Australian
Bureau of
Statistics

Information Paper

Death Registrations to Census Linkage Project – Methodology and Quality Assessment

Australia

2011-2012

3302.0.55.004

www.abs.gov.au

ABS Catalogue No. 3302.0.55.004

Produced by the Australian Bureau of Statistics.

INQUIRIES

For further information about these and related statistics, contact the National Information and Referral Service on 1300 135 070.

EXECUTIVE SUMMARY

Life expectancy is a broad measure of a population's long-term health and wellbeing. The Council of Australian Governments (COAG) has set a target of closing the life expectancy gap between Aboriginal and Torres Strait Islander and other Australians within a generation (see *Overcoming Indigenous Disadvantage: Key Indicators 2011*, available at www.pc.gov.au). So that progress toward this target can be more accurately measured, COAG has funded the Australian Bureau of Statistics (ABS) to undertake an ongoing program of work to improve the quality of life expectancy and other mortality estimates for the Aboriginal and Torres Strait Islander population.

This paper presents information on the methodology and quality of the statistical data integration project that linked death registrations to the 2011 Census of Population and Housing. This project, which has been referred to as the Indigenous Mortality Project, was conducted as part of the 2011 Census Data Enhancement program and built on the foundation of the first such study conducted in conjunction with the 2006 Census. The primary aim of the linkage was to assess the consistency of the identification of Indigenous status as reported in death registration and Census data, and thereby provide input into the compilation of life tables and life expectancy estimates for Aboriginal and Torres Strait Islander people.

Death registrations are provided to the ABS by State and Territory Registrars of Births, Deaths and Marriages. This project used information from deaths that were registered during the Census processing period in 2011-12 when Census name and address were available as linking variables. Probabilistic linking methods were used to bring the datasets together and identify the best match. This process involves comparing several variables common to both files and generates a single numerical measure of how well two particular records match.

In addition to advances in technology and data linking software, one of the main improvements was the allocation of greater resources to clerical review. This enabled a high level of quality control through manual checking to the point where virtually all links assigned in this project are assessed as true links; that is, the death registration and Census record belong to the same individual.

At the completion of the linkage, 93% of death registrations had been linked to a Census record. The raw linkage rate for Aboriginal and Torres Strait Islander deaths was 80%, a considerable improvement over the corresponding figure of 74% in the 2006 Census study.

While the Census aims to count every person in Australia on Census Night, inevitably some people are missed. It could be expected, therefore, that not all death registrations will in fact have a corresponding record in the Census file. After applying an adjustment factor to account for people who were missing from the Census, the linkage rate for Aboriginal and Torres Strait Islander deaths rises to about 90% compared with an adjusted linkage rate of 96% for non-Indigenous deaths.

This information paper is the first report to be released from the 2011 cycle of the Indigenous Mortality Project. Two other reports will present statistics from the project including differences in the identification of Indigenous status between death registrations and the Census. One examines the characteristics of linked and unlinked records for Aboriginal and Torres Strait Islander deaths in more depth (*Death registrations to Census linkage project - Key Findings for Aboriginal and Torres Strait Islander peoples, 2011-12*, cat. no. 3302.0.55.005) and the other presents mortality statistics including life expectancy estimates by Indigenous status (*Life Tables for Aboriginal and Torres Strait Islander Australians*, cat. no. 3302.0.55.003). These reports are due for release in November 2013.

CONTENTS

SECTION 1: BACKGROUND

SECTION 2: DATA LINKING PROCESS

SECTION 3: RESULTS

ADDITIONAL INFORMATION

ACKNOWLEDGMENTS

The 2011 Indigenous Mortality project was funded by the Council of Australian Governments (COAG). The project enables reporting against the COAG target “to close the life expectancy gap within a generation”. The ABS acknowledges the provision of Death Registration data from State and Territory Registries of Births, Deaths and Marriages, and ongoing efforts to improve the identification of Indigenous status within death registrations. The ABS also acknowledges the provision of data from all Australians, including Aboriginal and Torres Strait Islander peoples, as part of the 2011 Census of Population and Housing.

We thank the members of the ABS Advisory Group on Aboriginal and Torres Strait Islander Statistics (AGATSIS) and the ABS Aboriginal and Torres Strait Islander Demographic Statistics Expert Advisory Group (ATSIDSEAG) for their input and guidance.

INTRODUCTION

The most recent cycle of the Death registrations to Census linkage project (otherwise referred to as the Indigenous Mortality Project) was conducted as part of the 2011 Census Data Enhancement (CDE) program. The CDE program included a number of linkage projects that brought together data from the 2011 Census of Population and Housing with other specified datasets. This project involved linking the Census with death registrations to examine differences in the reporting of Indigenous status across the two datasets in order to apply adjustment factors to mortality and life-expectancy estimates. See also *Census Data Enhancement Project: An Update, October 2010* (cat. no. 2062.0).

This information paper is the first of three reports from the Indigenous Mortality project. It describes the methodology and quality features implemented in the process and presents an assessment of linkage quality. Adjusted mortality rates for the Aboriginal and Torres Strait Islander population together with 2011 Census-based life expectancy estimates will be released in *Life Tables for Aboriginal and Torres Strait Islander Australians* (cat. no. 3302.0.55.003) on 15 November 2013. At that time, *Death registrations to Census linkage project - Key Findings for Aboriginal and Torres Strait Islander peoples, 2011-12* (cat. no. 3302.0.55.005)—an examination of social and demographic characteristics associated with differences in Indigenous status identification between the two datasets —will also be released.

This project linked 20,928,304 records from the 2011 Census with 153,455 death registrations where the death occurred between the 10th of August 2011 and the 27th of September 2012 inclusive. The time period of death registrations was selected in order to capture as many deaths as possible of people who were counted in the 2011 Census (which took place on the 9th of August 2011). Due to lags between occurrence and registration, the number of deaths registered during this period would be lower than the number of deaths that occurred, particularly for those deaths occurring towards the end of the reference period.

The 2011 Census processing period, like the 2006 Census, provided the opportunity to link deaths with the full Census using name, address and other variables common to both datasets. This is referred to by the ABS as a *Gold linkage* method*.* It should be noted that all names and addresses collected in the Census were destroyed at the conclusion of Census processing on 7 December 2012, setting an end point to the Gold linkage period.

In detail, the aims of this project were to:

- assess the consistency of Indigenous status as reported in death registration and Census data
- estimate measures of undercoverage of deaths of Aboriginal and Torres Strait Islander people by state/territory and remoteness areas of Australia
- investigate the feasibility of applying adjustment factors for Aboriginal and Torres Strait Islander deaths output data
- provide input into the compilation of life tables for Aboriginal and Torres Strait Islander people, life expectancy estimates and Aboriginal and Torres Strait Islander/non-Indigenous differences in life expectancy and other mortality measures that are consistent with population estimates based on the adjusted 2011 Census of Population and Housing.

The 2011 Death registrations to Census linkage project expanded on the methods used in the 2006 cycle as described in the research paper *Linking Census Records to Death Registrations* (Cat. No. 1351.0.55.030). The main enhancements implemented for the 2011 project included:

- improvements to linking software and hardware
- improvements to data cleaning and standardisation – particularly names
- refinements to the blocking and linking strategy
- use of the Expectation-Maximisation algorithm for estimating linkage model parameters
- increased and targeted clerical review on record pairs where the death registration was for a person identified as being of Aboriginal and Torres Strait Islander descent.

SECTION 1: BACKGROUND

The purpose of this project was to update the previous cycle conducted in conjunction with the 2006 Census, which provided estimates of under-identification of Aboriginal and Torres Strait Islander deaths during the Census processing period in 2006-07. These estimates were used in the calculation of the 2006 Census-based series of life expectancy estimates for Aboriginal and Torres Strait Islander people (see *Indigenous Mortality Quality Study, 2006*, ABS cat. no. 4723.0, and *Experimental Life Tables for Aboriginal and Torres Strait Islander Australians*, 2005-07, ABS cat. no. 3302.0.55.003).

Collection of Death registrations data for Aboriginal and Torres Strait Islander people

Death registrations data from the State and Territory Registries of Births, Deaths and Marriages are used by the ABS to produce estimates of Aboriginal and Torres Strait Islander deaths. The information relates to all registered deaths including those referred to a Coroner. Prior to 2007, Indigenous status was recorded by the ABS based only on information supplied on Death Registration forms; that is, as reported by a relative, or other person acquainted with the deceased, or by an official of the institution where the death occurred. While there is some variation in practice among the jurisdictions, information supplied on both the Death Registration form and the Medical Certificate of Cause of Death (completed by medical practitioners) has been used where available to derive Indigenous status since 2007. Estimates of Aboriginal and Torres Strait Islander deaths are used as an input for calculating Aboriginal and Torres Strait Islander population and life expectancy estimates.

Collection of Census data for Aboriginal and Torres Strait Islander people

The Census is usually completed by a responsible adult answering for themselves or on behalf of another person present in the dwelling on Census night. In the standard Census form, Indigenous status is reported by the person completing the form and in some instances may not be answered. By contrast, Interviewer Household Forms are used in discrete Aboriginal and Torres Strait Islander communities and are available for people in other areas. These forms are completed by a trained interviewer, who is recruited from the local community wherever possible. The interviewer collects Census information verbally from a household member/s, resulting in a lower non-response rate for all items, including Indigenous status.

2.1 – DATA LINKING INFRASTRUCTURE

Linking software

The linking software chosen for use in this project was FeBRL (Christen, Churches, and Hegland, 2004). The version used was FeBRL v0.3, released under an open source licence. The ABS made considerable changes to the FeBRL software to integrate it within the ABS information technology environment and improve both capacity and performance.

The ABS also spent considerable resources to develop supporting infrastructure such as a secure remote linking environment and the Generic Clerical Review (GCR) system. The GCR is an end to end program management tool to facilitate the clerical review of linked records and efficiently manage and measure the quality of the process. The GCR incorporated customisable layouts and Work Health and Safety features, which were important given the large amount of clerical review that was undertaken.

Linking infrastructure also utilised the analytical software SAS. SAS was used to perform a number of functions, such as standardisation of variables, preparation and creation of files to be used in FeBRL and the GCR, and in the analysis of linkage results.

Data Security and Hardware

Confidentiality of data used for linking was maintained in accordance with the *High Level Principles for Data Integration* using Commonwealth data endorsed by Commonwealth Secretaries in 2010. Operational arrangements for managing data flows within the ABS included restricting access to information through the *functional separation* of roles. These roles ensured segregated access to data and separated those requiring identifiable information for linking and clerical review from groups performing data analysis. For more information refer to the separation principle on the National Statistical Service website (www.nss.gov.au).

All data preparation, linking and clerical review was performed using password-protected remote virtual servers located within the Census Data Processing Centre (DPC) environment. These servers were designed to cater for the memory-intensive requirements of FeBRL, with each server consisting of eight processors with 72 GB RAM, 800 GB hard disk, and running a Windows 7 operating system.

Table 2.1 shows the number of record pairs generated within each pass of the 2006 and 2011 projects, along with the associated computing times. While the number of comparisons contributes to the overall computing time required for each pass, the amount of concurrent server activity is another source of variance for the run times provided. The more robust linking infrastructure utilised for the 2011 Death registrations to Census linkage project resulted in much faster computing times than in the 2006 cycle for the large number of record pairs that were generated.

Table 2.1 – NUMBER OF RECORD PAIRS AND COMPUTING TIMES, By pass number–2006 and 2011

	NUMBER OF RECORD PAIRS GENERATED		COMPUTING TIME	
	2006	*2011*	*2006*	*2011*
Pass 1	9 588 669	18 518 697	1 hr 32 mins	2 hrs 22 min
Pass 2	91 052 466	96 857 621	12 hrs 42 mins	3 hrs 27 min
Pass 3	4 309 542	5 140 165	1 hr 20 mins	42 min
Pass 4	164 267 699	119 930 890	37 hrs 44 mins	9 hrs 4 min
Pass 5	. .	282 410 550	. .	1 hr 47 min

. . Not applicable.

2.2 - DATA LINKING METHODOLOGY

The statistical linking methodology applied in this project is called probabilistic linking (Felligi & Sunter, 1969). This method links records from two datasets using several variables common to each. A key feature of the methodology is the ability to handle a variety of linking variables and record comparison methods to produce a single numerical measure of how well two particular records match. This allows ranking of all possible links and optimal assignment of the link or non-link status (Solon and Bishop, 2009).

The probabilistic linking methodology used here can be generalised into the following steps:

- standardisation of data
- blocking
- record pair comparison
- decision model.

2.2.1 – Standardisation

Before records on the two datasets are compared, the contents of the two datasets need to be standardised to facilitate comparison. This includes a number of steps such as verification, recoding and reformatting fields, and parsing text fields (i.e. separating text fields into their components). Additionally, some fields require substantial repair.

Some variables differ between the two datasets in a predictable way, and an adjustment is required to negate this difference. Some variables are coded differently at different points in time, and concordances may be necessary to create variables which align on the two datasets. Variables may also be recoded or aggregated in order to obtain a more robust form of the variable. This set of procedures is collectively termed *standardisation*. Standardisation takes place in conjunction with a broader evaluation of the dataset, in which potential linking variables are identified.

The standardisation procedure for the Death registrations to Census linkage project involved coding imputed and invalid values for selected variables to a common missing value. These variables included day of birth, month of birth, year of birth, age, sex, year of arrival and marital status. Entire imputed records created for persons known to exist but from whom no Census form had been received, were removed from the pool of Census records prior to linkage.

The following is a description of further standardisation techniques that were performed on variables for this project:

First name

For the Census data, the original names were first subjected to repair processes at the DPC. First names were compared against a master name index, which allowed for names that were misread by the DPC Optical Character Recognition (OCR) software to be parsed and repaired. Standardisation of first names included removal of non-alphabetical characters and titles (e.g. Ms, Dr).

First names were then compared against a nickname concordance, ensuring that different variations would be grouped into a common name for the purposes of linkage. For example, the names 'Bradley' and 'Brad' may both be standardised to 'Bradley'. Any first names that could not be matched to a nickname retained their original form.

Name data on the death registrations were of considerably better quality than those on the Census, and as such were not required to go through a repair process. However the remainder of the First name standardisation process for death registrations was consistent with the Census.

Surname

Census surnames underwent repair processes at the Census DPC. Surnames that were repaired were subject to further standardisation prior to linkage; otherwise the original stated surname was used.

For both Census and death registrations, non-alphabetical characters were removed from surnames. Records with multiple surnames that had not stated a first name had the first part of the surname substituted into the final First name field.

Initial 4

The variable 'Initial 4' was derived by concatenating the first two letters of the standardised first name with the first two letters of the standardised surname. If either the standardised first name or standardised surname was missing, then initial 4 was set to missing. This variable was used to group names into common categories.

Sex

Census records that contained an imputed value for sex but had provided a first name were compared against a name index in an attempt to determine if the name was commonly given to males or females. If the Census name matched to a name on the index, then the relevant sex was applied to the Census record. If the Census name could not match to any name on the index, then the value for sex was coded to missing.

Address (Street number, Street name, Suburb, Postcode)

Linking was conducted based on the usual residential address of Census records and death registrations. Census addresses were also repaired using the output from Census address coding. Death registrations where only a residential title was supplied (e.g. nursing home, hospital etc.) underwent additional coding.

Mesh Block

Mesh Blocks are the smallest geographical area defined by the ABS. The 2011 Australian Statistical Geography Standard (ASGS) contains 347,627 Mesh Blocks covering the whole of Australia without gaps or overlaps.

The standardised Mesh Block variable was based on the usual residential address of a record. Instances where a Mesh Block could not be assigned or the respondent usually resided overseas were recoded to missing.

Age

Age was standardised to three digits and top-coded to a maximum value of 115. For death registrations, age in months under one year was recoded to zero.

Year of birth

Year of birth values that were either invalid or had only two digits were amended using age information, when possible. For example, where a record had only stated '07' as the year of birth, this value would be recoded to either '1907' or '2007', depending on supplementary age information that had been provided.

Birthplace

A two-digit Birthplace was created in order to minimise disagreement when linking records belonging to people born outside Australia. This allowed for records to agree using broader regions rather than specific countries where information might disagree (e.g. 'Northern Europe' instead of 'England', 'Norway', etc.).

Year of arrival in Australia

Records that did not state a Year of arrival between 1896 and 2011 but had stated an age had a derived value created in the same manner as had been done for year of birth. Records that stated a Year of arrival and also stated they were born in Australia did not have the Year of arrival recoded to missing, as the birthplace may have been misreported.

Overall, the quality of both datasets was reasonably high, with most key variables containing less than 10% of values that were missing or invalid. There were some exceptions to this general finding, however, including the following (Note that the figures relate to post-standardised versions of the variables):

- 13% of all Death registrations (20,103 records) did not have a valid street number
- of Death registrations belonging to persons born overseas, 57% (27,036 records) did not have a valid Year of arrival.

In some cases, supplementary information was available to support the linkage, for example:

- of Death registrations that did not have a valid Street number, 71% (14,250 records) had stated a building name (such as a nursing home)
- of the 109 Death registrations without a valid street name, 85% (93 records) had stated a building name.

Table 2.2 displays the rate of missing values for standardised blocking and linking variables from Deaths and Census data. Note that the table does not account for other data quality issues such as unrepaired names and addresses.

Table 2.2 – MISSING DATA FOR KEY LINKING VARIABLES, Death registrations and Census

	DEATHS		CENSUS	
	no.	%	no.	%
Geographic information				
Street number	20 103	13.35	496 937	2.37
Street name	109	0.07	152 514	0.73
Suburb	115	0.07	143 295	0.68
Mesh Block	3	< 0.01	189 145	0.90
Personal information				
First name	5	< 0.01	85 032	0.41
Surname	2	< 0.01	125 868	0.60
Initial 4	7	< 0.01	135 222	0.65
Personal characteristics				
Sex	0	0.00	200 148	0.96
Age	1	< 0.01	109 332	0.52
Day of birth	5	< 0.01	2 008 275	9.60
Month of birth	5	< 0.01	2 009 294	9.60
Year of birth	5	< 0.01	98 834	0.47
Birthplace	385	0.25	626 072	2.99
Year of arrival(a)	133 199	86.80	15 874 099	75.85
Marital status(b)	4 001	2.61	4 179 771	6.85

(a) Includes persons born in Australia.

(b) Includes persons aged under 15.

2.2.2 – Blocking

Once data files have been standardised, record pairs (consisting of one record from each file) can be compared to see whether they are likely to be a match, i.e. belong to the same person. However, if the files are even moderately large, comparing every record on File A with every record on File B is computationally infeasible. Blocking reduces the number of comparisons by only comparing record pairs where matches are likely to be found – namely, records which agree on a set of blocking variables. Blocking variables are selected based on their reliability and discriminatory power. For instance, sex is partially useful as it is typically well reported, however it is minimally informative as it only divides datasets into two blocks, and is thus used in conjunction with other variables.

The process of blocking reduces the computational intensity of data linking. However, comparing only records that agree on a particular set of blocking variables means a record will not be compared with its match if it contains missing, invalid or legitimately different information on a blocking variable. To mitigate this, the linking process is repeated a number of times, using a range of different blocking strategies. For example, on the first pass, a block by a low level of geography (Mesh Block) was used to capture the majority of Death registrations that had matching addresses with their corresponding Census records. This means, however, that those Death registrations that had moved since being enumerated in the Census were not compared. Records which failed to link in the first pass proceeded to the next pass, in which a different set of blocking variables was used. For the second pass, by blocking on date of birth rather than geography, the Death registrations of people who had moved or who had missing or invalid address information were able to be compared.

Table 2.3 presents the blocking variables used for each pass. The strategy employed was similar to the approach used in the 2006 cycle, with some minor adjustments being made to the first four passes of the linkage run. Refer to *Linking Census Records to Death Registrations* (Cat. No. 1351.0.55.030) for the 2006 blocking and linking strategy.

Table 2.3 – BLOCKING VARIABLES, By pass number

Pass 1	**Mesh Block**
Pass 2	**Sex, Initial 4**
Pass 3	**Day, month and year of birth**
Pass 4	**Sex, Postcode**
Pass 5	**Indigenous status**

A more significant change to the 2011 blocking and linking strategy was the inclusion of a fifth pass, which involved linking any remaining unlinked Aboriginal and Torres Strait Islander deaths to the Census. In this pass, a modified Indigenous status variable was used for blocking, which enabled Aboriginal and Torres Strait Islander people on both datasets to be compared. This run was computationally feasible as it excluded all non-Indigenous Census records.

2.2.3 – Record pair comparison

Within a blocking pass, records on the two files which agree on the specified blocking variables are compared on a number of linking fields. Each linking field has associated *field weights,* which are calculated prior to comparison. Field weights indicate the amount of information (agreement, disagreement, or missing values) a linking field provides about whether the records belong to the same or a different person (true match status). Field weights are based on two probabilities associated with each linking field: first, the probability that the field values agree on a record pair given that the two records belong to the same person (match); and second, the probability that the field values agree on a record pair given the two records belong to different persons (unmatch). These are called m and u probabilities (or match and unmatch probabilities) and are defined below.

$$m = P(\text{fields agree} \mid \text{records belong to the same entity})$$

$$u = P(\text{fields agree} \mid \text{records belong to different entities})$$

Given that the m and u probabilities require knowledge of the true match status of record pairs, they cannot be known exactly, but rather must be estimated. The ABS uses a number of techniques to estimate m and u probabilities. For the series of 2011 linking projects, the Expectation Maximisation (EM) algorithm was used (see Samuels, 2012). In some instances the EM algorithm is deemed unsuitable, or fails to converge on an estimate, and in such cases m and u probabilities are based on those of similar linking projects. Note that m and u probabilities are calculated for each pass, conditional on agreement on the specified blocking fields, as all records compared will agree on blocking variables.

As a new feature to the suite of 2011 Census Data Enhancement projects, m and u probabilities for missing data on a linking field were calculated. These capture the probability that a pair belonging to the same individual (match) and a pair belonging to two different individuals (unmatch) are missing on either dataset (or both datasets) for a linking field. The m and u probabilities used in this project are presented in Appendix C: Linking m and u probabilities for each pass.

Match (m) and unmatch (u) probabilities are then converted to agreement, disagreement and missing field weights. The formulae to convert m and u probabilities to field weights are a small extension of the Fellegi and Sunter (1969) linking methodology to now provide weights for missing data.

They are as follows.

$$\text{Agree} = \log_2\left(\frac{m}{u}\right)$$

$$\text{Missing} = \log_2\left(\frac{m_{\text{missing}}}{u_{\text{missing}}}\right)$$

$$\text{Disagree} = \log_2\left(\frac{1 - m - m_{\text{missing}}}{1 - u - u_{\text{missing}}}\right)$$

These equations give rise to a number of intuitive properties of the Fellegi–Sunter framework. First, in practice agreement weights are always positive and disagreement weights are always negative. Second, the magnitude of the agreement weight is driven primarily by the likelihood of chance agreement. That is, a low probability of two random people agreeing on a field (for example, Date of Birth) will result in a large agreement weight applied when two records do agree. The magnitude of the disagreement weight is driven by the stability and reliability of a variable. That is, if a variable is well-reported and stable over time (for example, Sex) then disagreement on the variable will yield a large negative weight. For each record pair comparison, the field weights from each linking field are summed to form an overall record pair comparison weight.

Before calculating m and u probabilities for some variables it is first necessary to define what constitutes agreement. Typical comparison functions include:

- exact match (e.g. Sex). Agreement occurs only when the two field values are identical. This criterion is used for most linking fields
- approximate string comparison (e.g. Name). Two strings may be said to agree in spite of a certain proportion of missing, differing, or transposed characters, allowing for misspellings, transcriptions of poor handwriting, etc. Approximate string comparators allow for partial agreement if the strings being compared are similar but do not exactly match, and can be used to ensure that both identical and similar string pairs are defined to agree
- numeric difference (e.g. Age). A pair may be defined to agree if their field values differ by an amount less than or equal to a specified maximum difference.

For further details on comparison functions used for linkage, see Christen & Churches (2005).

Alternatively, near or partial agreement may be factored into the linking process by converting m and u probabilities to weights. For example, a person's age on equivalent records will frequently be an exact match, and the m and u probabilities are calculated based on this definition. During linkage, however, a partial agreement weight was given for ages within two years difference.

Table 2.4 displays the comparator types and tolerances applied to linking fields in this project. Comparator types were changed and tolerances were relaxed for some linking fields in later passes of the linkage, in order to broaden the search for remaining unlinked records.

Blocking fields, linking fields, comparator types, and *m* and *u* probabilities are input to linking software. Records which agree on the blocking variable(s) are compared on all linking fields.

Table 2.4 – KEY LINKING FIELDS, By comparator type and tolerance

	Comparator type and tolerance
Geographic information	
Street number	**Exact String**
Street name	**Approximate String, threshold value=0.85**
Suburb	**Approximate String, threshold value=0.85**
Personal information	
First name	**Approximate String, threshold value=0.85**
Surname	**Approximate String, threshold value=0.85**
Personal characteristics	
Sex	**Exact String**
Day of birth	**Exact String (Passes 1 & 2), Numeric Comparison with Absolute Tolerance ±2 (Passes 4 & 5)**
Month of birth	**Exact String**
Age	**Numeric Comparison with Absolute Tolerance ±1 (Passes 1 & 2) , ± 2 (Passes 4 & 5)**
Birthplace	**Exact String**
Year of arrival	**Numeric Comparison with Absolute Tolerance ±1 (Passes 1, 2 & 3) / ± 2 (Passes 4 & 5)**
Marital status	**Exact String**

2.2.4 – Decision model

A decision rule determines whether the record pair is linked, not linked or considered further as a possible link. The first phase of this process is automated, in which a record is assigned to its best possible pairing. This process is known as one-to-one assignment. Ideally (and often true in practice) each record has a single, obvious best pairing, which is its true match.

Linking projects in the ABS have typically used an auction algorithm to assign optimally one record on the first dataset to one record on the second dataset. The auction algorithm maximises the sum of all the record pair comparison weights through alternative assignment choices, such that if a record A1 on File A links well to records B1 and B2 on File B, but record A2 links well to B2 only, the auction algorithm will assign A1 to B1 and A2 to B2, to maximise the overall comparison weights for all record pairs.

The second phase of the decision rule stage takes the output of one-to-one assignment and decides which pairs should be retained as links, and which should be rejected as non-links. This is done by defining cut-off weights against which record pair comparison weights are evaluated. The simplest decision rule uses a single cut-off such that all record pairs with a weight greater than or equal to the cut-off are assigned as links, and all those pairs with a weight less than the cut-off are assigned as non-links. A more sophisticated decision rule was used in the Death registrations to Census linkage project and employs lower and upper cut-offs. Record pairs with a weight above the upper cut-off are declared links while those with a weight below the lower cut-off are declared non-links. The record pairs with weights between the upper and lower cut-offs are designated for clerical review.

Note that even where the original data is of very high quality, the information on equivalent records may not be identical across all the blocking and linking variables. For this reason, several 'passes' are used to optimise the opportunity for equivalent records to be linked, with different combinations of blocking and linking variables for each pass. Records on each dataset not linked on one pass are included in the pool of possible links for the next pass.

In clerical review, each record pair is manually inspected to resolve its match status. A clerical reviewer is often able to utilise information which cannot be captured in the automated comparison process, such as variations in names and common transcription errors (e.g. 1 and 7). Reviewed records are either accepted as links or rejected as non-links.

In order to establish the upper and lower cut-off values, a sample of the record pairs is clerically reviewed. This enables an estimate of the number of false links. In the 2011 Death registrations to Census linkage project the upper cut-offs were set at a weight value such that no false links were detected above the cut-offs. In the fifth pass neither sampling nor one-to-one assignment was used. Rather, all potential links for the remaining unlinked Aboriginal and Torres Strait Islander deaths were manually reviewed. In all passes, any record pair that included an Aboriginal and Torres Strait Islander death and had a link weight below the lower cut-off was also subjected to clerical review, regardless of the link weight.

Thus considerable resources were assigned to clerical review to ensure greater control over quality. This achieved:

- a reduction in the amount of false links–since a high upper clerical cut-off weight could be chosen before automatically assigning record pairs as links
- tailored clerical review–allowing for specific sub populations, such as potential Aboriginal and Torres Strait islander links, to be targeted.

Quality Assurance of clerically reviewed record pairs

Quality assurance (QA) techniques were applied to clerical review, which involved having a sample of the clerical record pairs reviewed a second time by a different reviewer. If a decision made to a QA record pair conflicted with the decision made in the original clerical review, this was identified as an *adjudication* pair.

Performing QA on clerically reviewed record pairs enabled a basic measure of quality, referred to as a *clerical review consistency rate (CR),* to be obtained. This rate is calculated by dividing the number of adjudication pairs against the total number of record pairs that were quality assured. Note that the CR is not strictly an estimate of clerical review accuracy, rather it is a measure of the level of consistency with which different coders applied decisions to record pairs. Neither the QA or Adjudication results were used to supplement the final linked results. Nevertheless, the fact that adjudication identified only 37 of the 3,000 record pairs that were quality assured (CR of 96.8%) gives a very positive indication of the accuracy of the clerical review process.

Table 2.5 – QUALITY ASSURED RECORD PAIRS, By pass number

	Pass 1	*Pass 2*	*Pass 3*	*Pass 4*	*Pass 5*	*Total*
Originally reviewed pairs	**12 088**	**6 058**	**6 717**	**1 500**	**1 738**	**28 101**
Quality assured pairs	**1 290**	**590**	**760**	**80**	**280**	**3 000**
Adjudicated pairs	**34**	**46**	**11**	**3**	**3**	**97**
Clerical review consistency rate (%)	**97.4**	**92.2**	**98.6**	**96.3**	**98.9**	**96.8**

Record pairs that were automatically assigned or clerically confirmed as links during a pass were not able to be linked again in any later passes. However, records from pairs that were deemed to be non-links were available to be linked again in later passes.

Linkage accuracy

Not all links are matches. That is, even where name and address are available, not all pairs assigned in a statistical linkage exercise result in a record pair belonging to the same individual. While the methodology is designed to ensure that the vast majority of links are true, some are nevertheless false.

In the 2006 cycle, statistical methods based on tolerances set in the linkage and clerical review process were used to estimate the proportion of false links. The false link rate for total deaths was estimated as 1.2%. Corresponding rates by Indigenous status were not calculated (see *Linking Census Records to Death Registrations,* cat. no. 1351.0.55.030).

As previously noted, greater resources were directed to clerical review in 2011 than in 2006 and higher clerical cut-offs were set. Cut-off weights were based on outcomes from clerically reviewing a random sample of record pairs. The upper cut-offs were chosen such that the number of estimated false links in each pass was zero. Thus the sampling estimate of link accuracy would be 100%. In reality there will be false links because:

- sample sizes were not large enough to detect the very small number of false links present
- some record pairs would have been wrongly assigned as matches in clerical review–while the clerical review staff made decisions in a highly consistent manner, even the small degree of inconsistency observed may have led to some false links being assigned.

While the number of false links is not able to be quantified precisely, the proportion is expected to be very small.

3.1 – LINKAGE RESULTS

At the completion of the linkage process, 142,697 out of the 153,455 death registrations in 2011-12 were linked to a 2011 Census record, comprising 1,884 Aboriginal and Torres Strait Islander deaths, 140,037 non-Indigenous deaths and 776 deaths where Indigenous status was not stated. Overall, 10,758 death registrations remained unlinked, of which 461 were identified as Aboriginal and Torres Strait Islander deaths (Tables 3.1 and 3.2).

Table 3.1 shows the number of records linked at each pass, and whether they were automatically assigned (above the upper cut-off) or confirmed through clerical review. Note that for all tables and analyses explored in this section, it is assumed that Indigenous status as reported in the Death Registration data is correct.

Table 3.1 – LINKAGE RESULTS, By Indigenous status and pass number

	Pass 1	*Pass 2*	*Pass 3*	*Pass 4*	*Pass 5(a)*	*Total*
ABORIGINAL AND TORRES STRAIT ISLANDER						
Automatically assigned links	1 161	360	13	0	0	1 534
Clerically reviewed record pairs	363	430	497	493	1 738	3 521
Confirmed as true links	146	112	38	25	29	350
Rejected as false links	217	318	459	468	1 709	3 171
Total links	1 307	472	51	25	29	1 884
NON-INDIGENOUS						
Automatically assigned links	106 927	22 853	851	31	0	130 662
Clerically reviewed record pairs	11 580	5 588	6 161	996	0	24 325
Confirmed as true links	4 127	2 659	2 335	254	0	9 375
Rejected as false links	7 453	2 929	3 826	742	0	14 950
Total links	111 054	25 512	3 186	285	0	140 037
TOTAL(b)						
Automatically assigned links	108 678	23 335	868	31	0	132 912
Clerically reviewed record pairs	12 022	6 058	6 714	1 500	1 738	28 032
Confirmed as true links	4 307	2 784	2 384	281	29	9 785
Rejected as false links	7 715	3 113	4 199	786	1 709	17 522
Total links	**112 985**	**26 119**	**3 252**	**312**	**29**	**142 697**

(a) Only death registrations for Aboriginal and Torres Strait Islander people were examined in Pass 5 of the linkage.

(b) Includes death registrations where the person's Indigenous status was not stated.

Of all links, about 76% were automatically assigned in the first pass of the linkage. Almost 7% of all links were confirmed through clerical review. Although numerically much smaller, the proportion of linked Aboriginal and Torres Strait Islander deaths confirmed through clerical review (19% or 350) was higher than that for linked non-Indigenous deaths (7% or 9,375). This is likely due to the fact that Aboriginal and Torres Strait Islander persons have higher rates of missingness and inconsistency on key linking variables, and are more geographically mobile. For more information, refer to *National best practice guidelines for data linkage activities relating to Aboriginal and Torres Strait Islander people, 2012*, available on the National Statistical Service website (www.nss.gov.au).

3.2 – CHARACTERISTICS OF LINKED AND UNLINKED DEATH REGISTRATIONS

The distribution of linked death registrations by age, sex, jurisdiction and remoteness were generally well aligned with those in the total death registration file. There were, however, some differences between linked and unlinked deaths and these resulted in a small amount of variation between linked and total deaths. Aboriginal and Torres Strait Islander male deaths were slightly under-represented in the linked file compared with the total file (52.6% compared with 54.6%), as were those aged under 50 years. In contrast, Aboriginal and Torres Strait Islander deaths for people aged 70 years and over were over-represented in the linked file (33.3% compared with 29.3% in the total file). There was some variation by jurisdiction, but overall the proportionate distribution within the linked file was within 2 percentage points of the corresponding distribution for each jurisdiction in the total file.

Table 3.2 – CHARACTERISTICS OF LINKED AND UNLINKED DEATH REGISTRATIONS, By Indigenous status

	ABORIGINAL AND TORRES STRAIT ISLANDER			NON-INDIGENOUS			TOTAL(a)		
	Linked (%)	*Unlinked (%)*	*Total (%)*	*Linked (%)*	*Unlinked (%)*	*Total (%)*	*Linked (%)*	*Unlinked (%)*	*Total (%)*
Sex									
Male	52.6	62.9	54.6	50.4	60.5	51.1	50.5	60.7	51.2
Female	47.4	37.1	45.4	49.6	39.5	48.9	49.5	39.3	48.8
Age group (years)									
0-19	4.4	7.8	5.1	0.6	1.4	0.6	0.6	1.7	0.7
20-29	4.6	10.2	5.7	0.8	3.5	0.9	0.8	3.9	1.0
30-39	6.8	16.3	8.7	1.2	5.1	1.4	1.3	5.6	1.6
40-49	12.6	21.5	14.4	2.6	7.5	2.9	2.7	8.1	3.1
50-59	18.5	16.3	18.1	5.9	11.2	6.2	6.1	11.5	6.4
60-69	19.7	15.0	18.8	11.8	14.7	12.0	11.9	14.7	12.1
70 and over	33.3	13.0	29.3	77.2	56.5	75.8	76.6	54.5	75.0
State / Territory of usual residence									
NSW	32.1	24.7	30.6	34.6	37.2	34.8	34.6	36.6	34.8
Vic.	3.7	4.8	3.9	25.2	23.3	25.1	24.8	22.3	24.6
Qld	27.6	19.1	25.9	18.0	19.4	18.1	18.2	19.6	18.3
SA	5.5	7.4	5.8	9.1	6.4	8.9	9.0	6.4	8.8
WA	14.6	22.3	16.2	8.8	9.1	8.8	8.9	9.7	8.9
Tas.	np	np	np	np	np	np	2.8	2.3	2.8
NT	13.9	21.3	15.4	0.3	1.1	0.3	0.5	2.0	0.6
ACT	np	np	np	np	np	np	1.2	1.1	1.2
Remoteness Area									
Major capital city	28.1	28.0	28.1	65.8	64.9	65.8	65.4	63.2	65.2
Inner regional	17.1	13.4	16.4	22.2	19.4	22.0	22.2	19.2	22.0
Outer regional	23.7	22.3	23.5	9.3	9.4	9.3	9.5	9.9	9.6
Remote	11.7	14.3	12.2	0.9	1.2	0.9	1.0	1.8	1.1
Very remote	15.4	17.1	15.7	0.3	0.3	0.3	0.5	1.0	0.5
Total(b) (%)	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**
Total(b)	**1 884**	**461**	**2 345**	**140 037**	**10 201**	**150 238**	**142 697**	**10 758**	**153 455**

np Not available for publication but included in totals where applicable, unless otherwise indicated.

(a) Includes Death registrations where the person's Indigenous status was not stated.

(b) Includes death registrations belonging to usual residents of 'Other Territories' and overseas visitors.

3.2.1 – Reasons for unlinked death registrations

There were two main reasons why death registrations were not linked to a Census record:

1. Records belonging to the same individual were present in the Death Registration and Census datasets but these records failed to be linked because they contained missing or inconsistent information.
2. A link was not possible because there was no Census record corresponding to the death registration as the person was missed from the Census.

Missing and/or inconsistent information

In these cases, the true record pair was present in the pool of possible links but it was not identified because there was a high level of inconsistency between information on the death registration and Census record or key linking fields were missing. As a special case, while a resident in a nursing home or hospital may have been included on a Census summary form for that institution, there would be insufficient information on such a form to establish a link to a death registration.

Address information was crucial in narrowing the search for and differentiating between true and false links. The fact that a person could relocate in the time period between Census enumeration and death increased the difficulty of linking the equivalent records to each other, and added uncertainty to clerical review.

Address information from the pool of linked records provides some insight into why the Aboriginal and Torres Strait Islander deaths may have been more difficult to link than non-Indigenous deaths. Linked Aboriginal and Torres Strait Islander death registrations had a higher rate of disagreement for geographic areas such as Mesh Block (30%) and SA2 (17%) than linked non-Indigenous death registrations (21% and 11% respectively).

Table 3.3 – LINKED DEATH REGISTRATIONS, Agreement and disagreement on geographic area of usual residence–By Indigenous status

	ABORIGINAL AND TORRES STRAIT ISLANDER		NON-INDIGENOUS		TOTAL(a)	
	no.	%	no.	%	no.	%
Mesh Block						
Agrees	**1 322**	**70.2**	**111 308**	**79.5**	**113 254**	**79.4**
Disagrees(b)	**562**	**29.8**	**28 729**	**20.5**	**29 443**	**20.6**
SA2						
Agrees	**1 573**	**83.5**	**124 702**	**89.0**	**126 966**	**89.0**
Disagrees(b)	**311**	**16.5**	**15 335**	**11.0**	**15 731**	**11.0**
State / Territory						
Agrees	**1 848**	**98.1**	**138 895**	**99.2**	**141 512**	**99.2**
Disagrees(b)	**36**	**1.9**	**1 142**	**0.8**	**1 185**	**0.8**
Total	**1 884**	**100.0**	**140 037**	**100.0**	**142 697**	**100.0**

(a) Includes Death registrations where the person's Indigenous status was not stated.

(b) Includes instances where the value was missing from one of the datasets.

No Census record

There are a number of reasons a person may not have completed or been enumerated on a Census form. The common reasons include:

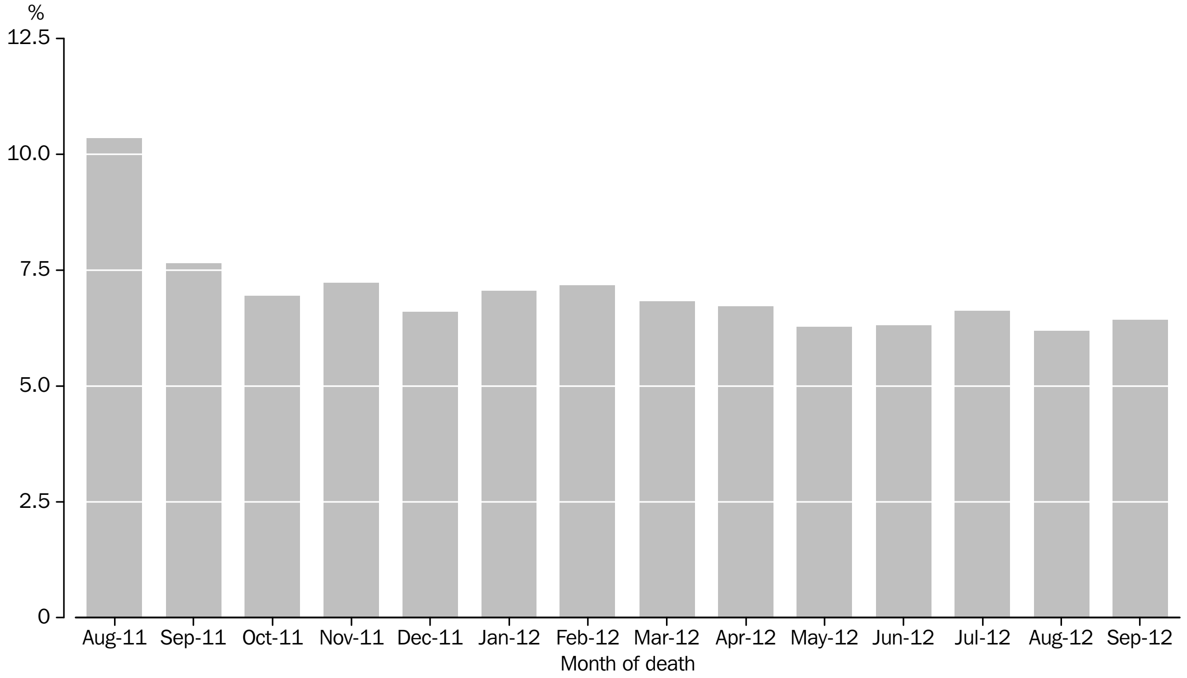
- they were travelling and were difficult to contact
- they mistakenly thought they were counted elsewhere
- there was insufficient space on the Census form in the household where they were staying and they did not obtain additional forms
- the person completing the form thought that certain people, for example, young babies, the elderly or visitors, need not be included
- they did not wish to be included due to concerns about confidentiality or a more general reluctance to participate
- the dwelling in which they were located was missed because it was difficult to find (e.g. in a remote or non-residential area)
- the dwelling in which they were located was mistakenly classed as unoccupied.

Some of these reasons may be more applicable than others for people who were alive at the time of the Census but subsequently died in the ensuing year. In a small number of cases, the absence of a Census form could be the result of the person being overseas at the time of the Census but subsequently dying in Australia and the death registered during the linkage reference period.

Rolling enumeration procedures for Census in remote Aboriginal and Torres Strait Islander communities may have increased the likelihood of an equivalent Census record not existing for deaths occurring around the time of the Census. Rolling enumeration involves conducting the Census over an extended period of four weeks. In these instances it is possible that a resident who moved may have been missed and therefore a corresponding Census record would not exist, or they may have passed away after Census Night (9 August 2011) but before Census enumeration was conducted in their residential area.

The pattern of linkage by month of death indicates that people who are close to death may well be missed, perhaps due to illness, or included in summary forms only on the Census. The months that contained the highest rate of unlinked records were those closest to Census Night: August (10%) and September (8%).

Graph 3.1 – UNLINKED DEATH REGISTRATIONS, By month of death.



Specially targeted linkages based only on address matching identified a small number of instances where a potential link to the death registration was found either on a Census summary form or on a Census household form but listed as a 'person temporarily absent' (PTA). A confirmed link could not be assigned in these instances, however, since insufficient personal information was available.

3.3 – RAW AND ADJUSTED LINKAGE RATES

The proportion of total death registrations that were linked to a Census record can be referred to as a raw linkage rate. Whilst it gives an initial estimate of overall linkage results, it does not take into account the number of death registrations for which there was no equivalent Census record.

Although it is impossible to decide on a case by case basis whether or not a Census record exists for a particular death registration, one way to model the overall likelihood is to use the ratio between the actual Census count and Estimated Resident Population (ERP). Under the assumption that people represented in the death registration file were subject to the same likelihood of being counted or missed in the Census as were the general population, this ratio can be applied to the total number of death registrations to provide an estimate of the number who had an equivalent Census record. The result can be referred to as expected links. The ratio between actual and expected links then provides an overall adjusted linkage rate.

To calculate the adjustment factor, the ratio of Census count to ERP by Indigenous status was applied to the death registration data by sex, age group and state/territory, as these variables are known to be critical in Census undercoverage. Due to the small number of Aboriginal and Torres Strait Islander deaths in Tasmania and the ACT, these jurisdictions were combined in the calculation and no separate adjusted linkage rates are available.

After applying the adjustment factor, the number of expected links was estimated at 150,266 (3,189 less than the total number of death registrations available for linking). Of the expected links, there were 2,099 that were for Aboriginal and Torres Strait Islander persons. For more information on the calculation of adjusted linkage rates for the 2011 Death registrations to Census linkage project, see Appendix A: Calculating Adjusted Linkage Rates.

For total death registrations in the 2006 and 2011 projects, the corresponding raw and adjusted linkage rates were remarkably similar. Note that a considerably higher number of death registrations were linked in 2011 than in 2006 (142,697 compared with 98,898).

In 2011, raw linkage rates were higher for non-Indigenous than Aboriginal and Torres Strait Islander deaths, but this difference was reduced after adjustment.

The raw linkage rate for Aboriginal and Torres Strait Islander deaths was higher in 2011 than 2006 (80% compared with 74%). The adjusted rate for Aboriginal and Torres Strait Islander deaths in 2011 was almost ten percentage points higher than the raw linkage rate (90% compared with 80%). Adjusted linkage rates were not calculated by Indigenous status in 2006.

Looking at this another way, although in total 461 death registrations for Aboriginal and Torres Strait Islander people remained unlinked at the end of the linkage process, the difference between expected and actual links was much lower at 215. Therefore, of unlinked Aboriginal and Torres Strait Islander deaths, just over half were likely to have remained unlinked because the Census record was present but not found and just under half because there was no Census record to link to.

Reflecting the pattern of Census undercount, the adjustment had a much larger effect on linkage rates for Aboriginal and Torres Strait Islander persons (nine percentage points) than non-Indigenous deaths (three percentage points).

Table 3.4 – RAW AND ADJUSTED LINKAGE RATES, 2006 and 2011, By Indigenous status

	ABORIGINAL AND TORRES STRAIT ISLANDER		NON-INDIGENOUS		TOTAL(a)	
	2006	*2011*	*2006*	*2011*	*2006*	*2011*
Total Death Registrations	**1 800**	**2 345**	**103 987**	**150 238**	**106 945**	**153 455**
Expected links(b)	**. .**	**2 099**	**. .**	**145 248**	**103 198**	**150 266**
Linked death registrations	**1 327**	**1 884**	**96 531**	**140 037**	**(c)97 696**	**142 697**
Unlinked death registrations	**473**	**461**	**7 456**	**10 201**	**8 047**	**10 758**
Raw linkage rate (%)	**73.7**	**80.3**	**92.8**	**93.2**	**92.5**	**93.0**
Adjusted linkage rate(d) (%)	**. .**	**89.7**	**. .**	**96.4**	**94.7**	**95.0**

. . Not applicable

(a) Includes Death registrations where the person's Indigenous status was not stated.

(b) Expected links for 2006 were calculated at different population levels relative to 2011. Care should be taken when making direct comparison between 2006 and 2011 adjusted rates.

(c) Excludes 1,202 linked death registrations in 2006 that were estimated to be false links.

(d) Referred to as a 'match-link rate' in 2006. See *Linking Census Records to Death Registrations* (cat. no. 1351.0.55.030) for more information.

For Aboriginal and Torres Strait Islander deaths, the largest adjustments were in the younger age groups (especially 20-29 years and 30-39 years) and for the Northern Territory. Similarly, for non-Indigenous deaths, the adjustment for the Northern Territory was higher than that for other jurisdictions.

While results by Indigenous status for Tasmania and the ACT are not published in order to protect confidentiality, high raw linkage rates were obtained for both Aboriginal and Torres Strait Islander and non-Indigenous deaths in these jurisdictions. Detailed results for Aboriginal and Torres Strait Islander deaths by sex, age group and jurisdiction are presented in Appendix B: Raw linkage rates for Aboriginal and Torres Strait Islander Death registrations.

Raw linkage rates for Aboriginal and Torres Strait Islander deaths were reasonably consistent across remoteness areas, although slightly higher in major cities and regional areas than in remote areas. Adjustment factors by remoteness were not available at the time of writing for calculation of adjusted linkage rates.

Table 3.5 – SELECTED CHARACTERISTICS FOR LINKED RECORDS, By raw and adjusted linkage rates–By Indigenous status

	ABORIGINAL AND TORRES STRAIT ISLANDER		NON-INDIGENOUS		TOTAL(a)	
	Raw linkage rate (%)	*Adjusted linkage rate (%)*	*Raw linkage rate (%)*	*Adjusted linkage rate (%)*	*Raw linkage rate (%)*	*Adjusted linkage rate (%)*
Sex						
Males	77.4	87.9	92.0	95.3	91.7	93.7
Females	83.9	92.0	94.5	97.6	94.4	95.3
Age group (years)						
0-19	69.7	82.2	84.6	86.5	83.0	84.0
20-29	64.7	82.7	74.5	82.5	73.6	81.2
30-39	63.2	75.9	76.1	80.4	75.0	78.7
40-49	70.6	82.9	82.6	85.5	81.8	84.4
50-59	82.3	94.1	87.7	90.7	87.5	90.2
60-69	84.3	92.1	91.7	94.7	91.5	94.0
70 and over	91.3	94.6	94.9	98.1	94.9	95.9
State or Territory of usual residence						
NSW	84.1	91.4	92.7	96.2	92.6	95.0
Vic.	75.8	83.1	93.7	96.4	93.7	95.0
Qld	85.5	94.5	92.7	95.7	92.5	94.1
SA	75.2	85.1	95.1	97.8	94.9	96.2
WA	72.8	83.6	93.0	97.5	92.3	95.7
Tas.(b)	np	. .	np	. .	94.3	. .
NT	72.8	85.3	77.8	85.5	75.7	83.8
ACT(b)	np	. .	np	. .	93.6	. .
Remoteness Area(b)						
Major capital city	80.4	. .	93.3	. .	93.2	. .
Inner regional	83.9	. .	94.0	. .	93.9	. .
Outer regional	81.3	. .	93.2	. .	92.7	. .
Remote	76.9	. .	90.8	. .	88.3	. .
Very remote	78.6	. .	93.2	. .	85.8	. .
Total(c)	**80.3**	**89.7**	**93.2**	**96.4**	**93.0**	**95.0**

. . Not applicable.

np Not available for publication but included in totals where applicable, unless otherwise indicated.

(a) Includes death registrations where the person's Indigenous status was not stated.

(b) Separate adjusted linkage rates not calculated for Tasmania, Australian Capital Territory and Remoteness Areas.

(c) Includes death registrations from Tasmania, Australian Capital Territory, Other Territories and death registrations that did not report an age, sex or state of residence.

REFERENCES

Australian Bureau of Statistics (2008) *Census Data Enhancement – Indigenous Mortality Quality Study, 2006-07*, cat. no. 4723.0, ABS, Canberra.

(2009) *Experimental Life Tables for Aboriginal and Torres Strait Islander Australians, 2005-07,* cat. no. 3302.0.55.003, ABS, Canberra.

(2010) *Census Data Enhancement Project: An Update, Oct 2010*, cat. no. 2062.0, ABS, Canberra.

(2012) *Census of Population and Housing: Details of Undercount, August 2011,* cat no. 2940.0, ABS, Canberra.

(2013) *Australian Demographic Statistics, Sep 2012*, cat. no. 3101.0, ABS, Canberra.

Australian Institute of Health and Welfare and Australian Bureau of Statistics (2012). *National best practice guidelines for data linkage activities relating to Aboriginal and Torres Strait Islander people,* cat. no. IHW 74. AIHW, Canberra.

Cross Portfolio Statistical Integration Committee (2010), *High Level Principles for Data Integration Involving Commonwealth Data for Statistical and Research Purposes*, CPSIC, Canberra.

Christen, P., Churches, T., and Hegland, M. (2004) "Febrl – A Parallel Open Source Data Linkage System", *Proceedings of the 8th Pacific-Asia Conference, PAKDD 2004, Sydney, Australia*, pp. 638-647.

Christen, P. and Churches, T. (2005) *Febrl 0.3 Documentation*, (last viewed on 8 April 2013), <http://cs.anu.edu.au/~Peter.Christen/Febrl/febrl-0.3/febrldoc-0.3/>

Conn, L. and Bishop, G. (2006) "Exploring Methods for Creating a Longitudinal Census Data Set", *Methodology Advisory Committee Papers*, cat. no. 1352.0.55.076, Australian Bureau of Statistics, Canberra.

Fellegi, Ivan P. and Sunter, Alan B. (1969) "A Theory for Record Linkage", *Journal of the American Statistical Association*, 64(328), pp. 1183–1210.

Samuels, C. (2012) "Using the EM Algorithm to Estimate the Parameters of the Fellegi–Sunter Model for Data Linking", *Methodology Advisory Committee Papers*, cat. no. 1352.0.55.120, Australian Bureau of Statistics, Canberra.

Steering Committee for the Review of Government Service Provision (2011), *Overcoming Indigenous Disadvantage: Key Indicators 2011*, Productivity Commission, Canberra.

Wright J. (2010) "Linking Census Records to Death Registrations" *Methodology Research Papers*, cat. no. 1351.0.55.030, Australian Bureau of Statistics, Canberra

APPENDIX A: CALCULATING ADJUSTED LINKAGE RATES

In order to produce indicative linkage rates that take account of death registrations for which there was no equivalent Census record, a Census undercount factor was applied to death registrations. This factor was based on the ratio between Census count and Estimated Resident Population (ERP) and applied by Indigenous status, age group, sex and state/territory (where Tasmania and the ACT were combined due to the small number of deaths). The result was an estimate of the (maximum) number of links that could be expected from the death registration data.

A number of adjustments are made to the Census count to produce ERP by Indigenous status. These are informed by the results of the Census Post Enumeration Survey conducted about six weeks after the Census and include an adjustment for net undercount, the imputation of Indigenous status where it is not stated in the Census (both item and unit non-response[1]), a component for mis-classification of Indigenous status and an adjustment for Australian residents temporarily overseas at the time of the Census (RTOs). For more information, see *Census of Population and Housing: Details of Undercount, August 2011* (cat no. 2940.0.) and *Australian Demographic Statistics, September 2012* (cat. no. 3101.0).

For this project, two minor modifications were made to the ERP prior to calculating the linkage adjustment factor. Both of these slightly decrease the magnitude of the linkage adjustment factor from what it would otherwise have been based on the actual ERP.

First, for total deaths, the RTO adjustment was removed from the ERP. That is, it was assumed that, on balance, people temporarily overseas at the time of the Census were unlikely to be represented in the death registration data. While there may well have been a small number of instances in which this actually occurred, it was assumed that the full RTO component would overstate the level of adjustment required.

An additional modification was made to the ERP when the ratio was applied to deaths by Indigenous status. The number of imputed Census records for which Indigenous status was not stated (item non-response only) was removed from the ERP. A not stated category for Indigenous status is included in both death registration and Census data but not the ERP. Therefore, the ratio of Census to ERP by Indigenous status may overstate the amount of undercount in death registrations if the imputation for item non-response were not removed.

The expected number of links from the death registration data was calculated as follows:

$$\text{Expected links}=\text{Death registrations } \times \frac{\text{Census count}}{\text{Modified ERP}}$$

The adjusted linkage rate is then calculated as the number of links as a proportion of expected links:

$$\text{Adjusted linkage rate} = \frac{\text{Linked death registrations}}{\text{Expected links}}$$

[1] Item non response refers to people for whom a Census form was returned but Indigenous status was not stated. Unit non-response refers to Census records imputed for people in households that were identified as occupied at the time of collection but from whom no Census form was returned.

Note that this method assumes that the Census undercount rates (by Indigenous status, age group, sex and state of residence) for the general population are the same as for the population who died in the period 10 August 2011 to 27 September 2012.

As expected, the adjustments to death registration data were broadly aligned with estimates of Census undercount. Based on the results of the Post Enumeration Survey, the 2011 Census net undercount was estimated at 17.2% for the Aboriginal and Torres Strait Islander population and 6.2% for non-Indigenous persons. Comparable figures for the increase of actual death registrations over expected links were 10.5% for Aboriginal and Torres Strait Islander deaths and 3.3% for non-Indigenous deaths. Differences between the two sets of figures relate both to the minor modification to the ERP used for the adjustment factors, and the different population structure for death registrations compared with the total population.

Table A.1 – ADJUSTMENT OF DEATH REGISTRATIONS, By Indigenous status

	Aboriginal and Torres Strait Islander	*Non-Indigenous*	*Total(a)*
2011 PES population estimate(b)	**662 335**	**21 216 926**	**21 879 261**
2011 Census count	**548 147**	**19 898 127**	**(c)21 504 721**
Difference (no.)	**114 188**	**1 318 799**	**374 540**
Difference – Census undercount (%)	**17.2**	**6.2**	**1.7**
Death Registrations	**2 345**	**150 238**	**153 455**
Expected links	**2 099**	**145 248**	**150 266**
Difference (no.)	**246**	**4 990**	**3 189**
Difference (%)	**10.5**	**3.3**	**2.1**

(a) Includes death registrations that did not state an Indigenous status

(b) Population as estimated by the Post Enumeration Survey (PES) as of Census night (9th of August 2011)

(c) Includes imputed persons in non-responding dwellings

While the calculation of adjusted linkage rates performed here is similar to the method used in 2006, other assumptions could be made about the likelihood of death registrations lacking a corresponding Census record. The ABS is undertaking further research to examine the effects of differential reporting of Indigenous status on raw and adjusted linkage rates.

APPENDIX B: RAW LINKAGE RATES FOR ABORIGNAL AND TORRES STRAIT ISLANDER DEATH REGISTRATIONS

The tables presented below provide a demographic breakdown of the raw linkage rates for Aboriginal and Torres Strait Islander persons. Aggregated tables for New South Wales, Victoria, Queensland, South Australia, Western Australia and the Northern Territory have also been provided. Tables for Tasmania and the Australian Capital Territory have not been published in order to protect confidentiality. Refer to section 3.3 of this paper for a description of raw linkage rates.

Table B.1 – RAW LINKAGE RATES FOR ABORIGINAL AND TORRES STRAIT ISLANDER DEATH REGISTRATIONS, By sex and age groups–Australia.

	Total Death registrations	*Linked Death Registrations*	*Raw linkage rate*
	(no.)	(no.)	(%)
MALES			
Age group (years)			
0-19	71	50	70.4
20-29	94	61	64.9
30-39	128	73	57.0
40-49	202	132	65.3
50-59	234	191	81.6
60-69	251	206	82.1
70 and over	301	278	92.4
Total	1 281	991	77.4
FEMALES			
0-19	48	33	68.8
20-29	39	25	64.1
30-39	76	56	73.7
40-49	135	106	78.5
50-59	190	158	83.2
60-69	189	165	87.3
70 and over	387	350	90.4
Total	1 064	893	83.9
Persons	**2 345**	**1 884**	**80.3**

Table B.2 – RAW LINKAGE RATES FOR ABORIGINAL AND TORRES STRAIT ISLANDER DEATH REGISTRATIONS, By sex and age groups– New South Wales.

	Total Death registrations	*Linked Death Registrations*	*Raw linkage rate*
	(no.)	(no.)	(%)
	MALES		
Age group (years)			
0-39	68	41	60.3
40-59	118	89	75.4
60 and over	184	162	88.0
Total	370	292	78.9
	FEMALES		
0-39	44	36	81.8
40-59	97	83	85.6
60 and over	207	193	93.2
Total	348	312	89.7
Persons	**718**	**604**	**84.1**

Table B.3 – RAW LINKAGE RATES FOR ABORIGINAL AND TORRES STRAIT ISLANDER DEATH REGISTRATIONS, By sex and age groups–Victoria.

	Total Death registrations	*Linked Death Registrations*	*Raw linkage rate*
	(no.)	(no.)	(%)
	MALES		
Age group (years)			
0-49	14	8	57.1
50 and over	33	28	84.8
Total	47	36	76.6
	FEMALES		
0-49	15	9	60.0
50 and over	29	24	82.8
Total	44	33	75.0
Persons	91	69	75.8

Table B.4 – RAW LINKAGE RATES FOR ABORIGINAL AND TORRES STRAIT ISLANDER DEATH REGISTRATIONS, By sex and age groups–Queensland.

	Total Death registrations	*Linked Death Registrations*	*Raw linkage rate*
	(no.)	(no.)	(%)
	MALES		
Age group (years)			
0-39	78	57	73.1
40-59	116	94	81.0
60 and over	162	148	91.4
Total	356	299	84.0
	FEMALES		
0-39	33	23	69.7
40-59	67	60	89.6
60 and over	152	138	90.8
Total	252	221	87.7
Persons	**608**	**520**	**85.5**

Table B.5 – RAW LINKAGE RATES FOR ABORIGINAL AND TORRES STRAIT ISLANDER DEATH REGISTRATIONS, By sex and age groups–South Australia.

	Total Death registrations	*Linked Death Registrations*	*Raw linkage rate*
	(no.)	(no.)	(%)
	MALES		
Age group (years)			
0-39	18	9	50.0
40 and over	61	48	78.7
Total	79	57	72.2
	FEMALES		
0-39	13	8	61.5
40 and over	45	38	84.4
Total	58	46	79.3
Persons	**137**	**103**	**75.2**

Table B.6 – RAW LINKAGE RATES FOR ABORIGINAL AND TORRES STRAIT ISLANDER DEATH REGISTRATIONS, By sex and age groups–Western Australia.

	Total Death registrations	*Linked Death Registrations*	*Raw linkage rate*
	(no.)	(no.)	(%)
	MALES		
Age group (years)			
0-39	50	26	52.0
40-59	63	40	63.5
60 and over	80	69	86.3
Total	193	135	69.9
	FEMALES		
0-39	27	15	55.6
40-59	64	46	71.9
60 and over	95	80	84.2
Total	186	141	75.8
Persons	**379**	**276**	**72.8**

Table B.7 – RAW LINKAGE RATES FOR ABORIGINAL AND TORRES STRAIT ISLANDER DEATH REGISTRATIONS, By sex and age groups–Northern Territory.

	Total Death registrations	*Linked Death Registrations*	*Raw linkage rate*
	(no.)	(no.)	(%)
	MALES		
Age group (years)			
0-39	67	45	67.2
40-59	83	58	69.9
60 and over	58	45	77.6
Total	208	148	71.2
	FEMALES		
0-39	38	25	65.8
40-59	60	47	78.3
60 and over	54	42	77.8
Total	152	114	75.0
Persons	**360**	**262**	**72.8**

APPENDIX C: LINKING *m* AND *u* PROBABILITIES FOR EACH PASS

The following tables present the m and u probabilities that were applied to linking fields in each pass of the 2011 Death registrations to Census linkage project. Refer to section 2.2.3 of this paper for an explanation of *m* and *u* probabilities.

Table C.1 – KEY LINKING VARIABLES, By *m* and *u* probabilities for linking fields–Pass 1.

	AGREE		DISAGREE		MISSING	
	m	*u*	*m*	*u*	*m*	*u*
First name	0.839	0.006	0.158	0.989	0.003	0.005
Surname	0.954	0.007	0.041	0.987	0.006	0.007
Sex	0.966	0.512	0.017	0.476	0.017	0.012
Day of birth	0.903	0.030	0.039	0.890	0.058	0.080
Month of birth	0.923	0.077	0.019	0.844	0.058	0.080
Age	0.975	0.051	0.019	0.942	0.006	0.007
Street number	0.751	0.112	0.047	0.595	0.202	0.293
Street name	0.881	0.389	0.118	0.573	0.001	0.038
Birthplace	0.900	0.525	0.015	0.414	0.085	0.061
Year of arrival	0.079	0.002	0.020	0.036	0.901	0.962
Marital status	0.879	0.293	0.069	0.534	0.052	0.173

Table C.2 – KEY LINKING VARIABLES, By *m* and *u* probabilities for linking fields–Pass 2.

	AGREE		DISAGREE		MISSING	
	m	*u*	*m*	*u*	*m*	*u*
First name	0.939	0.350	0.061	0.650	0.000	0.000
Surname	0.993	0.109	0.007	0.891	0.000	0.000
Day of birth	0.922	0.030	0.030	0.881	0.048	0.089
Month of birth	0.937	0.076	0.015	0.835	0.048	0.089
Age	0.979	0.024	0.017	0.972	0.004	0.004
Street number	0.634	0.010	0.103	0.824	0.262	0.165
Street name	0.767	0.001	0.228	0.992	0.005	0.007
Suburb	0.871	0.001	0.122	0.992	0.007	0.006
Birthplace	0.901	0.551	0.014	0.418	0.084	0.030
Year of arrival	0.074	0.001	0.019	0.027	0.907	0.971
Marital status	0.880	0.241	0.069	0.565	0.051	0.194

Table C.3 – KEY LINKING VARIABLES, By *m* and *u* probabilities for linking fields–Pass 3.

	AGREE		DISAGREE		MISSING	
	m	*u*	*m*	*u*	*m*	*u*
First name	0.861	0.008	0.135	0.989	0.004	0.003
Surname	0.957	0.001	0.037	0.993	0.006	0.005
Sex	0.979	0.495	0.242	0.493	0.015	0.013
Street number	0.634	0.010	0.135	0.824	0.262	0.165
Street name	0.752	0.001	0.018	0.993	0.006	0.006
Suburb	0.856	0.001	0.014	0.992	0.008	0.007
Birthplace	0.895	0.460	0.072	0.492	0.090	0.048
Year of arrival	0.074	0.003	0.005	0.038	0.908	0.959
Marital status	0.874	0.411	0.103	0.534	0.055	0.055

Table C.4 – KEY LINKING VARIABLES, By *m* and *u* probabilities for linking fields–Pass 4.

	AGREE		DISAGREE		MISSING	
	m	*u*	*m*	*u*	*m*	*u*
First name	0.909	0.007	0.087	0.990	0.004	0.003
Surname	0.931	0.002	0.063	0.993	0.006	0.005
Day of birth	0.923	0.142	0.024	0.765	0.053	0.092
Month of birth	0.925	0.076	0.023	0.832	0.053	0.092
Age	0.983	0.029	0.012	0.966	0.005	0.005
Street name	0.812	0.015	0.185	0.981	0.003	0.003
Birthplace	0.898	0.510	0.016	0.466	0.086	0.024
Year of arrival	0.079	0.002	0.015	0.043	0.906	0.955
Marital status	0.877	0.220	0.072	0.568	0.051	0.212

Table C.5 – KEY LINKING VARIABLES, By *m* and *u* probabilities for linking fields–Pass 5(a).

	AGREE		DISAGREE	
	m	*u*	*m*	*u*
First name	0.909	0.008	0.091	0.992
Surname	0.931	0.002	0.069	0.999
Sex	0.976	0.491	0.024	0.509
Day of birth	0.923	0.142	0.077	0.858
Month of birth	0.925	0.076	0.075	0.924
Age	0.983	0.029	0.017	0.971
Street number	0.605	0.011	0.395	0.989
Street name	0.758	0.001	0.243	0.999
Suburb	0.868	0.001	0.132	0.999
Birthplace	0.898	0.467	0.102	0.533
Year of arrival	0.079	0.004	0.921	0.996
Marital status	0.877	0.22	0.123	0.78

(a) m_{missing} and u_{missing} not applied to the linking fields used in this pass (i.e. missing weights set at 0).

FOR MORE INFORMATION . . .

www.abs.gov.au the ABS website is the best place for data from our publications and information about the ABS.

INFORMATION AND REFERRAL SERVICE

Our consultants can help you access the full range of information published by the ABS that is available free of charge from our website. Information tailored to your needs can also be requested as a 'user pays' service. Specialists are on hand to help you with analytical or methodological advice.

POST Client Services, ABS, GPO Box 796, Sydney NSW 2001

FAX 1300 135 211

EMAIL client.services@abs.gov.au

PHONE 1300 135 070

FREE ACCESS TO STATISTICS

All ABS statistics can be downloaded free of charge from the ABS web site.

WEB ADDRESS www.abs.gov.au

Produced by the Australian Bureau of Statistics