Australian Government
Australian Institute of Health and Welfare

Australian Bureau of Statistics

National best practice guidelines

for data linkage activities relating to Aboriginal and Torres Strait Islander people

2012

Australian Government
Australian Institute of Health and Welfare

Australian Bureau of Statistics

National best practice guidelines

for data linkage activities relating to Aboriginal and Torres Strait Islander people

2012

Australian Institute of Health and Welfare
Australian Bureau of Statistics
Canberra

AIHW Cat. No. IHW 74

ISBN 978-1-74249-309-1

Suggested citation
Australian Institute of Health and Welfare and Australian Bureau of Statistics 2012. National best practice guidelines for data linkage activities relating to Aboriginal and Torres Strait Islander people.
AIHW Cat. No. IHW 74. Canberra: AIHW.

David Kalisch
Director/CEO
Australian Institute of Health and Welfare

Brian Pink
Australian Statistician
Australian Bureau of Statistics

Any enquiries about or comments on this publication should be directed to:
Communications, Media and Marketing Unit
Australian Institute of Health and Welfare
GPO Box 570
Canberra ACT 2601

Tel: (02) 6244 1032

Email: info@aihw.gov.au

Published by the Australian Institute of Health and Welfare and Australian Bureau of Statistics.
Cover art by Azeam Fraser, *Krambruk (landing place)*, 2011.

Please note that there are related publications to this report. Please check the online version at <www.aihw.gov.au> for any updates and additional material.

Contents

Preface

Accurate data about Aboriginal and Torres Strait Islander people are needed to guide policy formulation, program development and service delivery, towards closing the gap in disadvantage between Aboriginal and Torres Strait Islander people and non-Indigenous Australians. Progress is difficult to measure accurately because many Aboriginal and Torres Strait Islander people are not consistently identified as such in key data sets.

Data linkage offers a cost-effective approach to enhancing the completeness and consistency of Indigenous status information on key data sets, for purposes of statistical reporting. There are, however, no nationally agreed approaches on how to deal with missing or inconsistent Indigenous status reporting across data sets. This leads to different methods being used and difficulties in interpreting findings, particularly when comparing results across studies.

To ensure that a consistent and informed program of data linkage work is carried out across Australia, the Council of Australian Governments (COAG) tasked the Australian Institute of Health and Welfare (AIHW) and the Australian Bureau of Statistics (ABS) to develop national best practice guidelines for linking data relating to Indigenous people.

To inform the preparation of the Guidelines, COAG also tasked the AIHW and ABS to conduct a review of past, ongoing and planned data linkage studies in Australia and overseas that have an Indigenous focus. The review studies are published separately, as attachments to these Guidelines. The AIHW and ABS not only worked closely but also consulted widely in the development of these Guidelines.

The review of past, current and planned data linkage studies points to two main ways in which data linkage is used in studies related to Aboriginal and Torres Strait Islander people. Data linkage has been used to enhance the quality of Indigenous status information on key data sets, especially where Indigenous status is missing or inconsistently reported across data sets. It has also been used to add value to data sets by bringing together data items from multiple data sets, with a view to carrying out analysis that would be impossible to undertake if the research were based on the individual data sets. In both cases, data linkage has not been used to alter Indigenous status information on source data sets.

The Guidelines focus on six key aspects of data linkage. These are the values and ethics in human research relating to Aboriginal and Torres Strait Islander people, transparency and accountability, the quality of the Indigenous status variable on key data sets, the quality of the variables that are used for the linkage, the quality of the linkage itself, and the methods and algorithms used to derive Indigenous status where Indigenous status varies across the individual data sets in the linked data set.

The Guidelines also discuss a number of options that can be used to derive Indigenous status where Indigenous status varies across data sets. These options are related to the quality, scope, content and characteristics of the input data sets, and the purposes for which the input data sets are collected. The Guidelines recommend that for each statistical reporting or research, analysts should, as best practice, always compare the impact of various methods of deriving Indigenous status on the estimated measures and indicators.

Many of the methods and algorithms for deriving Indigenous status are explored in a separate study, *Getting Our Story Right*, designed to provide the evidence base and guidance to analysts for choosing methods to derive Indigenous status from linked data. The *Getting Our Story Right* study was undertaken as a joint project between the ABS, West Australian Department of Health and the Telethon Institute for Child Health Research.

David Kalisch	Brian Pink
Director/CEO	Australian Statistician
Australian Institute of Health and Welfare	Australian Bureau of Statistics

Acknowledgments

Authors for this publication are Tetteh Dugbaza and Brendan Scott from the Australian Institute of Health and Welfare, and Sybille McKeown from the Australian Bureau of Statistics.

The contributions of Fadwa Al-Yaman, Teresa Dickinson, Anthony Cowley, Phil Anderson, Tenniel Guiver, of the AIHW; Sarah Macgregor, previously of the AIHW; Brent Bufton and Jeffrey Wright of the ABS; and Glenn Draper of the Western Australian Department of Health/ABS, are gratefully acknowledged.

Support, advice and comments on the Guidelines were received from several individuals and institutions. These include Dr Lee Taylor, of the New South Wales Department of Health and the Centre for Health Record Linkage (CHeReL), on methods for deriving Indigenous status from linked data.

Useful comments on earlier drafts of the Guidelines were received from Professor Steven Zubrick, Dr Francis Mitrou, Dr Daniel Christensen, Dr David Lawrence, Dr Carrington Shepherd, Dr Amanda Langridge and Brilliana von Katterfield, of the Telethon Institute for Child Health Research, as well as from Dr Lucy Firth, Ying Chen and Mary Sullivan of the Department of Health, Victoria; Dr Fenglian Xu of the University of New South Wales/AIHW; Dr Judy Katzenellenbogen of Curtin University; Professors Tom Briffa and Pamela Bradshaw of the University of Western Australia; Dr David Roder of the University of South Australia; Professor Joan Cunningham, Dr John Condon and Associate Professor Sue Sayers of the Menzies School of Health Research; Steve Guthridge of the University of Adelaide; Dr Ron Somers of the Cancer Registry, South Australia; Professor Paul Memmott of the University of Queensland, and Professor Robyn McDermott of the University of South Australia/SA-NT Data Linkage.

Support for the Guidelines, as well as comments and advice on dissemination of the Guidelines, were provided by Dr Merran Smith of the Population Health Research Network, while Geoff Davis and Di Rosman, of the Western Australian Department of Health provided very useful comments and advice on various aspects of the Guidelines.

The authors are also grateful to George Truman of CHeReL for the insight he provided into data linkage protocols at CHeReL.

Abbreviations

ABS	Australian Bureau of Statistics
ACT	Australian Capital Territory
AIHW	Australian Institute of Health and Welfare
CDE	Census Data Enhancement
CHeReL	Centre for Health Record Linkage
COAG	Council of Australian Governments
EOC	episode of care
ID	identification
MLK	master linkage key
NHMRC	National Health and Medical Research Council
SA-NT DataLink	South Australia-Northern Territory Data Linkage Unit
SLK	statistical linkage key
WAAHIEC	West Australian Aboriginal Health Information and Ethics Committee
WADLS	West Australian Data Linkage System

The Principles and Guidelines

Principle 1: Values and ethics in Aboriginal and Torres Strait Islander research

The conception, design and conduct of all Aboriginal and Torres Strait Islander data linkage activities for statistical purposes should be guided by the core values and ethics of Aboriginal and Torres Strait Islander human research.

Principle 2: Quality of Indigenous status information in data collections

The quality of Indigenous status information within data sets included in the linkage study should be considered before analysis.

Principle 3: Quality of linkage variables

Linkage variables should be assessed before linkage to gauge their accuracy, completeness and comparability, and to ensure they are of sufficient quality to support the purposes of the linkage study.

Principle 4: Assessment of quality of data linkage

The quality of data linkage should be assessed and understood. Any limitations arising from the quality of the data linkage should be taken into account in the analysis of the linked data.

Principle 5: Methods for deriving Indigenous status

Analysts should investigate multiple methods for deriving Indigenous status and select those that best fit the purpose of the analysis. Where possible, analysts should also explore and report the impact of using various methods to derive Indigenous status on health and wellbeing measures and indicators.

Principle 6: Transparency

All relevant aspects of the data linkage activity, including data linkage quality assessment, analysis of the linked data, and methods for deriving Indigenous status, should be fully documented and publicly reported.

1

Background to the Guidelines

The purpose of the Guidelines is to create a foundation from which data linkage activities focussing on outcomes for Aboriginal and Torres Strait Islander people will be undertaken consistently and be fit for purpose.

1.1 Background

In 2008, the Council of Australian Governments (COAG) agreed to a set of high-level targets for 'Closing the Gap' in disadvantage between Indigenous and non-Indigenous Australians. These targets are in the areas of life expectancy, infant and child mortality, early childhood education and employment (COAG 2008).

Accurate data about Aboriginal and Torres Strait Islander people are needed to guide policy formulation, program development and service delivery, as well as to monitor and evaluate the success of Government and community programs in reaching the 'Closing the Gap' targets.

Currently, progress is difficult to measure accurately for targets relying on administrative data sets because many Aboriginal and Torres Strait Islander people are not consistently identified as such in these collections. This results from various factors, ranging from data collection practices, systems design and individual choices concerning disclosure of Indigenous status. Much is being done to improve the collection of data about Aboriginal and Torres Strait Islander people, which will lead to improvements in Indigenous identification in the future (AIHW 2010a; COAG 2008; Wood 2012).

In the meantime, data linkage offers a cost-effective approach to enhancing the completeness and consistency of Indigenous status information on key data sets, for purposes of policy formulation, service delivery, program evaluation, statistical reporting and research. For example, by linking data sets where Indigenous status is missing or is inconsistently reported, Indigenous status in the respective data sets can be compared, reviewed and enhanced. There are, however, no nationally agreed approaches on how to deal with missing or inconsistent Indigenous status reporting across data sets when linking data. This leads to different methods being used and difficulties in interpreting findings, particularly when comparing results across studies.

To ensure that a consistent and informed program of data linkage work is carried out across Australia, COAG tasked the Australian Institute of Health and Welfare (AIHW) and the Australian Bureau of Statistics (ABS) to develop national best practice guidelines for linking data related to Indigenous people. COAG requested that the guidelines cover linkage methods and protocols, privacy protocols, quality standards, and procedures (see page F-82, National Indigenous Reform Agreement (COAG 2008)). This report (the Guidelines) is the result of this collaboration.

These Guidelines were developed to be used for preparing indicators and outcome measures at the aggregate level, to improve monitoring and reporting. The Guidelines do not recommend that the Indigenous status of individuals be altered on source data sets. Indeed, they adopt the position that self-reporting in response to the standard Indigenous status question is the most accurate means of ascertaining a client's Indigenous status (AIHW 2010b).

1.2 Purpose of the Guidelines

The Guidelines are designed to assist multiple users who are involved in various stages of statistical reporting and research, and who use linked data relating to Aboriginal and Torres Strait Islander people. Primarily, users include statisticians and analysts who produce the statistics, measures and indicators used by state and federal governments in formulating policies, designing programs and delivering services to improve the wellbeing of Aboriginal and Torres Strait Islander people.

Secondary users include researchers who may use linked data to evaluate and monitor the performance of government and community programs to improve the wellbeing of Aboriginal and Torres Strait Islander people, as well as researchers who use linked data to study the cross-sectoral factors associated with the health and wellbeing of Aboriginal and Torres Strait Islander people.

The Guidelines aim to:

- provide options that can be used to derive Indigenous status, where Indigenous status is missing, inconsistently reported across data sets, or considered unreliable, and where data linkage is seen as a viable option for improving the reporting and monitoring of outcomes relating to Aboriginal and Torres Strait Islander people
- describe how data quality and the quality of the linkage process are important factors for analysts to consider when undertaking their analysis and statistical reporting
- provide information and assurances to data custodians, data collection agencies, Aboriginal and Torres Strait Islander people and the general public, of the privacy and data confidentiality protections involved in data linkage protocols
- recognise that data linkage projects that focus on outcomes for Aboriginal and Torres Strait Islander people should be inclusive of, and will benefit from the guidance and knowledge of, Aboriginal and Torres Strait Islander people.

The approaches recommended in the Guidelines with regard to data management, the data linkage process and the options for deriving Indigenous status from linked data will help to introduce a level of consistency and comparability to data linkage projects.

While the Guidelines have been designed primarily for use by analysts, researchers and those who use the findings of linked data studies, they may also benefit data custodians, data collection agencies, data linkage institutions, ethics committees and institutions that review applications for access to data for research or statistical reporting.

Scope

The Guidelines cover six domains, specifically relating to linking data on Aboriginal and Torres Strait Islander people. The Guidelines contain information on:

- Aboriginal and Torres Strait Islander values and ethics as applied to data linkage studies as well as to social and human health research
- the quality of Indigenous status reporting within datasets, and how they can impact on analysis
- the quality of data linkage variables across data sets, and how it can impact on the data linkage process
- the quality of the data linkage process, and how it can impact on the analysis of linked data
- methods for deriving Indigenous status where Indigenous status is either missing or is recorded inconsistently across data sets
- transparency, accountability, data security and the protection of privacy and data confidentiality.

To inform the Guidelines, reviews of past, current and planned data linkage studies in Australia and overseas, that focus on Indigenous statistical reporting, have been undertaken. The reviews are published separately as attachments to the Guidelines. These studies have been reviewed with respect to:

- the objectives of the studies and their benefit to Aboriginal and Torres Strait Islander Australians
- the data sets and analytical methods used
- issues of Indigenous under-identification, and how they were considered and addressed in various studies
- the impact on outcome measures of the various methods used for deriving Indigenous status from multiple data sets.

These Guidelines are essentially about enhancing the quality of Indigenous status information on data sets used in statutory reporting and research. Although the Guidelines also review the quality of other variables that may be used for data linkage, the emphasis in these Guidelines is on the quality of the Indigenous status variable.

1.3 Context of the development of the Guidelines

Although some aspects of the Guidelines may be applicable to general data linkage, the Guidelines have been developed specifically for data linkage studies relating to Aboriginal and Torres Strait Islander people, and have been informed by various considerations, including the following:

- Aboriginal and Torres Strait Islander values and ethics as outlined in *Values and Ethics: Guidelines for Ethical Conduct in Aboriginal and Torres Strait Islander Health Research* (NHMRC 2003)
- Aboriginal and Torres Strait Islander priority research themes as outlined in *The Road Map: A Strategic Framework for Improving Aboriginal and Torres Strait Islander Health Through Research* (NHMRC 2002) and *Road Map II: A Strategic Framework for Improving Aboriginal and Torres Strait Islander Health Through Research* (NHMRC 2010)
- ethical conduct in human research as detailed in *National Statement on Ethical Conduct in Human Research* (NHMRC 2007)
- data quality issues specific to the Aboriginal and Torres Strait Islander population
- past and on-going data linkage studies focusing on Aboriginal and Torres Strait Islander people, Indigenous identification and methods for deriving Indigenous status
- *Best practice guidelines for collecting Indigenous status in health data* (AIHW 2010a)
- data linkage protocols operating in various state and territory data linkage institutions throughout Australia
- *High Level Principles for Data Integration Involving Commonwealth Data for Statistical and Research Purposes* (Cross Portfolio Statistical Integration Committee 2010a, 2010b).

1.4 Organisation of the Guidelines

The Guidelines bring together knowledge from data linkage studies relating to Indigenous statistical reporting in Australia and overseas, as well as a synthesis and analysis of known issues, in order to provide a complete picture of the issues relevant for linking and analysing data relating to Aboriginal and Torres Strait Islander people. These Guidelines have been organised into seven chapters (see Figure 1.1).

The remainder of Chapter 1 discusses relevant data linkage concepts, as well as protocols of data linkage. The Guidelines are not a technical publication about data linkage, and only provide brief descriptions of the protocols, models and methods of data linkage. Readers should refer to references for more details on these topics.

Chapter 2 stresses the importance of conducting data linkage studies about Aboriginal and Torres Strait Islander people in accordance with the core values and ethics of Aboriginal and Torres Strait Islander people, as recommended by the National Health and Medical Research Council.

Chapter 3 discusses the problem of missing and inconsistent Indigenous status reporting on data sets, and how this impacts on the quality of outcome measures. Data linkage is seen as a viable option for enhancing Indigenous status information on data sets.

Chapter 4 shows that the quality of the data linkage is affected by the quality of other variables on the data set. Some of these data quality issues are specific to data relating to Aboriginal and Torres Strait Islander people.

Chapter 5 shows that the quality of data linkage affects the quality of the analysis based on the linked data. It is therefore important to assess the quality of the data linkage, so that limitations arising from the data linkage process are taken into account during the analysis.

Chapter 6 proposes various methods and algorithms that can be used to derive Indigenous status. It proposes that the choice of methods and algorithms will be influenced by the purpose of the study or analysis, the quality and characteristics of the data sets on which the data linkage is based, as well as on the quality of the data linkage itself. It also proposes that analysts explore various algorithms for any given project and then assess the impact of the algorithms on the outcome measures based on the linked data.

Chapter 7 looks at ways of ensuring that data linkage projects are transparent in terms of project objectives and project design, data sets used, data linkage protocols and full documentation of the data linkage protocols.

The organisation of the Guidelines is summarised in the diagram on page 7.

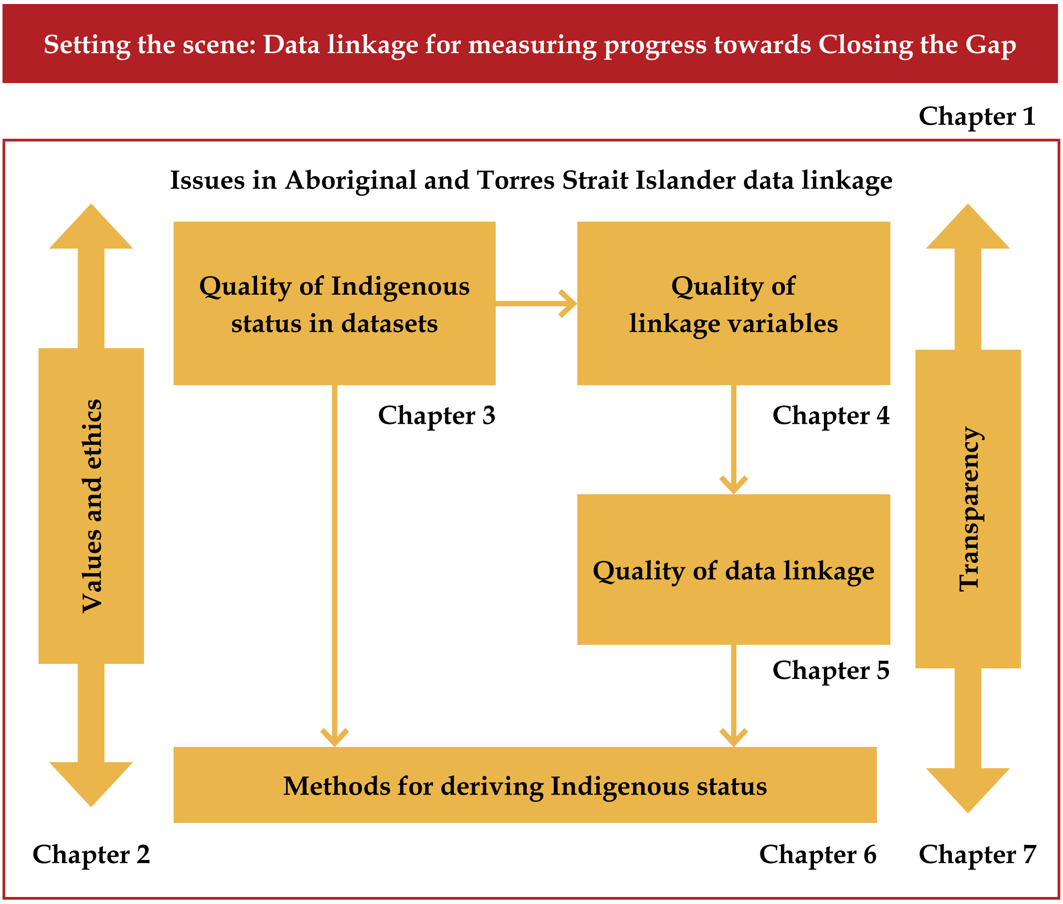

Setting the scene: Data linkage for measuring progress towards Closing the Gap
Chapter 1
Issues in Aboriginal and Torres Strait Islander data linkage
Values and ethics
Quality of Indigenous status in datasets
Quality of linkage variables
Transparency
Chapter 3
Chapter 4
Quality of data linkage
Chapter 5
Methods for deriving Indigenous status
Chapter 2
Chapter 6
Chapter 7

Figure 1.1: Data linkage for measuring progress towards 'Closing the Gap'

1.5 Key concepts in data linkage

This section aims to provide an overview of key data linkage concepts relevant to later chapters of the Guidelines. For more detailed descriptions of these concepts, referenced source material should be consulted.

1.5.1 What is data linkage?

The term data linkage or record linkage was first used by H.L. Dunn in 1946 (Gill & Baldwin 1987) to express the concept of collating health-care records into a cumulative personal file, starting with birth and ending with death.

Data linkage is the process of bringing together two or more sets of information belonging to the same person, event or place, into a single record of information. The sets of information may exist on different databases or in different parts of the same database (e.g. a person may have more than one record for services they have accessed (e.g. hospital admission) or a type of event they have experienced (e.g. the birth of children). By bringing this information together, connections between the information can be made and relationships established.

1.5.2 Uses of data linkage

Data linkage can be used to improve information on Indigenous status in cases where information on Indigenous status is either missing in data sets or inconsistently recorded across and/or within data sets. For example, where the Indigenous status of a person is recorded on two data sets and one is known to be more reliable, then by linking the two data sets, the more ‘reliable’ data set can be used to enhance the Indigenous status information recorded in the other. Similarly, where Indigenous status is inconsistently reported across a number of data sets, linking the data enables the analyst to cross-check the consistency of the information and derive more reliable Indigenous status information.

More complete and reliable Indigenous identification on health data sets, for instance, can help researchers and policy makers better understand Aboriginal and Torres Strait Islander people’s preferences in the health services they access, and which services and policy interventions produce the best outcomes to overcome Indigenous disadvantage and improve health outcomes (AIHW 2010b).

More generally, one key objective of linking data sets is to supplement data available on one data set with those on another to increase detail and accuracy. The linked data set can then be used to develop more reliable measures, as well as to establish relationships between data items that previously did not exist together in the same data set.

Data linkage has wide applications in health research, social research and policy development, and enables researchers to look at large volumes of data from various sources and make links or study variables within the various data sets. Consider the example where a researcher wants to examine whether the low birth weight of Aboriginal and Torres Strait Islander babies may have a long-term impact on their educational outcomes. The data required for this analysis may exist on two different datasets. Data on birth weight may exist on Midwives Data Collections held by state and territory health departments, while data on, for example, Year 12 scores, may exist on state and territory education department databases. This investigation may not be possible unless the two data sets are brought together.

Another use of data linkage involves linking a series of records taken at various times pertaining to the same entity. For example, a person's death records could be linked with his or her in-patient hospital separation records to enable researchers to study the relationship between age at death or cause of death and history of admitted patient care, over an extended period of time.

1.5.3 Data linkage methods

Data linkage methods usually fall across a spectrum between deterministic and probabilistic methods. A combination of linkage methods may be used in any one project, but the choice of method depends on the types and quality of linkage variables available on the data sets to be linked. The key linking variables may include identification (ID) numbers, full names, full date of birth, sex, and geographic variables such as residential address. These can be supplemented by socio-economic variables such as country of birth, as well as dates of specific events such as date of separation from hospital or date of death.

Deterministic linkage

Deterministic linkage ranges from simple joining of two or more datasets by a reliable and stable key to sophisticated stepwise algorithmic linkage. Simple deterministic linkage uses a single identifier or linkage key to join two or more data sets. A high degree of certainty is required in deterministic linkage. This high degree of certainty can be obtained if there is a unique entity identifier, such as a personal identification number, which uniquely identifies an individual across data sets. If this unique identifier exists on all the data sets to be linked, then it can be used to link an individual's records across those data sets.

Because deterministic linkage is based on exact matches, variables used in deterministic linkage need to be accurate, robust, stable over time and complete. Stable variables are fixed or tend to change very little over time. Examples are sex, date of birth, first name and, to some extent, last name for males. Alternatively, a combination of attributes, including last name, first name, sex and date of birth, can be used to create a linkage key which is then used to match records that have the same linkage key value (Christen & Goiser 2007). This linkage key is known as a statistical linkage key or SLK (see section 1.5.5).

A more sophisticated form of deterministic linkage is stepwise deterministic record linkage, which has been developed in response to variations that often exist in the attributes that are used in creating the linkage keys for deterministic linkage. Stepwise deterministic record linkage uses auxiliary information on the data sets to provide a platform from which variation in the reported linkage key or SLK information can be considered. This differs from simple deterministic linkage that relies on an exact, one-to-one character matching of linkage keys across two or more data sets.

Another variant of deterministic linkage is "rules-based linkage" where a set of rules can be used to classify pairs of records as matches or non-matches. Such rules can be more flexible than using a linkage key, but their development is labour intensive and highly dependent on the data sets to be linked (Christen & Goiser 2007).

Probabilistic linkage

Probabilistic linkage may be undertaken where there are no unique entity identifiers or statistical linkage keys, or where the linking variables and/or entity identifiers are not as accurate, stable or complete as are required for deterministic linkage.

In such cases, matching and linking will depend on achieving the closest approach to unique identification by using several identifying variables. Each of these variables is only a partial identifier, but, in combination, they provide a match that is sufficiently accurate for the intended purpose of linking two or more data sets. Probabilistic linkage has a greater capacity to link records with errors in their linking variables (NCSIMG 2004). In the traditional probabilistic linkage approach, pairs of records are classified as matches if their common attributes predominantly agree or, as non-matches, if they predominantly disagree. (Readers interested in more information on this may wish to refer to Fellegi and Sunter 1969).

Owing to the invariable presence of errors and variations in the recording of demographic data, probabilistic methodologies can lead to a much better linkage of records from separate data collections than simple deterministic methodologies for 'statistical' linkage purposes (NCSIMG 2004). While linkage can be supplemented by socio-economic variables such as marital status and country of birth, the main linking variables for probabilistic linkage remain full names, full date of birth, sex and residential address.

1.5.4 The data linkage process

Data linkage can be project-based *(ad hoc)* or systematic. Systematic data linkage involves the maintenance of a permanent and continuously updated master linkage file and a master linkage key. Project data linkage involves the linkage of two or more data sets for a specific project, and does not involve the maintenance of a master linkage file and master linkage key. The ABS and AIHW undertake project data linkage for specific projects involving specific data sets. State and territory data linkage units such as the Western Australian Data Linkage Branch and the NSW Centre for Health Record Linkage operate systematic programs of data linkage.

The master linkage file holds a discrete and select range of linkage variables (e.g. name, date of birth, sex, address) across multiple datasets, while the Master Linkage Key is a system of continuously updated links within and between core datasets.

The data linkage process may vary, depending on whether the linkage model is project-based or systematic, and the linkage method is deterministic or probabilistic. There are however four key steps that are common to both data linkage models. These are data cleaning and data standardisation, blocking and searching, record pair or record group comparisons, and a decision model (Guiver 2011). These are briefly described below. However, while data cleaning and standardisation are common to both deterministic and probabilistic linkage, the other three processes are more relevant to the probabilistic method.

Data cleaning and standardisation

The data files to be linked may often have been collected by different agencies and coded differently. For example, the response categories for some data sets may have numerical codes while others may have alphabetical codes. The order of the response categories may also be different. Some files may also contain errors. Cleaning and standardisation identify and remove the errors and inconsistencies in the data, and parses the text fields (such as residential address) so that the data items in each data file are comparable (Guiver 2011).

Blocking

Data linkage often involves large data sets. When two data sets are linked, the number of possible comparisons equals the product of the number of records in the two data sets. Blocking reduces the number of comparisons needed, by only comparing record pairs where links are more likely to be found. Blocking involves selecting sets of blocking attributes, such as sex, date of birth, last name or components of first and last name, and only comparing records with the same attributes (Christen & Goiser 2007).

Record pair comparisons

During the comparison stage, record pairs are compared on each linkage field, and the level of agreement is measured. Field comparison weights are assigned to each linkage field for each record pair. The field comparison weights are then summed over the linkage variables to form a record pair comparison weight.

Decision model

Record pair comparison weights help data linkers decide whether a record pair belongs to the same entity. This decision can be based on a single cut-off weight or on a set of lower and upper cut-off weights. Under the single cut-off weight approach, all record pairs with a comparison weight equal to or above the cut-off weight are assigned as links and all those below the cut-off weight are assigned as non-links. Under the lower and upper cut-off weights approach, all record pairs with a comparison weight below the upper cut-off are assigned as links and those with weights below the cut-off are assigned as non-links. Record pairs with comparison weights between the upper and lower cut-offs are assigned as possible links, and designated for clerical review. In clerical review, data linkers manually inspect all the variables available for the record pairs whose link status cannot be automatically determined, and then decide whether the record pairs belong to the same entity.

1.5.5 Statistical linkage keys

The statistical linkage key (SLK) is a code that replaces a person's first name and last name, to protect the person's identity. It may be defined as "a derived variable used to link data for statistical and research purposes that is generated from elements of an individual's personal demographic data and attached to de-identified data relating to the services received by that individual" (NCSIMG 2004).

Generally, most SLKs are constructed from last name, first name, sex and full date of birth. SLKs protect privacy and data confidentiality because they serve as an alternative to a person's names and dates of birth being on the data sets to be linked.

A commonly used statistical linkage key is the SLK581, which is now used in a number of community services data collections. It consists of the concatenation of the 2nd, 3rd and 5th letters of the family name, the 2nd and 3rd letters of the given name, date of birth as a character string of the form '*ddmmyyyy*', followed by the character '1' for male and '2' for female. This is known as a '581' SLK, because it comprises five (5) characters from a person's first and last names, eight (8) characters from date of birth and one (1) character representing the person's sex, a total of 14 characters (see (Karmel 2005) for a more detailed description of the 581 SLK).

Data linkage using an SLK is commonly deterministic (Karmel et al. 2010), but this requires the variables used in constructing the SLKs to be accurate, complete and as exact as possible. This means that names must be spelt correctly, precisely and consistently across all databases. First name and last name must also not be transposed. Difficulties with SLK construction may be experienced where variations of names are used on different databases (e.g. Dick/Richard; Smith/Smythe; Thompson/Thomson; Bernice/Bernadette/Bernie etc.). Non-common, ethnic or Aboriginal and Torres Strait Islander names may be spelt wrongly or inconsistently across multiple databases. Last names may also change, for example due to marriage or divorce.

There are two kinds of errors associated with SLKs. Firstly, there may be incomplete or missing data items on an individual's record, which means that the SLK will be incomplete.

Secondly, errors in the source data may lead to the generation of multiple SLKs for the same individual, or to multiple individuals sharing similar identifying information, and as a consequence, multiple individuals will share the same SLK (Bass & Garfield 2002). SLKs are likely to be more problematic for Aboriginal and Torres Strait Islander people because of known errors with names and dates of birth (see section 3.2 and Bass & Garfield 2002).

To overcome these problems, the AIHW has developed a more sophisticated stepwise deterministic method which uses auxiliary information on the data sets to provide a platform from which variation in the reported SLK information can be considered (Karmel et al. 2010).

Additionally, where a statistical agency or data linking institution recognises potential problems with their SLK, they may combine deterministic linkage and linkage with SLKs with probabilistic linkage, which requires less exacting standards of accuracy, stability and completeness (Bass & Garfield 2002; Karmel et al. 2010).

1.6 Protocols of data linkage

There are several data linkage units operating in Australia, each with its own protocols (that is, processes and procedures that govern how data linkage is undertaken and by whom). These are intended to ensure:

- secure management of data
- that privacy and data confidentiality are protected
- that data linkage is conducted in an open and accountable way.

They also serve to assure the public that data linkage projects are subject to various rigorous approval processes, legislative provisions and protocols to protect privacy and data confidentiality.

Data linkage units with a focus on health and welfare currently operating in Australia include:

- Australian Institute of Health and Welfare (AIHW)
- Australian Bureau of Statistics (ABS)
- Western Australia Department of Health
- Centre for Health Record Linkage (CHeReL) in NSW/ACT
- Victoria Data Linkages
- SA-NT DataLink
- Queensland Centre for Health Data Services, Health LinQ
- Menzies Research Institute/Tasmanian Data Linkage Unit
- Centre for Data Linkage at Curtin University.

Each of these data linkage units has enacted strict protocols to protect privacy and confidentiality. Details about the protocols of the various institutions can be found at the websites below.

AIHW

<http://www.aihw.gov.au/WorkArea/DownloadAsset.aspx?id=6442454736>

The following references may also be consulted: (AIHW 2004, 2006; Karmel 2005)

For more information about confidentiality, see the Confidentiality Information Series:
<http://nss.gov.au/nss/home.NSF/pages/Confidentiality+Information+Sheets>

The Commonwealth

High-level principles for data integration involving Commonwealth data for statistical and research purposes

<http://www.nss.gov.au/nss/home.nsf/NSS/00FB7E20E1D56B96CA2577F20016C3DB?opendocument>

NSW and ACT (CHeReL)

<http://www.cherel.org.au/hrec.html>

<http://www.cherel.org.au/CHeReL_flyer.pdf>

<http://www.cherel.org.au/HowCHeReLworks.pdf>

Victoria (Victoria Data Linkage Unit)

<http://www.health.vic.gov.au/vdl/downloads/introduction_to_vdl.pdf>

Queensland (Queensland Centre for Health Data Services, Health LinQ)

<http://www.healthlinq.org.au/health-linq>

<http://www.healthlinq.org.au/privacy>

Western Australia (Western Australian Data Linkage System)

<http://www.datalinkage-wa.org/privacy-and-security>

<http://www.datalinkage-wa.org.au/sites/default/files/DLB%20Access%20Policy%20Dec2010%282%29.pdf>

South Australia and the Northern Territory (SA-NT DataLinkage)

<https://www.santdatalink.org.au/FAQs>

<https://www.santdatalink.org.au/security>

Tasmania (Tasmanian Data Linkage Unit)

<http://www.menzies.utas.edu.au/article.php?Doo=ContentView&id=1055>

Population Health Research Network (PHRN)

<http://www.phrn.org.au/about-us/what-is-data-linkage>

<http://www.phrn.org.au/about-us/privacy-and-security>

Centre for Data Linkage

<http://www.phrn.org.au/participants/centre-for-data-linkage>

<http://www.phrn.org.au/about-us/privacy-and-security>

2

Values and ethics in Aboriginal and Torres Strait Islander research

Principle

The conception, design and conduct of all Aboriginal and Torres Strait Islander data linkage activities for statistical purposes should be guided by the core values and ethics of Aboriginal and Torres Strait Islander human research.

Guidelines

2.1 Data linkage activities and research relating to Aboriginal and Torres Strait Islander people should be conducted in accordance with NHMRC guidelines regarding Aboriginal and Torres Strait Islander values and ethics in health research.

2.2 Members of the project team should engage with Aboriginal and Torres Strait Islander people and seek relevant community support and input into the design of the project, especially in determining its aims and objectives, governance, risks and overall appropriateness to achieve the intended outcomes.

2.1 Background

The results of data linkage studies may have a direct impact on the wellbeing of Aboriginal and Torres Strait Islander people through, for example, the formulation of policies and programs, changes to service delivery, or funding decisions. Such data linkage studies must therefore be undertaken in a way that is consistent with Aboriginal and Torres Strait Islander values and ethics. Without input from Aboriginal and Torres Strait Islander people, there is the risk that projects will not achieve their desired outcomes.

2.2 Aboriginal and Torres Strait Islander core values

Aboriginal and Torres Strait Islander communities are not homogeneous (NHMRC 2003). Each culture has its own 'established and respected values and protocols … There are however six common values that have been identified as being important to all Aboriginal and Torres Strait Islander people' (NHMRC 2005).

These core values are summarised below, and full details can be found in the NHMRC documents *Values and ethics: Guidelines for ethical conduct in Aboriginal and Torres Strait Islander health research* (NHMRC 2003) and *Keeping research on track: A guide for Aboriginal and Torres Strait Islander people about health research ethics* (NHMRC 2005).

Reciprocity: Reciprocity refers to an obligation among Aboriginal and Torres Strait Islander people to achieve an equitable distribution of resources, including benefits from research.

Respect: Respect includes trust, cooperation and respect for human dignity. Respectful research relationships acknowledge and affirm the right of people to have different values, norms and aspirations. In the research context, respect also includes consultation and engagement with Aboriginal and Torres Strait Islander people, and valuing their knowledge and contribution to the research effort.

Equality: Equality refers to 'equal value of people'. Equality does not mean 'sameness'. In the research context 'equality' refers to a 'commitment to distributive fairness and justice'.

Responsibility: Underlying the value of responsibility is the obligation to do no harm. In the research context, researchers must ensure that the research is beneficial and not harmful to Aboriginal and Torres Strait Islander people.

Survival and protection: Refers to the determination of Aboriginal and Torres Strait Islander people to protect their cultures and identity from erosion by external forces. Research must not be used to undermine Aboriginal and Torres Strait Islander culture, solidarity or distinctiveness. Research must not be used to exploit Aboriginal and Torres Strait Islander people or to contribute to discrimination and derision of Aboriginal and Torres Strait Islander people just for the sake of knowledge.

Spirit and integrity: Spirit and integrity are the overarching values that bind the other five values into a coherent whole. Spirit and integrity demonstrate the continuity of Aboriginal and Torres Strait Islander culture and core values over time; research should not be used to harm or destroy Aboriginal and Torres Strait Islander culture or its core values over time.

2.3 Project approval

In addition to obtaining approval from data custodians and a human research ethics committee, some data linkage studies may also require approval from an Aboriginal and Torres Strait Islander Ethics Committee. Researchers may be asked to demonstrate:

- that they have engaged with Aboriginal and Torres Strait Islander people and sought appropriate community support for the study
- how Aboriginal and Torres Strait Islander people can contribute to the study, in terms of knowledge, wisdom and experience
- that they have taken into consideration Aboriginal and Torres Strait Islander input into the design of the project, especially in determining its aims and objectives, governance, risks, sharing of benefits and acknowledgement
- how they intend to disseminate and facilitate the translation of the results of the study to Aboriginal and Torres Strait Islander communities, representative bodies and organisations.

In most cases, researchers will be directed to the appropriate ethics committee when applying for linked data. For example, researchers proposing to use the Western Australian Data Linkage System to undertake data linkage projects that focus on Aboriginal and Torres Strait Islander outcomes are required, in addition to all other requirements, to submit their proposal to the Western Australian Aboriginal Health Information and Ethics Committee (WAAHIEC). Such requirements are in place to ensure that research projects that focus on Aboriginal and Torres Strait Islander people meet the requirements as set out by the NHMRC (NHMRC 2003).

Further information on Aboriginal and Torres Strait Islander human research ethics committees as well as the requirements for submitting human research proposals involving Aboriginal and Torres Strait Islander people may be found by exploring the following links.

National

<http://www.nhmrc.gov.au/_files_nhmrc/publications/attachments/e52.pdf>

<http://www.indigenoushealthethics.net.au/hrec>

New South Wales

Aboriginal Health & Medical Council Research Committee of New South Wales

<http://www.ahmrc.org.au/Ethics%20and%20Research.htm>

University of Sydney Human Ethics Committee

<http://sydney.edu.au/research_support/ethics/human/>

<http://www.indigenoushealthethics.net.au/nsw_ethics>

Ethics Committees with Indigenous members

<http://www.indigenoushealthethics.net.au/nsw_hrecs>

Victoria

Victoria Aboriginal Ethics Project

<http://www.onemda.unimelb.edu.au/docs/VicAboriginalEthicsProjectReport.pdf>

Queensland

Indigenous Health Ethics Network

<http://www.indigenoushealthethics.net.au/qld_ethics>

Western Australia

Western Australian Aboriginal Health Information and Ethics Committee (WAAHIEC)

<http://www.aboriginal.health.wa.gov.au/ethics/index.cfm>

<http://www.research.murdoch.edu.au/ethics/hrec/Policies/WAAHIEC%20guidelines.pdf>

<http://www.aboriginal.health.wa.gov.au/docs/WAAHIEC_TOR.pdf>

South Australia

South Australia Aboriginal Health & Research Ethics Committee

<http://www.ahcsa.org.au/media/docs/ahrec_info.pdf>

The Flinders University of South Australia and Flinders Medical Centre Social and Behavioural Research Ethics Committee

<http://www.flinders.edu.au/about_research_files/Documents/Info%20for%20Research/Ethics%20and%20Biosafety/SBREC/TOR_SBREC.pdf>

Northern Territory

Human Research Ethics Committee

<http://www.menzies.edu.au/about-us/board-committees/human-research-ethics-committee>

ACT

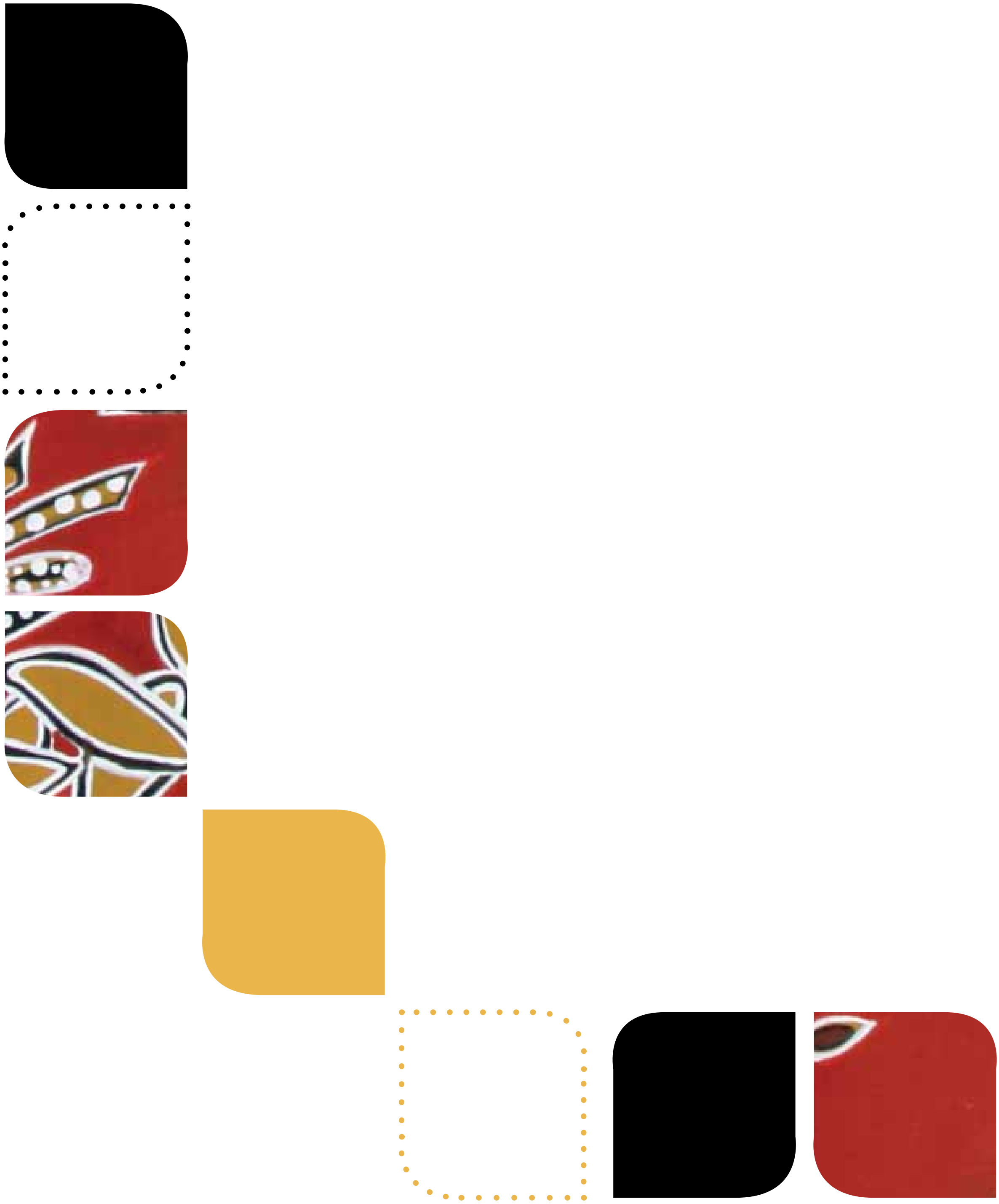
Australian Institute of Aboriginal and Torres Strait Islander Studies Research Ethics Committee

<http://www.aiatsis.gov.au/research/ethicsreview.php>

Tasmania

Human Research Ethics Committee (Tas Network)

<http://www.utas.edu.au/research/integrity-and-ethics/human-ethics>



3

Quality of Indigenous status information in data collections

Principle

The quality of Indigenous status information within datasets included in the linkage study should be considered before analysis.

Guidelines

3.1 Analysts should investigate the data collection practices and data collection environment for each data set involved in the linkage study, in order to understand quality limitations with the Indigenous status variable.

3.2 Analysts should consult with data custodians in investigating the quality of the Indigenous status variable.

3.1 Background

Indigenous status, that is, whether or not a person identifies as Aboriginal and/or Torres Strait Islander, may be incorrectly or inconsistently recorded within and across data sets. In a data linkage study, an individual's Indigenous status can be derived from the information available across all of the linked data sets. The challenge for analysts is how to define an individual's Indigenous status for the purpose of the study given the multiple sources of information available, which may have varying degrees of quality, and may be inconsistent or incomplete. The first step in addressing this issue is to understand the factors affecting the quality of Indigenous status in the contributing data sets.

The aim of this chapter is to provide analysts with information about factors to consider when exploring the quality of Indigenous status within a data set.

3.2 Factors affecting the quality of Indigenous status information

In an administrative data setting, a range of factors influences the recorded Indigenous status of an individual. These factors fall into two main domains:

- the data collection environment
- the respondent environment.

These factors should be explored to help analysts understand why Indigenous status may be inconsistent or incomplete across or within datasets. Data custodians will be well placed to help analysts in this regard.

3.2.1 The data collection environment

The data collection environment comprises data collection systems and practices, protocols for collecting information on Indigenous status, including training given to staff to collect that information, and quality control.

Importance of Indigenous status information to the funding and operation of data providers and data collection agencies

The collection, processing and analysis of Indigenous status information are fundamental to the funding, operation and strategic direction of particular service providers, agencies and institutions. Where this is the case, the agency or institution concerned may be more likely to attach a lot of importance to the collection of accurate, consistent and reliable information on Indigenous status.

Such institutions are likely to follow consistent and quality assured standards and protocols, such as the ABS (1999) and AIHW (2010a) standards and protocols, on the collection of Indigenous status information, to have their staff properly trained, monitored and supervised in the collection of Indigenous status information, and to also carry out regular audits and quality assurance checks on the quality of Indigenous status information collected.

Data sets from these institutions are likely to be considered 'trusted' data sets for purposes of choosing appropriate methods for deriving Indigenous status from linked data sets.

Variation across collection points in how the Indigenous status question is framed

There is a national standard for how Indigenous status information should be collected and recorded (ABS 1999). Adherence to the standard, both in terms of the question asked and how it is coded within systems, supports comparability and consistency across data collections. Analysts should consider the impact of deviations from the standard. The *National Best Practice Guidelines for collecting Indigenous status in health data sets* (AIHW 2010a) provides further information about how Indigenous status information should be collected and recorded in health settings. Principles in these guidelines are largely transferable to other sectors.

The recommended format of the question is:

Q1. [Are you] [Is the person] [Is (name)] of Aboriginal or Torres Strait Islander origin?

(For persons of both Aboriginal and/or Torres Strait Islander origin, mark both 'Yes' boxes.)

No □

Yes, Aboriginal □

Yes, Torres Strait Islander □

While most agencies follow the above format, some do not. This leads to inconsistencies and lack of comparability in the value of the resulting Indigenous status information.

Variation across collection points in how the Indigenous status question is asked

If there are variations across agencies and service providers in how the Indigenous status question is asked, then there will be variations and inconsistencies in the quality of the resulting information. The official protocol is to ask the Indigenous status question, in the same way, of all persons accessing a service. This protocol is often not consistently applied across agencies or even within the same agency.

Recorders of information at various service or data collection points may assign one Indigenous status to a person at one time and a different Indigenous status to the same person at another time within the same facility or service point. Recorders in some facilities may be more thorough and consistent than others, in asking, probing, eliciting and recording responses to the question on Indigenous status. Other recorders may not ask a person for his or her Indigenous status, and may, instead, assume the person's Indigenous status from the way the person presents.

Variation across collection points in how information on Indigenous status is recorded

Some agencies may record an Indigenous status for a person whenever the person presents for a service, while other agencies may update a person's previous Indigenous status with the one obtained from the current contact with the person. Other agencies may not ask the question at each contact with the person, and may thus have outdated information.

Where an Indigenous status is recorded at each contact, then it is possible to have multiple Indigenous status information for the same client at different periods of time, and to be able to observe the trend or changes in Indigenous identification over time.

Regular audits and quality checks

Some agencies may institute regular audits or quality checks to evaluate the quality of the Indigenous status information collected. Such audits or quality checks may include whether the question on Indigenous status is asked consistently of all clients and properly recorded according to the agency's own protocols. They may also check for consistency between the recorded Indigenous status on the agency's databases, and Indigenous status as obtained from periodic audits. The nationwide hospital audit project conducted by the AIHW (2005) is an example of this. In this project, a sample of hospital patients were interviewed with regard to their Indigenous status, and this information was compared against the already recorded Indigenous status of the same patients in the hospital databases.

In addition to being useful for analysts in assessing the quality of Indigenous status, information on these audits and quality checks, even where conducted by the agency itself, could be used in improving the agency's data collection systems and practices, especially in the area of staff training, documentation, data collection and data processing.

Data collection quality assurance

Analysts should consider the quality assurance practices implemented in collecting Indigenous status information. This may include whether or not incomplete Indigenous status information is routinely followed up, or whether or not quality audits are undertaken.

3.2.2 The respondent environment

The respondent environment comprises the person accessing a service, and third parties, comprising family members, friends, data compilers or service providers, who are sometimes called upon to provide information on the Indigenous status of their clients. Service providers include both private operators and public servants, including teachers, midwives, Centrelink staff, medical doctors, funeral directors etc.

Respondent variation

In some instances, Indigenous status is not collected directly from the individual. This may impact on the consistency of the information when comparing across and within data collections.

For example, a person's Indigenous status may be:

- provided by a relative, as is the case for death registrations or school enrolment
- recorded based on the assumptions made by those responsible for collecting the data, rather than asking the Indigenous status question directly
- derived from other information, such as in birth registration data where the Indigenous status of the baby is usually based on the Indigenous status of the parents.

Inconsistencies and errors can sometimes occur because the person whose information is being reported is not always the one asked the Indigenous status question.

Individual choice

Given that Indigenous status information is generally based on self-identification, the factors affecting individual choice to respond to the question should also be considered when assessing the Indigenous status variable. An individual may legitimately choose to report different Indigenous status across data collection points and some people may refuse to respond because they consider it irrelevant to the service being provided. For Aboriginal and Torres Strait Islander people, other barriers may impact their choice to identify, including experiences of past discrimination and the data collection setting.

Confusion between non-response and being 'non-Indigenous'

The quality and reliability of Indigenous status information on data sets is impacted by the proportion of non-responses, or "don't know" records on the data sets. Some non-Indigenous clients or respondents may confuse non-response with being ' non-Indigenous', and may leave the question blank or unanswered instead of marking the 'non-Indigenous' response. In data processing and data analysis however, a blank data field to the question of Indigenous status is not the same as being non-Indigenous.

3.3 Checklist for understanding the quality of Indigenous status information

The following checklist aims to summarise the issues analysts should consider when assessing the quality of Indigenous status within data collections.

Table 3.1: Summary of issues to consider when assessing the quality of Indigenous status within data collections

Understanding the collection process	
Asking the question	Is the standard question used? Are the standard response categories used? Is the information collected directly from the individual? Can an individual's response be updated or revised? How have procedures changed over time?
Quality assurance procedures	Are staff trained in why and how to ask the question and record responses? Are follow-up procedures in place for non-response?
Assessing the quality	
Not-stated responses	What is the level of not-stated response? Has the level of not-stated responses changed over time?
Quality assessments	Do objective quality assessments exist? (e.g. audits, previous data linkage studies) Are the results of quality assessments available? Do other opportunities for objective quality assessment exist?

4

Quality of linkage variables

Principle

Linkage variables should be assessed before linkage to gauge their accuracy, completeness and comparability, and to ensure that they are of sufficient quality to support the purposes of the linkage study.

Guidelines

4.1 Analysts should be particularly mindful of the impact of data quality on linkage quality and outcome measures based on linked data sets.

4.2 Analysts should consult with data linkers and custodians to obtain relevant information on how both parties have:

1. evaluated, cleaned and standardised data linkage variables to ensure that linkage variables are accurate, stable, complete and comparable
2. evaluated the quality of the linking variables before deciding on the choice of linkage method.

4.3 Analysts should also work with data custodians and data linkers to put in place a system where they provide feedback and work in consultation with data linkers and data custodians to continuously improve the quality of data for purposes of data linkage.

4.1 Background

In data linkage, records are linked on the basis of common identification data, also known as linkage variables. The types and quality of the linkage variables will determine what type of linkage will be undertaken. In Canada and the United States, for instance, a unique identification number, such as a Social Security or Medicare number, may exist, allowing records to be matched across multiple data sets (Manitoba Centre for Health Policy 2006). Where there is no such unique identification number, the linkage is more complex and will be undertaken on the basis of groups of personal, social, demographic and geographic variables. Linkage will then occur depending on how closely the linkage variables match across the data sets being linked.

Common types of variables used in data linkage include the following:

- personal variables (e.g. names, including first and last names, nicknames, abbreviations and for Aboriginal and Torres Strait Islander people, clan or nation names, if available)
- demographic variables (e.g. sex and full date of birth)
- geographic variables (e.g. usual residential address or part of address, such as suburb or postcode and residential address of service provider, such as an aged care home)
- dates of specific events (e.g. date of admission to or separation from hospital or date of death). (Dates of events are particularly important in data linkage involving hospital data where names are often unavailable as linkage variables. The combination of date of birth and date of specific events, such as date of death, often provides a unique identification that enables records to be linked.)
- other socio-economic variables (e.g. marital status, religion, language spoken at home).

The quality of linkage variables has a direct impact on the quality of the linkage and on any analysis based on the linked data set. Assessing the quality of the linkage variables is therefore an essential component of any data linkage. This chapter aims to provide an overview of the issues associated with the quality of linkage variables, with a focus on issues more prevalent for the Aboriginal and Torres Strait Islander population. More technical detail is available in Herzog, Scheuren and Winkler (2007). Should the quality of datasets or linkage variables be of significant concern in the datasets being linked, the merit of undertaking the data linkage activity should be carefully considered.

Ideally, data linkers and analysts should work together with data custodians and data providers to ensure uniformity in the collection and processing of key data items that are required for data linkage activities relating to Aboriginal and Torres Strait Islander people, and to make these variables available for data linkage.

4.2 Quality issues for linkage variables

As with all variables on an administrative data set, linkage variables are subject to errors in data entry, such as spelling, recording and transposition errors (for example, incorrect spelling of a name, or transposition and incorrect recording of birthdates) as well as omissions or missing values. Linkage variables are also subject to additional quality issues.

The comparability of linkage variables across data sets should also be considered and may be affected by a range of factors including:

- variations in the spelling of names across multiple data sets (e.g. John Smith, Jon Smythe) or the use of nicknames and multiple forms of names (e.g. Thomas, Tom, Tommo; Richard, Dick; Alfred, Fred or Alf, in which case the initial may be ‘A’ or ‘F’)

- differences due to changes over time (e.g. residential address may be different across data sets, depending on when the addresses were last recorded) and changes in name, for example, due to marriage
- differences in the use of standards and classifications to record information (e.g. place of birth may be reported as 'country', 'town' or 'community' on one data set, and as full residential address on another, or different categories of marital status in use on different data sets, such as 'single', 'married', 'divorced' and 'widowed', on one data set, compared to 'never married', 'married', 'de facto', 'separated', 'divorced' and 'widowed', on another).

In light of these considerations, data linkers first evaluate, clean and standardise the linkage variables before using them for data linkage. This involves evaluating likely errors and the extent to which these errors can affect the quality of the data linkage, and whether data exist elsewhere to allow the errors in the input data sets to be corrected. Where the coding of linking variables varies between input data sets, it is standardised to the same format on all the input datasets before they are used for linkage.

There is evidence that name, date of birth and address variables may be subject to more variation and be less consistently reported among Aboriginal and Torres Strait Islander Australians than among other Australians. If these variables are particularly subject to error or variation, then other less discriminating variables, such as socio-economic variables, may be needed as linkage variables. These issues are described in more detail below.

4.2.1 Name changes

First and last names are critical linkage variables in both probabilistic and deterministic data linkage (see 1.5.3). Variations in the spelling of names can affect the quality of the linkage. This may be particularly the case for the spelling of traditional Aboriginal and Torres Strait Islander names which may have a different structure to European-type names, with its own set of nicknames, aliases and diminutions. These issues can affect the quality of linkage, since names are usually the key variables used in data linkage.

Sayers et al. (2003) examined an Aboriginal birth cohort of mobile subjects belonging to diverse cultural and language groups in the Northern Territory and found that Aboriginal children had multiple names relating to kinship, clans and relationships with family groups, and that name changes often occurred following the death of another community member.

Sayers et al. (2003) found that, out of a sample of 686 Aboriginal mother-child pairs living in the Top End of the Northern Territory, by the age of four years:

- 30% of children had changed their names at least once
- 18% had changed address once
- 2% had had three different name changes
- 2% had had four different addresses.

If name changes are prevalent in remote Aboriginal and Torres Strait Islander communities, they could have an impact on the stability of the name variable used in data linkage. This issue is further complicated by the diversity in Aboriginal and Torres Strait Islander communities, with a range of practices with respect to alternate names between different groups, between people with a more Western lifestyle and those with a more traditional lifestyle, and between urban and remote areas. Direct community consultations are likely to afford the most accurate information about local practices such as naming conventions.

The difficulty of linking individuals with different names recorded, combined with the prevalence of name changes for Aboriginal and Torres Strait Islander people, makes it important for those linking data to actively develop strategies to address this issue. Depending on the data linkage activity being undertaken, the impact of not successfully addressing multiple names for individuals can be to over-represent the number of Aboriginal and Torres Strait Islander people in linked and merged datasets. This occurs through not identifying the different named records as the same individual and thereby recording individuals more than once. Alternately, where the linking of records is used to enhance records in one data set, there may be a greater number of Aboriginal and Torres Strait Islander records unlinked.

In certain instances, comprehensive name registers exist and are used to check, correct and standardise the spelling of names before they are used for data linkage. These registers often contain mostly European-type names. It may become necessary in the future to build such a database exclusively for Aboriginal and Torres Strait Islander names.

The problem of alternate names or aliases also exists in other data sets such as homelessness data, so solutions could possibly be imported across from these. The solution may be at the source of data collection in some cases, such as explicitly asking respondents what other names they are or have been known by. However, attempts to change names may result in two genuinely different people being matched, and creating a false match.

4.2.2 Date of birth

There is evidence that some older Aboriginal and Torres Strait Islander people, particularly those living in remote communities, have had difficulties providing date of birth information. In the Northern Territory, for instance, date of birth was not recorded on death registration forms until 1994; only age at death was recorded in prior years. A high proportion of older Indigenous people in the Northern Territory do not know their exact age and have only an approximate year of birth (Condon et al. 2004; Condon et al. 1998). In some cases, recorded age is calculated from approximate date of birth.

Even where date of birth information has been reported, they have been known to suffer from data quality problems that might impact on the quality of data linkage or on the quality of analysis based on that information. These data quality problems include inaccurate, incomplete or missing components of the date of birth information, date of birth information in one data set not matching that provided in another data set, and date of birth information that is not consistent with other demographic information, such as parity, school participation or residence in a residential aged care home.

The AIHW enhanced mortality database project also found many Aboriginal and Torres Strait Islander records that contained information that was inadequate for data matching. In the Residential Aged Care dataset, for instance, the linkage process revealed that up to 26% of death records relating to Aboriginal and Torres Strait Islander people had dates of birth that were likely to be imputed or estimated, such as 1 January or 30 June. A similar observation was made with regard to the National Hospital Morbidity dataset, where during the matching process, about 11% of Indigenous hospital records were identified as having imputed or estimated birth dates (AIHW: Choi et al. 2012).

Date of birth information on some data sets may therefore be incomplete (it may only include year of birth), or inaccurate (it may only include approximate year of birth or approximate date of birth calculated from approximate age). These quality issues affect the quality of data linkage.

4.2.3 Mobility and levels of geographic reporting

Research has found high levels of mobility among remote Aboriginal and Torres Strait Islander communities, which may result in different levels of reporting for the address variable (Memmott et al. 2004). For instance, address may be recorded as:

- town
- suburb
- community region.

Consequently, in some datasets the full address may be recorded, while in others, community name may be the only address information available.

Mobility also affects the stability and completeness of reporting of the address variable, thereby limiting its use as a stable variable for linkage. In some data collections, such as the Census, information on address in the past 12 months and past 5 years is collected. Information on previous address can then be used, where necessary, to supplement information on current address.

For example, analysis of the 2001 and 2006 Censuses showed that between the 2001 and 2006 Censuses, about 43% of Aboriginal and Torres Strait Islander people changed their usual place of residence. Of these, 14% had moved to a different state. There were also large movements between remoteness areas. Between 2001 and 2006, 12% of Aboriginal and Torres Strait Islander people (aged 5 and over) had moved between remoteness areas (ABS 2009b, 2010).

These movements are age-related. In 2006, Aboriginal and Torres Strait Islander people aged 5–19 accounted for 43% of net movements between remoteness areas, and were most likely to move to inner regional areas. This age group were also most likely to leave remote and very remote areas, accounting for 45% and 57% of the movement out of these areas (ABS 2009b, 2010).

In addition to these movements, there may also be shorter-term movements that may not be captured by the Census. This may often involve repeat movements of less than 12 months at a time, which may be even more frequent than the medium and longer-term movements recorded in Census data.

All these movements may have implications for data linkage, especially where the movements involve accessing services such as education and health, as well as income, employment and housing support. Frequent mobility and change of address mean that a person's recorded address may vary between services accessed and between data sets. Under these circumstances, a person's address may not be a reliable variable for data linkage, as it will lack stability and consistency. The use of address as a blocking or linking variable should therefore be a last resort, even though other variables such as name and date of birth are also known to have data quality issues.

In addition to the high level of mobility and frequent changes of residential address, data linkers and analysts may also need to be aware of issues regarding the types of households that Aboriginal and Torres Strait Islander people live in, and the impact these will have on data linkage. Not all Aboriginal and Torres Strait Islander people live in standard households where the household occupies one place of residence. According to Memmott et al. (2004), there is strong evidence in remote Aboriginal communities of linked households or clustered households that are characterised by an extended family group dispersed across a number of places of residence. People may move between several residences and not regard themselves as having a single usual place of residence (Memmott 2011). The use of a single address as a personal identifier is the norm in many data sets in Australia, including the Census, but this may be a poor fit for the cultural practice of clustered households.

4.3 Impact of quality of linkage variables on data linkage

There are specific data quality issues that affect choice of linkage method, quality of data linkage and outcome measures based on linked data that are particularly relevant to Aboriginal and Torres Strait Islander people. It is important that data linkers and analysts are aware of these issues and take them into account in designing their data linkage strategies and interpreting any analyses based on linked data. These issues are discussed broadly below, with references provided for more detail.

A number of studies have examined the impact of variations in the quality of key data linkage variables on linkage quality and outcome measures estimated from linked data sets (Bass & Garfield 2002), (Karmel et al. 2010) and (AIHW 2011). These studies found that match accuracy and match accuracy rate, including the proportion of false-positives, as well as outcome measures, such as number of days in hospital and relative risk of death, varied considerably, depending on the quality of the linking variables and the linking method used.

As a result of their analysis, Bass and Garfield concluded that linkage of data for Aboriginal and Torres Strait Islander people was more difficult than linkage involving other cultural groups. The authors attributed these difficulties to frequent name changes and relatively poor recording of dates of birth and other demographic details (Bass & Garfield 2002).

To overcome some of the problems associated with inconsistently reported linkage variables, the AIHW has developed a sophisticated stepwise deterministic method which uses auxiliary information on the data sets to provide a platform from which variation in the reported SLK information can be considered (Karmel et al. 2010).

Although the stepwise deterministic linkage strategy was not specifically developed for Aboriginal and Torres Strait Islander data linkage, it is particularly useful in this case because of the considerable variability in Aboriginal and Torres Strait Islander names and dates of birth, as well as missing or incomplete values of other linkage variables across data sets. This variability can affect the quality of SLK-based deterministic linkage involving Aboriginal and Torres Strait Islander people. The 'stepwise deterministic' linkage offers an approach to enhancing deterministic or SLK-based data linkage involving Indigenous persons.

The stepwise deterministic linkage strategy takes account of variation in the statistical linkage keys, as well as additional variables on the data sets to be linked, and uses these additional variables to enhance the simple deterministic linkage (see section 1.5.3). Readers interested in reading more about the stepwise deterministic linkage strategy should refer to Karmel et al. (2010) and AIHW (2011).

5

Assessment of quality of data linkage

Principle

The quality of data linkage should be assessed and understood. Any limitations arising from the quality of the data linkage should be taken into account in the analysis of the linked data.

Guidelines

5.1 Analysts should understand the quality of the data linkage before the linked data are analysed or used to derive Indigenous status.

5.2 Analysts should be able to understand and interpret the results of the quality assessment of the data linkage.

5.3 Analysts should be proactive in requesting specific quality checks on the linked data, taking into account the data sets to be linked and the nature of the planned analysis. These may include sensitivity and specificity tests as well as tests involving the internal consistency of data items on linked data sets.

5.1 Background

This chapter aims to highlight the importance and features of data linkage quality assessments. The ultimate goal of such quality assessments is to ensure that limitations arising from the data linkage process are taken into account during analysis. The types of measures used to evaluate data linkage quality will depend on the size and characteristics of the input data sets, the processes and software used in undertaking the data linkage, and what other information is available to be used in the quality assessment. As such, this chapter is intended as a guide only, and analysts are encouraged to discuss these issues with their data linkage institutions.

The quality of the data linkage will affect the choice of methods that can be used to derive Indigenous status from a linked data set. For example, a high false-positive rate in the data linkage can increase the risk of misclassification of non-Indigenous people as Aboriginal and Torres Strait Islander if an ‘ever-Indigenous’ approach is used in deriving Indigenous status (the ‘ever-Indigenous’ approach classifies a person as ‘Indigenous’ if they are identified as ‘Indigenous’ at least once in any record or data set).

In addition, information obtained from the quality assessment of the linkage can be fed back to data custodians and data collecting agencies to be used as input into their data collection, data processing and data improvement strategies. Information from quality assessment of data linkage can also be used to improve data linkage methods and strategies.

5.2 Determinants of linkage quality

Key determinants of linkage quality include the quality of the statistical linkage keys, in the case of deterministic linkage or the quality of the blocking and linkage variables (see 1.5.4), in the case of probabilistic linkage. Where the linkage method is deterministic, and statistical linkage keys, rather than unique identifiers are used to compare the records, then the quality of the linkage will depend on the quality of the variables used in constructing the SLKs as well as the additional variables used to assist linking. These variables, as already noted, include first and last name, sex and date of birth.

In addition to having accurate, reliable and consistently reported blocking and linking variables, the quality of data linkage also depends on the blocking and linking strategy adopted (see section 5.2.2).

5.2.1 Quality of blocking and linking variables

Blocking and linking variables include full name, sex, full date of birth and residential address. Blocking and linking variables are used in constructing statistical linkage keys, in the case of deterministic linkage, or in linking the data sets, in the case of probabilistic linkage.

Poor quality blocking and linking variables could lead to some records not being linked or to some records being linked to the wrong records. Chapter 4 'Quality of linkage variables' (section 4.2.1) shows that there may be quality issues associated with how Aboriginal and Torres Strait Islander names, date of birth and geographical variables are recorded in data collections.

All these errors can affect the quality of the linkage and the types of analysis that the linked data can support. In addition, some records are unable to be linked because other key blocking and linking variables are missing, incomplete, inaccurate or are inconsistently reported on the data sets to be linked.

It is important to note that Indigenous status should not be used as a blocking or linking variable because in most cases, it is the measure that is intended to be determined from the linkage. It is also unsuitable because it is known to be incomplete or missing in some data sets or inconsistently reported across data sets.

5.2.2 Blocking and linking strategy

The quality of the blocking and linking variables affects the blocking and linking strategy adopted for the linkage. A blocking and linking strategy refers to the variables used as blocking and linking variables, the number of iterations in which they are used to compare records, and how the blocking and linking variables are combined and used during each iteration.

Although analysts may not have access to some of the linking variables on the linked data sets they receive from the data linker (e.g. names and date of birth), it is important to discuss the blocking and linking strategy with the data linker, as part of the assessment of the quality of linkage. This is particularly relevant for linkage involving Aboriginal and Torres Strait Islander people, because key blocking and linking variables may change frequently, or they may be incomplete or inconsistently reported (e.g. name, date of birth and address) (Sayers et al. 2003).

Some records may not be able to be linked if the blocking and linking strategies are inappropriate and do not take into account the quality of the blocking and linking variables. An inappropriate blocking strategy could lead to block sizes being too big or unbalanced, or some records not falling into valid blocks. For example, where a blocking variable has many missing records or does not satisfy the criteria of completeness, accuracy and stability, then it could lead to some records not being able to fall into valid blocks for comparison.

Probabilistic data linking also requires the estimation of some parameters, such as m and μ probabilities, based on the data to be linked. While a discussion of these parameters is beyond the scope of this document, it is worth noting that the accuracy of these parameters will also affect the quality of the data linkage (Samuels 2012).

5.3 Assessing the quality of data linkage

There are two main ways in which the quality of data linkage can be assessed, and this depends on the data linkage model, that is, whether the linkage is a project (ad-hoc) linkage or is part of a systematic linkage program usually undertaken by data linkage institutions. These two data linkage quality assessment regimes are described in the following sections. Both models can use either deterministic or probabilistic methods.

Systematic data linkage quality assessment

Data linkage institutions that operate a data linkage system involving the creation and storage of master linkage keys constructed from multiple data sets often undertake a systematic program of quality assurance validation of their master linkage keys. The quality assurance checks aim to validate the linkages and to confirm that linked record pairs or groups of records do indeed belong to the same individuals.

Master linkage keys, including SLKs, require regular validation to ensure that errors in the data keys are detected and corrected so that the quality and integrity of research, policies, programs, services and policy reviews based on linked data are maintained.

Institutions such as CHeReL and WADLS that have well-established data linkage programs, also have in place quality assurance procedures to regularly review and validate their linkage keys to detect false links. These procedures include internal consistency checks of the data items used in developing the master linkage and statistical linkage keys.

A description of CHeReL's quality assurance checks can be found at:
<www.cherel.org.au/media/19151/2011_qa_report.pdf>

Analysts of linked data should consult with their data linkage institutions about the range and nature of checks undertaken. Analysts must be proactive in commissioning quality checks from their data linker, or in studying the results of quality checks, to ensure that possible errors are corrected, or that any analysis carried out on the linked data, or interpretation of results of analysis based on the linked data, are informed by the quality of the linked data.

The following are examples of quality checks undertaken by CHeReL, the principles of which can be applied to other linkage models, whether dynamic or ad hoc.

1. Dates of birth differ between records, and other personal details are also different.
2. Date of birth of baby is different between data sets.
3. Mother's age at birth of baby is less than 13 years.
4. Person ID contains more than 5 baby records and mother's name on Registrar of Births, Deaths and Marriages data set is different from mother's name on the Perinatal Data Collection.
5. Person ID contains more than 5 baby records and baby's records differ between data sets.
6. Date of birth of baby is earlier than date of confinement of mother.
7. Date at death is earlier than date of admission to hospital.
8. Date at death differs between records by 2 or more days.
9. Postcode varies between records, and day of birth varies by 2 or more days, and year of birth also varies by 5 or more years.
10. Person ID contains multiple death records, and dates of death are different.

Any record pairs or groups of records that fulfil any of the above criteria undergo further clerical assessment.

Quality assessment of ad hoc linkage

Quite often, linkage may be carried out on a project basis, specifically for a given project, and comprising specific data sets that are not part of a dynamic data linkage system. The linkage process and the linked data will therefore not be subject to the same dynamic quality assurance checks that are employed by data linkage institutions that have a dynamic data linkage system in place.

Institutions that carry out ad hoc data linkage must therefore employ their own quality assessment protocols that are consistent with the types of linkage they undertake and the types of data sets they link.

There are many ways in which the quality of data linkage and of the linked data set can be assessed. These include, but are not limited to, the following:

- review of quality of linkage through clerical assessment
- estimating and reviewing measures of quality of data linkage
- assessing the characteristics of unlinked records
- edit checks that analysts should perform before using the linked data.

Some of these checks may be best carried out by the data linker, while others may be best carried out by the analyst of the linked data. For example, the analyst generally does not have access to the full range of demographic or personal variables that were used in creating the linked data set. It is important that analysts understand these quality assessment options and explore as many of them as possible in collaboration with data linkage providers.

5.4 Approaches to assessment of data linkage quality

There are three main approaches to the assessment of the quality of data linkage. These are:

- clerical assessment
- comparison with a "truth file" or gold standard
- simulation.

The focus of the discussion in the Guidelines will be on clerical assessment, with only a passing mention of the other two approaches. This is because assessment of the quality of data linkage is much easier to undertake through the use of clerical assessment than through the other approaches, and also because much of the assessment of the quality of data linkage is based on clerical assessment rather than on comparison with a truth file or gold standard, or through the use of simulation.

5.4.1 Measuring linkage quality through clerical assessment

Clerical review is mostly used to determine the link status of potential links or record pairs whose link status cannot be automatically determined from their linkage weights or linkage probabilities. Clerical review can also be used, post-linkage, to assess the quality or accuracy of the link status assigned to record pairs whose link status has already been determined. This is referred to as clerical assessment, to distinguish it from clerical review to determine link status. This review may be undertaken by the data linker because only the data linker may have the full range of demographic and personal data items to be able to undertake this review.

Errors in the blocking and linking variables could lead to false links and missed links. For example, errors in date of birth, or variations in the spelling of names, such as have been observed among Aboriginal and Torres Strait Islander people, could lead to record pairs being declared as links when, in fact, they may belong to two separate individuals. Similarly, missing as well as errors and variations in linkage variables, could result in record pairs that belong to the same individual being declared as non-links or as not able to be linked.

These errors may have implications for the methods used in deriving Indigenous status from linked data, the reliability of Indigenous status derived from linked data. Such errors may also affect the reliability of outcome measures, such as morbidity and mortality rates, based on such linked data. For this reason, it is important that the quality of the data linkage and of the linked data are clerically assessed, wherever possible, before methods for deriving Indigenous status are chosen, and also before the data are analysed and outcome measures based on the linked data are estimated.

Clerical reviews and clerical assessments are subjective, and rules to support more objective analysis should be developed and used, bearing in mind the special characteristics of data involving Aboriginal and Torres Strait Islander people.

Sample-based and full clerical assessments

Clerical assessment may be carried out on all or a sample of the records in a data linkage system, or on all or a sample of the records linked in an ad hoc data linkage project. In a dynamic data linkage system, the number of records to be assessed may be very large, often running into several millions of records. Resources, including time, may not always be available to clerically assess all the master linkage keys on a yearly basis, in which case a sample-based assessment may be undertaken. Some ad hoc linkage projects may also involve several data sets, each containing thousands, tens of thousands, or even millions of records. Clerically assessing all the links may be cumbersome, expensive and time-consuming. Under these situations, a sample-based, rather than a full, clerical assessment may be undertaken.

There are rules regarding how to determine the sample size for a sample-based clerical assessment. The size of the sample is scientifically selected to ensure that the results of the sample-based clerical assessment are valid and applicable to the whole file from which the sample is taken. When a sample-based clerical assessment is undertaken, rules may also need to be set regarding the level of error that may be considered acceptable, and the level of error that may automatically trigger the need for a full clerical assessment. For example, it may be decided that if the number of links assessed as false links or false non-links is above a certain threshold (e.g. 5 per cent), then there will be a full clerical assessment, that is, all the links in the linkage file or linkage system will be clerically assessed.

5.4.2 Other approaches to assessing the quality of linkage

There are other ways of assessing linkage quality that are not described here. These include comparing the results of the linkage against a 'gold standard' or a 'truth' file, and the use of simulation. SimRate is a good example of simulation. SimRate uses the observed distribution of data in matched and unmatched record pairs to generate a large simulated set of record pairs, assigns a match weight to each pair based on specified match rules, and uses the weight curves of the simulated pairs for error estimation. SimRate has the ability to examine different thresholds, allowing the user to monitor both the sensitivity and specificity of the decision rule for selecting linked pairs. Readers interested in further details on this topic may wish to consult Winglee et al. (2005).

5.5 Measures of quality of data linkage

One may choose to measure the result of the clerical assessment of data linkage quality in terms of the proportion of previously declared links that are confirmed as true positive links, and the number of previously declared links that are assessed as false positive links, as a result of the clerical assessment. True positive links and false positive links are referred to in other terminologies as 'true matches' and 'false non-matches', respectively.

Similarly, the result of a clerical assessment of previously declared non-links could be measured in terms of the number of previously declared non-links that are confirmed as true negative links, and the number of previously declared non-links that are now assessed as false negative links, as a result of the clerical assessment.

Post-linkage or post-hoc data linkage quality assessment is best undertaken by the data linker because it is fundamental to the data linkage process, and the data linker may also have the data from the linkage to enable this review to be undertaken.

Measures that may be used to assess data linkage quality include *accuracy, sensitivity, specificity, precision* and the *false-positive rate*. These measures are based on basic data that are simply defined from the results of data linkage (see Table 5.1). Some of these measures are, however, difficult to calculate and are therefore not recommended (Bishop & Khoo 2007; Christen & Goiser 2007). Figure 5.1 provides information on how these measures are defined and determined.

Not all of these measures are easily calculated, because their calculation depends on knowing the number of true non-matches or true negatives (TN). These are often unknowable or difficult to calculate, except in balanced classification problems or where one of the data sets being matched is a subset of the other (for example, people injured in car accidents (data set A) and all hospital patients (data set B).

Figure 5.1: Classification of matches and links

		Match status (True)		
		Matches	Non-matches	
Link status (assigned by computer)	Links	True Links or True Positives (TP) (Matches that are linked)	False Links or False Negatives (FP) (Non-matches that are linked)	**Total Links**
	Non-links	Missed links or False Non-links (FN) (Matches that are not linked)	True Non-links or True Negatives (TN) (Non-matches that are not linked)	**Total Non-links**
		Total matches	**Total Non-matches**	**Total record pairs**

Source: (Bishop & Khoo 2007)

Even where it is possible to determine the number of true negatives, they tend to have very large values which dominate calculations of quality measures based on the number of true negatives, thereby making the resultant rates difficult to interpret (Christen & Goiser 2007).

As a result, quality measures like accuracy, specificity and the false negative rate, which use the number of true non-matches or the number of true negatives in their calculation, are unreliable and difficult to calculate, and are not recommended.

Thus, the most widely used, and recommended, quality measures are sensitivity or the true positive rate, and precision or the positive predictor value. These measures are briefly described below. Further details about these quality measures and how they can be calculated can be found in Christen and Goiser (2007) and Bishop and Khoo (2007).

- **Sensitivity or true-positive rate:** This is the proportion of matches that are correctly classified as matches. It may be defined as the proportion of all records in a file or database with a match in another file that is correctly accepted as links (true links). This measure is calculated as: *TP/(TP + FN)*.
- **Precision or positive predictor value** is the proportion of all classified links that are true links or true positives. This measure is calculated as: *TP/(TP + FP)*.

5.5.1 Other measures of quality of linkage from clerical assessment

Where information from a clerical assessment is available, as described above, then quality measures such as *accuracy*, *specificity* (true negative rate) and the false positive rate can also be estimated, in addition to *sensitivity* and *precision*.

As all three measures include the number of 'true negatives' or TN, they are affected by the large size of the TN, with the TN tending to dominate the formula (Christen & Goiser 2007). These three measures are therefore not widely used because of the difficulty in calculating them.

Readers interested in further details about the measurement of data linkage quality may consult Christen and Goiser (2007) who provide full details about these quality measures, including their calculation and their limitations.

5.6 Characteristics of unlinked records

After each data linkage, some records will not have been linked. The substantial majority of these cases occur when individuals are represented in only one of the datasets.

Whether or not records are linked will depend on several factors, including the following:

- records that exist in one data set but are missing in the alternative data sets
- the quality of the input data sets, such as the blocking and linking variables
- the quality of other data items on the input data sets that can be used in clerical review
- the blocking and linking strategy adopted.

The characteristics of unlinked records should be examined to understand why they could not be linked, as this information is closely related to the quality of the input data sets, the linkage strategy adopted and the usefulness of the linked data set. For example, in 2007, the ABS undertook the Indigenous Mortality Quality study as part of the ABS Census Data Enhancement program (ABS 2008a). The aim of the quality study was to link Census records with death registration records to examine differences in the reporting of Indigenous status across the two datasets. Records from the 2006 Census were linked with records of deaths that were registered between 9 August 2006 and 30 June 2007, except for Victoria where death registrations were only available up to mid-March 2007. The registered deaths in question were of persons who were alive at the time of the Census but subsequently died.

Assessment of the quality of the data linkage showed that whereas 92.8% of people of non-Indigenous origin on the death registration form were able to be linked to their Census records, the linkage rate for people of Indigenous origin was 73.7%, that is, as many as 26% of persons of Indigenous origin on the death registration form could not be linked to any corresponding Census records (ABS 2008a).

Analysis also showed that the linkage rate varied considerably by age, and marginally by sex, with linkage rates being higher for older age groups and for females than for males or younger persons. The quality assessment showed that a key reason why death records could not be linked to Census records was that an equivalent Census record did not exist for linking, and that this could be due to the large net undercount for Aboriginal and Torres Strait Islander persons in the Census (ABS 2008a, 2009a). If linkage rates are lower for younger than for older age groups, then the effect of this would be to underestimate Indigenous mortality and over-estimate Indigenous life expectancy. Further assessment of the quality of the Indigenous Mortality Quality Study and its impact on estimates of Indigenous life expectancy have been published by the ABS (2008a, 2009a).

In particular, the following examinations could be carried out on the unlinked records:

- the size of the unlinked records as a proportion of the expected number of links (e.g. if all the records on one data set are expected to be on the other data set, then the expected number of links is the number of records on the smaller data set)
- the number of blocking and linking variables that are missing on the unlinked records
- characteristics of unlinked records with missing data on blocking and linking variables
- reasons why the unlinked records could not be linked, e.g.
  - unlinked records are out of scope or not present in one or more data sets
  - unlinked records have missing or poor quality blocking and linking variables
- the representativeness of unlinked records in relation to the source data set: for example, are the unlinked records more heavily represented by particular demographic groups, e.g. young Aboriginal and Torres Strait Islander adult males living in remote areas
- comparison of basic measures based on the linked data set and unlinked records: for example, if some records are not able to be linked because of missing or poor quality blocking and linking variables, then one could compare basic measures based on the linked and unlinked data sets. This analysis may only be carried out in respect of personal and demographic variables that are on the linked data set. For example, for linkage involving the deaths data set, one could compare median age at death in the linked and unlinked data sets.

While the importance of these quality assessments is not in doubt, it is less clear who should carry them out. Data linkers are often in a better position to carry out these quality assessments, because they may have access to both the linked and unlinked records. However, while data linkers may be able to examine and compare the demographic characteristics of the linked and unlinked records, such as sex, age distribution, country of birth, and address, they may not have any content information such as age at death, dates of admission and/or separation from hospital, to enable them to assess the potential effect of the unlinked records on outcome measures estimated from the linked data.

While analysts may have the content information, they often do not request information on unlinked records because they may not know about quality assessment of data linkage or the uses to which information on unlinked records may be put. In addition, if access to this information is not specifically included in the ethics application for access to data, then it will not be possible for the data linker to provide this information to the analyst. It is therefore important that during the planning and design phase of a data linkage project, data linkers and analysts discuss the issue of quality assessment of data linkage as well as the types of data that are necessary for this assessment, and who should carry out quality assessment.

Data custodians also have a very good understanding of the data, and in most cases, a better understanding of the data than the data linker or analyst. Data custodians could provide assistance in reviewing the quality of the input variables, in terms of coding, variable values, and missing values. They could also assist in clerical review and quality assessments, and in the interpretation of linkage results and why some records could not be linked. Feedback of linkage results to data linkers may also assist in future improvements to the quality of the data, in terms of form design, staff training, as well as in data collection and data processing protocols.

5.7 Edit checks that should be performed before analysing linked data

Data linkage is not perfect, neither are the data sets on which the linked data sets are based. Analysts should therefore undertake a series of edit checks before they start analysing data.

CHeReL has provided a list of edit checks that analysts must perform on linked data sets before the linked data sets are ready to be used for the purpose for which they were created. The list of edit checks can be found at:
<www.cherel.org.au/further-resources>

Some of the edit checks that analysts must perform are briefly described below.

5.7.1 Familiarisation with the data collections in the study

Analysts should read and understand the data dictionaries, and be aware of changes in coding of variables over time, and differences in the coding of variables across data sets.

Validation studies may also exist that may have previously reviewed the quality of data items in the linked data sets. It is important that analysts familiarise themselves with the findings of these validation studies.

Examples of some of these validation studies can be found at:
<www.cherel.org.au/media/13588/validation-studies-april-2009.pdf>

Researchers should also be aware of records that may be missing from linked data sets due to the scope of the various data sets that are included.

5.7.2 Basic frequency analyses of individual data sets before merging data

Frequency checks must be carried out on the individual data sets before they are merged to form the linked data set. They must also be carried out on the merged file before it is analysed. The two sets of frequencies could then be compared for discrepancies.

5.7.3 Knowledge of quality of input data sets and quality of linked data set

Analysts must be aware of the types of errors that exist on the respective data sets in the linked data file. Some of these errors may relate to inconsistencies in the values of data items across and within data sets due to data entry errors, respondent errors or recorder errors in obtaining and recording the correct responses.

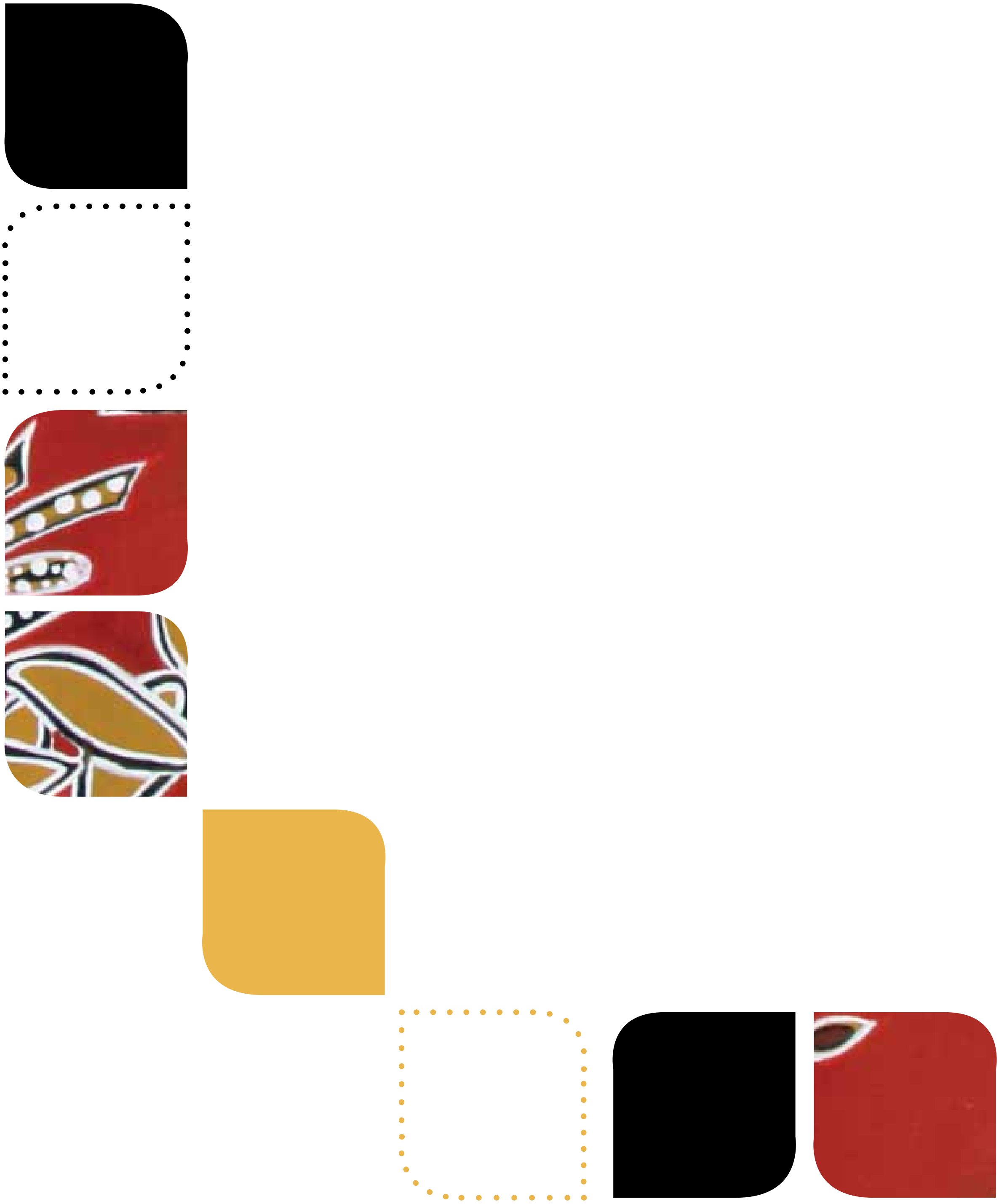
Linked data may contain linkage errors, leading to false-positive links. It is important that analysts understand the quality of the linkage and how this will impact on the analysis they plan to undertake.

5.7.4 Logic or internal consistency checks on the linked data set

The analyst should perform logic and internal consistency checks to decide if the data are internally consistent, that is, if the different values of data items in one data set are consistent with each other and with the other data sets. For example, researchers may want to be sure that males are not being confined to have babies, that 13 year old girls are not mothers with five children, that people are not admitted to hospital after they have died etc.

Some of the consistency checks proposed by CHeReL include:

- unlikely values (e.g. a person aged 130 years or a woman having given birth to 12 children etc.)
- "missing" records (a person being recorded on the hospital data sets as having been separated from hospital through death with no matching entry in the death data)
- illogical grouping of events (e.g. female with prostate cancer).



6

Methods for deriving Indigenous status

Principle

Analysts should investigate multiple methods for deriving Indigenous status and select those that best fit the purpose of the analysis. Where possible, analysts should also explore and report the impact of using various methods to derive Indigenous status on health and wellbeing measures and indicators.

Guidelines

6.1 Analysts should investigate the suitability and sensitivity of various methods for deriving Indigenous status, in the context of the characteristics of their input data sets. The suitability of various methods for deriving Indigenous status will depend on the number, characteristics and quality of data sets involved in the linkage, as well as the quality of the linkage process.

6.2 Analysts should discuss in their reports the impact on their estimates and indicators of using alternative methods to derive Indigenous status.

6.3 Where possible, analysts should validate the results of their methods for deriving Indigenous status against external data sets.

6.1 Background

Not all records of Aboriginal and Torres Strait Islander people appearing in data sets are identified as such. Indigenous status may be missing or may be inconsistently recorded across and within data sets. Data linkage offers a way of dealing with some of these errors and inconsistencies. The challenge for analysts of linked data relating to Aboriginal and Torres Strait Islander people is how to deal with the inconsistent or incomplete Indigenous status information in analysis.

There are many possible methods for determining Indigenous status (Taylor 2010). The choice of method must be guided by the purpose of the analysis, the characteristics of the underlying data sets and the quality of the linkage process. The aim of this chapter is to provide guidance as to what methods could be used to derive Indigenous status, the factors impacting on the choice of methods and the importance of investigating the sensitivity of various methods on estimates and measures. An online attachment (see Forthcoming publications) discusses a number of studies that have explored multiple methods for deriving Indigenous status, and have assessed the sensitivity of specified methods on the results of the analyses.

The methods proposed here are intended only as a guide to analysts in how they may derive Indigenous status from linked data, where Indigenous status is missing or is inconsistently reported across data sets. The Guidelines do not recommend any specific methods because the evidence for this is still being assembled and also because the performance of the methods depends on the quality and characteristics of the data sets on which the methods are based.

Currently, the ABS, the Western Australian Department of Health and the Telethon Institute for Child Health Research are testing several methods on 11 administrative and survey data sets, containing up to 29 million records, to determine the impact of various methods on outcome measures based on the methods. The results of the study, known as the *Getting Our Story Right* project, will be published as a companion document to the Guidelines. The report will offer guidance as to how different methods perform with respect to various data sets.

In this chapter, methods for deriving Indigenous status from linked data have been grouped under two broad headings – algorithms and aggregate methods. Algorithms refer to approaches that use information about Indigenous status at the individual level and make a determination for each individual about their Indigenous status. Aggregate methods refer to approaches that combine information at the individual level with information at an aggregate level and do not make a determination for each individual about their Indigenous status. Rather, an adjustment is made for under-or over-identification of Aboriginal and Torres Strait Islander people at an aggregate level.

6.2 General issues to consider when choosing a method

The selection of any methods for deriving Indigenous status will depend on the purpose of the study, as well as the characteristics of the data sets involved and the linkage approach used in the study. Issues to consider include:

- the number of data sets involved in the study
- characteristics of the data sets involved in the study, including
  - the scope and coverage of the data sets
  - the quality of Indigenous status on each of the data sets
  - the data collection approach to Indigenous status e.g. whether previous entries for Indigenous status are updated or over-written by successive or more recent entries
  - whether there are single or multiple entries for Indigenous status on each data set
  - the independence of the Indigenous status information on the respective data sets
  - whether there is a unique entity identifier to enable a person’s identification to be traced over time and across facilities

- the quality of the linkage variables
- the quality of the linkage process in terms of the proportion of missed and false links in the linked data set.

There are some generic issues that are common to most methods for deriving Indigenous status from linked data. These are discussed below.

Numerator–denominator bias

Many methods are susceptible to numerator-denominator bias. This occurs where the numerator and denominator are derived from different populations.

For example, consider mortality rates where the number of Aboriginal and Torres Strait Islander deaths is the numerator and the estimated Aboriginal and Torres Strait Islander population is the denominator. The numerator is obtained from death registration information compiled by the Registrars of Births, Deaths and Marriages, while the denominator is taken from the estimated resident population compiled from Census and other data. In this case, the Indigenous status in the numerator and denominator may not align and the resulting estimates can be affected by numerator-denominator bias. The magnitude and seriousness of this problem will vary between data sets, analysis variables, and the methods used for deriving Indigenous status. Researchers should take this into account when interpreting the results of their analyses.

For more information about numerator-denominator bias refer to (Ajwani et al. 2003; Blakely et al. 2002a; Blakely et al. 2002b; Hill et al. 2007).

Changes in quality of Indigenous identification over time

Not only does the quality of Indigenous identification vary between data sets, it also varies over time. It may thus be inappropriate to use methods that combine or compare data sets covering different historical periods or that cover a long period of time.

Indeed over the past few years, there have been well-publicised attempts by various levels of government to improve the quality of Indigenous identification on data sets. These attempts have included the following:

- standardising the wording of the question on Indigenous status
- standardising the coding of the responses to the question on Indigenous status
- explaining the benefits of correct Indigenous identification and encouraging Aboriginal and Torres Strait Islander people to identify as such
- training of data collection agencies on how to collect Indigenous status information across their data sets.

It is likely that these measures have led to improvements in the quality of Indigenous identification across data sets in recent years. This needs to be taken into account when choosing methods for deriving Indigenous status.

Analysts may want to consider having a cut-off period for the data sets from which they will be deriving Indigenous status. The cut-off period may vary between data sets, and may be based on knowledge of the changes in the quality of the data sets over time.

Misclassification of Indigenous status

The proportion of Aboriginal and Torres Strait Islander people on most national, state or territory data sets is very small. This is due to the size of the Aboriginal and Torres Strait Islander population relative to the whole population (ranging from 0.7% in Victoria to 30.4% in the Northern Territory). As a result, the number of people identified as Aboriginal and Torres Strait Islander can be sensitive to misclassification error (that is, non-Indigenous people incorrectly coded as being Aboriginal and Torres Strait Islander and vice-versa).

Consider the case where Aboriginal and Torres Strait Islander people contribute 2–4% of records in a dataset. A 1% misclassification of non-Indigenous as Aboriginal and Torres Strait Islander will result in 20–33% of the identified Aboriginal and Torres Strait Islander records actually being non-Indigenous records.

This issue can be exacerbated when combining information across a number of data sets.

It is therefore important that analysts are aware of this issue when selecting a method for deriving Indigenous status.

6.3 Algorithms for deriving Indigenous status

Broadly speaking, the algorithms that can be used to derive Indigenous status for the purpose of analysis can be divided into two categories:

- simple methods involving mostly single entry data sets
- complex methods involving multiple entry data sets across time and across facilities.

A selection of the most common algorithms are summarised in Table 6.1 and discussed in further detail below. It is important to note that these algorithms should only be used for the purpose of statistical analysis. They should not be used to change a person's Indigenous status on source data sets.

Table 6.1: Algorithms for deriving Indigenous status

Algorithm	Description	Limitation
A: Simple algorithms		
1. Ever Indigenous	A person is assigned as being of Aboriginal and/or Torres Strait Islander origin if they are recorded as such on at least one data set.	• May be prone to over-count depending on the level of under-identification • Greater risk of numerator-denominator bias • May not be suitable where there is known to be significant error in collecting Indigenous status information in the linked data sets or where the linkage process results in a significant rate of false-positives.
2. Frequency-based methods		
(i) Aboriginal and/or Torres Strait Islander on more than one data set.	A person is assigned as being of Aboriginal and/or Torres Strait Islander origin if they are recorded as such on two or more data sets.	• May be difficult to implement where there are only two data sets or where the rate of non-response to the Indigenous status question is high across data sets.
(ii) Aboriginal and/or Torres Strait Islander on at least x% of data sets.	A person is assigned as being of Aboriginal and/or Torres Strait Islander origin if they are recorded as such on x% or more datasets (e.g. 50% or more).	• The number of persons identified as Indigenous may be limited where the rate of non-response to the Indigenous status question is high across data sets.
3. Single source methods		
(i) The most trusted data set.	A person is assigned as being of Aboriginal and/or Torres Strait Islander origin if they are recorded as such on the 'most trusted' data set.	• It may be difficult to define "most trusted".

(continued)

Table 6.1 (continued): Algorithms for deriving Indigenous status

Algorithm	Description	Limitation
B: Complex algorithms		
1. The current Indigenous status entry or record	If a person has multiple entries of Indigenous status on either the same data set or across multiple data sets, then the person's Indigenous status is based on the most recent or current entry or data set.	• Indigenous status may refer to different time periods for different individuals depending on which data sets a person appears on and when information on Indigenous status was collected for different data sets. • May be subject to varying levels of quality if the most recent entry varies considerably between individuals, and the quality of Indigenous status reporting has changed over time.
2. Within and across data sets approach	This approach combines Indigenous status over time across multiple entries within a data set and over time across multiple data sets.	• Complex to implement and explain to users.
3. 'Weighted' data set approach	Data sets are ordered and assigned weights or scores, according to their level of quality. A person's Indigenous status is derived by the cumulative weight or scores obtained from the data sets on which they are recorded as being of Aboriginal and/or Torres Strait Islander origin.	• Complex to implement and explain to users. • May require subjective judgments of level of quality and what level of score should be accepted as indicating that a person is considered to be of Aboriginal and/or Torres Strait Islander origin for the purposes of analysis.

6.3.1 Simple algorithms

The simple algorithms comprise the 'ever Indigenous,' frequency-based and single source methods. Table 6.2 illustrates the application of various simple algorithms for deriving Indigenous status, to five data sets involved in linkage. The table shows that a person's Indigenous status varies depending on which algorithm is used.

Table 6.2: Illustration of selected algorithms for deriving Indigenous status

Record ID	Data sets					Algorithm and result		
	A	B	C	D	E	Ever-Indigenous	More than one data set	Majority of data sets
AAB1	Indigenous	Not stated	Non-Indigenous	Non-Indigenous	Non-Indigenous	Indigenous	Non-Indigenous	Non-Indigenous
AAB2	Indigenous	Non-Indigenous	Non-Indigenous	Indigenous	Non-Indigenous	Indigenous	Indigenous	Non-Indigenous
AAB3	Indigenous	Indigenous	Non-Indigenous	Not stated	Indigenous	Indigenous	Indigenous	Indigenous
AAB4	Indigenous	Not stated	Not stated	Not stated	Not stated	Indigenous	???	???
AAB5	Indigenous	Indigenous	Non-Indigenous	Non-Indigenous	Not stated	Indigenous	Indigenous	???

'Ever-Indigenous' algorithm

The 'ever-Indigenous' algorithm is the simplest of all the methods for deriving Indigenous status. For the purposes of the study, a person is assigned as being Aboriginal and/or Torres Strait Islander if they are recorded as such on any one of the data sets included in the study, even if the person is recorded as non-Indigenous on all other data sets. The 'ever- Indigenous' approach is the most widely used method for determining Indigenous status because of its simplicity and minimal data requirements, particularly where there are only two data sets to link.

The algorithm has a number of limitations that should be taken into account. Inconsistencies in a person's Indigenous status are not accounted for in this approach, such as for record AAB2 in Table 6.2, for instance, where a person is recorded as being of Aboriginal and/or Torres Strait Islander origin on two data sets and as non-Indigenous on three data sets, and record AAB1, where the individual is recorded as Aboriginal and/or Torres Strait Islander on one data set, as non-Indigenous on three data sets and as 'not stated' on a fifth data set.

The algorithm is also prone to over-count, arising from both errors in data processing and the quality of the linkage process. For example, errors in the coding of Indigenous status may cause some non-Indigenous people to be inadvertently coded as being of Aboriginal and/or Torres Strait Islander origin. These over-counts could have a substantial impact on indicators and outcome measures based on the derived population of Aboriginal and Torres Strait Islander persons.

Another drawback of the 'ever-Indigenous' approach is related to the quality of the linked data used in deriving an individual's Indigenous status, especially where the linked data set is created from probabilistic linkage. Probabilistic linkage will rarely be completely accurate, and usually results in a certain level of missed links and false links (false-positives). The higher the number of false links, the higher the likelihood that a person will be incorrectly deemed to be of Aboriginal and/or Torres Strait Islander origin.

Frequency-based algorithms

Aboriginal and Torres Strait Islander on more than one data set

For purposes of the study, one simple algorithm involves assigning a person the status of 'Aboriginal and/or Torres Strait Islander' if they are recorded as such on more than one of the data sets in the linkage study. Thus, if there are five data sets and a person is recorded as Aboriginal or Torres Strait Islander on more than one data set, the person is assigned as such for the analysis even if the person's Indigenous status is recorded as 'missing' or 'non-Indigenous' on a majority of data sets. This approach allows for validation of a person's Indigenous status from at least one other data set. It is widely used because of its simplicity and minimal data requirements.

In this example, the situation becomes problematic, if in a linkage involving five data sets, a person is recorded as being of Aboriginal and/or Torres Strait Islander origin on one data set and as 'not-stated' on the remaining four data sets (see Record AAB4 in Table 6.2). While the 'ever Indigenous' approach may assign the status of 'Indigenous' to the individual, the situation is less clear for this algorithm. The analyst therefore needs to set clear guidelines as to how to treat missing or not stated cases when particular algorithms are being used. In most cases, if an algorithm cannot determine the Indigenous status of a record with missing or not stated Indigenous status, then that record should be excluded from the denominator or the population at risk.

Aboriginal and Torres Strait Islander on x% of data sets

For this algorithm, a person is assigned as being Aboriginal and/or Torres Strait Islander, for the purposes of a study, if they are recorded as such on at least x% of data sets. The analyst chooses the cut-off, and a commonly used cut-off is more than 50% (i.e. the majority).

Complications arise where there are four data sets and a person is recorded as Aboriginal and/or Torres Strait Islander on two data sets and as non-Indigenous on the other two, or where there are five data sets and a person is recorded as Indigenous on two data sets, non-Indigenous on one data set and has a 'not stated' response or a 'missing' entry on two data sets (see records AAB4 and AAB5 in Table 6.2).

Clear guidelines should be set in such situations to ensure that a consistent approach to deriving Indigenous status is adopted. What some analysts may do in such situations is to amend the majority rule to refer to '50% or more of the data sets' or 'if the number of times a person is recorded as Aboriginal or Torres Strait Islander is equal to or greater than the number of times the person is recorded as non-Indigenous'. The analyst should also make clear from the outset how to treat missing or 'not stated' cases when particular algorithms are being used.

Single source methods

Indigenous status on most 'trusted' data set

This algorithm takes the Indigenous status from a single data set which is assessed to have high quality information about Indigenous status and therefore to be considered as having a higher level of trust than others. This trust may be based on a qualitative evaluation of the data set, involving the following considerations:

- reasons why the data collecting institution or agency collects information on Indigenous status, and whether Indigenous status is critical to the funding or operation of the agency
- training given to staff on how to collect information on Indigenous status
- how consistently the question on Indigenous status is asked

- whether the method of collecting information on Indigenous status is consistent with the *National Best Practice Guidelines for collecting Indigenous status in health data sets* (AIHW 2010a)
- whether there are regular audits of the reliability and usefulness of responses to the question on Indigenous status, and whether results of the audits are used in modifying the data collection and data processing protocols of the data collecting agencies.

Some of these issues are described in Chapter 3 (Quality of Indigenous status in data collections).

A drawback of this method is that it is often difficult to assess or to define 'quality', 'reliability' or 'level of trust' objectively.

6.3.2 Complex algorithms

In data sets where a person could have multiple entries over time and across facilities or service points, more complex rules for deriving Indigenous status may be possible. Consider the following data sets, for instance, which may involve a combination of single and multiple entry records, for the same person, over time and across facilities:

- national, state and territory admitted patient care and emergency department data sets, involving multiple records or multiple episodes of care across hospitals and across time
- National Disability Support Services Data set, involving multiple entries/episodes of care
- (Permanent) Residential Aged Care data set, involving one entry/episode of care
- National Perinatal Data Collection, involving multiple births over time
- birth registrations data, involving multiple births over time.

Given these characteristics, a linkage study involving these data sets may consider the following more complex algorithms.

Current (or most recently recorded) Indigenous status

This approach uses the most recently recorded entry of Indigenous status without consideration of previous entries or records. Although this method appears simple, it is complex to implement because the most recent record or entry may be on different data sets for different individuals and may relate to different events for different individuals.

Data sets are often related to the life cycle (e.g. births, school enrolment, admitted patient care, residential aged care and death etc.). The most recent entry for different individuals may therefore relate to dates of events that may be many years apart. For example, the most recent entry for one person could be the date of death (e.g. 2010), while for another, it could be when the person had her last baby (1989) or was last admitted to hospital (2002).

If the quality of Indigenous status information has been changing over time, then this will affect the reliability and comparability of Indigenous status based on the current or most recent data set algorithm.

Furthermore, application of the 'current or most recent data set' algorithm depends on how the individual data sets are compiled. Different service providers and data collection agencies may adopt different methods in the compilation of multiple episodes of care (EOC) records.

The current or 'most recent data set' approach to deriving Indigenous status may thus be applied in the following situations, depending on the structure of the data sets available:

- where there are multiple records per person, and Indigenous status is updated any time the person presents for service (e.g. National Disability Support Services Data set)
- where the data sets are multiple entry data sets and there are multiple entries or records per person on each or some of the data sets, covering a period of time (e.g. Births registrations data).

In the example in Table 6.3, the same person has been recorded differently, across four data sets and over a 12-year period, as 'Indigenous', 'Non-Indigenous' and 'Not stated'. However, the most recent entry (data set B, 24/10/2010) records the person as 'Indigenous' and therefore the algorithm would assign this person the status of Aboriginal and/or Torres Strait Islander.

Table 6.3: Illustration of selected algorithms for deriving Indigenous status for Person X

Data set A National Hospital Morbidity Data set		**Data set B** Non-Admitted Patients (Emergency)		**Data set C** National Perinatal Data Collection – Births[1]		**Data set D** National Disability Support Services Data set	
Date of EOC	Indigenous status	Date of EOC	Indigenous status	Date of EOC	Indigenous status	Date of EOC	Indigenous status
08/09/2010	Indigenous	24/10/2010	Indigenous	02/10/1989	Non-Indigenous	16/08/2010	Not stated
11/12/2006	Indigenous	11/11/2006	Not stated	16/08/1986	Indigenous	22/11/2006	Indigenous
22/06/2006	Not stated	26/09/2005	Indigenous	16/03/1982	Non-Indigenous	14/04/2005	Indigenous
14/03/2004	Not stated	18/06/2003	Indigenous	01/12/1969	Not stated	06/10/2004	Not stated
19/10/2003	Non-Indigenous	02/04/2002	Not stated	22/04/1966	Not stated		
26/08/1998	Indigenous	18/08/1998	Non-Indigenous				

[1] The Indigenous status information in the National Perinatal Data Collection pertains to episodes when Person X delivered a baby, and not when Person X was born.

In a slight modification to the 'current or most recent data set' algorithm, where an entry exists for each service contact or episode of care, a researcher could examine the trend in the reporting of Indigenous status in each data set, and decide whether or not the trend within each data set points to the person being Aboriginal and/or Torres Strait Islander. In the example above, the person would be identified as being of Aboriginal and/or Torres Strait Islander origin as the trend in all but the perinatal data set points to this being the case.

Within and across data sets approach

A more comprehensive method of deriving Indigenous status is to combine information from both within and across services or data sets. This approach examines how a person is identified over time within a data collection and over time across a number of data collections. This approach is applicable where personal details are collected during each service contact or episode of care and are used in constructing the data set. It takes into account all the evidence from all the data sets available to be linked.

The algorithm first assigns Indigenous status for an individual for each data set, based on applying one of the previously described algorithms to the records within that data set (e.g. most recent or frequency-based algorithm etc.). The algorithm then assigns an Indigenous status to an individual, based on applying one of the previously described algorithms across the data sets, using the Indigenous status decision for each individual data each data set.

For example, Table 6.4 shows episode of care entries for Person X in four data sets. The table shows that Person X has five entries on data set A, four entries each on data sets B and C, and one entry on data set D. Using a frequency-based algorithm, a decision is made about the person's Indigenous status for each data set. In this example, a person is assigned as 'Indigenous' if the majority of records within the data set record the person as such. Using this approach, the person is considered to be Indigenous for three of the four data sets. Applying the same frequency-based algorithm across the data sets results in the person being recorded as Indigenous overall, as the majority of data sets classify the person in this way.

Table 6.4: Algorithms based on multiple entries in multiple data sets for Person X

Data set A		Data set B		Data set C		Data set D		Overall decision
Entry/ Episode of care	Indigenous status	Entry/ Episode of care	Indigenous status	Entry/ Episode of care	Indigenous status	Entry/ Episode of care	Indigenous status	
1	Indigenous	1	Non-Indigenous	1	Indigenous	1	Not stated	
2	Indigenous	2	Indigenous	2	Indigenous			
3	Non-Indigenous	3	Indigenous	3	Not stated			
4	Not stated	4	Not stated	4	Indigenous			
5	Indigenous							
Within data set decision based on frequency algorithm (majority)	Indigenous		Indigenous		Indigenous		Not stated	Indigenous

Weighted data sets

In this approach, data sets are ordered and weighted according to the level of 'trust' placed in them, based on an assessment of their relative levels of quality. The quality assessment may take into account the care with which the data are collected and processed, whether or not the data sets are the result of a statutory, financial or programmatic requirement, whether or not the data collecting agency has a good record of collecting information on Indigenous status, and whether or not the data sets are audited with respect to the quality of recording of Indigenous status.

Data sets could be weighted and ranked according to their perceived level of trust, among the data sets to be linked, as shown in Table 6.5. The weighting will be informed by the analyst's knowledge of the quality of the input data sets and their perceived level of trust, as described above. Despite the weighting being based on objective criteria, the actual weighting may be subjective and may vary between reviewers. The weights may range from one (1) to a number representing the total number of input data sets. Thus if there are five data sets to be linked, for example, then the weights may range from one to five, with '1' representing the lowest weight, and '5', representing the weight for the most trusted data set in the hierarchy of data sets.

A person's Indigenous status may then be based on the sum of the weights of the data sets on which the person is recorded as being of Aboriginal and/or Torres Strait Islander origin. The minimum requirement for a person to be assigned as Aboriginal and/or Torres Strait Islander origin is that the person is recorded as such on the most trusted data set and on one other data set, or on any combination of data sets with a pre-determined cumulative score.

This approach is similar to the 'most trusted data set' approach, and is based on the premise that all data sets are not equal, and that data sets differ with regard to their quality, purpose, the skill and dedication of the staff compiling the data set, and the statutory importance of the data set, as well as their importance to the operation of the agency collecting them. This is illustrated in Table 6.5.

Table 6.5: Weighting of data sets to derive Indigenous status

Record ID	Data set A	Data set B	Data set C	Data set D	Data set E	Cumulative score	Determination
	Weight = 5	Weight = 4	Weight = 3	Weight = 2	Weight = 1	15	Indigenous if score => 6
X1	Indigenous	Not stated	Non-Indigenous	Non-Indigenous	Indigenous	6	Indigenous
X2	Indigenous	Not stated	Non-Indigenous	Non-Indigenous	Non-Indigenous	5	Non-Indigenous
X3	Non-Indigenous	Non-Indigenous	Non-Indigenous	Indigenous	Indigenous	3	Non Indigenous
X4	Non-Indigenous	Non-Indigenous	Indigenous	Indigenous	Indigenous	6	Indigenous
X5	Non-Indigenous	Not stated	Indigenous	Not stated	Indigenous	4	Non-Indigenous
X6	Indigenous	Indigenous	Indigenous	Non-Indigenous	Not stated	12	Indigenous

It is apparent from Table 6.5 that although a person may be recorded as Indigenous on the most trusted data set (Record X2) or as Indigenous in two or more data sets, (Records X3 and X5), the person's cumulative score is such that the person is assigned the status of non-Indigenous. In the 'weighted data set approach', a person has to be recorded as being of Aboriginal and/or Torres Strait Islander origin on at least one trusted data set or on several other data sets to be assigned as such for the purpose of analysis.

The above example provides a simple illustration from which an analyst could use the 'weighted data sets' method. Difficulties in applying the 'weighted data sets' method include the degree of subjectivity that is used in first evaluating the data quality, assigning values to the data sets and then arriving at a cumulative score of data quality.

6.4 Aggregate methods for deriving Indigenous status

'Aggregate methods' refer to approaches that use data linkage to inform adjustments to Indigenous status at a more aggregate level, rather than making adjustments at the individual level.

For example, Zubrick et al. (2006) linked the WA Aboriginal Child Health Survey data to the birth register and WA Midwives Notification System. The linked data were used to explore whether children whose identification status was different between the data sources were on average different from those whose Indigenous status matched between the two sources. A modelling approach was used to develop a correction factor to be applied to the historical birth register data.

Another example is the ACT Hospital Data Linkage project that uses the results of data linkage to assess and correct for Indigenous under-identification in the ACT Hospital Morbidity Data set (AIHW, 2010a).

An aggregate method was also used by the ABS for the purposes of estimating Aboriginal and Torres Strait Islander life tables for the 2005-2007 period, based on the 2006 Census Data Enhancement Indigenous Mortality Quality Study. In this study, deaths which occurred and were registered in the 11 months following the 2006 Census (except for Victoria where only nine months of deaths data was available) were linked with the deceased's corresponding records from the 2006 Census when they were still alive. The linkage study was used to estimate the level of under-identification of Aboriginal and Torres Strait Islander deaths, relative to the Census. For the purposes of producing life tables, aggregate death counts were then adjusted for the under-identification in death registrations. A further adjustment was applied to ensure Indigenous status was consistent between adjusted deaths and population estimates in order to minimise numerator-denominator bias (ABS 2009a).

6.5 Quality checks for derived Indigenous status

Once appropriate methods to derive Indigenous status have been applied, the analyst should undertake some quality assessments of the results. For example, the analyst may compare the characteristics of individuals who were identified as being of Aboriginal and/or Torres Strait Islander origin only after the application of the algorithm with those consistently identified as such. Other checks may aim to explore possible errors arising from the linkage process.

For instance, using the 'ever-Indigenous' algorithm, the analyst may choose to check whether country of birth is Australia. Although Aboriginal and Torres Strait Islander people can be born overseas, this assessment, in combination with other data items (such as place of birth of parents) may be a useful quality check to attempt to identify if there is a probable high rate of false-positives.

In the case of the weighted method, the analyst may choose to consider the consistency of Indigenous status across data collections and revise their weightings where consistency is not reflected in the cumulative score on account of low weighting for some of the data sets. Consistency of Indigenous status across collections can be considered an indicator of quality.

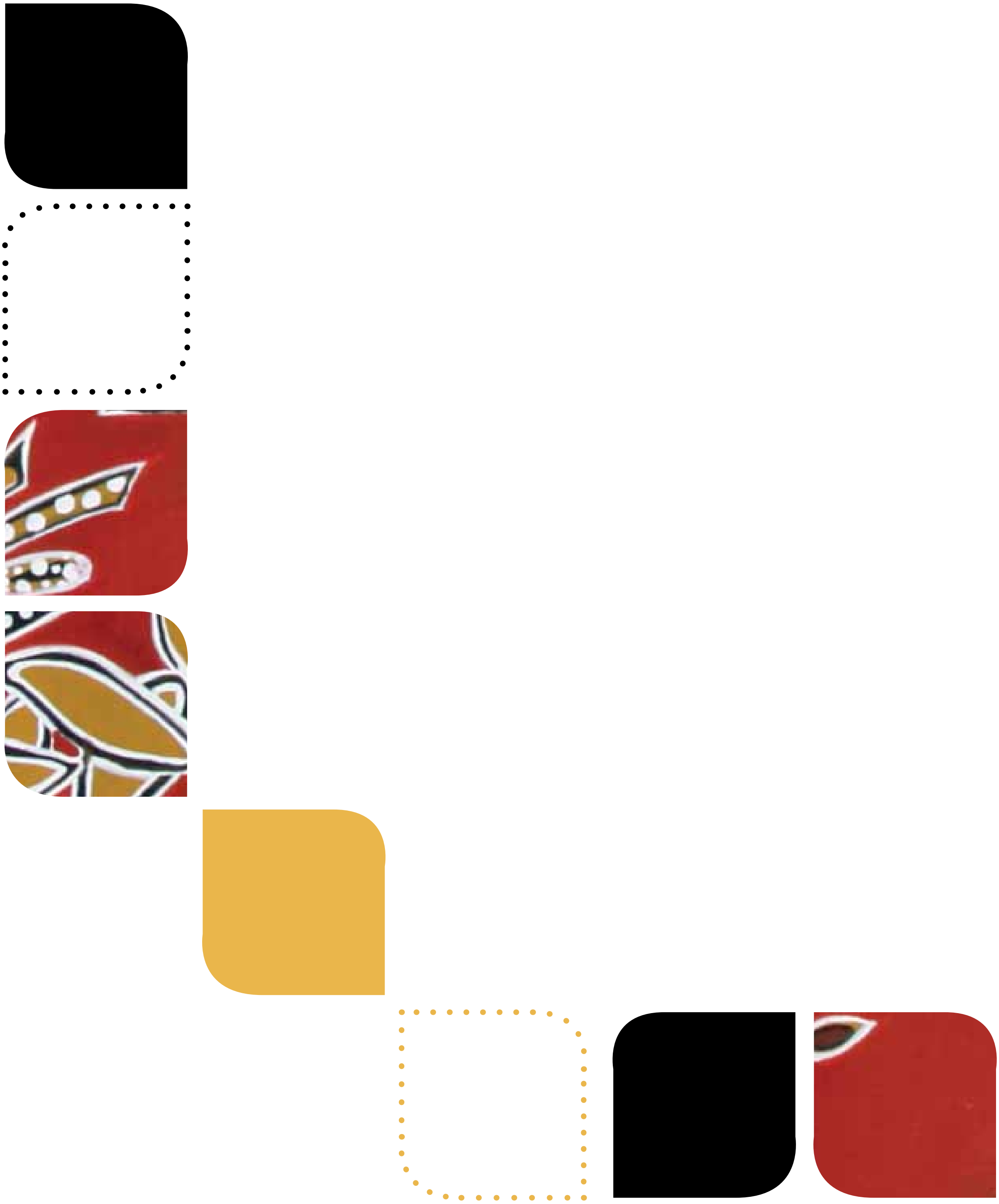
The analyst should consider the derived Indigenous status carefully and consider whether there are any occurrences within the application of the algorithm that may require special adjustment or rejection of the method. It should be noted that an algorithm cannot derive an Indigenous status with absolute certainty or that it will derive Indigenous status 'accurately' in all instances. Instead, it has to attempt to derive Indigenous status based on what is most probable.

In some instances, analysts will have access to data sets that only relate to Aboriginal and Torres Strait Islander people, such as native title registers. Depending on the scope and nature of the study, analysts may be able to validate the reliability of the derived Indigenous status from data linkage against purposely-created Indigenous population registers, such as native title population registers (Draper et al. 2009).

It is also recommended that wherever possible, analysts explore multiple methods to derive Indigenous status and to prepare outcome measures for each classification of Indigenous status derived. These outcome measures could then be compared to determine the sensitivity or impact of the various algorithms for deriving Indigenous status on the estimated outcome measures.

Very few studies have explored the sensitivity of various algorithms on outcome measures based on these algorithms (see for example, ABS 2008b and AIHW: Choi et al. 2012). The few studies that have explored multiple algorithms to derive Indigenous status include those by Draper, Pilkington and Thompson (Draper et al. 2009), Mak and Watkins (Mak & Watkins 2008) and Kennedy, Howell and Breckell (Kennedy et al. 2009a; Kennedy et al. 2009b).

In these examples, the authors use multiple algorithms to derive Indigenous status. They then undertake their analyses and prepare multiple outcome measures for each classification of Indigenous status derived from the various algorithms. The resultant estimates, measures or indicators are then compared to determine the sensitivity or impact of each of the specific algorithms for deriving Indigenous status on the derived estimates. These studies are described in more detail in related publications attached to this report (see the section on 'Forthcoming publications' for more information).



7

Transparency

Principle

All relevant aspects of the data linkage activity, including data linkage quality assessment, analysis of the linked data and methods for deriving Indigenous status, should be fully documented and publicly reported.

Guidelines

7.1 Project proposals should contain sufficient detail to ensure that ethics committees/authorising institutions and data custodians understand how data will be linked, merged and analysed.

7.2 Analysts publishing the results of analysis of linked data should include details of the linkage process and linkage quality sufficient to allow other analysts to replicate the linkage or to make decisions regarding the level of confidence to be placed on the results.

7.3 Analysts should publish as much detail about methods for deriving Indigenous status, and their impact on estimates and measures based on the linked data, as would allow other analysts to evaluate critically the relative performance of the algorithms for deriving Indigenous status.

7.4 If legislation allows, results and findings from the analysis of linked data should be made available to the custodians of the data prior to wider publication.

7.5 Analysts should maintain documentation about core details of each Aboriginal and Torres Strait Islander data linkage project to aid quick and efficient reporting about their project.

7.1 Why is transparency important?

Transparency is closely related to integrity, openness, accountability and communication. In data linkage, transparency refers to trustworthiness and openness in communicating which data sets are accessed during data linkage, why the data sets are accessed, integrity in using the data sets only for approved purposes, openness in communicating risks to privacy and confidentiality in the use of the data, and accountability to data custodians and persons whose personal information is being accessed, in ensuring that their personal information is protected.

Transparency at all phases of a data linkage project is essential to:

- ensure findings are interpreted correctly, and that results, where appropriate, inform decision-making
- allow results and processes of Aboriginal and Torres Strait Islander data linkage projects to be appropriately compared with other projects, and to maximise repeatability
- demonstrate to the public that data linkage protects privacy and confidentiality and reduces the use of name-identified data for research and for statistical reporting
- ensure complete and informed feedback by returning information to the public or individuals that provided the original input data, for example, through appropriate dissemination strategies
- inform the public of the uses to which the data have been put and the benefits accruing from the data linkage
- enable ethics committees to assess fully the ethical risks of these projects before they are approved
- enable data custodians to understand the results of analysis based on data they have provided to the analysts
- enable a full understanding of how the results of data analysis may have changed as a result of using linked data
- enable limitations in Indigenous identification to be fully understood
- prevent misinterpretation of the results of the project.

Additionally, complete feedback and engagement with the Aboriginal and Torres Strait Islander community has advantages in terms of being able to benefit from the knowledge and experience of Aboriginal and Torres Strait Islander people in matters relating to data quality, data interpretation and relationship between data items.

Adherence to guidelines in this chapter will provide as much opportunity as possible for the evidence base for best practice in Aboriginal and Torres Strait Islander data linkage to be expanded.

7.2 What should be disclosed, by whom and to whom?

7.2.1 Before the project starts

In project proposals, ethics committees, data custodians, data linking institutions and other relevant authorising institutions should be fully informed by investigators about:

- the purpose of the project, and whether the study is specifically about Aboriginal and Torres Strait Islander people
- what data sets will be used in the project, and how they will be stored and accessed
- how privacy and data confidentiality will be protected
- the benefits of the project to Aboriginal and Torres Strait Islander people.

As described in previous chapters, proposals should contain sufficient detail to ensure that ethics committees, authorising institutions and data custodians understand exactly what will happen in the course of the data linkage project.

It is important to prepare documentation covering all aspects of the project, including the data linkage process and all outputs from the data linkage. This output may include confidentialised unit record linked data, data cubes, research papers, as well as release practices covering the preparation and publication of metadata, as well as other information about the data linkage project and the suite of products that are generated from the project.

To further promote transparency, project owners may also develop a suite of documentation for the project, including (but not limited to) project objectives, governance, protocols of data linkage, as well as protocols for data storage, data security, data transfer, data release and protocols for the protection of privacy and confidentiality. A tailored summary of the documentation, suitable for reading by the wider community, could be prepared and made available through websites and pamphlets.

7.2.2 After the project has been completed

Quality of linkage

It is good practice for analysts to describe the quality of the linkage process in reports that are publicly available. This enables readers to interpret the estimates derived from the linkage study appropriately.

There are various measures of linkage quality in use by data linkage specialists. Much can be learnt from these, but they often require extremely specialised knowledge to interpret. Comparability of data linkage projects will be optimised if the measures that are reported are broadly applicable, and simple to implement and explain.

Details of the quality of the data linkage that may be reported by investigators may include the following:

- data linkage method used
- variables used in data linkage
- extent and characteristics of records that were unable to be linked
- quality measures of linkage, e.g. sensitivity or precision.

Methods for deriving Indigenous status

The level of detail to be provided will vary between studies, and some suggested outputs have been provided for consideration below. At a minimum, however, all studies and reports should provide details of the source of Indigenous status data and the methods used to derive Indigenous status (including the rationale for choosing those methods and any identified limitations). It will also be useful to undertake and report on some level of impact analysis (e.g. comparison of results based on multiple algorithms for deriving Indigenous status).

The following elements could accompany data linkage findings:

- full description of the methods used, rationale for their use and their limitations
- number of records identified as Aboriginal and Torres Strait Islander, before and after linkage
- the characteristics of people, disaggregated by Indigenous status, before and after linkage (this is considered particularly important if there is likely to be bias associated with the linkage process, e.g. by age or some other demographic characteristic).

It must also be made clear that the purpose of applying these Guidelines is to enhance Indigenous status information on data sets that have been created for the purpose of statistical reporting and research. The Guidelines should not be used to change individuals' Indigenous identification on source data sets.

Results

If legislation allows, results from the analysis of linked data should be provided to data custodians prior to publication. This recognises that data custodians have a degree of control over the use of their data, and gives them the opportunity to provide input into the interpretation of any findings generated from the use of their data. Feedback from data custodians prior to general publication is also useful for analysts, as data custodians may have more insight into the underlying reasons for the results. Feedback to data custodians on the quality of the Indigenous status variable on their data sets will help them improve the quality of their data collection and data processing.

It will also be useful if analysts and the institutions overseeing the data linkage project prepare documentation, detailing the analytical uses to which the linked data have been applied, key findings from the analyses, and the benefits that have flowed from the project.

Technical documentation

It is good practice to prepare technical documentation covering the actual linkage process, or to refer to documentation already in the public domain, where it exists. The documentation could include information on:

- the purpose and scope of the data linkage project overall
- the methodologies for constructing linkage maps and generating linked data sets, for managing and releasing data, and for protecting privacy
- what linked data sets have been created and the users to whom they have been released.

7.2.3 Documentation considerations for data linkage institutions

The sections above have focussed on disclosure guidelines applicable to analysts of linked data. This section provides considerations for data linkage institutions, focussing on documentation that would support analysts in meeting their disclosure commitments. Ideally, data linkage institutions should document their procedures in publicly available data management plans, and make available to analysts details of the linkage process and the results of the assessment of the quality of the linkage.

There is also a role for data linkage institutions in publicly disclosing information about linkage quality. Data linkers are often in the best position to report on the quality of linkage since they, not the analysts of linked data, will most often have overseen the physical linkage of the data, and will have access to information not available to analysts. Data linkers may also be constantly linking data from the same sources, for different investigators with different research interests, and so are in a better position than analysts to co-ordinate information about the quality of data linkage.

Glossary

Aboriginal people People who identify or are identified as being of Aboriginal origin.

Aboriginal and Torres Strait Islander people People who identify or are identified as being of Aboriginal and/or Torres Strait Islander origin. See also *Aboriginal people and Torres Strait Islander people*.

Ad hoc data linkage This involves the linkage of two or more data sets for a specific purpose and a specific, often non-ongoing, project, using a specific set of input data sets. *Ad hoc* data linkage does not involve the maintenance of a master linkage file and master linkage key.

Adjustment factors See *correction factors*.

Administrative data Information that is collected for the purpose of, or in the process of, service delivery, such as providing health care (National Hospital Morbidity Database), responding to the legal requirements of registering particular events (births and deaths registration data) or providing a particular service (Residential Aged Care Data set).

Aggregate methods Approaches that use data linkage to inform adjustments to Indigenous status at an aggregate level, rather than making adjustments at the individual level.

Algorithm A process or set of rules used for calculation or problem-solving. In these Guidelines, 'algorithm' is used to refer to a set of rules that is used to determine Indigenous status of an individual based on a linked data set.

Blocking In data linkage, blocking reduces the number of comparisons needed by only comparing record pairs where links are more likely to be found. Records on each file are placed into blocks so that only record pairs that agree on certain data items are compared.

Blocking variables Variables used in partitioning records into blocks. Only records having the same value in a blocking variable are compared. Blocking variables must be stable, accurate and available on all the files to be linked. Examples of blocking variables are first and last name, components of first and last name, sex, components of date of birth (e.g. month of birth or year of birth) and components of usual place of residence.

Clerical review A manual review of record pairs whose link status cannot be automatically determined from their linkage weights or linkage probabilities. Clerical review helps determine the link status of these record pairs. Clerical review can be also be used to obtain a quality assessment of a linkage.

Clerical assessment A manual review of the validity or accuracy of the link status assigned to record pairs. The result of this assessment will assess whether the linked record pairs are true links or false links, true non-links or false non-links.

Closing the Gap in Indigenous disadvantage An initiative of the Council of Australian Governments (COAG) to close the gap between Indigenous and non-Indigenous Australians in the areas of life expectancy, infant and child mortality, early childhood education, reading writing and numeracy achievement, year 12 attainment, and employment outcomes.

Confidentiality Treatment of information about an individual or entity in a manner that will not disclose the identity of that individual or entity.

Comparison record pair Any pair of records from two data sets that are being compared to determine whether or not they belong to the same person or entity.

Correction factors Statistical values which are used to correct or adjust a number, rate, ratio or another statistical value that is understated, overstated or that has an error in it. See also *adjustment factors*.

Coverage The extent to which a data set captures the population in scope.

Data cleaning The process of editing data to remove duplicate records and errors such as illogical and out-of-scope values, and data entry errors, such as typographical errors and transposed values. In data linkage, data cleaning may also encompass data standardisation. See also *data standardisation*.

Data standardisation The process of making different data sets comparable and compatible, and conform to the same quality rules, in terms of structure of data set, scope, completeness, coding, structure and spelling of variable names, and range and format of data values.

Data custodian The authority, body or person responsible for the safe custody, transport and storage of data, and implementation of business rules regarding use of the data. Data custodians may either have collected the data themselves or they may have legal and administrative custody of it on behalf of the owner or collector of the data.

Data linkage The process of bringing together two or more sets of information belonging to the same person, event or place, into a single record of information. See *Record Linkage*.

De-identification Processes for removing identifying information from datasets, most commonly to protect the privacy of individuals.

Deterministic linkage Deterministic linkage ranges from simple joining of two or more data sets by a reliable and stable key to sophisticated stepwise algorithmic linkage. See simple *deterministic linkage* and *stepwise deterministic linkage*.

Dummy birth dates Birth dates assigned to an individual if actual date of birth information is incomplete (e.g. month of birth is missing) or is not available.

Dynamic data linkage system A system of data linkage that involves the ongoing linkage of core data sets and the permanent maintenance of a master linkage file and master linkage key. The master linkage file holds a discrete and select range of linkage variables (e.g. name, date of birth, sex, address) across multiple data sets, while the Master Linkage Key is a system of continuously updated links within and between core data sets.

False-negative link A pair of records belonging to the same individual or entity that is incorrectly assigned as non-matches or as not belonging to the same individual or entity.

False-negative rate The proportion of all record pairs belonging to the same individuals or entities that are incorrectly assigned as non-links.

False-positive link A pair of records belonging to two different individuals or entities that are incorrectly assigned as links.

False-positive rate The proportion of all record pairs belonging to two different individuals or entities that are incorrectly assigned as links.

Indigenous person A person who identifies, or is identified, as being of Aboriginal and/ or Torres Strait Islander origin. See also **Aboriginal people**, **Aboriginal and Torres Strait Islander people**, and **Torres Strait Islander people**.

Indigenous status The name of the variable that describes whether or not a person identifies, or has been identified, as being of Aboriginal and/or Torres Strait Islander origin.

Indigenous under-identification This may occur if Indigenous status is not correctly collected and recorded for all clients. While this can also lead to over-identification, the tendency has often been for Aboriginal and Torres Strait Islander to be recorded as non-Indigenous or for their Indigenous status not to be recorded at all.

Life expectancy The average number of additional years a person of a given age and sex might expect to live if the age-specific death rates of the given period continued throughout his/her lifetime.

Link A decision that two records correspond to the same person or entity.

Linked The status of a record that has passed through the data linkage process and was linked to a record from the other file.

Linking variables Variables that are common to the data files being linked, and are used for comparing records. Examples of linking variables include first name, last name, sex, full date of birth, usual place of residence and country of birth. Some linking variables can also be used as blocking variables. See also *blocking variables* and *matching variables*.

Master Linkage Key The codes created and stored by a data linkage unit that can be used to group records that refer to the same person or entity.

Match A record pair that contains information that relates to the same unit. See also *Link*, *Non-link*, *Non-match*.

Match accuracy rate is the proportion of all record pair comparisons that are true positives (TP) or true negatives (TN). The denominator for this rate is the number of all record pair comparisons, while the numerator is the number of record pairs that are correctly classified as true matches or false matches.

Matching variables See *Linking variables*.

Native title population registers Population registers that only contain names and information about Aboriginal and Torres Strait Islander people. Information in the register is collected specifically to administer Indigenous-specific programs. The Indigenous status of persons in the register is verified using information from official records, historical records, archival and genealogical sources as well as oral histories.

Non-link A decision that two records do not correspond to the same person or entity.

Non-match A record pair that contains information that relates to different people or entities.

Numerator-denominator bias A bias arising where the numerator and denominator of a rate or ratio are derived from different populations. This may occur when different data sources are used in the numerator and denominator and which are collected and/or compiled under different conditions and for different purposes. An example is mortality rates where the numerator is the number of deaths compiled by the Registrars of Births, Deaths and Marriages, while the denominator is the estimated resident population compiled from Census and other data.

Overcount More Aboriginal and Torres Strait Islander people are counted in a data set than would be present if the data set was perfect. This may be partly a consequence of misreporting, and partly a consequence of other factors, such as double counting of Aboriginal and Torres Strait Islander people, or poor quality adjustment/correction factors.

Pass One iteration of a record linkage, using a particular set of blocking and matching variables. See *Blocking*, *Blocking variables*, *Matching variables*.

Precision or positive predictor value The proportion of all classified links that are true links as opposed to classified links that are false links. It is calculated by dividing the number of links that are ascertained as true, by the total number of classified links.

Privacy The right of a person or group of people to keep their lives and personal affairs out of public view, and to control the flow of information about themselves.

Probabilistic linkage A method of record linkage that uses the probabilities of agreement and disagreement between a range of linkage variables.

Record linkage The process of bringing together two or more sets of information belonging to the same person, event or place, into a single record of information, in a way that protects individual privacy. See *Data linkage*.

Record pairs See *comparison record pairs*.

Sensitivity or true-positive rate The proportion of all records in a file or database with a match in another file that were correctly accepted as a link.

Simple (one-step) deterministic record linkage Simple linkage using a single identifier or linkage key to join two or more data sets.

SimRate An approach that uses the observed distribution of data in matched and unmatched pairs to generate a large simulated set of record pairs, assigns a match weight to each pair based on specified match rules and uses the weight curves of the simulated pairs for error estimation.

Specificity or true-negative rate The proportion of all records on one file or database that have no match in the other file that were correctly not accepted as a link.

Statistical linkage key (SLK) A code used in data linkage that replaces a person's first and last name to protect the person's identity. It is generated from elements of an individual's personal demographic data and attached to de-identified data relating to the services received by that individual.

Stepwise deterministic data linkage Deterministic data linkage that uses auxiliary information on the data sets to provide a platform from which variation in the reported linkage key or SLK information can be considered. This differs from simple deterministic linkage that relies on an exact, one-to-one character matching of linkage keys across two or more data sets.

Torres Strait Islander people People who identify or are identified as being of Torres Strait Islander origin.

True-positive link Two records that truly do correspond to the same person or entity See *Link, Non-link, True non-match*.

True-positive link Two records that truly do not correspond to the same unit (e.g. two different people). See also Link, Non-link, True match.

Undercount Fewer Aboriginal and Torres Strait Islander people are counted in a dataset than truly would be present if the dataset was perfect. This is partly a consequence of misreporting, and partly a consequence of poor coverage of Aboriginal and Torres Strait Islander people.

Under-identification When used in relation to Aboriginal and Torres Strait Islander people, “under-identification” is a consequence of misreporting that results in fewer Aboriginal and Torres Strait Islander people being identified in a dataset than are truly present.

Unique record identifier A variable that uniquely identifies a person, place, event or other unit.

Unlinked The status of a record that has passed through the data linkage process and was not linked to a record from the other file.

References

ABS (Australian Bureau of Statistics) 1999. Standards for Statistics on Cultural and Language Diversity. ABS cat. no. 1289.0. Canberra: ABS.

ABS 2008a. Information Paper: Census Data Enhancement—Indigenous Mortality Quality Study. ABS cat. no. 4723.0. Canberra: ABS.

ABS 2008b. Deaths, Australia, 2007. ABS cat. no. 3302.0. Canberra: ABS.

ABS 2009a. Experimental Life Tables for Aboriginal and Torres Strait Islander Australians 2005–07. ABS cat. no. 3302.0.55.003. Canberra: ABS.

ABS 2009b. Estimates and Projections, Aboriginal and Torres Strait Australians, 1991–2021. ABS cat. no. 3238.0. Canberra: ABS.

ABS 2010. Population Characteristics, Aboriginal and Torres Strait Islander Australians, 2006 ABS cat. no. 4713.0. Canberra: ABS.

AIHW (Australian Institute of Health and Welfare) 2005. Improving the quality of Indigenous identification in hospital separations data. Cat. no. HSE 101. Canberra: AIHW.

AIHW 2006. Data linkage and protecting privacy: a protocol for linking between two or more data sets held within the Australian Institute of Health and Welfare. Canberra: AIHW.

AIHW 2010a. National best practice guidelines for collecting Indigenous status in health data sets. Cat. no. IHW 29. Canberra: AIHW.

AIHW 2010b. Indigenous identification in hospital separations data: quality report. Health Service Series No 35, Cat. no. HSE 85. Canberra: AIHW.

AIHW 2011. Comparing an SLK-based and a name-based data linkage strategy: an investigation into the PIAC linkage. Data linkage series no. 11. Cat. no. CSI 11. Canberra: AIHW.

AIHW: Choi C, Hyndman R, Smith L, Kun Z & Dugbaza T 2012. An enhanced mortality database for estimating Indigenous life expectancy: A feasibility study. Cat. no. IHW 75. Canberra: AIHW.

Ajwani S, Blakely T, Robson B, Atkinson J & Kiro C 2003. Unlocking the numerator-denominator bias III: adjustment ratios by ethnicity for 1981–1999 mortality data. The New Zealand Census-Mortality Study. NZ Med J 116:U456.

Bass J & Garfield C 2002. Statistical linkage keys: How effective are they? Proceedings of Symposium on Public Health. Available online at: <http://www.publichealth.gov.au/pdf/reports_papers/symposium_procdngs_2003/bass.pdf>.

Bishop G & Khoo J 2007. Methodology of Evaluating the Quality of Probabilistic Linking. ABS Cat. no. 1351.0.55.108. Canberra: ABS.

Blakely T, Kiro C & Woodward A 2002a. Unlocking the numerator-denominator bias. II: adjustments to mortality rates by ethnicity and deprivation during 1991–94. The New Zealand Census-Mortality Study. New Zealand medical journal 115:43–7.

Blakely T, Robson B, Atkinson J, Sporle A & Kiro C 2002b. Unlocking the numerator-denominator bias. I: Adjustments ratios by ethnicity for 1991–94 mortality data. The New Zealand Census-Mortality Study. New Zealand medical journal 115:39–42.

Christen P & Goiser K 2007. Quality and complexity measures for data linkage and deduplication. Quality Measures in Data Mining: 127–51.

COAG (Council of Australian Governments) 2008. National Indigenous Reform Agreement (Closing the Gap), Schedule F. COAG.

Condon J, Barnes A, Cunningham J & Smith L 2004. Demographic Characteristics and Trends of the Northern Territory Indigenous Population 1966 to 2001. Darwin: Cooperative Research Centre for Aboriginal Health.

Condon J, Williams DJ, Pearce MC & Moss E 1998. Northern Territory hospital morbidity dataset: validation of demographic data 1997. Darwin: Northern Territory Health Services.

Cross Portfolio Statistical Integration Committee 2010a. High Level Principles for Data Integration Involving Commonwealth Data for Statistical and Research Purposes. Viewed 5 May 2010, <http://nss.gov.au/nss/home.NSF/pages/High+Level+Principles+for+Data+Integration+-+Content?OpenDocument>.

Cross Portfolio Statistical Integration Committee 2010b. Data Integration Involving Commonwealth Data for Statistical and Research Purposes: Governance and Institutional Arrangements. Viewed 5 May 2010, <http://nss.gov.au/nss/home.nsf/NSS/00FB7E20E1D56B96CA2577F20016C3DB?opendocument>.

Draper GK, Somerford PJ, Pilkington AS & Thompson SC 2009. What is the impact of missing Indigenous status on mortality estimates? An assessment using record linkage in Western Australia. Australian and New Zealand Journal of Public Health 33:325–31.

Fellegi I & Sunter A 1969. A theory of record linkage. Journal of the American Statistical Association 64:1183–210.

Gill L & Baldwin J 1987. Methods and technology of record linkage: Some practical considerations. In: Baldwin J, Acheson E & Graham W (eds). Textbook of Medical Record Linkage. New York: Oxford University Press.

Guiver T 2011. Sampling-based clerical review in probabilistic linking. ABS Cat. no. 1351.0.55.034. Canberra: ABS.

Herzog TN, Scheuren F & Winkler WE 2007. Data quality and record linkage techniques. Springer Verlag.

Hill K, Barker B & Vos T 2007. Excess Indigenous mortality: are Indigenous Australians more severely disadvantaged than other Indigenous populations? International Journal of Epidemiology 36:580–9.

Karmel R 2005. Data linkage protocols using a statistical linkage key. AIHW Cat. no. CSI 1. Canberra: AIHW.

Karmel R, Anderson P, Gibson D, Peut A, Duckett S & Wells Y 2010. Empirical aspects of record linkage across multiple data sets using statistical linkage keys: the experience of the PIAC cohort study. BMC Health Services Research 10:41.

Kennedy B, Cornes S, Martyn S, Wills R & Breckell C 2009a. Measuring Indigenous perinatal outcomes—should we use the Indigenous status of the mother, father or baby? Brisbane: Health Statistics Centre, Queensland Health.

Kennedy B, Howell S & Breckell C 2009b. Indigenous identification in administrative data collections and the implications for reporting Indigenous health status. Brisbane: Health Statistics Centre, Queensland Health.

Mak DB & Watkins RE 2008. Improving the accuracy of Aboriginal and non-Aboriginal disease notification rates using data linkage. BMC Health Services Research 8:118.

Manitoba Centre for Health Policy 2006. Concept: Record Linkage. Manitoba: University of Manitoba. Viewed 5 January 2011, <http://mchp-appserv.cpe.umanitoba.ca/viewConcept.php?conceptID=1209>.

Memmott P 2011. Personal communication. 5 October 2011.

Memmott P, Long S, Bell M, Taylor J & Brown D 2004. Between places: Indigenous mobility in remote and rural Australia. AHURI Positioning Paper No. 81. Melbourne: Australian Housing and Urban Research Institute, Queensland Research Centre, RMIT University.

NCSIMG (National Community Services Information Management Group) 2004. Statistical Data Linkage in Community Services Data Collections: A report prepared by the Statistical Linkage Key Working Group. Canberra: AIHW.

NHMRC (National Health and Medical Research Council) 2002. NHMRC Road Map: Strategic Framework for Improving Aboriginal and Torres Strait Islander Health Through Research. Canberra: NHMRC.

NHMRC 2003. Values and Ethics: Guidelines for Ethical Conduct in Aboriginal and Torres Strait Islander Healath Research. Canbera: NHMRC.

NHMRC 2005. Keeping research on track: A guide for Aboriginal and Torres Strait Islander peoples about health research ethics. Canberra: NHMRC.

NHMRC 2007. National Statement on Ethical Conduct in Human Research. Canberra: NHMRC, Australian Research Council, Australian Vice-Chancellors' Committee.

NHMRC 2010. The NHMRC Road Map II: A strategic framework for improving the health of Aboriginal and Torres Strait Islander people through research. Canberra: NHMRC.

Samuels C 2012. Using the EM Algorithm to Estimate the Parameters of the Fellegi-Sunter Model for Data Linking. Research Paper, Methodology Advisory Committee. ABS cat. no. 1352.0.55.120. Canberra: ABS.

Sayers SM, Mackerras D, Singh G, Bucens I, Flynn K & Reid A 2003. An Australian Aboriginal birth cohort: a unique resource for a life course study of an Indigenous population. A study protocol. BMC International Health Human Rights 3:1.

Taylor L 2010. Algorithms for deriving Indigenous status: Personal communication. 25 August 2010.

Winglee M, Valliant R & Scheuren F 2005. A Case Study in Record Linkage. cat. no. 12-001 Statistics Canada.

Wood C 2012. Personal communication. 1 March 2012.

Zubrick SR, Silburn SR, De Maio JA, Shepherd C, Griffin JA, Dalby RB et al. 2006. The Western Australian Aboriginal Child Health Survey: Improving the educational experiences of Aboriginal children and young people. Perth. Viewed 26 October 2011, <http://www.ichr.uwa.edu.au/waachs/publications/volume_three>.

List of tables

List of figures

Forthcoming publications

There are three companion documents or attachments to the Guidelines which are currently being prepared. These will be web publications and can be accessed at both the AIHW and ABS websites when they become available. These publications are:

- *Review of past, current and planned data linkage projects with an Indigenous focus*: This review analysed various Australian and overseas studies based on linked data relating to Aboriginal and Torres Strait Islander people, in terms of whether the purpose of the study was to enhance the value of Indigenous status information across data sets or to add value to data for purposes of undertaking research that cannot be undertaken using data from only one source. The review also examined the core themes of the studies, as well as the data sets and data linkage methodology used or intended to be used in the studies, data quality issues encountered or anticipated, problems with the quality of the Indigenous status variable on the various data sets, the method of analysis, and what algorithms or methods were used or are planned to be used in deriving Indigenous status, if Indigenous status was missing or was inconsistently reported across the various input data sets. In particular, the review also investigated whether the researchers explored or intend to explore the impact of various algorithms or methods for deriving Indigenous status on the estimated outcome measures or indicators.

  When it becomes available, this publication can be accessed at the AIHW website: <www.aihw.gov.au>

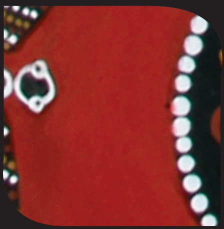
- *Project list of past, current and planned data linkage projects with an Indigenous focus*: This is a listing and description of past, current and planned data linkage studies relating to Aboriginal and Torres Strait Islander people. The publication provides a brief listing of the name of the project, the names of the investigators, the date of the study, the jurisdiction where the study is based, the data sets used in the study, the core issue or theme of the study, the method of analysis, and the method or algorithms used or intended to be used to derive Indigenous status information, if required.

  When it becomes available, this publication can be accessed at the AIHW website: <www.aihw.gov.au>

- *Getting Our Story Right*:
  This publication is the result of a collaborative research effort between the ABS, the Western Australian Department of Health, and the Telethon Institute for Child Health Research, to test and report on several methods for deriving Indigenous status where Indigenous status is missing or inconsistently reported across data sets. The study is based on 11 administrative and survey data sets, containing up to 29 million records.

  When it becomes available, this publication can be accessed at the ABS website: <www.abs.gov.au>

In 2008, the Council of Australian Governments (COAG) agreed to a set of targets for 'Closing the Gap' in disadvantage between Aboriginal and Torres Strait Islander people and non-Indigenous Australians. Currently, progress is difficult to measure accurately because Indigenous status is either missing or inconsistently reported across data sets. Data linkage has been identified as a potential way to improve reporting against the COAG targets. To ensure a consistent approach, COAG directed the Australian Institute of Health and Welfare and the Australian Bureau of Statistics to develop national best practice guidelines for linking data related to Aboriginal and Torres Strait Islander people. This report offers guidance on how to derive Indigenous status when it is missing or inconsistently reported, and covers guidelines in areas such as: values and ethics in Aboriginal and Torres Strait Islander research; quality of Indigenous status information in data collections; quality of linkage variables; assessment of quality of data linkage; methods for deriving Indigenous status; and transparency.



Cover artwork: **Azeam Fraser,**
Krambruk (landing place), 2011.