



Australian Government

**Australian Institute of
Health and Welfare**

An enhanced mortality database for estimating Indigenous life expectancy

**A feasibility study
2012**



Australian Government

**Australian Institute of
Health and Welfare**

*Authoritative information and statistics
to promote better health and wellbeing*

An enhanced mortality database for estimating Indigenous life expectancy

A feasibility study

2012

Australian Institute of Health and Welfare
Canberra

Cat. no. IHW 75

The Australian Institute of Health and Welfare is a major national agency which provides reliable, regular and relevant information and statistics on Australia's health and welfare. The Institute's mission is authoritative information and statistics to promote better health and wellbeing.

© Australian Institute of Health and Welfare 2012



This product, excluding the AIHW logo, Commonwealth Coat of Arms and any material owned by a third party or protected by a trademark, has been released under a Creative Commons BY 3.0 (CC-BY 3.0) licence. Excluded material owned by third parties may include, for example, design and layout, images obtained under licence from third parties and signatures. We have made all reasonable efforts to identify and label material owned by third parties.

You may distribute, remix and build upon this work. However, you must attribute the AIHW as the copyright holder of the work in compliance with our attribution policy available at <www.aihw.gov.au/copyright/>. The full terms and conditions of this licence are available at <<http://creativecommons.org/licenses/by/3.0/au/>>.

Enquiries relating to copyright should be addressed to the Head of the Communications, Media and Marketing Unit, Australian Institute of Health and Welfare, GPO Box 570, Canberra ACT 2601.

A complete list of the Institute's publications is available from the Institute's website <www.aihw.gov.au>.

ISBN 978-1-74249-315-2

Suggested citations

Australian Institute of Health and Welfare 2012. An enhanced mortality database for estimating Indigenous life expectancy: A feasibility study. Cat. no. IHW 75. Canberra: AIHW.

Australian Institute of Health and Welfare

Board Chair

Dr Andrew Refshauge

Director

David Kalisch

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

**Please note that there is the potential for minor revisions of data in this report.
Please check the online version at <www.aihw.gov.au> for any amendments.**

Contents

Acknowledgments.....	iv
Abbreviations.....	v
Symbols.....	vi
Summary	vii
1 Introduction.....	1
2 The data.....	3
Mortality Extract Data set (MED)	4
Death registration.....	4
Residential aged care data	5
National Hospital Morbidity Data set.....	6
National Perinatal Data Collection	6
The collection and recording of Indigenous status information across data sets	7
3 The method.....	9
The linkage process.....	9
Method of enhancing information on Indigenous status obtained from linked data set ..	10
Numerator-denominator bias.....	10
4 Results of the linkage process	12
5 Discussion.....	19
Conclusions	23
Appendix A: The data sources.....	24
Appendix B: Setting up the Mortality Extract Data set (MED).....	30
Appendix C: Developing the ‘pass’ structures and the validation study	32
Appendix D: The linkage processes	35
Appendix E: Classification of excellent, good and acceptable matches	43
Appendix F: Life expectancy at birth, calendar years 2001 to 2005, for NSW, Qld, WA, NT and Australia, with 95% confidence lower and higher limits.....	44
References	45
Glossary.....	47
List of tables	50
List of figures	51

Acknowledgments

The authors of this study are Ching Choi of the Australian Demographic and Social Research Institute at the Australian National University, Rob J Hyndman of the Department of Econometrics and Business Statistics at Monash University, Len Smith of the Australian Demographic and Social Research Institute at the Australian National University, Kun Zhao of Health Registers and Cancer Monitoring Unit, the Australian Institute of Health and Welfare, and Tetteh Dugbaza from the Social and Indigenous Group, at the Australian Institute of Health and Welfare.

This research project was commissioned by the Australian Institute of Health and Welfare (AIHW) and was funded by the Office for Aboriginal and Torres Islander Health of the Australian Government Department of Health and Ageing. The AIHW Indigenous Mortality Data Advisory Group was set up to provide advice to the research team. The members of the advisory group were:

- Fadwa Al-Yaman (Chair – AIHW)
- Patrick Corr (Australian Bureau of Statistics)
- Yin Paradies (Melbourne University)
- Mark Short (Australian Institute of Health and Welfare)
- Peter McInnes (Australian Government Department of Health and Ageing)
- Kirrily Harrison (Australian Government Department of Health and Ageing)
- Peter Brandt (NSW Department of Health).

We are grateful to the AIHW National Perinatal Statistics Unit at the University of New South Wales for the perinatal death data, the AIHW Population Health Unit for the National Mortality Database, the Data Linkage Unit for the National Death Index, the AIHW Hospital Unit for the National Hospital Morbidity Database, and the AIHW Ageing and Aged Care Unit for the Residential Aged Care Data set, which is made up of a collection of tiles from the Aged and Community Care Management Information System. We are also grateful to state and territory registrars of births, deaths and marriages, state and territory health departments and the Australian Government Department of Health and Ageing for permission to use their data for the linkage work. The authors are further grateful to the Australian Bureau of Statistics for providing the ABS Mortality CODURFs, on which the AIHW National Mortality Database is based.

The contributions from Brendan Scott, Michelle Gourley, Phil Anderson, George Bodilsen, Peter Braun, Karen Bishop, Katrina Burgess, Lisa Hilda, Rosemary Karmel, Christine Sturrock, Liz Sullivan, Mary Sullivan, Lee Taylor and members of the Advisory Group are very much appreciated.

Abbreviations

ABS	Australian Bureau of Statistics
ACAT	Aged Care Assessment Team
ACCR	Aged Care Client Record
ACT	Australian Capital Territory
AIHW	Australian Institute of Health and Welfare
CDE	Census Data Enhancement
COAG	Council of Australian Governments
COD	Cause of death
CODURF	Cause of Death Unit Record File
DOB	Date of birth
DOD	Date of death
DoHA	Department of Health and Ageing
DRS	Death Registration Statement
EMD	Enhanced Mortality Database
ERP	Estimated Resident Population
MCCOD	Medical Certificate of Cause of Death
MED	Mortality Extract Data set
NDI	National Death Index
NHMD	National Hospital Morbidity Database
NMD	National Mortality Database
NPDC	National Perinatal Data Collection
NPESU	National Perinatal Epidemiology Statistics Unit
NT	Northern Territory
RAC	Residential Aged Care
RBDM	Registrar of Births, Deaths and Marriages
ROC	Receiver Operating Characteristic
SLA	Statistical Local Area
SPARC	System for the Payment of Aged Residential Care

Symbols

—	nil or rounded to zero
..	not applicable
n.a.	not available
n.p.	not publishable because of small numbers, confidentiality or other concerns about the quality of the data

Summary

In 2008, the Council of Australian Governments (COAG) agreed to reduce the gap in life expectancy between Indigenous and non-Indigenous Australians. Robust measures of Indigenous mortality are needed to monitor the success, or otherwise, of government and community efforts to reduce the gap.

There is imperfect identification of Indigenous deaths in the key data sets used in preparing Indigenous life tables and mortality estimates. As a result, the true mortality characteristics of Indigenous Australians cannot be reliably measured.

The aim of the Enhanced Mortality Database project is to improve information on Indigenous status on the registered deaths data set by linking it to several additional data sources that contain information on Indigenous deaths and Indigenous identification. The enhanced data are expected to enable more accurate estimates of Indigenous mortality, including life expectancy, to be made. This study will also show whether such data linkage work is feasible and whether the results are credible.

The additional data sets are the Residential Aged Care (RAC) data set, the National Hospital Morbidity Database (NHMD) and the National Perinatal Data Collection (NPDC). The quality of these data sets varies between states and territories.

An 'ever-Indigenous' approach was used in determining Indigenous status from the various data sets. The approach accepts that the deceased was Indigenous if indicated by any of the data sets.

The death registration data set (2001–2006) used for this project contained 10,547 deaths listed as Indigenous. The linkage of the additional data sets to the death registration data set identified 1,081, or 10.2%, more deaths than the 10,547 originally recorded on the death registration data set as 'Indigenous'. Of the additional Indigenous deaths, 74.5% were from the National Hospital Morbidity Database, 24% from the Residential Aged Care data set and 1.5% from the National Perinatal Data Collection.

Using these linked data sets produces national estimates of Indigenous expectation of life at birth, of 66.6 years for males and 72.7 for females for the period 2001–2006. These estimates are close to estimates of 67.2 years for Indigenous males and 72.9 for Indigenous females, prepared by the ABS from the Mortality Quality Study, which was conducted as part of the 2006 Census Data Enhancement Study.

Although data from all states and territories were linked and used in preparing the Indigenous life tables for Australia, separate life tables were only prepared for New South Wales, Victoria, Queensland, Western Australia, South Australia and the Northern Territory. This was because the number of Indigenous deaths in Tasmania and the Australian Capital Territory were too few to enable reliable life tables to be constructed.

1 Introduction

The gap between Indigenous and non-Indigenous life expectancy in Australia is of great concern. It is government policy to reduce this gap. Robust measures of Indigenous mortality are needed to monitor the success or otherwise of government and community efforts to reduce the gap.

Estimates of Indigenous mortality rely on accurate identification of Indigenous deaths in the death registration system and in population censuses. Deaths are registered by state and territory registrars of births, deaths and marriages (RBDMs), and estimates of the size and demographic characteristics of the Indigenous population are made by the ABS using the population census as the base. Current estimates show that Indigenous identification in mortality data is incomplete.

While estimates of the size of the Indigenous population have improved through better population census counts and estimates of under-coverage, improvements in death data coverage have been slower. The imperfections in the death and population data vary between states and territories. Results from the 2006 ABS Mortality Quality Study (which was conducted as part of the overall ABS Census Data Enhancement or CDE project) indicated that coverage of Indigenous deaths in death registration data is about 85% nationally, ranging from 76% in New South Wales to 99% in the Northern Territory (ABS 2008d). The study involved linking 2006 Census records with deaths registered from 9 August 2006 to 30 June 2007 for all states and territories except Victoria, where death registration records were only available to mid-March 2007.

While the CDE Mortality Quality Study indicated that coverage rates for Indigenous deaths are much higher than previously thought, the estimated coverage rates are subject to certain limitations, including a very restricted time frame of 11 months, from early August 2006 to the end of June 2007. ABS estimates that there was a net undercount of Indigenous people at the 2006 Census of over 11%, including up to 24% in WA. Obviously, the people who were not counted in the census could not be linked to any corresponding death records.

The undercount is one reason why a very high percentage of Indigenous death records (26%) could not be linked to a corresponding census record. The proportion of unlinked Indigenous death records was as high as 35% in Western Australia and 40% in the Northern Territory (ABS 2008c, 2008d).

This research attempts to enhance death data through the joint use of a number of sources containing information on Indigenous deaths and identification. This study aims to show whether such data linkage work is feasible and whether the results are credible.

The enhanced death data will be used to create the AIHW Enhanced Mortality Database (EMD). The EMD will enable more accurate estimates of Indigenous mortality, including life tables and life expectancy estimates. The availability of additional data sources that contain information on Indigenous deaths and Indigenous identification allows enhancements to be made to the deaths data for particular age groups or states and territories.

The period selected for this research is 1 July 2001 to 30 June 2006. For this project, Indigenous mortality rates and life tables were calculated using the estimated deaths in the EMD and Indigenous population estimates published by the ABS. ABS population estimates as at 30 June 2006 were based on the 2006 Population Census, while estimates for previous years were reverse survived or 'back-cast' populations from the 30 June 2006 estimates.

The life tables created from the EMD are for research purposes only and are not meant to replace or duplicate the official ABS Experimental Life Tables for Aboriginal and Torres Strait Islander Australians. The purpose of the EMD Database is to help improve methodologies for the production of more reliable Indigenous mortality measures, including life expectancy. Though both the ABS and AIHW are attempting to use data linkage to enhance the quality of Indigenous status information on death records, there are fundamental differences between the approaches adopted by the respective agencies.

The starting point for both methods is death registration data compiled by the state and territory RBDMs. While the ABS linked this data set to the Census data to derive an estimate of the coverage of Indigenous deaths in the deaths registration data, the AIHW linked the data to three administrative data sets (see Chapter 2).

The ABS method uses deaths data over an 11-month period and produces estimates of Indigenous life expectancy every five years. The AIHW approach, on the other hand, uses 5 years deaths data and is able to produce yearly estimates of Indigenous life expectancy in the 5-year period. It is hoped that this study will lead to the production of a long-term database that can be used for Indigenous mortality research and can be used to construct a time series of Indigenous mortality measures.

It should be noted that while annual life expectancy estimates are produced for the jurisdictions with relatively large Indigenous populations, life expectancy estimates for only aggregated years (for example, 2001–2006) are produced for jurisdictions with relatively small Indigenous populations (such as Victoria and South Australia). Separate life expectancy estimates are not produced for Tasmania and the Australian Capital Territory because of their much smaller Indigenous populations.

As the aim of this project is to produce enhanced death data on a long-term basis for research purposes, it should be possible in future to produce rolling 3-year grouped life expectancy estimates for the smaller jurisdictions, which could be released annually.

Approval to undertake this study and to access relevant data sets was obtained from the AIHW Ethics Committee as well from all Commonwealth, state and territory custodians of all the data sets used for the study. The AIHW Ethics Committee operates in accordance with National Health and Medical Research Council Guidelines.

The study was designed and implemented in accordance with all relevant Commonwealth and state and territory privacy legislation and the AIHW data linkage protocols.

2 The data

The data sets used for the linkage are sourced from death registration, the National Hospital Morbidity Data set (NHMD), the Residential Aged Care (RAC) data set and the National Perinatal Data Collection (NPDC). The deaths data used in the estimation of the Indigenous life expectancy estimates are extracted from the National Death Index and the National Hospital Morbidity Data set. Deaths that occurred between 1 July 2001 and 30 June 2006 were selected and used for this linkage study.

Admitted hospital patients account for about 52–54% of all deaths in Australia. These deaths are included in the NHMD (AIHW 2011). Separations from residential aged care as a result of death account for about 31–33% of all deaths in Australia. This information is included in the RAC data set (AIHW 2008c, 2009). Together, these two sources may account for the large majority of all deaths in Australia.

There could be a small amount of overlap in the recording of separations from hospital and residential aged care through death, whereby the same death may be recorded in both data sets.

These overlaps may occur because some residential aged care residents, who have been temporarily transferred to hospital and subsequently die in hospital, may be recorded in both the NHMD and the RAC as having separated due to death. This may be especially likely to occur if the temporary transfer happens on the same day as discharge from the residential aged care facility due to death (Broad 2012). These overlaps, however, will not affect the enhanced mortality database, because any duplicated records will be identified during the data linkage process.

The perinatal death data are restricted to early infant deaths. Additional sources of Indigenous infant death data are important for the enhanced mortality database because Indigenous infant deaths are likely to be under-identified when they are registered.

These three additional data sources together give very good, although not exhaustive, coverage of deaths in Australia. It must be pointed out, however, that death registration records were only matched to Indigenous-identified deaths in the additional data sets. Because inconsistencies in Indigenous status reporting are not limited to death registration, it is possible that not all Indigenous records would have been identified on the additional data sets. It is also possible that some records were incorrectly identified as Indigenous. Such errors contribute to false positives and false negatives in matching and are common to all linkage studies.

Another potential source of information on Indigenous deaths and Indigenous identification is the medical certificate of cause of death. All jurisdictions have enacted legislation that requires the attending physician or the doctor who certifies a patient's death to also separately notify the RBDM within a specified period (usually 48 hours) after the death. Information provided to the Registrar by the attending or certifying physician is in the form of a medical certificate of cause of death (MCCOD). This requirement excludes coroner-certified deaths that are covered by separate reporting arrangements. In addition to information on the cause of death, the MCCOD also includes demographic information on the deceased, including Indigenous status.

In some jurisdictions, Indigenous information on the MCCOD has already been incorporated into death registration records. Where it is not incorporated, this information has not been computerised and is therefore not available for this project.

Indigenous death records from the other three additional data sources were linked to records of deaths registered by state and territory RBDMs.

The sections below provide a brief description of the respective data sets used for this project. Appendix A gives details of the quality of each of the data sets.

Mortality Extract Data set (MED)

The MED is the primary data file on which the data linkage is based. It is created from extracts of registered death data from the AIHW National Mortality Database (NMD) and the National Death Index (NDI). The NMD is based on the ABS Cause of Death Unit Record File (CODURF). Both the ABS CODURF and the NDI are compiled from deaths registration data obtained from the RBDMs in the various jurisdictions.

The NDI contains names; the NMD does not. The linking of extracts of data from the NDI to an extract of data from the NMD therefore results in the Mortality Extract Data set that contains edited and validated records with names. The deaths registration data are described in the section below.

Data fields extracted from the NDI for inclusion in the MED were full names, sex, date of birth, date of death, state of registration, year of registration, geographic variables, Indigenous status, and a unique record identifier. Data fields extracted from the NMD comprised the same set of variables as were extracted from the NDI, with the exception of names and address, which are not included on the NMD. The MED contains only the variables required for linking to the other data set, including names. It does not contain any clinical or health care information.

Of the three data sets linked to the MED, only the RAC contains names which could be used as a linking variable. The other data sets do not contain information on names; however, date of birth and date of death create a relatively unique combination of variables that could be used for linking, in addition to other linkage variables such as sex and geographic variables (house number, street name, postcode, Statistical Local Area (SLA), where available).

Death registration

Registration of deaths is the responsibility of state and territory RBDMs. When a death occurs, some jurisdictions require that the funeral director, or any other person who arranges for the disposal of the remains, must complete a Death Registration Statement (DRS) within 7 days of the disposal of the remains, to inform the RBDMs of the fact of death. In some jurisdictions, a DRS must be completed within 14 days of death. The DRS also includes other sociodemographic information on the deceased, including full name, sex, date of birth, date of death, full usual residence information, and Indigenous status, as well as information on date of death and place of death.

At the same time, a doctor who was responsible for a person's medical care immediately before death, or who examines the body of a deceased person after death must within 48 hours after the death give written notice of the death to the RBDM, including particulars required by regulation. These particulars include the cause of death, with some jurisdictions also providing information on the Indigenous status of the deceased.

Where a death is subject to coronial enquiry, then the coroner must, as soon as practicable, notify the RBDM of that fact. The information provided by the coroner may include information on the Indigenous status of the deceased.

The question asked to ascertain Indigenous status on both the Death Registration Statement and the MCCOD is the standard question:

Was the deceased (or deceased person) of Aboriginal or Torres Strait Islander origin?

- No
- Yes, Aboriginal
- Yes, Torres Strait Islander
- Yes, both Aboriginal and Torres Strait Islander (not all states).

Death registration data are provided to the ABS to compile and publish annual death statistics. After de-identification and coding of causes of death, the ABS creates a file called the ABS Cause of Death Unit Record File (CODURF). The AIHW uses these data to maintain the AIHW NMD.

The AIHW also receives from each registrar monthly 'fact of death' files that include the names and demographic characteristics of the deceased, including their full name, sex, date of birth, date of death, place of usual residence, Indigenous status and a unique record identifier. This monthly data collection is used to compile the NDI, which contains a listing of all deaths that have occurred in Australia since 1980.

The availability of full names of the deceased on the NDI enabled the linkage of the Mortality Extract Data set with the RAC data set which also contains names (see Appendix B). This linking of named registered deaths to named residential aged care deaths provide the basis for the construction of the linkage algorithm for linking registered deaths to other data sets (see Appendices B and C).

Residential aged care data

The provision of residential aged care is the responsibility of the Australian Government. Residential aged care data are collected by the Aged Care Assessment Team (ACAT) at the time a person is assessed for government-subsidised aged care, and by the aged care home while they live there.

The data provided to the AIHW for this project are from the System for the Payment of Aged Residential Care (SPARC), administered by the Australian Department of Health and Ageing (DoHA). SPARC contains information gathered through a number of instruments (AIHW 2008c). Of these, the one that relates to this project is the Aged Care Client Record (ACCR) (previously Aged Care Application and Approval form).

The ACCR is filled in at the time of assessment for government-subsidised aged care by persons applying for admission (or their carer), as well as by the ACAT. Data on Indigenous identification, date of birth, sex and usual place of residence at the time of assessment are recorded on this form.

The question to ascertain Indigenous identification on the Aged Care Client Record is the standard question (with different order of answer categories):

Is the client of Aboriginal or Torres Strait Islander origin?

- Yes, Aboriginal
- Yes, Torres Strait Islander
- Yes, both
- No, neither.

Each year DoHA provides a copy of the residential aged care data to the AIHW for statistical analysis and publication. The data provided contain basic sociodemographic information about each resident, including full name, sex, Indigenous status, date of birth, date of death, usual residence, and address of the facility where the death occurred.

National Hospital Morbidity Data set

Hospital data used in this project relate to inpatient separations from both public and private hospitals throughout Australia. Hospital patients who are not admitted (for example, emergency department patients, and outpatients) are not included. The exclusion of non-admitted deaths should not have a large impact on the project. In 2005–06, 1,620 patients died in emergency departments without being admitted, representing 2% of all hospital deaths. The admitted patients or their carers fill in sociodemographic details at the time of admission, and these include full name, age, sex, date of birth, full information on usual residence, and Indigenous status.

Indigenous identification is ascertained through the standard question:

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

- No
- Yes, Aboriginal
- Yes, Torres Strait Islander.

Hospital data are compiled and published by the AIHW from data provided by state and territory health authorities. Unit records of each separation are provided to the AIHW without names, but with detailed demographic details. These include the personal and sociodemographic information collected on admission, as well as additional data that may now have become available, such as date of death and address of the facility where the death occurred. These demographic details are suitable for use in data linkage

National Perinatal Data Collection

The National Perinatal Data Collection (NPDC) is a national data collection of pregnancy and childbirth, managed by the AIHW National Perinatal Epidemiology Statistics Unit (NPESU) at the University of New South Wales. Midwives and other staff complete notification forms for each birth, using information obtained from mothers, hospital admission and inpatient records, and antenatal pregnancy records. Data from the various sources are collected and collated by the health departments of the respective jurisdictions. The collated data are sent to the NPESU. The NPESU then compiles the collated data into the National Perinatal Data Collection, or NPDC. Information is included in the NPDC on both live births and stillbirths of at least 400 grams birthweight or at least 20 weeks gestation.

Only perinatal deaths, comprising fetal and neonatal deaths, are included in the NPDC. A fetal death refers to the birth of a fetus weighing at least 400 grams (or where birthweight is unavailable of at least 20 or more completed weeks of gestation) which shows no signs of life. Fetal deaths are commonly known as stillbirths. Neonatal deaths are deaths of a liveborn baby within 28 days of birth.

The NPDC collects Indigenous status information of only the mother of the baby, and not that of the father or the baby. For this reason, the baby's Indigenous status has tended to be

based on the Indigenous status of the mother. The standard question to ascertain Indigenous status is not uniformly used in all the data sources for the NPDC. While hospitals use the standard form of the question recommended by the ABS (1999), other collections, such as antenatal pregnancy records, may not.

As the NPDC collects Indigenous identification only of the mother, neonatal deaths of babies with Indigenous fathers and non-Indigenous mothers may not be identified as Indigenous.

Unit records of each neonatal death were provided to this project without names, so linkage of this data set to the MED was based on sex, date of birth, date of death and the mother's usual address. Without names, the combination of date of birth and date of death provides an almost unique identifier for purposes of data linkage.

Although both the ABS and NPESU use the same definition of a perinatal death, there are some differences between the perinatal deaths data in the NPDC and those in the ABS CODURF. Firstly, the ABS CODURFs are compiled from deaths registration data provided by the RBDMs, while the NPDC data are based on information from mothers and hospital or other records provided to the perinatal data collection in each state and territory. The NPDC may not capture some perinatal deaths if they occurred outside the setting where data are collected. For example, some perinatal deaths may be missed if the birth occurred in a very remote area and the child dies before it comes into contact with the health system.

Secondly, while the NPDC includes fetal deaths, the ABS CODURF does not, although both data sets include information on neonatal deaths. Thirdly, the NPDC does not contain information on babies dying 28 days or more after birth.

Fourthly, while the NPDC only provides information on the Indigenous status of the baby's mother, the Registration data set records the Indigenous status of the deceased. Lastly, the perinatal deaths data in the ABS CODURF are based on year of registration data while those in the NPDC are based on year of occurrence.

The collection and recording of Indigenous status information across data sets

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. The *National Best Practice Guidelines for collecting Indigenous status in health data sets* (AIHW 2010b) provides further information about how Indigenous status information should be collected and recorded in health settings.

The recommended format of the question is:

Are you [Is the person] of Aboriginal or Torres Strait Islander origin?

(For persons of both Aboriginal and Torres Strait Islander origin, both “Yes” boxes should be marked).

No ☐

Yes, Aboriginal ☐

Yes, Torres Strait Islander ☐

It could be seen from the description of how information on Indigenous status was obtained across the various data sets that variations of the standard question were used, including variations in the response categories. It is not clear, however, if and how these may affect the quality of the Indigenous status information on these data sets.

There may also be inconsistencies in the way information on Indigenous status is derived, especially in data sets such as the NHMD where there are multiple entries or contacts per person. Some agencies may update a person’s Indigenous status at every contact, whereas others may simply accept the recorded Indigenous status as accurate. It is also not clear if and how these inconsistencies may affect the quality of the Indigenous status information in the registered death data sets.

3 The method

The method involved linking death records in the MED to Indigenous death records in other identified data sources that contain substantial numbers of Indigenous deaths, and comparing information on Indigenous status across the linked data sets.

The method can be divided into four main processes. These are:

- creating the MED
- linking the MED to the additional data sets
- comparing and enhancing the Indigenous status of records in the linked data sets and
- using the enhanced mortality data set to prepare life expectancy estimates for the Indigenous population.

The linkage process

The objective of the linkage was to enhance the Indigenous status information on death records contained in the MED. To achieve this, records in the MED were linked to Indigenous death records from each of the additional data sources. This was done by separately linking each of the records in the MED (whether Indigenous or not) to Indigenous death records from the RAC, NHMD and NPDC. These additional data sets comprised Indigenous deaths only, and were matched to records in the MED using linkage fields such as names, sex, date of birth, date of death, and geographic variables. Where there were no names recorded, as in the NHMD and NPDC, the combination of the other linkage variables, including the dates of birth and death, provided enough information. The complete information from each of the linked records was then added to the MED, which now becomes the Enhanced Mortality Database or EMD.

The linkage process was undertaken in four stages. In the first, a set of variables were extracted from the NDI and linked to an extract of data from the NMD to construct the MED. Data fields extracted from the NDI and NMD to form the MED were the names and address of usual residence of the deceased (NDI only), their unique record identifier, sex, date of birth, date of death, and Indigenous status. Details of this linkage phase are in Appendix B.

The second stage involved developing criteria, or 'pass structures' for linking the MED to the other data sets. These pass structures feature a sequence of linkage iterations or passes where record pairs match on different 'blocks' or combinations of the names (where available), sex, date of birth, date of death and geographic variables. The RAC data set contains names, and linkage involved using names and other personal characteristics. Because the NHMD and NPDC do not contain names, sets of different matching criteria not involving names were constructed. These criteria, or pass structures, were validated using the linkage of aged care records to records from the MED. Names are available on both the RAC data set and the MED, and linking the two data sets provides the gold standard against which the other linkage processes are validated. Details of the 'pass structures' and the validation study are in Appendix C.

The third stage is the actual linkage of the various data sets to the MED. This was done individually because the identifying information is different in each data set. Rules were developed to determine if record pairs being compared in each pass were considered a

match (belonging to the same individual). These rules generally reflect the extent to which details of two records have enough similarities to represent the same death. Details of this stage of the linkage process and the data items used for matching are in Appendix D.

In the fourth stage, the linked data sets were used in creating the EMD, which contains Indigenous status information for each death record across five data sets. The Indigenous status of the death records across the data sets was then compared, and the probable Indigenous status of each record decided, to be then used for preparing mortality and life expectancy estimates. The Indigenous deaths data were then analysed and used in developing the Indigenous life tables.

Method of enhancing information on Indigenous status obtained from linked data set

The EMD has five Indigenous status fields from the five data sets that were used to create it. A death record is considered to be Indigenous if it is recorded as Indigenous in any of the five data sets. This approach can be described as the ‘ever-identified’ approach. It accepts the Indigenous identification in any of the data sets as accurate.

There are other approaches to the way Indigenous status may be derived from a linked data set. These include determining that a record is Indigenous if it recorded as such in: two or more data sets, in a majority of data sets, in the most recent data set, in the most trustworthy data set, in a gold standard or ‘truth’ file, or if a person achieves a certain weighted score or probability of being Indigenous, based on the person’s Indigenous status across multiple data sets.

It is sometimes argued that the ‘ever-identified’ approach can overestimate the true number of Indigenous deaths, as a non-Indigenous person could be wrongly identified as Indigenous in any of the linked records (false-positive). This is particularly a problem if the linking relates to events that can occur to an individual many times over a period, such as hospital admissions. In such cases, there will be more chances of wrongly identifying a person as Indigenous. As this project deals with death records, this problem exists but is not compounded, as death occurs only once.

However, while it is true that more Indigenous deaths may be identified under the ever-Indigenous than under other approaches, this does not necessarily mean that the ever-Indigenous approach will always overestimate the number of Indigenous deaths. The ever-Indigenous approach will only overestimate the number of Indigenous deaths if the number of true non-Indigenous people recorded as Indigenous (false positives) is greater than the number of true Indigenous people recorded as non-Indigenous (false negatives).

Numerator-denominator bias

One potential issue that arises when using administrative data sets to prepare population-based rates is the problem of numerator-denominator bias, where the numerator (the events of interest) are not drawn entirely from the population included in the denominator. This may arise where the numerator and denominator are collected by different agencies, for different purposes, using different processes of data collection, and often relating to different reference periods.

For example, in the estimation of birth, death, morbidity and hospital separation rates, the numerator, such as the number of deaths, births, hospital admissions and separations, diagnoses for particular diseases etc., is based on the number of persons experiencing a certain event or accessing a particular service over a defined period. Conversely, the denominator is a synthetic estimate of the population in the middle of the reference period, based on information collected at the census. The population of events estimated by the numerator is not necessarily a subset of the population estimated by the denominator – because the numerator events occur at a known point in time, and the denominator is an estimate.

The problem of numerator-denominator bias also exists in other rates, such as marriage and divorce rates, school participation rates and hospital utilisation rates, where the numerator is based on administrative data on the number of persons participating in the activity, and the denominator is an estimate of the number of people in the middle of the year.

Putting aside any issues of bias that are not specific to the Indigenous context, in this study numerator-denominator bias will occur if the deaths identified as Indigenous at the end of the linkage process do not represent all deaths in the Indigenous population; that is, the extent to which deaths within that population are not identified as Indigenous (false negatives) or deaths of people who are not from the population are identified as Indigenous (false positives), and the balance between the two.

The degree of numerator-denominator bias has not been estimated in this project.

4 Results of the linkage process

Matched records were classified into excellent quality, good quality, or acceptable matches. The classification system varies between the linked data sets, as linkage variables used in each linkage are different. In general, matches that meet strict criteria such as having exact or near exact matches on name, date of birth and date of death are classified as excellent quality matches. Those that meet less strict criteria, such as having some minor differences in name, date of birth and date of death are classified as good quality matches. Those that meet even less strict but nevertheless acceptable criteria (such as matching on only one component of the date of birth but matching on names or addresses) would be classified as acceptable matches.

Appendix E contains details of this classification system.

As Table 4.1 shows, in total, 86% of the additional death records were able to be linked to the MED. The linkage rates differ between data sets. The best linkage rate was obtained for the RAC data set, where names were included in the linkage. The NPDC had the lowest linkage rates. This may indicate that some neonatal deaths were not registered and thus not able to be linked to the MED containing death registration data. It may also reflect the lack of identifying details for neonatal deaths (for example, names).

Table 4.1: Quality of matches for each data set

	NHMD	RAC	NPDC	All data sets
Original number of records	5,610	990	290	6,890
Not matched to MED	770	84	115	969
All matches to MED	4,840	906	175	5,921
% matched	86	92	60	86
Excellent & good matches	4,024	822	142	4,988
Excellent or good matches as a % of all matches	83	91	81	84

Source: Enhanced Mortality Database.

In total, 84% of all matches were considered excellent or good quality matches, and 16% were considered acceptable matches. Because the proportion of acceptable matches is relatively small, all subsequent analyses of results in this report refer to all matches.

Table 4.2 below shows broad results of the linkage process for each data set by state and territory. As expected, the contribution from the NHMD is the largest, accounting for 75% of all additional records. The RAC data set came next, contributing 24%.

Table 4.2: Additional Indigenous deaths identified through linkage, by state and territory

State/territory	In MED	NHMD	RAC	NPDC	All added records	Added records as percentage of NDI/NMD records
NSW	2,464	303	102	4	409	17
Vic	368	121	31	3	155	42
Qld	2,780	198	48	6	252	9
WA	1,885	93	37	2	132	7
SA	700	35	20	0	55	8
Tas.	112	45	11	0	56	50
ACT	65	0	4	0	4	6
NT	2,200	12	5	1	18	1
Australia	10,574	807	258	16	1,081	10

Source: Enhanced Mortality Database.

The percentage of additional Indigenous deaths identified for the Northern Territory is relatively low, probably reflecting the traditionally better Indigenous death registration in that jurisdiction. For Tasmania, 56 records were added on to a small base of 112 registered deaths in the MED, representing a 50% increase. A very small number of Indigenous deaths are involved in the ACT; 4 deaths are added to a base of 65.

There are limitations to the results shown in Table 4.2, which may be related to the use of the ever-Indigenous approach in deriving the Indigenous status of records with missing and non-Indigenous status in the MED. The limitations relate to the fact that if some records in any of the additional data sets are wrongly classified as Indigenous, then use of the ever-Indigenous approach could overstate the number of records deemed to have been added to the MED through data linkage. The extent to which this has been the case is being evaluated in the second phase of the Enhanced Mortality Database project now under way. Table 4.3 shows further details of the extra Indigenous deaths identified through data linkage.

Table 4.3: Details of Indigenous deaths from additional sources

Linkage result	Records
Indigenous in none	649,259
Indigenous in one source only	7,287
Indigenous in MED	
<i>Indigenous in MED only</i>	6,264
<i>Indigenous in MED and one other source</i>	
Indigenous in MED & Hospital	3,547
Indigenous in MED & Residential Aged Care	420
Indigenous in MED & Perinatal	31
<i>Total MED and one other source</i>	3,998
<i>Indigenous in NMD and two other sources</i>	
Indigenous in MED Hospital & Residential Aged Care	213
Indigenous in MED Hospital & Perinatal	99
<i>Total MED and two other sources</i>	307
<i>Indigenous in MED and three other sources</i>	5
Total Indigenous in MED	10,574

(continued)

Table 4.3 (continued): Details of Indigenous deaths from additional source

Linkage result	
Extra Indigenous Deaths	
<i>Extra — one source only</i>	
Extra — Hospital only	749
Extra — Residential Aged Care only	258
Extra — Perinatal only	16
<i>Total one source only</i>	1,023
<i>Extra — two sources</i>	
Extra — Hospital & Perinatal	16
Extra — Hospital & Residential Aged Care	41
<i>Total two sources</i>	57
Total extra	1,081
Total Indigenous in any source	11,655
Total in NMD	10,574

Source: Enhanced Mortality Database.

Of the nearly 661,000 deaths in the MED in the period, 10,574 were originally identified as Indigenous, and of these, 6,264 were not Indigenous in any of the other linked sources. A further 3,488 were Indigenous in both the MED and the NHMD. No deaths were identified as Indigenous in all four sources, which was expected, since the RAC and NPDC cover entirely different age groups. Only five deaths were identified as Indigenous in all possible sources (all except either RAC or NPDC).

Of the deaths not identified as Indigenous in the MED, 1,081 were Indigenous according to at least one of the other sources. Of these, 1,023 were identified in only one source (749 in the NHMD only, and 258 in the RAC collection). Only 57 deaths were identified in two sources, and only one was identified in three sources.

Table 4.4 shows the number of additional deaths identified by broad age groups and sex. Relatively few deaths occur at ages 5-44 years, and the number of added deaths in these age groups is also small, both in absolute numbers and in relative terms. High percentage additions occurred in the very young ages (0-4) and the older ages (65+). This may reflect, in part, the availability of the NPDC and the RAC data set for linkage, but this may also reflect poor identification of Indigenous deaths at these ages.

The poor registration of perinatal deaths may be due to different legislative requirements in different jurisdictions. In some jurisdictions (for example, Western Australia) legislation requires the registration of both still births and deaths resulting from live births. Other jurisdictions only have the legislative requirement to register deaths resulting from live births. This could mean some deaths that occur during or soon after confinement are not registered. On the other hand, where these deaths occur in hospital, then they may be registered in hospital and midwives or perinatal data collections. This could lead to discrepancies between the death registrations data sets and perinatal data sets when these two data sets are compared or linked.

Table 4.4: Number of records added through linkage by age group and sex

Age	Males				Females			Persons	
	No. added	NMD males (per cent)	Total NMD Males	No. added	NMD females (per cent)	Total NMD Females	No. added	NMD persons (per cent)	Total NMD persons
0–4	41	10	430	48	16	301	89	12	731
5–19	13	6	233	3	2	138	16	4	371
20–44	49	3	1,771	43	5	934	92	3	2,705
45–64	117	6	1,980	109	8	1,405	226	7	3,385
65+	312	19	1,654	346	20	1,728	658	19	3,382
All ages	532	9	6,068	549	12	4,506	1,081	10	10,574

Source: Enhanced Mortality Database.

While the results of the analysis shown in Tables 4.3 and 4.4 pertain to Australia as a whole, a similar analysis has also been carried out for the jurisdictions, and the results are available on request.

Life table analysis

The life table is essentially a summary measure of age specific mortality rates. Because mortality differs considerably between the two sexes, the life table is usually calculated for males and females separately. The quality of the various life table estimates is directly affected by the quality of the input death and population data. For this project, life tables for the intercensal period 2001–2006 were compiled.

It is known that the census count of the Indigenous population suffers from net undercount. The ABS has estimated the levels of net undercount in the censuses and has used these to produce adjusted estimates of the Indigenous population. The net undercount of the Indigenous population has been estimated by the ABS at 6.1% and 11.5% of the Indigenous population, respectively, in the 2001 and 2006 population censuses (ABS 2007, 2008c).

It is also known that the identification of Indigenous status in the population census was not stable and changed from one census to another, resulting in unexplained growth that could not be explained by births and deaths. However, for the period 2001 to 2006, the amount of unexplained growth was much smaller than in past intercensal periods, indicating some stabilisation of identification (ABS 2008a). Nevertheless, any inaccuracy in the estimates of Indigenous estimated resident population (ERP), both in total numbers and in the age and sex of the population, will have a direct effect on life table values.

The death data used to prepare these life tables come from the EMD that was created by enhancing the Indigenous status information in the linked data set and identifying the extra Indigenous deaths from the additional data sources. While the data sets used for the linkage cover a high proportion of all deaths, they are nevertheless not exhaustive. More deaths could be identified if additional data sets were available for linkage. For example, deaths of adults are only partially covered by the NHMD and RAC data sets; they are not covered by the NPDC. How many Indigenous deaths are still missing from the EMD is not known, but it is certain that its coverage of Indigenous deaths is not 100%. This effectively means that the life table estimates presented in this report represent the lower limits of Indigenous mortality and the upper limits of Indigenous life expectancy.

As discussed in the previous section, the adoption of an ever-identified approach may result in false-positive Indigenous identification. Given that the linkage is to death records and not to events that can happen more than once, it is believed that this problem is not large. No adjustments have been made to account for this.

The previous section also noted that 14% of death records identified as Indigenous in the alternative data sets were unable to be linked to the MED. These records have poor identifying information and are not concentrated in particular age groups or in any particular sex category. To take the unlinked records into account, a separate set of life expectancy estimates was made, by assuming that the unlinked records have the same propensity as the linked ones for being Indigenous records and being missed; they could therefore be added to the enhanced database.

The estimated number of extra records added by the inclusion of unlinked records is 193, increasing the total number of added records to 1,274, or 12.1%, of MED records. Life expectancy estimates using linked records only, and using both linked and unlinked records, are presented in Table 4.5.

The direct estimation method is used in the compilation of single year age life tables to age 85+ (Chiang 1984). The compilation of single year age life tables for the period 2001–2006 requires single year age death data for the period as well as single year age population data for 30 June 2001 and 30 June 2006. Single year age Indigenous death data for the period are in the enhanced linked death data set. Up-to-date Indigenous population data (estimated resident population or ERP) by single year of age for 30 June 2006 were based on the 2006 Population Census and provided by the ABS.

The ABS has revised Indigenous ERPs for years before 2006, back to 1986. These revised estimates were compiled using a back-cast method that did not rely on the results of previous population censuses (ABS 2009b). But these revised estimates are only available by 5-year age groups. For this project, the required single year age ERP for 30 June 2001 was calculated by disaggregating the back cast ABS 5-year age ERPs into single year of age. The disaggregation was based on the single year age distributions of Indigenous ERPs for 30 June 2001 that were compiled by the ABS before the 2006 Population Census. This set of Indigenous ERPs was estimated based on the 2001 Population Census.

Infant mortality is normally calculated by relating infant deaths to births of the same year. However, because of late registrations as well as under-registration of Indigenous births, infant mortality rates for this project are calculated by relating deaths to the population rather than to births. In addition, and as detailed earlier, there are cases where Indigenous fathers are not identified on the birth registration form. In such cases, only mothers' details are recorded on the form, leading to under-identification of Indigenous births.

To overcome irregularities in the death rates caused by the small numbers of deaths in many age categories, the death rates were smoothed. This was achieved by using partially monotonic penalized regression splines on the $\log m_x$ values in the life tables (Hyndman & Ullah 2007). Monotonicity was assumed for death rates above 30 years of age.

Confidence intervals for life expectancy estimates were calculated using a bootstrap approach (Hyndman & Ullah 2007). The population estimate for each age level was assumed to be normally distributed with standard deviation equal to 2.95%, based on the 2006 Census Post-Enumeration Survey (ABS 2007). The number of deaths for each age level was simulated using a Poisson distribution. In this way, a simulated life table could be constructed. The procedure was repeated 2,000 times to give 2,000 replicates of each life expectancy estimate. Confidence intervals were obtained using the 2.5% and 97.5% quantiles of the simulated life expectancies.

Table 4.5: Indigenous life expectancy at birth, at age 20, 45 and 65, based on enhanced death data, Australia 2001–2006^(a)

Age	Linked records only		Including unlinked records	
	Males	Females	Males	Females
Birth	66.8 [66.5– 67.2]	72.9 [72.5–73.3]	66.6 [66.3– 67.0]	72.7 [72.4– 73.1]
20 years	48.3 [47.9– 48.6]	54.0 [53.7–54.4]	48.1 [47.7–48.4]	53.9 [53.5–54.2]
45 years	27.2 [26.9– 27.6]	31.3 [31.0–31.7]	27.1 [26.8–27.5]	31.2[30.8– 31.6]
65 years	13.4 [13.1– 13.8]	16.1 [15.8–16.5]	13.3 [13.0–13.7]	16.0 [15.7– 16.4]

(a) 95% confidence intervals are in square brackets.

Source: Enhanced Mortality Database.

State and territory estimates of life expectancy at birth for 2001–2006 are in Table 4.6 for New South Wales, Victoria, Queensland, Western Australia, South Australia, and the Northern Territory. The numbers of Indigenous deaths in Tasmania and the Australian Capital Territory are considered too small for the construction of reliable life tables.

Table 4.6: Indigenous life expectancy at birth, based on enhanced death data, including unlinked records, selected States and the NT, 2001–2006^(a)

	Males	Females
NSW	69.4 [68.8–70.0]	74.7 [74.1–75.4]
Vic	69.2 [68.0–70.7]	77.2 [75.8–78.7]
Qld	66.5 [65.8–67.2]	72.6 [72.0–73.4]
WA	63.4 [62.5–64.3]	68.2 [67.4–69.1]
SA	62.7 [61.2–64.1]	68.7 [67.2–70.1]
NT	58.2 [57.3–59.1]	66.1 [65.1–67.1]
Australia	66.6 [66.3–67.0]	72.7 [72.4–73.1]

(a) 95% confidence intervals are in square brackets.

Source: Enhanced Mortality Database.

There are large differences in estimated life expectancy at birth between states and the territories. Life expectancy at birth for the New South Wales Indigenous population is highest for males – estimated at 69.4 years, 11 years longer than those for Northern Territory Indigenous males. Life expectancy at birth for Indigenous females was highest in Victoria, at 77.2 years (11 years longer than Northern Territory Indigenous females). The difference in life expectancy between males and females also varies from 4.8 years for Western Australia to 8 years for Victoria.

Table 4.6 also shows the confidence intervals associated with the life expectancy estimates. There are variations between jurisdictions in the width of the confidence intervals. On the whole, the confidence intervals are relatively small for the bigger jurisdictions with large Indigenous populations and reported deaths, and relatively large for the smaller jurisdictions with much smaller Indigenous populations and reported deaths.

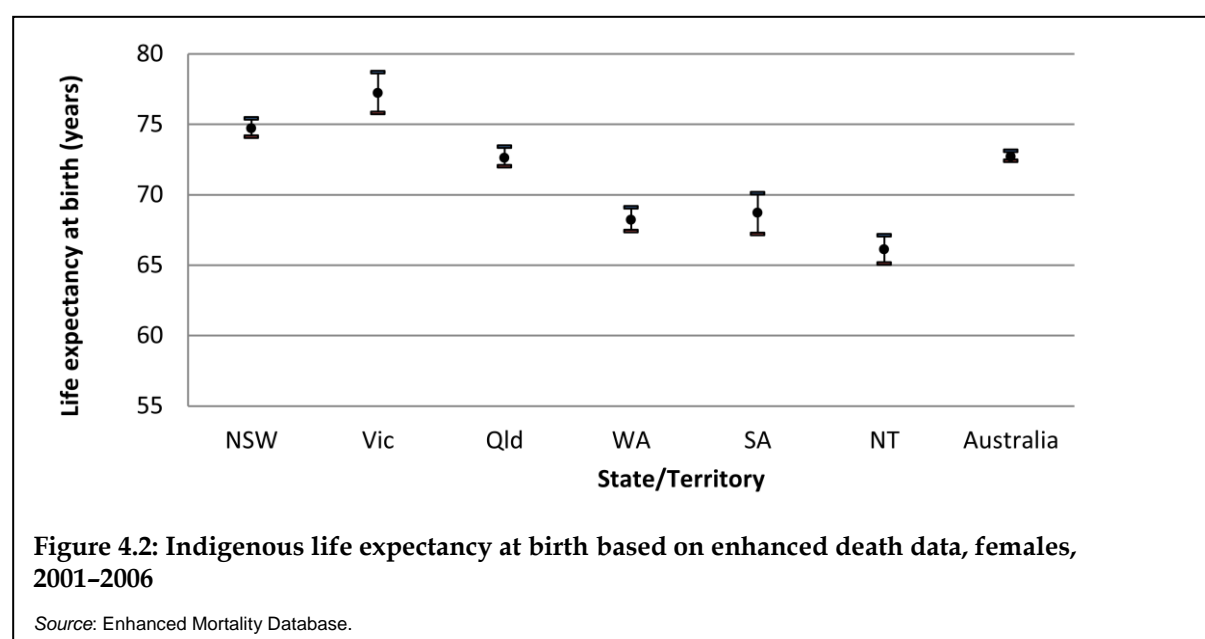
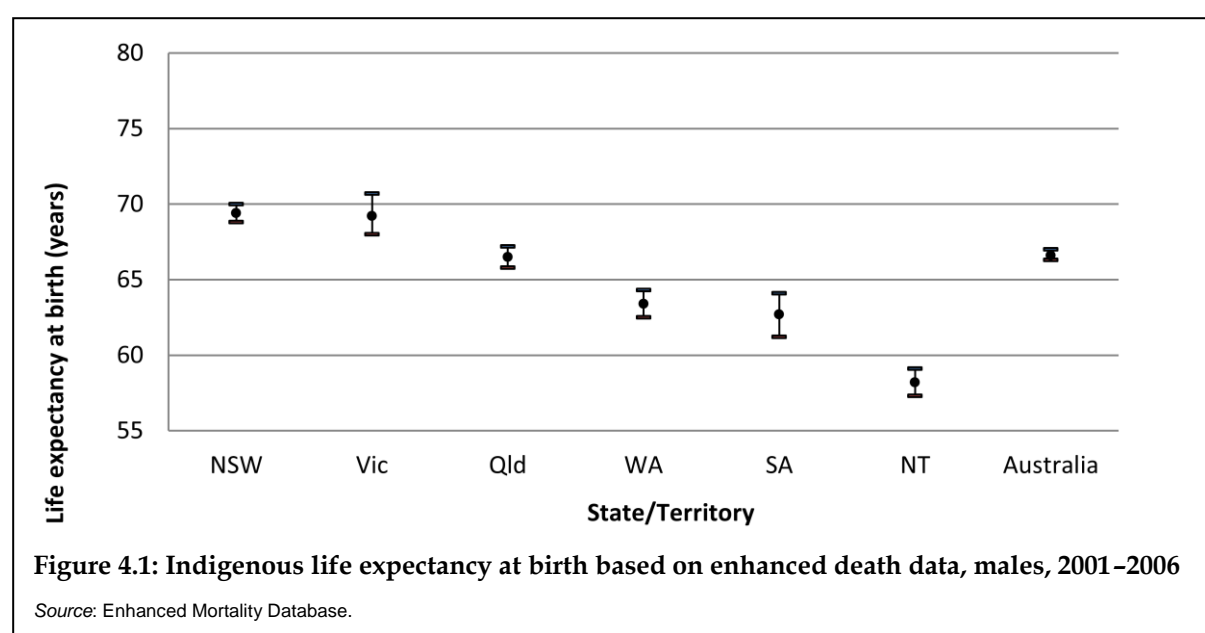
For example, the confidence intervals for males range between 1.2 years for New South Wales and 2.9 years for South Australia, with Victoria also having a relatively large

confidence interval of 2.7 years, on account of the much smaller reported Indigenous deaths in Victoria and South Australia, compared to New South Wales.

Similarly for females, the confidence intervals range from 1.3 years for New South Wales to 2.9 years for Victoria and South Australia. The confidence intervals are also small (1.4 years) for Queensland on account of its relatively large Indigenous population.

The confidence intervals are a pointer to the degree of uncertainty around the life expectancy estimates, and illustrate the relationship between population size and the degree of reliability of estimates. As a further illustration of this relationship, it could be seen that the confidence intervals for Australia, as a whole, are much smaller than for individual jurisdictions.

For clarity and easy reference, the life expectancy estimates and their respective confidence intervals are in Figures 4.1 and 4.2.



5 Discussion

The results compared

The life expectancy estimates presented above are the upper limits, as the data sets used for linkage are not exhaustive. They are optimistic estimates and are most likely to represent life expectancies that are higher than reality. If more data sets containing Indigenous deaths and Indigenous identification become accessible to be linked, the estimates of mortality could become higher and estimates of life expectancy could be lower.

In addition, two other data deficiencies affect the estimates. The first is that the data sets used in linkage do not have perfect Indigenous identification, and some Indigenous deaths would have been missed by all of the sources.

The data sources used also do not cover some categories of deaths. Deaths at home are not covered by the NHMD or RAC data. Deaths of babies with Indigenous fathers and non-Indigenous mothers are not covered by the National Perinatal Data Collection.

However, the more data sources used reduces the chance of Indigenous deaths being missed by all sources. As further data sources are added, the return (i.e. identifying more Indigenous death records) from these added data sets for linkage will diminish.

Table 5.1 below shows that the use of linked data has reduced the estimated Indigenous life expectancy at birth by 1.7 years (males) and 2.0 years (females). These reductions have come from the addition of 12% deaths (including records that could not be linked) to the EMD.

Indigenous life expectancy at birth in the period 2001–06 is estimated to be 66.6 years for males and 72.7 years for females, based on linkage with all data sets and with unlinked records included. The difference between Indigenous males and females is about 6.1 years, more than a year larger than the 4.9 years between males and females for all Australians. The estimated gap in life expectancy at birth between Indigenous Australians and all Australians is about 11.3 years for males and 10.2 years for females.

The ABS has estimated Indigenous life expectancy using the results of the CDE Mortality Quality Study. These results are in Tables 5.1 and 5.2. Even though the ABS and AIHW estimates are not based on the same methods, and use different data, there are similarities between the two sets of estimates, at least at the national level.

Table 5.1: Comparison of various life expectancy at birth estimates

Description	Males	Females
AIHW Indigenous 2001–06 unlinked records included	66.6	72.7
AIHW Indigenous 2001–06 unlinked records excluded	66.8	72.9
AIHW Indigenous 2001–06 NMD data only	68.3	74.7
2005–07 ABS CDE linkage adjusted	67.2	72.9
2001–06 ABS All Australians (averaged 2001 to 2006) ^(a)	77.9	82.9
2005–07 ABS All Australians (averaged 2005 to 2007) ^(a)	78.7	83.5

(a) ABS 2009a.

Source: Enhanced Mortality Database.

The Mortality Quality Study essentially linked the 2006 Population Census Indigenous records to deaths registered after the Census, to ascertain the quality of the Indigenous information on the registered death records. Based on this work and on the 2006 Census post-enumeration survey, adjustment factors were derived to correct death registration records for under identification of Indigenous deaths (ABS 2008c, ABS 2008a, ABS 2009c). Estimates of life expectancy at birth derived from the CDE project are 67.2 for males and 72.9 for females. These are close to the AIHW estimates given the normal statistical error margins and the difference in the data sources used.

The ABS has compiled life tables for New South Wales, Queensland, Western Australia and the Northern Territory. All ABS estimates of life expectancy at birth are higher than AIHW estimates, with the difference for the Northern Territory being the largest, at more than 3 years (Table 5.2). But the pattern of the state differences and those between males and females are similar.

Table 5.2: Estimates of life expectancy at birth by selected states and the NT

	Males				Females			
	NSW	QLD	WA	NT	NSW	QLD	WA	NT
AIHW 2001-2006	69.4	66.5	63.4	58.2	74.7	72.6	68.2	66.1
ABS 2005-2007	69.9	68.3	65.0	61.5	75.0	73.6	70.4	69.2
Difference	0.5	1.8	1.6	3.3	0.3	1.0	2.2	3.1

Source: Enhanced Mortality Database.

However, as stated earlier, the AIHW estimates are the upper limits. The real life expectancy at birth estimates would be lower if more, better data sets were available to be linked. As the AIHW approach is based on yearly deaths data, it can yield yearly life expectancy estimates, whereas the ABS approach (based on deaths data linked to the Census) can only provide life expectancy estimates every five years after each Census.

Changes over time

The EMD also allows the calculation of life tables for each calendar year for the years 2001 to 2005. Life tables for 2006 are not presented here. The database was based on deaths registered up to the end of December 2006, and a proportion of the deaths which occurred in late 2006 would not have been registered by this date. Life tables for 2006 based on this database will underestimate the mortality rates.

Indigenous life expectancies at birth for New South Wales, Queensland, Western Australia, the Northern Territory and Australia are in Table 5.3 below.

Table 5.3: Indigenous life expectancy at birth for NSW, Qld, WA, NT and Australia, 2001-2005

	2001	2002	2003	2004	2005
Males					
NSW	69.3	68.4	68.7	70.3	70.2
Qld	65.7	66.0	66.4	65.9	68.5
WA	63.0	62.8	63.9	63.2	63.9
NT	58.2	57.4	58.4	58.2	58.6
Australia	66.4	66.2	66.4	66.9	67.2

(continued)

Table 5.3 (continued): Indigenous life expectancy at birth for NSW, Qld, WA, NT and Australia, 2001-2005

	2001	2002	2003	2004	2005
Females					
NSW	74.2	73.5	75.3	75.8	74.8
Qld	72.2	71.8	72.5	72.1	74.7
WA	68.2	66.8	69.4	69.1	67.6
NT	66.3	64.4	67.3	66.9	65.6
Australia	72.7	71.5	73.2	72.8	73.5

Source: Enhanced Mortality Database.

The changes in life expectancy estimates over the 5 years 2001 to 2005 are not smooth. Improvements in some years are often not followed by another improvement in the next year. Movements for some years are much larger than other years. This is true even at the national level, where the number of deaths involved is much larger. As expected, the random error margins measured by the 95% confidence interval for calendar year estimates are higher than those for the period 2001-2006. Calendar year confidence limits at the 95% level range from about 1.3 years for New South Wales to 2 or more years for the Northern Territory. Details of these confidence limits are in Appendix F.

However, comparing 2001 life expectancies with those for 2005, a general increase is observed in male life expectancy at birth between 2001 and 2005 for all states and the Northern Territory (Table 5.4). They range from 2.8 years for Queensland males to 0.4 years for the Northern Territory. For females, Western Australia and the Northern Territory estimates have shown a decline.

Table 5.4: Changes in life expectancy at birth for Indigenous Australians and all Australians, 2001-2005

	NSW	Qld	WA	NT	Australia
Males					
Indigenous Australians	0.9	2.8	0.9	0.4	0.8
All Australians ^(a)	1.6	1.4	1.6	1.7	1.5
Females					
Indigenous Australians	0.6	2.5	-0.6	-0.7	0.8
All Australians ^(a)	0.9	0.9	1.9	1.7	0.9

(a) Derived from ABS 2009a. Deaths Australia 2008 (data cube).

Source: Enhanced Mortality Database.

With the exception of Queensland, the improvement in life expectancy at birth between 2001 and 2005 is smaller for Indigenous Australians than for all Australians. For Australia as a whole, the improvement in male life expectancy was 0.8 years for Indigenous Australians compared with 1.5 years for all Australians. The improvement in female life expectancy at birth was 0.8 years for Indigenous Australians and 0.9 years for all Australians.

Given the uncertainty and the random statistical errors associated with the Indigenous life expectancy estimates, it is premature to conclude that the male life expectancy gap between

Indigenous Australians and all Australians is widening. A longer time series of estimates is needed to overcome the difficulties in interpreting the statistical outliers in particular years and for particular states, such as the relatively high 2005 life expectancy estimates for Queensland. Estimates for further years will give more data points for an adequate time series analysis to be undertaken.

Nevertheless, these limited estimates for 2001 to 2005 give no indication that the life expectancy gap between Indigenous and non-Indigenous Australians is closing.

Quality of identifying information for record matching

An aspect of this study was how often it proved impossible to obtain matches, due to inadequate identifying information in some data sets for matching records. This is unfortunate because the general quality of the data sets is good and improving further. For example, the proportions of records with unstated Indigenous status have declined, and missing age and sex data are rare (see Appendix A). While we do not have details of all the quality control processes applied to the data sets, it is clear from the results that many records still contain inadequate information for data matching. There were considerable numbers of imputed 'date of birth' fields in all data sets, and there was a lack of precision in many names, where included, in the NDI and RAC data sets (see Appendix D). In part, this may reflect the fact that a certain level of error is unavoidable and to be accepted, in particular in data fields that are not used for the production of statistics. However, with the rapid increase in the use of health databases for linkage studies, this quality issue clearly needs to be addressed.

This project has not been able to utilise the Indigenous identity information that may exist in the NDI. The reasons are two-fold. Firstly, there are differences in the way states and territories are providing Indigenous identifier data to the AIHW. Not all states and territories are providing MCCOD Indigenous identifier data. Some provide both Indigenous identifiers from the Death Registration Statement and the MCCOD to the AIHW, while others provide only the identifier from the Death Registration Statement. Secondly, only a part of the content of the NDI is currently used by the AIHW for data linkage purposes, and this usage does not include Indigenous identifier information. Perhaps because of this, the quality and integrity of any such information that exist in the NDI have not been assessed and maintained.

Conclusions

This study has shown that improved Indigenous death data can be obtained through data linkage, and that an Enhanced Mortality Database can be constructed. Life tables generated from this study, while admittedly overestimating life expectancies are credible and plausible.

Comparison of life expectancy estimates from this study with those derived from the ABS Mortality Quality Study shows that the two independent estimates are comparable, at least at the national level. The results of the two studies support and give confidence to the parallel attempts by both the ABS and AIHW to improve Indigenous life expectancy estimates through data linkage.

There are, however, differences between the ABS and AIHW estimates at the state level, which may be related to data quality specificities of the various Commonwealth and state-level data sets used in the two studies.

Both the AIHW and ABS studies are at their infancy; the ABS has termed its life expectancy estimates 'experimental', while the AIHW has described its own study as a 'feasibility study'. Both agencies have shown considerable interest in the results of the other agency, and the ABS has published the results of the AIHW study in their annual deaths publication (ABS 2009c) for comparative purposes.

There is still considerable work to be done before the AIHW can be confident in the results of such linkage work. An important and interesting aspect of this feasibility study, however, is that the study relies on death registration, hospital, aged care and perinatal data. These data sets are available on a yearly basis and are used to prepare official yearly mortality, hospital and residential aged care separation rates, as well as yearly indicators of perinatal outcomes. It is feasible to apply the method developed in the feasibility study to these yearly data sets to prepare yearly Indigenous life expectancy estimates.

The key purpose of this feasibility study is to provide a permanent database containing enhanced deaths data that can be used for time series research on a yearly basis. Though there are at present issues related to the quality of data for this study, as more data sets of improved quality become available, the linkage will improve and the EMD will become a very valuable national Indigenous mortality researchable database.

Appendix A: The data sources

Registered deaths data set

Coverage

The ABS compiles and publishes annual national and sub-national deaths statistics, classified by various sociodemographic variables, including Indigenous identification. This is ascertained through the question on the Death Registration Statement: 'was the deceased (or deceased person) of Aboriginal or Torres Strait Islander origin?'.

The usual practice for the registration of a death is that the funeral director will complete a Death Registration Statement and forwards it to the RBDMs, on behalf of the family of the deceased. The form contains demographic and other details of the deceased person. The Death Registration Statement can also be provided directly by the family of the deceased. The Registrar will then issue the death certificate. The attending physician also completes a medical certificate of cause of deaths and forwards it to the Registrar. For deaths that are subject to coroner investigations, the coroner provides a certification to the Registrar.

It is generally believed that all deaths are reported and registered in Australia. In an early study of the accuracy of Australian census counts, national consistency by age and sex was obtained between census counts adjusted for under-enumeration and populations created by the use of births, deaths and net migration numbers (ABS 1983). However, a study that specifically investigates the completeness of death registration has never been conducted. The number of deaths registered in Australia has gradually increased, reflecting the increase and the ageing of the population.

Table A1: Registered deaths

	Total deaths registered in Australia	Number of Indigenous deaths registered	Indigenous status not stated	
			Number	Per cent
2001	128,544	2,072	5,706	4.4
2002	133,707	2,136	4,931	3.7
2003	132,292	2,079	3,739	2.8
2004	132,508	2,136	1,798	1.4
2005	130,714	2,141	1,527	1.2
2006	133,739	2,279	1,112	0.8
2007	137,854	2,421	1,421	1.0

Source: (ABS 2009a); Deaths, Australia, various years.

In recent years, the number of registered Indigenous deaths has been increasing and the number of deaths for which Indigenous status was not stated has declined. While these indicate some improvement in the coverage of registered Indigenous deaths, a large number of records are still missing information on Indigenous identification. There is also evidence that a considerable number of Indigenous deaths are still being wrongly registered as non-Indigenous. The extent of these errors in identification is the subject of this project, which aims to enhance the death registration data by the use of other data sets containing Indigenous deaths.

While there can be a delay in death registration, the overwhelming majority of deaths are registered within three months. For delayed registrations, the timing of registration is affected by the possible need for a coronial investigation, as is required for all deaths caused by external causes. The average time taken for registration to occur is about 1.1 months (ABS 1998), but is much shorter in some jurisdictions, such as Western Australia where most deaths are registered within 14 days (Brett Burns: 2011, personal communication).

ABS data showed that of deaths registered in 2001, 4.6% occurred the year before (presumably in the latter months of the previous year), and only 0.08% occurred two or more years before (ABS 2002). This pattern has not changed much. Of deaths registered in 2007, 4.8% occurred the year before and 0.17% occurred two or more years before (ABS 2008b).

Late registration is thought to be more prevalent for Indigenous deaths. Many Indigenous deaths occur in remote communities where registration can take longer. Compared with non-Indigenous Australians, a higher proportion of Indigenous deaths are caused by injuries and other external causes, requiring coronial investigations.

While the problem of late registration is not large, the data used in this project were extended to cover the period to December 2006, so that the vast majority of late registrations of deaths that occurred in the first half of 2006 would be included.

Data for matching

The variables used for matching the MED records with other records are sex, date of birth, date of death and various fields based on usual address. Names were also used to match with residential aged care records.

Data on sex of the deceased had few unknown or missing values, and were also generally consistent with the name of the deceased. However, data on date of birth are not of the same quality. While almost all death registration records have a date of birth or age (in 2007, only 11 records did not have a date of birth), for a considerable number of records the recorded date of birth seem to have been imputed or estimated. For Indigenous deaths records, up to 5% appear to have dummy birth dates such as 1 January or 1 July. This estimate of dummy records is based on a comparison between the dates of births stated in the MED and those on alternative data sources.

Residential aged care data

Coverage

The source of residential aged care data is the System for the Payment of Aged Residential Care, administered by the Australian Department of Health and Ageing. SPARC contains information gathered through a number of instruments (AIHW 2008c). Among these instruments, the one that relates to this project is the Aged Care Client Record (AACR) (previously Aged Care Application and Approval form). The AACR is filled in at the time of assessment for Australian Government subsidised aged care by persons applying for admission (or their carer), as well as by the Aged Care Assessment Team. Data on Indigenous identification, date of birth, sex and usual place of residence at the time of assessment are recorded on this form. Residential aged care is subsidised by the Australian Government and payment of subsidies to residential care homes are, in the main, based on the number of residents and the level of care the residents receive. SPARC is a payment system that is updated regularly. It is reasonable to assume that this system covers all residents in all residential aged care.

Separation from aged care homes through death is recorded in the SPARC system. Terminally ill residents transferred to acute care facilities before death are not counted as separations through death. However, for purposes of this study, residents on hospital leave (not separated) who subsequently did not return to the aged care home are assumed to have died, and the records are amended to record separations through deaths. These numbers are, however, very small. In any case, any double counting is picked up during linkage and eliminated.

In the period 2000–01 to 2006–07, there has been a gradual increase in the number of deaths in aged care homes, reaching 45,200 in 2006–07. These account for slightly over 30% of all deaths in Australia (32% in 2006–07).

The number of Indigenous deaths in aged care homes has also increased in recent years, from 169 in 2000–01 to 222 in 2006–07, accounting for less than 0.5% of all deaths in aged care homes.

Table A2: Deaths in residential aged care

Period	Total deaths in residential aged care	Indigenous deaths in residential care	Indigenous status not stated
2000–01	37,618	169	3,990
2001–02	37,978	165	3,602
2002–03	40,283	179	3,953
2003–04	41,722	170	3,489
2004–05	41,624	176	2,536
2005–06	44,235	185	2,013
2006–07	45,174	222	1,447

Source: AIHW special tabulations, RAC data set, various years.

The proportion of deaths without a stated Indigenous status was nearly 11 % in 2000–01, but this has gradually declined to just over 3% in 2006–07. The number of records with Indigenous status not stated is still very high, although it indicates some improvement in the Indigenous data in the RAC data collection.

Data for matching

There are no missing data on age and sex in the RAC data set. However, there are many cases where the date of birth is evidently incorrect, for example when the implied age is less than 10 (AIHW 2008c). There are also records that seem to have imputed date and month of births. For Indigenous deaths records, this is large. The linkage process has revealed that up to 26% have dates of birth that are likely, although not necessarily, to be dummy dates, such as 1 January or 30 June. These estimates of imputation come from comparisons between aged care records and death registration records. This presents a problem for matching with death registration records. Fortunately names are included on residential aged care records and on death registration records for data linkage.

The extent of missing or imputed data on the address of usual place of residence is much lower. For all residents, including both Indigenous and non-Indigenous, very few records do not have an address of usual residence — about 1% in June 2001, falling to 0.3% in June 2006 and 0.2% in June 2008. However, the address information is of variable quality. For some, full address information is included, but for others only local government names or postcodes are included. When address information is used for data matching, inspections of the data are done clerically.

Hospital data

Coverage

Hospital data are compiled and published by the Australian Institute of Health and Welfare (AIHW) from data provided by state and territory health authorities. Hospital records refer to hospital separations and not hospital patients.

Deaths in hospitals during the period 2001-2005 numbered about 70,000 each year, accounting for more than 53% of all registered deaths (about 130,000 a year). There seems to be a small but gradual decrease in the proportion of deaths in hospitals, but this has fluctuated. In 2006-07, 53.7% of all deaths were in hospitals (AIHW 2008b, 2011).

The number of Indigenous deaths in hospitals has increased slightly in recent years, from about 1,000 a year in the early-2000s to more than 1,100 in the mid-2000s. This may indicate an improvement in Indigenous identification. The number of Indigenous deaths in hospitals is large, almost half of the number registered by state and territory registrars of births, deaths and marriages. However, there are a large number of hospital death records with missing Indigenous status. These Indigenous not-stated cases outnumber the Indigenous stated cases, although it may be assumed that most of these are non-Indigenous deaths.

Table A3: Deaths in hospitals

	All hospital deaths	Indigenous	Indigenous status not stated	Hospital deaths as % of all registered deaths
2000-01	69,161	971	1,416	54.0
2001-02	70,671	1,013	1,638	54.3
2002-03	71,573	980	1,902	54.1
2003-04	71,932	1,044	1,817	54.0
2004-05	70,799	1,023	2,119	53.9
2005-06	71,122	1,140	1,473	53.1
2006-07	72,440	1,182	1,791	53.7

Source: AIHW Australian Hospital Statistics, various years.

An audit of the quality of Indigenous identification in public hospital separations data was conducted during 2007-2008 by the AIHW with the cooperation of state and territory health authorities (AIHW 2010a). This audit concluded that public hospitals data on Indigenous identification in New South Wales, Victoria, Queensland, South Australia, Western Australia, and the Northern Territory are now of sufficient quality for publication. About 93% of Indigenous persons in the audit had been correctly identified in the hospitals data. After adjusting for sample bias, it was estimated that about 89% of all Indigenous patients were correctly identified.

In addition, significant progressive improvements were identified in Indigenous identification in New South Wales and Victoria, and the data for these two states are of acceptable quality for publication from 2004-05 onwards.

Data for matching

There are very few missing age and sex data in the hospitals separation collection. The numbers in Table A4 are very small compared with the millions of separations each year. In 2006-07, there were 7.6 million separations from all hospitals. Nevertheless, it is surprising to find cases where the sex of the patient is missing, even if the numbers are small.

Table A4: Number of records with age and sex not stated in hospital separations

	Date of birth/age not stated	Sex not stated
2000–01	3	72
2001–02	76	158
2002–03	143	244
2003–04	2	77
2004–05	11	82
2005–06	52	90
2006–07	23	474

Source: AIHW Australian Hospital Statistics, various years.

However, existing age data are not all of acceptable quality. During the matching process, about 11% of Indigenous hospital records were identified as having birth dates that are likely to be dummy dates when compared with data from death registration records. This is surprising as date of birth, with name and sex, are important information for correct patient identification and health care.

Geographic information such as local statistical area or postcodes is used for data matching through clerical review. AIHW assigns a Statistical Local Area (SLA) code to hospital records using the address data provided, which may be full address, the name of a local council, or a postcode, or a combination of various address information. The AIHW is able to assign SLA codes consistently over the period 2000–01 to 2006–07 to over 99.5% of all separation records by use of the address information provided (AIHW 2008b).

The National Perinatal Data Collection

Coverage

The National Perinatal Data Collection (NPDC) is a national data collection of pregnancy and childbirth, managed by the AIHW National Perinatal Epidemiology Statistics Unit (NPESU) located at the University of New South Wales. The source data are from state and territory health departments. Midwives and other staff complete perinatal notification forms for each birth, using information obtained from mothers, hospital admission and inpatient records and antenatal pregnancy records etc. In some states data from the RBDM were also used. Data from the various sources are collated and edited by the health departments of the states and territories to form the NPDC and are sent to the AIHW NPESU for further quality control, analysis and publication. Information included in the NPDC is on both live births and stillbirths of at least 400 grams birthweight or at least 20 weeks gestation. From this data collection, only deaths within 28 days of live birth were selected for use in this project.

The standard question to ascertain Indigenous identification of the mother is not uniformly used in all the data sources for the NPDC. While hospitals use the standard question, other data sources such as antenatal pregnancy records do not.

An assessment of Indigenous status data quality in the NPDC was conducted in 2007 by the AIHW. It used questionnaires to determine how many hospitals obtained information on the Indigenous status of mothers, and if independently, and whether the information is validated or checked by midwives. This assessment showed that 37% of hospitals always obtained the information from hospital databases/admission forms, 39% by midwives independent from hospital records, 16% jointly, and 8% through other sources. It also

showed that 66% of hospitals always have the Indigenous information checked or validated by midwives (AIHW: Leeds et al. 2007).

Data for matching

The number of records where the Indigenous status of the mother is not stated is small, although there were increases in 2005 and 2006. This is perhaps the result of using multiple sources that supplement each other. In total, for the period 2001–2006, only 0.07% of mothers have Indigenous status not stated.

Table A5: Women who gave birth by Indigenous status, Australia (excluding Tasmania) 2001–2006

Year	Indigenous	Non-Indigenous	Not stated
2001	8,681	235,699	79
2002	8,822	236,194	117
2003	8,857	238,214	78
2004	8,904	238,466	88
2005	9,867	257,798	128
2006	10,183	266,628	625
2001–06	55,314	1,472,999	1,115

Source: AIHW: Australia's mothers and babies, various years.

Likewise, the numbers of mothers whose age or usual residence was not stated were relatively low. For the period 2001–2006, only 336 cases of not-stated age were reported, out of more than 1.5 million mothers. For the period 2004 to 2006 where data are available, 99 mothers did not have their usual residence stated (AIHW 2004a, 2004b, 2005, 2006, 2007, 2008a). These are very small numbers of not-stated cases, indicating that the data contained in the NPDC are adequate for statistical reporting purposes.

Appendix B: Setting up the Mortality Extract Data set

Step 1: Extraction of data from the National Death Index

All death records, including duplicate records, of all deaths between 2001 and 2006 (inclusive) were extracted from the NDI.

Variables extracted: the full names and demographic characteristics of the deceased, including their first name, middle name, surname, sex, date of birth, date of death, state of registration, year of registration, cause of death, SLA of usual residence, Indigenous status, and the unique record identifier, (which is created from the year of registration, state of registration and death registration number).

Step 2: Extraction of data from the AIHW National Mortality Database

A total of 785,086 unique records of all deaths between 2001 and 2006 (inclusive) were extracted from the AIHW NMD.

Variables extracted: sex, date of birth, date of death, place of usual residence, Indigenous status and a unique record identifier, (which is created from the year of registration, state of registration and death registration number). The unique record identifier serves as a key for linking the record with the NDI.

Step 3: Linkage of data extracts from the NMD to data extracts from the NDI to form the MED

Extracts of the two data sets were linked using the unique record identifier and the linkage variables (for example, sex, date of birth, date of death, and components of usual address, such as state and post code) included on the two data sets. Following this linkage, there were two types of NMD records:

NMD records that merged successfully with the NDI using the Mortality ID.

A single person could have duplicate death records on the NDI but not on the NMD, which ABS subjects to extensive quality control. This means that when the NMD was linked to the NDI, the total number of records increased due to duplicate NDI records. Where the duplicate records shared the same information on all the fields in the current study, one of the records was deleted. Where the duplicate records each had different information referring to a particular person (for example, one record had the given name 'Marguerita' and another record had the given name 'Rita') both records were kept, since either version could be linked to another data set.

Furthermore, although an NDI record and an NMD record could be linked using the Mortality ID, in some cases there were discrepancies between the two databases on fields such as sex, date of birth and date of death for a given record. Where there were four or more differences between the two databases, the link was rejected and all NDI information for the particular record was removed from the MED.

NMD records that were not merged with the NDI using the Mortality ID.

For various reasons, a number of records from the NMD could not be linked to the NDI using the Mortality ID. These NMD-only records were then merged with the remaining NDI-only records using a combination of the sex, date of birth, and date of death, usually a unique identifier. The Mortality ID was examined for these additional matches and the match was accepted where it was similar (for example, where there was one-digit difference, where two digits were transposed). However, this method did not work for all NMD-only records and some of these remained unlinked to the NDI.

Step 3: Finalising the MED database

The final MED database that was ready for linkage comprised a total of 798,324 records representing 785,086 unique cases. Of these records:

- 734,444 had been merged perfectly between the NDI and the NMD and had no duplicates
- 41,553 unique NMD records had merged perfectly with the NDI but their record had at least one duplicate. Including duplicates, this resulted in a total of 54,770 records
- 969 unique NMD records had been merged to the NDI using sex, date of birth, date of death and had similar (but non-identical) Mortality IDs. Including their duplicates, this resulted in a total of 990 records
- 7,936 NMD-only records failed to merge to the NDI in any way possible and so had no names or addresses
- 184 records had been merged perfectly between the NDI and NMD but the information on the two databases was so disparate that these links were rejected and the NMD-only records remained without names or addresses.

These different types of records are summarised in the table below (Table B1).

Table B1: Final MED database

Type of record	Source code	Fields available	No. records	No. unique records
Merged and unique	1	All NMD fields, all NDI fields	734,444	734,444
Merged with duplicates	2	All NMD fields, all NDI fields	54,770	41,553
NDI ¹ -only and NMD ² -only merged using sex and dates	3	All NMD fields, all NDI fields, two different Mortality IDs	990	969
Unlinked NMD ² -only	4	All NMD fields only	7,936	7,936
Originally source=1 but fields too disparate so NMD ² record remained unlinked	5	All NMD fields only	145	145
Originally source=2 but fields too disparate so NMD ² record remained unlinked	6	All NMD fields only	39	39
Entire MED			798,324	785,086

1. Records extracted from NDI only.

2. Records extracted from NMD only.

Source: Enhanced Mortality Database.

Appendix C: Developing the ‘pass’ structures and the validation study

The MED was linked to three other data sets, summarised in the table below (Table C1).

Table C1: Summary of data sets linked to the MED

Data set	Type of linkage	Linkage fields	Exclusions	No. unique records
Residential aged care	Probabilistic & clerical review	Surname, given names, birth date, death date, sex, state, SLA, postcode	No deaths from first half of 2001	990 (all Indigenous)
Perinatal	Deterministic & minimal clerical review	Birth date, death date, sex, state, SLA	No stillbirths. No deaths from 2006.	290 (all Indigenous)
Hospitals	Deterministic & minimal clerical review	Birth date, death date, sex, state, postcode	No deaths from first half of 2001 or second half of 2006. No deaths from ACT.	5,246 (all Indigenous)

Source: Enhanced Mortality Database.

Each linkage consisted of a series of passes in which there were a different set of variables used for linkage—called here ‘blocking variables’. These blocking variables differ between data sets linked. The major characteristics of the linkages differed depending on whether names were available.

Linkages where names were available

Blocking variables: The main blocking variables were surnames, given names, sex, components of the date of birth, and components of the date of death. The soundex phonetic algorithm was used to index names based on pronunciation rather than spelling, to enable matches of similar names with different spelling. The soundex phonetic algorithm encrypts names according to how they are pronounced rather than how they are spelled.

Weights: Each pair of records that were compared received a weight that reflected the quality of the match: the higher the weight, the higher the quality. This weight was based largely on the names and two contributing factors. The first was the frequency of the name in the MED (or that portion of it being used in the match). For example, a match of John to John received far less weight (about 6.5) than a match of Zbygniew to itself (about 20) because the former was much more likely to occur by chance. The second factor came into play when the names were not the same. An algorithm was used to determine how close the two names are (see Jaro 1989, McLaughlin 1993, and Winkler 1990). Names that were very similar received almost the same weight that would have been earned had they been the same. As the difference grew, the weight diminished until it reached a maximum disagreement of about 10. An agreement on sex and the year of birth also contributed towards a higher weight.

Clerical review: The record linkage software produced a list of potential matches for each pass based on a cut-off weight. Then a clerical review took place to examine each of the matches and a decision was made to accept or reject them. This clerical review took into account additional information that was not used for blocking, such as names (for passes where names were not blocked on), SLA and postcode of usual residence, state of registration, age at death, and cause of death. Based on the information available for each linkage data set, the clerical rules differed between linkages.

Linkages where names were unavailable

Blocking variables: The main blocking variables were components of the date of birth, components of the date of death, sex, state, postcode and/or SLA. Additional rules, such as a difference in the birth year by up to 10 years, or restrictions to allow dummy dates only, were imposed in poorer-quality passes to restrict the number of obviously false matches.

Development of pass structure: The pass structure was developed using the results from an earlier named linkage (the aged care linkage). Matched records were examined to identify the most common types of differences that existed for true matches.

Table C2: Most common types of differences between matched pairs of records

Types of differences between matched pairs	No. of records	% of matched records
All fields matching	485	53.5
SLA different	66	7.3
Birth year different	58	6.4
Birth day, month and year different	49	5.4
Death day different	47	5.2
Birth day different	32	3.5
Birth day and birth month different	29	3.2
Birth year and SLA different	22	2.4
Birth day and birth year different	13	1.4
Birth month different	13	1.4
Birth month and birth year different	12	1.3
Birth month, birth year and SLA different	10	1.1
Other types of differences	70	7.7

Source: Enhanced Mortality Database.

Based on a combination of these differences and previous linkage experience, a pass structure was developed that blocked on various combinations of the components of the birth date, components of the death date, state, postcode, and SLA. The passes were presented in a hierarchy of sub-linkages, so that the first sub-linkage comprised a pass where all 9 linkage fields were matching, the second sub-linkage comprised passes where 8 out of 9 fields were matching, the third sub-linkage comprised passes where 7 out of 9 fields were matching, and so on. After each sub-linkage, matched records were removed from both data sets to prevent them from being searched in later passes.

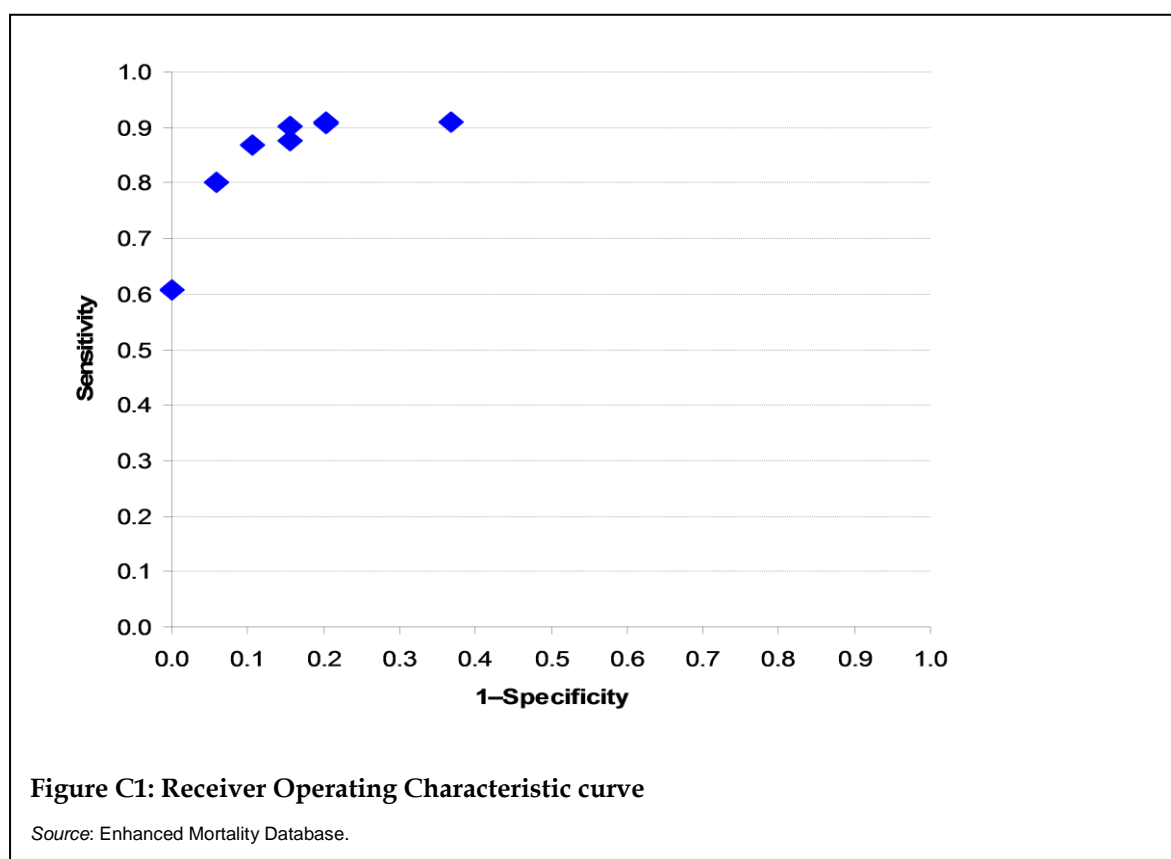
Validation study: The various sub-linkages developed from the previous step were used to match the aged care records to the MED. The results from each sub-linkage were then validated against the original accepted matches from the named aged care linkage. After

each sub-linkage, the accumulated matches were compared with those from the original results and the specificity and sensitivity were calculated. In general, the sensitivity increased and the specificity decreased with the progression from earlier, more specific sub-linkages (blocking on many fields) to later, more general sub-linkages (blocking on fewer fields).

The sensitivity and specificity combinations for each proposed linkage structure were plotted and the final combination was selected to maximise the area under the Receiver Operating Characteristic (ROC) curve (Figure C1). This combination then formed the linkage structure for the unnamed neonatal and hospitals linkages.

The ROC Curve shows a plot of the true positive rate (sensitivity) against the false-positive rate ($100 - \text{Specificity}$). Each point represents the sensitivity-specificity combination for a particular pass structure as tested in this project. In general, there is a trade-off between sensitivity and specificity; more precise pass structures will have a higher specificity (fewer false matches) but also a lower sensitivity (fewer true matches).

The point of perfect discrimination is at the upper left corner of the graph, where there is 100% sensitivity and 0% false-positive rate. Therefore, pass structures that optimise the linkage will have sensitivity-specificity combinations closer to this point.



Appendix D: The linkage processes

The following sections explain each of the four linkages in more detail.

Hospitals linkage

Characteristics of the hospitals data set:

- 5,246 records of Indigenous hospital deaths between July 2001 and June 2006 (inclusive), excluding Australian Capital Territory records (approval not received).
- Linkage fields: sex, date of birth, date of death, state of usual residence, state of hospital, SLA of usual residence, postcode of usual residence.

Limitations for matching:

- All records had no names.

Linkage:

The pass structure for the no-named hospitals linkage was created by developing different combinations of passes and testing these passes in a validation study using the named aged care records. The combination of passes that maximised the sensitivity and specificity in the aged care validation linkage then formed the final pass structure for the hospitals no-named linkage.

The final linkage structure is detailed in the table below (Table D1). The sensitivity and specificity of this pass structure from the aged care validation study was 0.90 and 0.85 respectively.

Passes were organised in hierarchical sub-linkages, where matched records were removed at the end of each sub-linkage to prevent them from appearing in later, poorer-quality passes.

Table D1: Hospitals and MED linkage structure

Pass	Blocking variables	No. potential matches	Additional restrictions	No. final matches
A1	sex, DOB, DOD, state, postcode or SLA	3,217		3,200
B1	sex, DOB, DOD, state	633		631
C1	sex, birthday, birth year, DOD, state, postcode or SLA	41		41
C2	sex, birth month, birth year, DOD, state, postcode or SLA	138		137
C3	sex, birthday, birth month, DOD, state, postcode or SLA	186	DOBs differ by no more than 10 years	186

(continued)

Table D1 (continued): Hospitals and MED linkage structure

Pass	Blocking variables	No. potential matches	Additional restrictions	No. final matches
C4	sex, DOB, death month, death year, state, postcode or SLA	121		120
C5	DOB, DOD, state, postcode or SLA	11		11
D1	sex, birth year, DOD, state, postcode or SLA	107		107
D2	sex, birthday, DOD, state, postcode or SLA	17	DOBs differ by no more than 10 years	17
D3	sex, birth month, DOD, state, postcode or SLA	38	DOBs differ by no more than 10 years	38
F1	Sex, DOD, state, postcode, SLA		1. DOBs differ by no more than 10 years 2. At least one of the DOBs is a dummy	80

Notes

1. The number of potential matches refers to the number of matches before duplicates and cross-matched records were dealt with.
2. A match on state occurred when either of the hospital state or state of residence on the hospitals data set matched to the MED state of death registration.
3. Sub-linkages are separated by grey rows. After each sub-linkage, the matched records were removed from the data sets to search, to prevent them from appearing in later sub-linkages.

Source: Enhanced Mortality Database.

RAC linkage

Characteristics of the aged care data set:

- 990 records of Indigenous deaths between 1 January 2001 and 31 December 2006.
- Available linkage fields: surname, given names, sex, date of birth, date of death, suburb of client, postcode of client, postcode of service, state of client, state of service, SLA of service, service address.

Limitations for matching:

- A large proportion of records had a dummy date of birth: 26% of records had a date of 1 Jan, 30 June, 1 July, 15 July compared with 5% of records on the MED.

Linkage:

The linkage was composed of three sub-linkages: A) which blocked on both birth and death dates and sex; B) which blocked on components of the death date only, and C) which blocked on components of the birth date only. The process was hierarchical so that matched records were removed from both data sets after each sub-linkage, to prevent them from appearing in proposed matches in later sub-linkages. Within each sub-linkage, the passes were mutually exclusive so that each matched pair could only appear in one pass.

Table D2 shows the clerical review rules for the passes in each sub-linkage. Note that these are general rules only – they could be overruled in exceptional cases based on human judgment.

Table D2: Aged care and MED linkage structure

Pass	Blocking variables	No. potential matches	Weight cut-off	Additional restrictions	Clerical review rules	No. final matches
A1	sex, DOB, DOD (as recorded on the NMD)	575			1. Where names are similar, accept. Otherwise, accept if addresses are the same. 2. Where one set of names and addresses are missing, accept if states are the same.	551
A2	sex, DOB, DOD (as recorded on the NDI)	13			1. Where names are similar, accept. Otherwise, accept if addresses are the same. 2. Where one set of names and addresses are missing, accept if states are the same.	13
B1	surnames, given names, sex, DOD	132			1. Accept all.	132
B2	DOD	152,628			1. Where both names are similar, accept. 2. Where only one of the names matches: accept where SLA/postcode and address are the same. 3. Where names and addresses are missing from the NDI: accept if SLAs/postcodes are the same or only the last digit is different and date of birth is the same, close or one is a dummy date.	126
B3	day and year of death	100,443	0		1. For weights of 10 or higher: where both names are similar and states match, accept. 2. For weights below 10: accept where both names are similar and addresses are the same. If dates of birth are more than 5 years apart, one of them must be a dummy date.	0

(continued)

Table D2 (continued): Aged care and MED linkage structure

Pass	Blocking variables	No. potential matches	Weight cut-off	Additional restrictions	Clerical review rules	No. final matches
B4	month and year of death	4,526,945	0		1. For weights of 10 or higher: where both names are similar and states match, accept. 2. For weights below 10: accept where both names are similar and addresses are the same. If dates of birth are more than 5 years apart, one of them must be a dummy date.	79
B5	day and month of death	46,464	0	1. DODs must be no more than 1 year apart	1. For weights of 10 or higher: where both names are similar and states match, accept. 2. For weights below 10: accept where both names are similar and addresses are the same. If dates of birth are more than 5 years apart, one of them must be a dummy date.	0
B6	year of death	728,583	10		1. Accept where both names are similar and SLAs/postcodes are the same. If birth dates are more than 5 years apart, one of them must be a dummy date.	3
B7	day of death	500,754	10	1. DODs must be no more than 1 year apart	1. Accept where both names are similar and SLAs/postcodes are the same. If birth dates are more than 5 years apart, one of them must be a dummy date.	0
B8	month of death	334,516	10	1. DODs must be no more than 1 year apart	1. Accept where both names are similar and SLAs/postcodes are the same. If birth dates are more than 5 years apart, one of them must be a dummy date.	0
B9	none	925,725	10	1. DODs must be no more than 1 year apart	1. Accept where both names are similar and SLAs/postcodes are the same. If birth dates are more than 5 years apart, one of them must be a dummy date.	0

(continued)

Table D2 (continued): Aged care and MED linkage structure

Pass	Blocking variables	No. potential matches	Weight cut-off	Additional restrictions	Clerical review rules	No. final matches
C1	surnames, given names, sex, DOB	0		1. DODs must be no more than 1 year apart	1. Accept all.	0
C4	DOB	4,783		1. DODs must be no more than 1 year apart	1. Where both names are similar, accept. Where only one of the names is similar, accept where SLA and address are the same. 2. Where names and addresses are missing from the NDI: if SLAs/postcodes are the same or only the last digit is different and DODs are close.	0
C5	day and year of birth	45,757	0	1. DODs must be no more than 1 year apart	1. For weights above 10: where both names are similar and states match, accept. 2. For weights below 10: accept where both names are similar and addresses are the same.	0
C7	month and year of birth	130,774	0	1. DODs must be no more than 1 year apart	1. For weights above 10: where both names are similar and states match, accept. 2. For weights below 10: accept where both names are similar and addresses are the same.	0
C9	day and month of birth	209,449	0	1. DODs must be no more than 1 year apart 2. DOBs must be dummy dates if they are more than 5 years apart	1. For weights above 10: where both names are similar and states match, accept. 2. For weights below 10: accept where both names are similar and addresses are the same.	0
C10	year of birth	71,498	10	1. DODs must be no more than 1 year apart	1. Accept where both names are similar and SLAs/postcodes are the same.	0

(continued)

Table D2 (continued): Aged care and MED linkage structure

Pass	Blocking variables	No. potential matches	Weight cut-off	Additional restrictions	Clerical review rules	No. final matches
C11	day of birth	115,273	10	1. DODs must be no more than 1 year apart 2. DOBs must be dummy dates if they are more than 5 years apart	1. Accept where both names are similar and SLAs/postcodes are the same.	1
C12	month of birth	318,915	10	1. DODs must be no more than 1 year apart 2. DOBs must be dummy dates if they are more than 5 years apart	1. Accept where both names are similar and SLAs/postcodes are the same.	1
C13	none	842,282	10	1. DODs must be no more than 1 year apart 2. DOBs must be dummy dates if they are more than 5 years apart	1. Accept where both names are similar and SLAs/postcodes are the same.	0

Notes

1. The number of potential matches refers to the number of comparisons made before weight cut-offs were applied.
2. Sub-linkages are separated by grey rows. After each sub-linkage, the matched records were removed from the data sets to search to prevent them from appearing in later sub-linkages.

Source: Enhanced Mortality Database.

Neonatal death linkage

Characteristics of the neonatal data set:

- 290 records of Indigenous neonatal deaths (deaths of live born babies within 28 days of birth) between 2001 and 2005 (inclusive).
- Linkage fields: date of birth, state of birth, state of mother's usual residence, postcode of mother's usual residence, sex, date of death, cause of death.

Limitations for matching:

- All records had no names.
- 15 records were missing at least one birth date component.
- 56 records had no death dates.
- In many cases, the cause of death did not match between the NMD and the neonatal data set, even if all the other fields matched correctly.
- 137 records belonged to infants who were born and died on the same day.

Linkage:

The pass structure for the no-named neonatal linkage was created by developing different combinations of passes and testing these passes in a validation study using the named aged care records. The combination of passes that maximised the sensitivity and specificity in the aged care validation linkage then formed the final pass structure for the neonatal no-named linkage (for more details, see Appendix C).

The final linkage structure is detailed in Table D3 below. The sensitivity and specificity of this pass structure from the aged care validation study was 0.87 and 0.89, respectively. It should be noted that the characteristics of the aged care and neonatal data sets are very different. An additional pass, E1, was also added to allow for matches to the large portion of neonatal records that had a missing date of death. This pass was unique to the neonatal linkage and replaced pass F1 in the hospitals linkage where entire dates of birth (rather than dates of death) were allowed to vary.

Passes were organised into hierarchical sub-linkages where matched records were removed at the end of each sub-linkage, to prevent them from appearing in later, poorer-quality passes.

Neonatal and MED linkage structure

Table D3: Neonatal and MED linkage structure

Pass	Blocking variables	No. potential matches	Additional restrictions	No. final matches
A1	sex, DOB, DOD, state, postcode	70		66
B1	sex, DOB, DOD, state	84		76
C1	sex, birthday, birth year, DOD, state, postcode	0		0
C2	sex, birth month, birth year, DOD, state, postcode	1		1
C3	sex, birthday, birth month, DOD, state, postcode	0	DOBs differ by no more than 10 years	0
C4	sex, DOB, death month, death year, state, postcode	6		6
C5	DOB, DOD, state, postcode	0		0
D1	sex, birth year, DOD, state, postcode	1		1
D2	sex, birthday, DOD, state, postcode	0	DOBs differ by no more than 10 years	0
D3	sex, birth month, DOD, state, postcode	0	DOBs differ by no more than 10 years	0
E1	sex, DOB, state, postcode	27	Neonatal DOD must be missing	25

Notes

1. The number of potential matches refers to the number of matches before duplicates and cross-matched records were dealt with.
2. A match on state occurred when either of the neonatal state of birth or state of mother's residence matched to the MED state of death registration.
3. Sub-linkages are separated by grey rows. After each sub-linkage, the matched records were removed from the data sets to search to prevent them from appearing in later sub-linkages.

Source: Enhanced Mortality Database.

Appendix E: Classification of excellent, good and acceptable matches

Table E1: Hierarchy of quality of matches

Linkage	Excellent quality matches (matchFlag=1)	Good quality matches (matchFlag=2)	Acceptable matches but some larger discrepancies in the fields exist (matchFlag=3)
Aged care linkage	A1, A2, B1, B2, C1, C4	B3, B4, B5, C5, C7, C9	B6, B7, B8, B9, C10, C11, C12, C13
Neonatal linkage	A1, B1	C1, C2, C3, C4, C5	D1, D2, D3, E1
Hospitals linkage	A1, B1	C1, C2, C3, C4, C5	D1, D2, D3, F1

Source: Enhanced Mortality Database.

Important note: This classification of match quality is based loosely on the number of fields matching for each pass, but it is still somewhat subjective. The ideal way of creating such a system would have been to assign a flag at the clerical review stage, where information on names, linkage fields, pass information and weight could all be assessed. Although this method categorises matches by their passes, it is still a fairly blunt instrument as it does not distinguish between matches within a pass.

Appendix F: Life expectancy at birth, calendar years 2001 to 2005, for NSW, Qld, WA, NT and Australia, with 95% confidence lower and higher limits

Table F1: Life expectancy at birth, calendar years 2001 to 2005^(a)

	NSW	Qld	WA	NT	Australia
Males					
2001	69.3 [68.0-70.6]	65.7 [64.4-67.2]	63.0 [61.5-64.7]	58.2 [56.5-59.9]	66.4 [65.6-67.2]
2002	68.4 [67.1-69.6]	66.0 [64.7-67.4]	62.8 [61.2-64.5]	57.4 [55.9-59.0]	66.2 [65.4-67.0]
2003	68.7 [67.5-69.9]	66.4 [65.1-67.8]	63.9 [62.2-65.5]	58.4 [56.8-60.0]	66.4 [65.6-67.1]
2004	70.3 [69.1-71.7]	65.9 [64.7-67.2]	63.2 [61.7-65.0]	58.2 [56.6-59.9]	66.9 [66.2-67.7]
2005	70.2 [69.0-71.6]	68.5 [67.1-70.0]	63.9 [62.3-65.4]	58.6 [57.0-60.2]	67.2 [66.4-68.0]
Females					
2001	74.1 [72.9-75.5]	72.2 [70.8-73.8]	68.2 [66.6-70.0]	66.3 [64.5-68.2]	72.7 [71.8-73.5]
2002	73.5 [72.2-74.8]	71.8 [70.4-73.2]	66.8 [65.1-68.5]	64.4 [62.8-66.3]	71.5 [70.6-72.3]
2003	75.3 [74.0-76.6]	72.5 [71.2-73.9]	69.4 [67.8-71.1]	67.3 [65.6-69.0]	73.2 [72.4-74.1]
2004	75.8 [74.5-77.2]	72.1 [70.8-73.5]	69.1 [67.6-70.7]	66.9 [65.3-68.7]	72.8 [72.0-73.6]
2005	74.8 [73.6-76.0]	74.7 [73.3-76.1]	67.6 [65.9-69.2]	65.6 [63.9-67.3]	73.5 [72.7-74.3]

(a) 95% confidence intervals are in square brackets.

Source: Enhanced Mortality Database.

References

- ABS (Australian Bureau of Statistics) 1983. Methods and procedures in the compilation of estimated resident population 1981 and the construction of the 1971-1981 time series. ABS cat. no. 3103.0. Canberra: ABS.
- ABS 1998. Quarterly births and deaths registration. Demography Working Paper 98/2. ABS cat. no. 3114.0. Canberra: ABS.
- ABS 1999. Standards for statistics on cultural and language diversity. Cat. no. 1289.0. Canberra: Australian Bureau of Statistics.
- ABS 2002. Deaths, 2001, Australia. ABS cat. no.3302.0. Canberra: ABS.
- ABS 2007. Census of Population and Housing – details of undercount, Aug 2006. ABS cat. no. 2940.0. Canberra: ABS.
- ABS 2008a. Census data enhancement – Indigenous Mortality Quality Study. ABS cat. no. 4723.0. Canberra: ABS.
- ABS 2008b. Deaths, 2007, Australia. ABS cat. no. 3302.0. Canberra: ABS.
- ABS 2008c. Discussion Paper: Assessment of methods for developing life tables for Aboriginal and Torres Strait Islander Australians, 2006. ABS cat. no. 3302.0.55.002. Canberra: ABS.
- ABS 2008d. Information Paper: Census Data Enhancement – Indigenous Mortality Quality Study. ABS cat. no. 4723.0. Canberra: ABS.
- ABS 2009a. Deaths, 2008, Australia. ABS cat. no. 3302.0. Canberra: ABS.
- ABS 2009b. Experimental estimates and projections, Aboriginal and Torres Strait Islander Australians. ABS cat. no. 3238.0. Canberra: ABS.
- ABS 2009c. Experimental life tables for Aboriginal and Torres Strait Islander Australians. ABS cat. no. 3302.055.003. Canberra: ABS.
- AIHW (Australian Institute of Health and Welfare) 2004a. Australia's mothers and babies 2001. Cat. no. PER 25. Canberra: AIHW.
- AIHW 2004b. Australia's mothers and babies 2002. Cat. no. PER 28. Canberra: AIHW.
- AIHW 2005. Australia's mothers and babies 2003. Cat. no. PER 29. Canberra: AIHW.
- AIHW 2006. Australia's mothers and babies 2004. Cat. no. PER 34. Canberra: AIHW.
- AIHW 2007. Australia's mothers and babies 2005. Cat. no. PER 40. Canberra: AIHW.
- AIHW 2008a. Australia's mothers and babies 2006. Cat. no. PER 46. Canberra: AIHW.
- AIHW 2008b. Australian hospital statistics 2006-07. Health services series no. 31. Cat. no. HSE 55. Canberra: AIHW.
- AIHW 2008c. Residential aged care in Australia 2006-07: a statistical overview. Aged care statistics series no. 26. Cat. no. AGE 56. Canberra: AIHW.
- AIHW 2009. Residential aged care in Australia 2007-87: a statistical overview. Cat no. AGE 66. Canberra: AIHW.
- AIHW 2010a (forthcoming). Indigenous identification in hospital separations data: quality report. Health services series. Canberra: AIHW.
- AIHW 2010b. National best practice guidelines for collecting Indigenous status in health data sets. Cat. no. IHW 29. Canberra: AIHW.

- AIHW 2011. Trends in palliative care in Australian hospitals. Canberra: AIHW.
- AIHW: Leeds K, Gourley M, Laws P, Zhang J, Al-Yaman F & Sullivan E 2007. Indigenous mothers and their babies, Australia 2001–2004. Perinatal statistics series no. 19. Cat. no. PER 38. Canberra: AIHW.
- Broad J 2012. Personal Communication (email): Overlap between deaths reported in residential aged care and hospital separation data sets. 17 February 2012, <Joanna Broad (Joanna.Broad@waitematadhb.govt.nz), Senior Research Fellow, Freemasons Department of Geriatric Medicine, University of Auckland, Auckland, New Zealand>.
- Chiang CL 1984. The life table and its application. Malabar, Florida, USA: Robert E Krieger Publishing Co.
- Hyndman RJ & Ullah MS 2007. Robust forecasting of mortality and fertility rates: A functional data approach. *Computational Statistics and Data Analysis*. 51:4942–56.
- Jaro MA 1989. Advances in record-linkage methodology as applied to matching the 1985 Census of Tampa, Florida. *Journal of the American Statistical Association* 84:414–20.
- McLaughlin G 1993. Personal communication of C-string comparison routine.
- Winkler WE 1990. String comparator metrics and enhanced decision rules in the Fellegi-Sunter model of record linkage. *Proceedings of the Survey Research Methods Section, American Statistical Association*: 354–9.

Glossary

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

Administrative data Information that is required by law or that is collected for the purpose 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).

Algorithm A process or set of rules used for calculation or problem-solving. In this report, 'algorithm' refers to a set of rules 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 (for example, 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.

Confidence interval (CI) A statistical term describing a range (interval) of values within which we can be confident that the true value lies, usually because it has a 95% or higher chance of doing so.

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*.

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

Ethics Committee A committee set up by a body or institution whose principal responsibility is to form an opinion of the acceptability or otherwise on ethical grounds of activities engaged in by the institution or body, with which it is associated. Membership of the AIHW Ethics Committee is in accordance with guidelines as specified by the National Health and Medical Research Council (NHMRC), and includes the Director of the Institute (or his/her representative) and eight others appointed to provide a range of expertise and experience in health and welfare research areas, including a representative of Registrars of Births, Deaths and Marriages.

Expectation of life An indication of how long a person can expect to live, depending on the age they have already reached. Technically, it is the number of years of life remaining to a person at a particular age if death rates do not change. The most commonly used example is life expectancy at birth. See also *life expectancy*.

Life table A representation of the probable years of survivorship of a defined population of subjects. It comprises any of various tables describing mortality and survival data for groups of individuals at specific times or over defined intervals. Life tables may summarize combined mortality experience by age over a brief period, usually one to three years (period life table) or may follow a cohort over time (cohort life table).

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

Fetal death (stillbirth): death prior to the complete expulsion or extraction from its mother of a product of conception of 20 or more completed weeks of gestation or of 400 grams or more birthweight. The death is indicated by the fact that after such separation the fetus does not breathe or show any other evidence of life, such as beating of the heart, pulsation of the umbilical cord, or definite movement of voluntary muscles.

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 identification The process of identifying or of being identified as being of Aboriginal and/or Torres Strait Islander origin.

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 An indication of how long a person can expect to live, depending on the age they have already reached. Technically, it is the number of years of life remaining to a person at a particular age if death rates do not change. The most commonly used example is life expectancy at birth. See also *expectation of life*.

Linked records Records that have passed through the data linkage process and were 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. Linking variables can also be used as blocking variables. See also *blocking variables* and *matching variables*.

Live birth: the complete expulsion or extraction from its mother of a product of conception, irrespective of the duration of the pregnancy, which, after such separation, breathes or shows any other evidence of life, such as beating of the heart, pulsation of the umbilical cord, or definite movement of voluntary muscles, whether or not the umbilical cord has been cut or the placenta is attached; each product of such a birth is considered liveborn.

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

Medical Certificate of Cause of Death (MCCOD) A document completed by a doctor who attended to a person before their death, or a document completed by a doctor who examined a person after their death, containing information on the underlying cause of death or the train of events leading directly to death. Information on the cause of death is coded

according to rules and conventions of the 10th revision of the International Classification of Diseases.

Neonatal death: death of a liveborn baby within 28 days of birth.

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.

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

Pass structure A set of passes or iterations of passes, in record linkage, using particular sets of blocking and matching variables. See *Blocking, Blocking variables*.

Postneonatal death: death of a liveborn baby after 28 days and within one year of birth.

Probabilistic linkage A method of record linkage that utilises 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*.

Separation The formal process by which an admitted patient in a hospital, resident in an aged care home or resident in any other facility providing care or treatment completes an episode of care or treatment, such as by being discharged, dying, transferring to another institution or facility or changing type of care.

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

Unlinked records Records that have passed through the data linkage process and were unable to be linked to a record from the other file.

List of tables

Table 4.1:	Quality of matches for each data set	12
Table 4.2:	Additional Indigenous deaths identified through linkage, by state and territory	13
Table 4.3:	Details of Indigenous deaths from additional sources	13
Table 4.4:	Number of records added through linkage by age group and sex	15
Table 4.5:	Indigenous life expectancy at birth, at age 20, 45 and 65, based on enhanced death data, Australia 2001–2006	17
Table 4.6:	Indigenous life expectancy at birth, based on enhanced death data, including unlinked records, selected States and the NT, 2001–2006	17
Table 5.1:	Comparison of various life expectancy at birth estimates	19
Table 5.2:	Estimates of life expectancy at birth by selected states and the NT	20
Table 5.3:	Indigenous life expectancy at birth for NSW, Qld, WA, NT and Australia, 2001–2005	20
Table 5.4:	Changes in life expectancy at birth for Indigenous Australians and all Australians, 2001–2005	21
Table A1:	Registered deaths	24
Table A2:	Deaths in residential aged care	26
Table A3:	Deaths in hospitals	27
Table A4:	Number of records with age and sex not stated in hospital separations	28
Table A5:	Women who gave birth by Indigenous status, Australia (excluding Tasmania) 2001–2006	29
Table B1:	Final MED database	31
Table C1:	Summary of data sets linked to the MED	32
Table C2:	Most common types of differences between matched pairs of records	33
Table D1:	Hospitals and MED linkage structure	35
Table D2:	Aged care and MED linkage structure	37
Table D2	Aged care and MED linkage structure	38
Table D3:	Neonatal and MED linkage structure	42
Table E1:	Hierarchy of quality of matches	43
Table F1:	Life expectancy at birth, calendar years 2001 to 2005	44

List of figures

Figure 4.1:	Indigenous life expectancy at birth based on enhanced death data, males, 2001–2006.....	18
Figure 4.2:	Indigenous life expectancy at birth based on enhanced death data, females, 2001–2006.....	18
Figure C1:	Receiver Operating Characteristic curve	34

The aim of *An enhanced mortality database for estimating Indigenous life expectancy: a feasibility study, 2012* is to improve reporting of information on Indigenous deaths by linking death registrations data with several additional data sets that contain information on Indigenous deaths.

Linkage of registered deaths data to the additional data sets identified 10.2 per cent more deaths that were not identified as Indigenous on the registered deaths data set. The enhanced data also showed that expectation of life at birth for Indigenous Australians over the period 2001–2006 was 66.6 years for males and 72.7 years for females. This report shows that data linkage is an effective tool for improving estimates of mortality and life expectancy for Indigenous Australians.