researching People with Complex Needs Using Data Linkage

This resource has been produced as part of the *Researching People with Problematic Alcohol Consumption and Mental Health Disorders and Cognitive Disabilities in the Criminal Justice System* project*,* funded by the Foundation for Alcohol Research and Education.

**Daren Fisher**
**Melissa Clarence**
**Leanne Dowse**
**Eileen Baldry**
**Ruth McCausland**

# Contents

# Table of Figures

## Introduction

The increasingly complex nature of contemporary social problems requires the development of sophisticated methods of inquiry. Particularly when problematic alcohol and other drug use are involved, the search for tailored and appropriate responses for those experiencing such problems is not a simple task. Growth in the capacity of human service agencies to collect information and data about the characteristics of their client groups and the interventions offered to them provides an opportunity to examine multiple and interacting individual experiences and agency interventions in a new way. Using data linkage to bring together multiple related sets of information allows researchers the possibility of simultaneously taking into account both the linear and multilayered experiences of individuals. This approach can move beyond static description, circumvent the need for long-term or longitudinal studies, and shed light on the dynamics of individual and multiple service interactions. Developing this kind of understanding can then assist in identifying more appropriate, responsive, and cost effective service models.

This Protocol and Training Resource draws on the experience of undertaking a data linkage study of a cohort of people with mental health disorders and cognitive disabilities (MHDCD) in the criminal justice system (CJS) by researchers at the University of New South Wales (UNSW). The key issues and processes we consider relevant to undertaking data linkage research with complex populations are identified in the hope that this will build research capacity in the health, criminal justice and human services sectors. While there is great potential in undertaking such research, it does require overcoming many challenges relating to limitations of data and differences across datasets - for example, that data have not been gathered by the various sources using the same variables, definitions and language. This Resource details the importance of thorough investigation, reflection and development of processes around planning, data management and data collection before proceeding to data linkage and analysis; the skill set required on the part of project personnel; as well as a number of key areas for consideration for agencies and researchers interested in engaging in data linkage research.

## Background: complex populations and data linkage research

Complex populations present a series of unique challenges for those conducting research. Beyond the difficulties associated with refining a useable and practical definition, in this case of those with a mental illness and other complex social needs including alcohol and other drug (AOD) use (Almedom, 2005:944), gaining access to those who are defined as 'complex individuals' has been a long observed problem (Weisner and Schmidt, 1993:824). The often coexisting problems of substance abuse, homelessness and mental illness (McDermott and Pyett, 1994:45), and the stigmatisation that these individuals experience (Hinshaw and Stier, 2008:268) frequently results in an inability to conduct direct systematic measurement (Weisner and Schmidt, 1993:824). Researching complex populations requires a range of considerations in order that the complicated interactions that these individuals experience with the range of service agencies can be mapped and understood. For example, the relationship between an individual's health, employment and housing status and involvement in the criminal justice system are multilayered and intertwined and gaining a coherent picture and understanding of these interactions is one of the primary difficulties in this field (Silins, Sannibale, Larney, Wodak, and Mattick, 2008:418).

Although every project that aims to explore complex populations differs, there is a growing recognition by those working in the field of the potential problems that may be faced, and the development of particular techniques that may be used to assist in this process. Gaining access to complex and often hidden populations has been a noted source of frustration for practitioners and researchers alike (Heckathorn, 1997:174). Difficulties gauging the population size, developing a coherent definition of a cohort, and accessing information on these groups or their own accounts of their experiences with services have been a significant source of frustration for past researchers on complex populations.

There is a need to develop systems that access these marginalised groups in a way that is non-intrusive and does not cause additional stress or trauma to already vulnerable individuals. The high rate of contact with health and social services presents an opportunity not only for estimating the size and composition of these groups of people, but also for a unique series of insights into their life-paths. The relatively low cost and potentially high return for researchers afforded by linking such data make them an important resource to be explored (Fisher and Rivard, 2010:548). Moreover, research utilising existing sources of data is beneficial for agencies keen to make maximum use of their available resources.

## This Resource

This Protocol and Training Resource aims to assist those contemplating data linkage research regarding complex populations by outlining key issues and processes for consideration. It is not intended as a step by step manual for the creation and maintenance of a database or for undertaking statistical analysis. Instead, this Resource sets out the key elements and stages of a data linkage research project based on the experience of a team of researchers at UNSW, and highlights techniques, skills and strategies that can assist in overcoming some of the challenges inherent in such research. Although this process is presented in terms of distinct stages, in practice it will be necessary to reflect on and revise earlier plans or decisions in light of issues that will inevitably arise. For example, inconsistencies in the way that data has been entered may not necessarily be evident until attempting to link individual datasets, and there may then be a need to revise the research strategy and/or request further data or information from agencies supplying data.

The methodology detailed in this Resource was refined during the development of the dataset for the MHDCD project, an ARC funded Linkage grant (2007-2010)[1] aiming to explore the over-representation of people with mental health disorders and cognitive disabilities across the criminal justice system. The dataset contains lifelong administrative information on a cohort of 2,731 persons who have been in prison in NSW and whose MHDCD diagnosis is known. The MHDCD dataset draws on data provided for the project by agencies including The Centre for Health Research in Criminal Justice, Justice Health NSW, NSW Department of Corrective Services, NSW Police, Legal Aid NSW, Housing NSW, Juvenile Justice NSW, Ageing, Disability and Home Care NSW, NSW Community Services, and NSW Health (via the CHeReL). Ethics approvals were gained to cover data gathering from all agencies and strict data confidentiality and privacy principles underpinned the project.

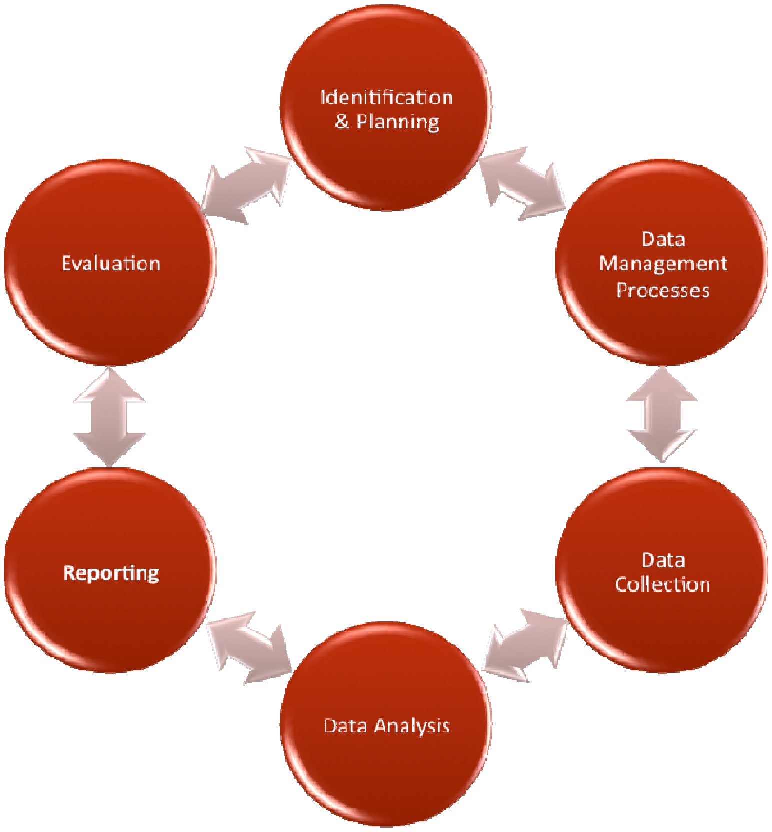
Drawing on the experience of developing the protocols, data management and analysis approach for that study, this Resource provides insight into the issues faced when attempting to link data across multiple sectors relating to a group of individuals with complex needs. Although this method was originally developed for a specific project, the practical guidelines and key areas for consideration are applicable to any data linkage project. It should also be noted that although this document references Microsoft SQL as the program used, other programs are available which provide similar functionality.

[1] ARC Linkage Project at UNSW *People with Mental Heath Disorders and Cognitive Disability in the Criminal Justice System in NSW* Chief Investigators: E. Baldry, L Dowse, I, Webster. Partner Investigators: T. Butler, S. Eyland, J. Simpson. Partner Organisations: Justice Health, NSW Corrective Services, NSW Police, Housing NSW, NSW Council for Intellectual Disability, Juvenile Justice NSW.

# Undertaking Data Linkage Research

The following diagram provides a model for the process of undertaking data linkage research:

**Figure 1: Circular Model of Data Linkage Research**



Useful information, relevant advice and questions for consideration that relate to and connect these various stages are outlined below.

## 1. Identification and Planning

When examining the possibility of undertaking a data linkage research project, the main areas of focus should be:

1. Honing the **research question(s)** in terms of the aims of the project and available data; and
2. The **capacity** of the research team to undertake the research in terms of skill set and resources.

Allocating adequate time at the outset of a project for these areas is crucial in terms of the overall long-term prospects of the project.

### *1.1 Research questions*

The project may start with defined research questions but these will usually need to be reviewed as the process of exploring the available data progresses. Some helpful questions to consider from the outset of a project may include:

- What do we want to find out and about whom?

- What does the current literature in the field say? What gaps exist in the research?
- What data is available that could address the research questions?
- Is there a relevant cohort of individuals for whom data exists that could be linked?
- Where is that data held and can it be released in a form that enables data linkage research?
- What are the key limitations of the data and how could these limitations impact on the research?
- What are the ethical and legislative constraints regarding access to the data?

A comprehensive literature review at the outset of a study will help identify important detail about the proposed cohort as well as methodological issues, data and agencies likely to be relevant in embarking on a particular data linkage project.

The questions driving the inquiry need to be made explicit, enabling an examination of why and how a data linkage project may best be able to address the aims of the research. In scoping the parameters of the proposed project, the implicit limitations of data linkage must then be considered. The type of data linkage being discussed in this document uses pre-existing datasets that have usually been collected for a purpose other than the proposed research; ie for administrative and reporting purposes rather than for analysis of individual and/or group experiences of service. The data available to researchers through agency collections is largely determined by the purposes that the data is intended to serve within the organisation. The processes by which data is collected may also vary. For instance, agencies may collect their data incidentally through screening processes or by conducting specified surveys and internal research.

The functional definitions of data elements collected and their relationships to each other are often embedded in organisation knowledge structures, and as such are not always overt or well recorded. Where they are recorded, the approach to capture and define the data may also vary. Furthermore, although classification systems and even terminology used may be similar to the academic literature on a topic, clarifying the operational use, functional definitions, and limits of key terms is an essential part of using this data. As a result, a realistic assessment of these extant datasets' capacity to provide insight into the key questions driving the inquiry is an essential first step.

A key means of understanding what data is available on a particular cohort and whether it is accessible for data linkage research will be the development of strong working relationships with key personnel in relevant agencies. This is discussed in more detail below.

Understanding the nature of the data that is available is critical in the planning phase of the research and requires extensive scoping of existing systems to be undertaken. Since data gathered reflects the core business of the agency that supplied it, it does not necessarily reflect what is not offered, who has been excluded from services, or the effectiveness of interventions offered. Not all data will be relevant to the linkage research, and each agency will carry data that will need to be filtered out of the dataset. Clarifying these requirements through discussion with each agency will help to limit the unnecessary inclusion of irrelevant data, and the potential exclusion of significant data. It is critical to manage perceptions of the research at this stage, as often agency representatives might restrict what data is available based on misperceptions of which data might be of interest to researchers. The study will be limited to conducting analysis on the data that is made available, so it is important to question whether there may be other existing data collections and variables not initially identified.

Although it may not be possible to be aware of or understand all of the systemic and institutional properties of the data sources beforehand, an awareness of the limitations of sources and datasets can help to clarify the scope of the research that is possible. Accessing any systems documentation that exists that defines the dataset and its constraints is useful in understanding these limitations.

Identifying the ethical and legislative constraints that protect data sources and any processes that need to be undertaken to enable access to each dataset is critical in the early phases of the research. Most datasets will have a data custodian who is responsible for approving any data that is extracted from the dataset. This is often a separate process with additional requirements to any ethical approval that is needed. Being aware of who can approve the research and under what conditions is essential to accessing the data. The fostering of a positive relationship early in the process and a shared understanding of the purposes and parameters of the research will be important for addressing issues that arise later in the process.

### *1.2 Capacity*

The capacity of an agency or researchers to undertake data linkage research in terms of access to personnel with the appropriate skill set as well as adequate resources is a crucial consideration at the outset of a project. Important questions include:

- Is the skill set available to undertake the range of tasks required in a data linkage research project?:
    - at the senior level, does the project have experienced, well connected Chief Investigators who can negotiate with data custodians and agency managers to obtain identifiable data?
    - at the technical level, does the project have access to personnel who have experience working with and manipulating large volumes of data, a commitment to attention to detail, and the ability to write code in a range of programs?
- Are sufficient resources available to ensure that:
    - appropriate personnel can be recruited or training undertaken if the research team do not have the skill set required?
    - systems can be established that enable extensive examination and cross checking to understand the quality and limitations of the data?
    - sufficient time can be taken to confirm that the research findings accurately represent the data?

If there is not sufficient capacity in terms of available personnel or resources, the integrity and timeframe of the research process and findings may be compromised.

## 2. Data Management

When embarking on a data linkage research project, the application of data management principles is critical to ensuring the appropriate and effective use of the large volume of data that will be gathered. The main areas of consideration at this stage should be:

1. **Research protocols** for the management of data;

2. **Ethics and privacy** issues pertaining to the use of the data;

3. **Documentation of systems and data.**

### *2.1. Research Protocols*

Developing a research protocol is a necessary step to establishing appropriate systems prior to the commencement of data collection. This document should comply with the National Health and Medical Research Council's *Australian*

*Code for the Responsible Conduct of Research* (http://www.nhmrc.gov.au/publications/synopses/r39syn.htm). A research protocol should cover a range of data management issues, including;

- how the data will be stored and secured;
- who can access the data;
- how access to the data will be managed;
- confidentiality and privacy restrictions of the data;
- ethical responsibilities of researchers;
- a confidentiality and privacy clause that all individuals must sign prior to accessing any data; and
- for how long the data will be kept and how it will be destroyed at the end of the project if required.

The Australian National Data Service provides resources for researchers to assist in data management (see http://ands.org.au/). These resources may provide guidance in developing data management systems around security, data storage, data documentation and during the data collection phase. Ethics and privacy issues will be discussed in more detail below. Researchers should get advice on their responsibilities regarding data retention and destruction in relation to their particular institution and jurisdiction.

Developing this document prior to data gathering ensures that these systems are established before the arrival of any data. It also ensures that clear and certain statements can be made in applications for data and ethics approval that establish the processes and systems used to safeguard the data during and after transference from its various sources and beyond the life of the project.

### 2.2. Ethics and Privacy

Having a clear understanding of the ethical and privacy issues and requirements of data linkage research projects is paramount, particularly when dealing with data relating to marginalised or vulnerable populations. Data linkage research often involves gaining access to data on individuals who have not provided consent for the researchers to do so. Ethical approval or oversight by one or more institutional Human Research Ethics Committees (HRECs) at the relevant university or human service agency will be required to allow the conduct of the study. Access to data may also include certain limitations on its use and on the variety of data that can be accessed.

Relevant material for review includes:

- National Health and Medical Research Council National Statement on Ethical Conduct in Human Research (http://www.nhmrc.gov.au/publications/synopses/e72syn.htm)
- Aboriginal Health and Medical Research Council Ethics Committee (http://www.ahmrc.org.au/Ethics%20and%20Research.htm)
- Federal Privacy Commissioner, Guidelines under Sections 95 and 95A of the *Privacy Act 1988* (http://www.privacy.gov.au/law/act/research)
- relevant state privacy legislation; and
- relevant university and/or agency HRECs.

One of the most common requirements of an Ethics Committee in relation to data linkage research is that the data be de-identified once the linkage has been completed to protect the identities of those who are in the data collection. Methods to do this include using the outside agencies to link data and returning linked datasets to the researchers with code IDs only, and methods not based on an identifiable identity (ID). Due to the sensitive nature of the data being collected, it will be necessary to store and/or dispose of the data in a secure manner. Being able to protect the data will be a requirement of the agencies whose data is being accessed, as well as a condition of any ethics approval. This will involve having a secure and protected server to store the data on and creating individual user access to prevent unauthorised individuals from accessing the data, as discussed in more detail below.

In addition to meeting the various ethics and privacy obligations, agencies will usually require sign off from the relevant data custodian to provide permission to release the data once ethical criteria have been met.

### *2.3. Documentation of Systems and Data*

It is important to have access to or to develop agency data dictionaries to fully understand the nature, relationships and limitations of datasets. Data dictionaries generally have the function of summarising the definitions and variety of classifications pertaining to the use of each variable, thereby providing clarity as to the scope and limits of the use of that variable. The information contained in a data dictionary enables the reliable and correct use of the data by others who have not been directly involved in the data linkage process. Data dictionaries should contain key information, including:

- the name of the table;
- the name of each of the fields that are included in the table;
- a brief description of the data that is included in each field;
- the type of data (integer, text, date);
- whether any constraints apply to each field; and
- the external tables and fields that can be linked to.

Many organisations do not have existing data dictionaries, rather the information is contained in a range of internal documents or is common knowledge amongst staff that work with the data. Developing a data dictionary that consolidates this information will then be an additional necessary task of the research team to ensure that the data is accurately analysed in the latter stages of the research.

An agency's protocols for data collection and storage will change over time, and the definitions that are used may not be in line with those in the academic literature. An example of this could relate to varying procedures across agencies in the collection of individuals' Indigenous status. There are clear Australian standards on the collection of this information, and many agencies use the codes established as part of these standards, however the data collection procedures vary greatly across agencies. Knowing the parameters of the data classification system used by each agency is vital to making sense of the data that is collected. As terminology will vary between sources, the creation of a standard language will help to organise and clarify the data. Any changes that occur during the labelling process should be recorded in the project data dictionary along with the definitions of these labels.

In addition, when dealing with large volumes of data from many sources, particularly if the data is longitudinal, it is critical that data constraints are documented. Data systems change over time, including definitions of variables, the way in which systems are operationalised and the importance of data (whether it is mandatory or not) impact on the data accuracy, completeness and consequently how it should be interpreted.

For large datasets sourced from multiple locations a visual data map can be a useful way to represent the data. A data map is a schema of the data and the relationships between tables, which sets out the extent of the data and the manner in which it can interact when queried.

Once data is collected it is likely that the data is not in a form that is useful for the research, necessitating the data manager to undertake a range of complicated transformations or restructuring of the data. Detailed documentation of these steps is necessary to ensure that the process undertaken from the point of obtaining the data to the final structure of the data can be understood in the absence of the person who originally transformed the data. It is also important in case a mistake is later identified that needs to be rectified. Whilst these decisions are detailed in any code used to transform the data, it should also be written in non-technical language to ensure that it is accessible to anyone wishing to understand the processes that have been implemented.

### *2.4. Intellectual Property*

Researchers will need to clarify ownership of intellectual property rights in the dataset with the various agencies from which are extracting data. It may be helpful to get legal advice relating to intellectual property, copyright and future uses of the dataset, and to include policies and processes relating to these areas in the project's Research Protocols document.

## 3. Data Collection

Compiling a linked dataset involves the following stages and responsibilities:

1. **Creating a database** in which to store the data;

2. **Extracting the data** from various sources, with a focus on matching individuals in the study across different data collections;

3. **Transferring the data** to the project database;

4. **Storing and securing the data**;

5. **Testing the data**; and

6. **Data retention or destruction**.

The parameters and processes for addressing these elements of the research are explored in detail in the following sections.

### *3.1. Creating a Database*

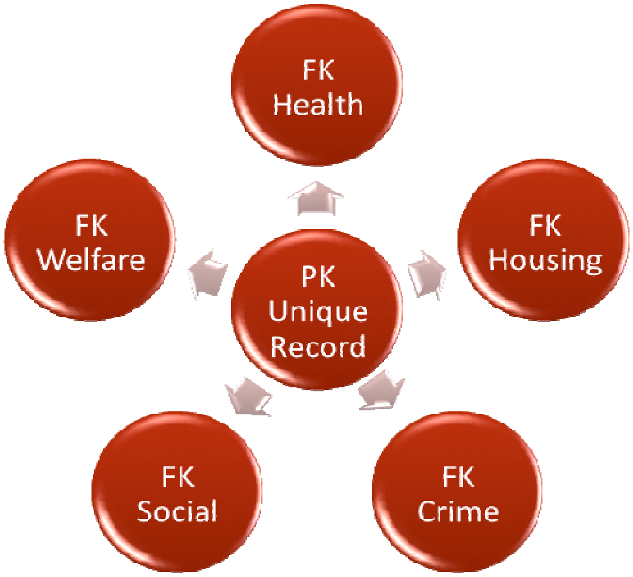
The identification and matching of project requirements and system capacity for compiling and storing a linked dataset is a key first step in data linkage research. The volume of data and the number of datasets to be linked should inform the decision about the required system resources, as well as the security and storage of potentially confidential data. The two main options available are:

- using flat files (i.e. csv, txt, xls) and joining them using procedures available in packages such as SPSS and SAS for analysis; and

- creating a database using available software such as MS Access, MySQL or SQL Server.

Considerations relevant to deciding between these options will vary depending upon agency resources and the amount of data. For example, using flat tables requires capacity to join tables using code and may require additional software to enable optimum use. Research oriented organisations are likely to have this software but such support is less likely to be available in service based agencies. The data needs to be safely and securely stored, referenced and duplicated. Creating a database can resolve many of these issues as databases have these tasks inbuilt.

Figure 2: Simple database structure



The volume of data that needs to be linked and access to the software are the main factors to consider in choosing the database. MySQL and SQL servers are powerful solutions that enable storage of information on a purpose built server, often with substantially superior processing capacity in comparison with a desktop computer where other solutions such as flat files and MS Access would typically be located. Using these services enables greater processing power which is beneficial, particularly when handling large volumes of data and running complex queries across many tables. Using a server with large processing capacity can reduce the processing time from hours to minutes on some complex queries. For these reasons a SQL server was chosen for the storage and linking of the MHDCD data.

In a relational database, such as a SQL server, data is stored in tables where the columns are called fields and the rows are called records. A relational database model is ideal for investigating complex populations as it enables the capture of interconnectivity of data across individual datasets. Relational databases allow tables to be stored on a database and, through the construction of queries, allow fields to be joined across many tables and records returned, without affecting the original table structure. The database acts as a repository for all the information in the research with multiple tables that can be joined. A simple strategy for managing the tables is to create a central table of identifiers for individuals in the study cohort, where a unique identifier can be created and all tables linked to this central table via the unique record. This schema is set out in Figure 2. Assigning a unique identifier for each individual and then including this unique identifier in each table allows data to be returned for any unique individual from all tables.

The use of primary keys and foreign keys provides an easy way to join data across tables. A primary key of a relational table uniquely identifies each record in the table. Every table should have a primary key to ensure that the table can easily be linked to other tables in the database and may contain additional foreign keys. A foreign key is a field in a table that exists as a primary key in another table. Therefore, a primary key can be inserted into other tables (called a foreign key in this instance) and used to join records across many tables by matching the primary and foreign keys. As each record has a unique primary key, values in the second table with the same value in the foreign key field will be returned for each record whenever two tables are joined using the primary and foreign keys.

Most databases enable relationships between tables to be defined, allowing easier programming. In larger scale projects, developing these relationships provides the capacity to develop a visual data map of the tables and data items used in the research.

Inherent in the design of SQL is the need to write queries using SQL syntax. The database is a repository for the data, so in order to manipulate the data, code needs to be produced that instructs what data is to be returned. This is done using the select statement. Basic SQL select statements require two lines of code which specify what fields are to be returned and from which table. The 'Select' line states the fields to be returned and the 'From' line specifies which table the fields are in. In addition, the number of rows to be returned can also be controlled using conditional joins or by restricting data in the 'Where' clause of a query. The 'Where' clause enables the number of rows returned to be restricted by imposing additional criteria on the data. Only rows that occurred in a specific time frame may be needed; this could be done by including a date range in the 'Where' clause.

Developing SQL skills is critical to understanding the data being examined, ensuring maximum flexibility when working with data in a SQL database. This will be necessary when dealing with complex data. There is a plethora of online training packages available to introduce and provide an explanation of the basic functions of SQL. Online courses provide a basic understanding of syntax and how a relational database operates, as well as often providing tasks to complete and confirm the application of these skills. Beyond these beginner courses providing the foundations of database management and syntax construction, there are few identifiable resources that document more advanced methods for use in research.

### *3.2. Extracting the Data*

Extracting the data requires firstly developing strategies to find and match individuals in each agency dataset and to transfer data to the researchers.

Difficulties in data matching are minimised if the data matching can be conducted centrally. However, it is likely that only a subset of individuals in each data system is included in the research. Data sensitivities and ethics constraints are unlikely to enable extraction of all data and matching conducted subsequent to this, so it is probable that data linkage will need to be conducted at the site where the data is held. This means that potentially different programming languages, different file structures and different rules need to be established at each participating data site.

#### *Matching Individuals*

Undertaking data matching on complex populations can be a difficult task, as many in the group may have criminal backgrounds, as is the case with the MHDCD study. Many may use one or more criminal alias(es) or due to extensive service use and contacts, have many administrative errors recorded in any single system. These serve to create possibilities for multiple records associated with an individual across possible data sources and within a single data source. Therefore, identifying suitable criteria to use for data matching needs careful consideration. Including a strategy which seeks to reduce false matches is particularly important where there are high rates of alias use. Taking steps to guard against false matches may in turn produce a higher rate of missed matches. Once such implications have been explored, a methodology for the matching can be decided.

There are two methods to undertake data matching, one using probabilistic and the other, deterministic techniques. The decision as to which of these methods to use should be based on a range of factors, including:

- The type of research being done and the importance of limiting false matches or not matching matches;

- The information available to undertake the data linkage on individuals;
- The capacity of the research team to undertake data linkage; and
- The information technology resources and software available to perform the data linkage.

### Probabilistic Matching

Probabilistic linkage uses statistical techniques, such as likelihood theory, to assign a probability to a match between two items. This allows the development of thresholds within which a decision about a match is made. The probability is based on the available data items that can be compared. This threshold allows automatic inclusion of records with a match value above a certain limit and excludes records that have low match value. The output will also include a selection of records for clerical review and decision. Administrative review of a sample of included and excluded data will allow an estimate of error rates in both accepted and rejected links. Figure 3 provides a few examples of slight variations in the data that will generate separate records unless a systematic matching process is used.

Specialist probabilistic data linkage software is required if this option is preferred. The advantage of probabilistic linkage is that it ensures that similar techniques can be used across many sites. However, the disadvantage is that these programs may require extensive computer resources when conducting matches across large datasets. Probabilistic linkage was trialled across a couple of sites for MHDCD project, and was successful when matching data in small agencies. However, with large datasets the capacity of the local computer resources was insufficient to complete the matching task. See FEBRL (http://cs.anu.edu.au/~Peter.Christen/Febrl/febrl-0.3/febrldoc-0.3/manual.html) – a free open source software package for probabilistic data linking.

### Deterministic Matching

Deterministic matching allows rule-based decisions to be established around a range of criteria. If two records meet these criteria exactly then they are considered a match, otherwise these records are not deemed a match. Deterministic matching is more coding intensive than probabilistic matching, with coding skill determining the capacity to identify records that should be matched as being matched. To illustrate this point, see Figure 2 where a decision about whether to match these two records needs to be made.

**Figure 2: Example of data matching problem**

Dataset 1: Peter J Brown 12/02/1970 2230

Dataset 2: Peter K Brown 02/12/1970 2230

Dataset 1: Paul MacDonald 12/02/1970 2203

Dataset2: Paul McDonald 12/02/1970 2230

If an exact match on first name, surname, date of birth and postcode were deemed necessary to match two individuals, then neither of these two records would match. The first decision needs to be made as to what is thought necessary to include as a match; ie. would these two sets of records be considered matched on the basis of this information? A few factors can affect this decision. Are the two datasets both small and do they contain a similar population so that it is likely that many of the individuals are contained in both sets? If so, then it is very likely that, in the instance of the first case, the transposed date of birth and opposing keys J and K were mistyped and that it is the same individual. This is implicitly making decisions about the importance of information. If the J was replaced with James and the K replaced with Kevin, would this decision still be made when the difference can no longer be assumed to be a mistyped key? Therefore the differential importance of contradictory information when the remaining information is similar needs to be considered. Then the decision is whether it is likely that a Peter Kevin Brown, born on the 2nd December 1970 exists in the same postcode as Peter James Brown born on the 12th February 1970 (Figure 2). It could also be that the date structure of the databases is different. Decisions about how to manage missing data also need to be determined when matching.

These decisions form the basis of the rules that determine what is matched. Once these decisions have been made, the question becomes how to make the matches. There are many strategies that can be used, as most programs have some form of fuzzy matching capacity or the capability to identify code that can be used to identify fuzzy matches. Soundex and Levenshtein distance are both examples of fuzzy matching algorithms that can be used to match similar records. These can be generated in MS Access or SQL server, though MS access has neither in-built. User forums are available on the internet that have this coding available to be implemented in Access by creating a module (old versions) or ribbon (2007 onwards) and a function for Levenshtein distance in SQL server.

Soundex retains the first initial and turns the remaining letters into a number, enabling McDonald and MacDonald to match as they both have a Soundex value of M235. In the instance of Jeffrey (J160) and Geoffrey (G160), because the first letter of the name is retained, this would not be a match. However, Levenshtein distance compares the keystrokes required to change one word into the other word, and consequently would return a difference of two keystrokes for the above example. If the threshold were set at a difference of less than three then, in this second example, records would be matched.

Soundex and Levenshtein distance are two of many types of fuzzy logic matching options available, but neither is perfect and therefore should not be relied upon entirely. Fuzzy matching options available within the package being used should be investigated and combined with deterministic string searches. Using string searches allows for the comparison of sections of the two values to be matched. For example, are the two names identical except for the initial letter - commonly seen in first names such as Cate and Kate. Using string searches that match on all criteria except for the some small variation will ensure that any matches that fall outside of those identified using fuzzy matches will be identified. Another example would be to examine if the first and middle name has been switched around as in the instance of hyphenated first names like Ann-Maree. When compiling a list of string searches to conduct, special consideration should be given to searches that might overcome the limitations of some of the fuzzy logic algorithms. The example of Cate and Kate, where only the initial value is different, is particularly relevant. Soundex would not match these two values. Soundex uses the first initial and then turns the remaining letters into a numeric representation of the sounds. Therefore C and K would not match. Using the Soundex value and excluding the requirement to match to the correct first letter could result in a very high false positive match rate. Writing a query that could identify these cases would be a better strategy.

### The Problem Of Aliases

Administrative error is to be expected in large databases, with small variations in names common. This is a stumbling block that can be overcome through the use of the algorithms noted above, such as Soundex and Levenshtein distance. The MHDCD study, which provides the model for the procedure outlined here, was complicated by the presence of large numbers of different identities (aliases) for the one person that could not be explained through administrative error. For the 2,731 individuals in the cohort there were around 30,000 aliases, making an average of 10 per person. However these were differentially distributed, with some individuals having 40 or 50 aliases. Aliases are commonly used by persons who have served time in prison (as with the MHDCD cohort), so while not unexpected, the large number associated with some individuals in the cohort required specific attention and exploration. These aliases were listed in the Police and Corrective Services databases but the aliases for an individual in one database were not necessarily the same in the other, and neither database ensured the listing of the person's real (legally given) name. Identifying details for a large number in the cohort had many variations (aliases) that could not be accounted for by small spelling variations.

Investigation of the incidence and pattern of aliases was further complicated during the final stages of the study by the discovery, during the matching process, that individuals were being matched to multiple distinct individuals in datasets who were evidently not the same person. These aliases/identities had to be eliminated, as their inclusion would have meant the study was tracking individuals other than the people in the cohort. Often these matches could be accounted for as a relative of the individual matched. However, in a range of cases there was no apparent connection to the individual. This raised the issue as to which individual was being tracked, and which individual should be matched in the agency dataset. The large number of aliases used, and the numerous dates of births provided, increased the chances of matching to a person in the community who was not the person in the cohort, especially when identities were common names in Australia. Additionally, there appeared to be deliberate attempts to use other people's identities, as often there would be an exact match with a person in the community where only one variation of that identity had been used.

A number of approaches to resolving the alias problem were employed, including the use of the matching methods above. In the event that these options were exhausted and non-validated matches were still found, researchers sought the assistance of each of the agencies providing data to assist in-house and within ethics approval, to match using other available identifiers such as address.

### *3.3. Transferring the Data*

Data security principles must be complied with during the matching and transferring of data with each dataset that is being linked. These principles should be clearly detailed in the project's research protocol document. In the process of providing data on individuals to be matched by other agencies, information could be being provided to other individuals about their client group that they are currently unaware of. It is therefore important that those undertaking the data matching exercise are aware of the data confidentiality requirements of the project. The transfer of the data should ensure that privacy and confidentiality principles are not risked by not transferring unsecured data across networks or email. Any transfer of data should use an encrypted password that is provided separately to the data. A password protected disk that is hand delivered is the most secure mechanism for transferring the data. Ensuring that a very strong password is used (a combination of upper and lower case letters, numbers and symbols of at least eight characters) will optimize the security of the data.

Ensuring identifying data is not transferred unnecessarily will also reduce any risks of data security breaches. When data is matched, the data provided back to the researchers should only include the unique identifier to be used to link data back to the research dataset. The provision of the identifying data that was used to match to the dataset should not be necessary.

### *3.4. Storing and Securing the Data*

Data should be stored in a secure location, with access restricted to approved individuals on a case-by-case basis. If using a database, it is likely that data is transferring across a network. It is critical to ensure that the data is not accessible across the internet or by anyone without authorised access. That the database is accessible only by members of the research team with permission is of utmost importance to guaranteeing data security and maintaining ethics and privacy obligations. Using a secure database is advantageous because:

- Access protocols on servers such as MS SQL Server are superior to storing data on a local computer in a password protected file; and

- Databases can utilise established protocols for backing up and restoring data.

However, using a database increases some risks which need to be managed. Personnel not directly involved in the research will become responsible for ensuring some elements of the data management plan for the research project are met, including ensuring;

- The physical security of the database;

- The data is backed up and can be restored if required;

- The data is adequately retained or destroyed depending on the protocols stated in the ethics agreement at the completion of the project or if there is a change in database server over the course of the project (i.e. if the database server is upgraded or changed the old database will need to be destroyed).

To ensure that ethical obligations are met, research staff should be involved in designing and implementing procedures to meet the required standards. All individuals who have access to the data should sign a data confidentiality agreement that clearly articulates their responsibilities to maintain the confidentiality of the data and any legislation that the data confidentiality is bound by. This should include any IT staff accessing the data as part of the database administration. Additionally, ensuring that research personnel are aware of the requirement for the database and data to be secure is important. Research personnel should be aware that it is important that data not be removed from the database and saved onto their local hard drives from the database.

### *3.5. Testing the Data*

Prior to using, analysing and reporting on data, it should be comprehensively tested to ensure that the data that was transferred was received and imported into the dataset. As data linkage projects will likely gather data across multiple systems and in many formats different from those required for the research dataset, transformations may need to be undertaken prior to being able to import the data into the database. Having the agencies that undertook any data linkage provide a summary of the number of rows and columns of data in each table is important for ensuring that the data that is imported is accurate and complete. This is particularly problematic when dealing with variables such as dates or addresses which both could contain different data formats to the data structures in the research database and which could become corrupted during the data import.

Checking for internal consistency and consistency across datasets being linked to ensure that the data is complete and coherent is important prior to use. This will confirm or identify any issues that may have arisen during the data matching and extraction stage. As the cohort in the MHDCD study was based on a sample of individuals who have been in prison, all the datasets that relate to or could contain information about an individual being in prison were interrogated for records confirming that the individual actually had been in prison. In this example, an individual had to have a record with the Department of Corrective Services indicating an incarceration, a court record and a police record.

As previously identified, using the cohort based on the criminal justice system raised issues around identification of individuals. In this instance, data testing was particularly crucial to identify possible errors and inconsistencies in the data matching.

### *3.6. Data Destruction*

Once all the data has been analysed and reports arising from the research have been produced, there may be an ethics requirement that the data is then destroyed within a specified period of time (this is variable depending on the legislation and jurisdiction). Data destruction, also known as data sanitisation, is not as simple as pressing the ‘delete’ button to remove the file from the server, USB or other storage device. Typically, pressing ‘delete’ merely archives the item and removes the visible link. As such, with the right software it may be recovered. To avoid potential recovery and misuse of sensitive data, the following process should be observed;

- Clearing - a method of overwriting on media. This process enables the specific targeting of part of a medium and is often low in cost;
- Degaussing - erasing the data collected on magnetic storage devices such as tapes and discs. This may also render the device inoperable as it may destroy basic formatting as well;
- Physical destruction - physically destroying and rendering the storage medium unusable.

If you are unsure whether you have successfully destroyed all data, many commercial companies specifically deal with this process and offer a variety of different techniques.

# 4. Data Analysis

The following two areas should be focused on in the data analysis stage of the process for data linkage projects:

1. **Documentation of tests and variables**

2. **Checking of results**

## *4.1 Documentation of tests and variables*

Throughout the process of data analysis, it is important to document each of the tests that are conducted and whether any variables have been manipulated in any way. As has been mentioned above, this will often require that variables are renamed, reorganised, or regrouped to make the operational definitions of a variable match the required outcomes of the research being conducted. If the data is required for another purpose, it may be necessary for the data to be returned to its original state.

Furthermore, careful documentation of the changes and tests performed on the data will help to minimise repetition of tasks, and ensures that there is a clear documented process for the manipulation of the data from its original structure to the final structure.

## *4.2 Checking of results*

Another important facet of the data analysis process is checking every query with a thorough attention to detail to ensure that it is returning the correct results. For example, it is easy to make small mistakes in writing the syntax of a query that will return subtly different results. By running a query multiple times, and checking that the results are coherent with the aim of the query, potential errors in syntax may be identified. While this process will not identify the nature of the problem, it will help to prevent errors from compounding and skewing the data that has been produced through the analysis. This is particularly critical when linking across tables as this can result in an explosion of records returned. Developing an understanding of the different types of joins is essential to ensuring the accuracy of the results of any queries that are executed. Joins are instructions that detail the way that two tables in a database should be joined to each other. In some cases this is simple, for example a primary key being joined by matching to a foreign key. This will return all results where the two values are the same. However, more complicated joins are possible such as a left or right join and inner and outer joins. Using different joins on the same tables with the same fields returned will provide extremely different results. This illustrates how crucial it is to have the appropriate skill set to undertake this kind of research.

The process of checking data is similar to that undertaken for all quantitative research, however in addition, when analysing linked data there are other steps required prior to conducting statistical tests, such as looking for mistakes in the code, examining whether the results reflect limitations in the data rather than being a significant result and checking for consistency with other results from the research and other studies.

# 5. Reporting

When reporting on the findings of a data linkage study, the following issues should be paramount:

1. **Accurate representation** of the data

2. **Maintaining confidentiality** of the data

## *5.1 Accurate representation*

Having a comprehensive understanding of the data limitations in a linkage study is important for interpreting and reporting results. The nature of the data that has been linked, the time period of the data collected, the accuracy of the matching, and the changes in data systems all impact on the limitations of the data and the extent to which you can report the result as a finding of the study instead of a manifestation of data limitations. Any results should first be examined to establish the role that the limitations of the data and how this could impact on the findings. When using administrative data there is likely to be limitations that should be considered in all results, as the data was not collected for research and may be neither consistent nor accurate.

Reporting for data linkage studies must be undertaken carefully to ensure that what is reported is an accurate representation of the data, with consideration to the limitations of the dataset. Ensuring consistency in analysis prior to reporting is a key step. If an interesting result is returned, the first consideration should be whether this is consistent with other results in the study or results from other similar studies. If it is not, it is critical to review the analysis looking for possible issues in the data integrity and the methodology, the program code, or whether other factors could explain the result. If none of these pertain, then the finding should be reported.

Having a large dataset with many different data items also presents the opportunity to run many statistical tests, consequently increasing the possibility of returning a false positive statistically significant result. As with any research, it is important that there is a well developed rationale for conducting statistical tests. This will constrain the number of tests conducted and reduce the chances of returning spurious results. Additionally, large datasets are very likely to have substantial statistical power to detect very small differences across groups, so it is also important to examine whether the statistical difference is a meaningful difference when reporting results.

### *5.2 Maintaining confidentiality*

When reporting findings on low frequency counts (such as a small sub group), care must be taken that individual units (such as a person or particular address) are not identifiable. This is crucial in order to avoid breaching the ethics obligations pertaining to the project, damage to any relationships with participating agencies, and of course, any potentially negative ramifications for individuals. This can be achieved by ensuring that details are not so specific as to allow (inadvertent) re-identification of the individual unit.

## 6. Evaluation

Evaluation should be a core part of any data linkage research project, with a focus on the following elements:

1. Ongoing **reflection and review** to inform the project's progress; and

2. Regular **documentation of processes, challenges and decisions**.

### *6.1 Reflection and review*

As depicted in relation to the circular model of data linkage research at Figure 1, evaluation should be an ongoing part of a project from its outset. Given the nature of this research and the need for constant reassessment of a project's aims, research questions, data sources and approach, there is great value in building in processes and time for reflection and review at various stages of the project.

Partner agencies involved in the project may also be prompted to reflect on their own internal methods of data collection, and to revise processes to ensure that data may be more accessible for such research in the future.

### *6.2 Documentation of processes, challenges and decisions*

As noted above, rigorous documentation should be a core part of a data linkage research project's methodology to ensure the validity of findings.

Also, as data linkage is a relatively new area of research, detailed documentation can be useful more broadly to the research team, partner agencies and others interested in undertaking future data linkage research. Documentation of the experiences and lessons learnt as a result of working on such a project can help inform future research, and more broadly, policy and practice in relation to people with complex needs.

## Summary of Key Issues for Successful Data Linkage Research

### *Research Design*

Understanding the data required for the investigation of a particular issue and where to access relevant data are crucial aspects of the research process. Deciding upon the sampling criteria for the study cohort and the time frame to be studied is the foundation of any investigation. Although this may be limited by the level of access granted to agency data, having a clear vision of what can be extracted from a dataset provides the framework for designing and conducting a data linkage project.

After the parameters of the dataset have been established, honing the research hypotheses, aims and questions is the next step. This will then inform the processes and information needed to undertake the study. Without knowing what is required to test the hypotheses or answer the questions, it is difficult to know in what format the data will need to be, or how to use software efficiently. Although research questions may have been developed before the study begins in earnest, it is usually necessary to review and modify them after the limitations of the dataset have been established.

### *Working with Stakeholders*

Maintaining a good working relationship with agencies that are contributing data to a linkage project is key to its success. The ability to clarify unclear terminology and gain access to procedural manuals and internal documentation will be essential to gaining a full understanding of the dataset being compiled.

Contributing agencies will have a vested interest in understanding and utilising their own data. Reporting back on preliminary findings and presenting the final results to participating agencies in a form that is useable for the agency will help to strengthen relationships, demonstrate the value of this method of analysis, and confirm that the commitment and effort on the part of the agency is well justified. It may also greatly assist the prospects of future data linkage projects.

### *Attention to Detail*

Attention to detail should be valued highly in personnel working with the dataset. The integrity of the project can rest on the capacity of individual members of the research team to understand and manipulate large volumes of data from a variety of sources. Fisher and Rivard (2010:546) have noted the importance of service based databases being designed primarily for use by officials in monitoring service delivery, rather than for research. When running any query, it is vital to confirm that the output generated is as expected. Slight changes in the syntax used to run a query may result in vastly different outputs. Checking the output of each query ensures that all of the required data are encapsulated in the query, and that there are no errors in the way that the syntax has been written.

For example, if a query is written that aims to include data within a particular date range, a simple exercise to check the data is to sort the results by the date and check that no data outside of this has been returned. It might be useful then to write the opposite code excluding these dates and see how many results are returned and whether any data in the date range has been returned. Adding the number of rows returned by these two different queries should then be equal to the total number of rows in the table if there are no criteria imposed on them. These simple tests can often identify small coding mistakes.

### *Managing Information and Analysis*

Using a database solution such as MySQL or SQL server requires Database Administration (DBA) skills to enable the creation of backup systems, develop security protocols and provide access to appropriate users. MS Access has inbuilt, easy to use features which can allow individual users to establish these systems, though the larger databases would require a DBA to effectively establish these systems.

As the sampling has most likely been undertaken by agencies providing data, knowledge of how people come into contact with agencies, and the process for this, is vital to understanding the cohort being investigated. By answering the following questions, many potential problems later in the process may be averted or foreseen;

- Can the same information on individuals be collected multiple times? For example, Indigenous status is recorded in numerous places in the MHDCD dataset.
- Can individuals exist in multiple categories simultaneously?
- Over what time period was the data collected?
- Are there any reasons why an individual may be excluded from a dataset?

Databases are dynamic and evolve over time. Managers modify them to deliver required organisational information, outcomes and tasks most efficiently and effectively. This may lead to the exclusion of certain data, for example information important for individually focussed unit-level research projects no longer seen as relevant to the organisation's needs. The completeness and accuracy of data collected may be diminished if it is not viewed as important by the collecting agency, and data of a certain age may no longer be deemed relevant and may be discarded. All these matters can be clarified through consultation with the agencies contributing data, and alternate solutions to gathering data may be suggested.

### *Assumptions*

Every dataset is different and will have different rules associated with its collection, storage, and dissemination. The following questions can be posed when attempting to assess the utility of any given dataset:

- How much missing data is there?
- What are the assumptions that are inherent in the data?
- What compromises have been made in the collection and storage of the data to make it practical for use?
- How was the data collected?
- For what purpose was the data collected, and what effects does this have?
- Is the data consistent across multiple measures?
- What else needs to be known before this data can be used for the purposes of the study?

The more that processes surrounding the production of data can be clarified early on, the more accurately the research team can assess the viability of undertaking a data linkage research project in relation to a particular cohort and research question.

## Bibliography

Almedom, A., 2005, Social capital and mental health: An interdisciplinary review of primary evidence, *Social Science and Medicine,* vol. 61, pp. 943-64.

Fisher, W. and Rivard, J., 2010, The research potential of administrative data from state mental health agencies, *Psychiatric Services,* vol. 61., no. 6, pp. 546-48.

Heckathorn, D., 1997, Respondent driven sampling: A new approach to the study of hidden populations, *Social Problems,* vol. 44, no. 2, pp. 174-99.

Hinshaw, S. and Stier, A., 2008, Stigma as related to cognitive disorders, *Annual Review of Clinical Psychology*, vol. 4, pp. 367-93.

McDermott, F. and Pyett, P., 1994, Co-existent psychiatric illness and drug abuse: A community study, *Psychiatry, Psychology and Law,* vol. 1, no. 1, pp. 45-52.

Silins, E., Sannibale, C., Larney, S., Wodak, A., and Mattick, R., 2008, Residential detoxification: essential for marginalised severly alcohol- and drug-dependent individuals, *Drug and Alcohol Review,* vol. 27, pp. 414-19.

Weisner, C. and Schmidt, L., 1993, Alcohol and drug problems among diverse health and social service populations, *American Journal of Public Health,* vol. 83, pp. 824-29.