1351.0.55.030

Australian Bureau of Statistics

Research Paper

Linking Census Records to Death Registrations

www.abs.gov.au

New Issue

Research Paper

Linking Census Records to Death Registrations

Jeffrey Wright

Analytical Services Branch

AUSTRALIAN BUREAU OF STATISTICS

EMBARGO: 11.30 AM (CANBERRA TIME) THURS 11 MAR 2010

ABS Catalogue no. 1351.0.55.030

Produced by the Australian Bureau of Statistics

INQUIRIES

The ABS welcomes comments on the research presented in this paper. For further information, please contact the Director, Analytical Services Branch on Canberra (02) 6252 6679 or email <analytical.services@abs.gov.au>.

CONTENTS

LINKING CENSUS RECORDS TO DEATH REGISTRATIONS

Jeffrey Wright
Analytical Services

ABSTRACT

In order to gain a better understanding of the extent of Indigenous identification in mortality data, the Australian Bureau of Statistics (ABS) linked 2006 Census data to death registrations to compare the reported Indigenous status from each dataset. Data linking was conducted by authorised ABS officers during the Census processing period when name and address were available to be used as linking variables. After Census processing, all Census names and addresses held by the ABS were destroyed. This data linking project is referred to as the Indigenous Mortality Quality Study, which forms part of the broader Census Data Enhancement project.

This paper builds on other papers already released about the Indigenous Mortality Quality Study, by elaborating on the probabilistic data linking methodology used to link the Census and death records. An evaluation of the linkage is also provided.

1. INTRODUCTION

In the lead up to the 2006 Census, the ABS initiated the Census Data Enhancement (CDE) project. The primary objective of the CDE Project is to enhance the value of the Census by combining it with future censuses and other datasets. The centrepiece of this project is the creation of a Statistical Longitudinal Census Dataset (SLCD), formed by selecting a 5% random sample of records from the 2006 Census, that will subsequently be combined with data from future Censuses. The aim is to bring these datasets together using statistical techniques that do not require name and address. It is envisaged that the SLCD will create a comprehensive picture of Australian society for researchers to investigate issues around improving family well being, employment, health, education outcomes and transitions.

Another key feature of the CDE project is to conduct *quality studies*, in which the full Census dataset is linked to other specified datasets, with the aim of improving ABS statistical outputs. One such quality study that has been conducted is the Indigenous Mortality Quality Study. The aim of the Indigenous Mortality Quality Study was to gain a better understanding of the extent of Indigenous identification in mortality data. This was achieved by linking 2006 Census data to death registrations, and then examining differences in reporting of Indigenous status across the two datasets. Some results and analysis from this quality study have already been released in other ABS publications (2008a, 2008b, 2009a), however the description of the data linking methodology was not comprehensive. This paper is written from a methodological perspective and provides a more thorough description of the methods and processes used to link the Census data to death registrations.

Further general information about the Census Data Enhancement project is available in ABS publications (2005a, 2005b, 2006a).

2. THE INDIGENOUS MORTALITY QUALITY STUDY

The quality of Indigenous mortality data is recognised as having significant limitations. While virtually all deaths are registered (non-Indigenous and Indigenous), it is believed that a significant percentage of Indigenous deaths are not identified as Indigenous on the death record. This problem of under-identification also varies for different states and territories, which further complicates and limits the usefulness of Indigenous mortality data for reporting and analysis, but also as a key input into population estimates and projections, and life expectancy estimates.

This quality study aimed to improve these key statistical outputs by gaining a better understanding of the extent of the under-identification issue, through linking Census records to death registrations and then comparing the reported Indigenous status from both datasets. Specific aims of the Indigenous Mortality Quality Study were to:

- assist in understanding the differences in recording of Indigenous status between death registration and Census data;
- assess the under-identification of Indigenous deaths in death registrations;
- identify factors that may be contributing to under-identification of Indigenous deaths in death registrations; and
- assess the feasibility of calculating and applying adjustment factors to improve estimates of Indigenous mortality.

The data linking was conducted during the 2006 Census processing period when name and address were available to be used as linking variables. After Census processing, all Census names and addresses held by the ABS were destroyed.

As mentioned in Section 1, results and analysis from this quality study have already been released in other ABS publications (2008a, 2008b, 2009a).

3. OVERVIEW OF THE LINKING PROCESS

3.1 General description of methodology

The aim of the data linking in this quality study was to bring together records that belong to the same person from the death registrations and the Census data. The linking methodology used to link the two files was probabilistic linking. This type of linking is used when there is partial identifying information, but no unique, error free, identifying key. The methodology links records from two files using variables common to both files.

A key feature of this methodology is its ability to handle a variety of linking variables (e.g. character string, numeric) and variable comparison methods to produce a single numerical measure of how well two particular records agree. This measure of agreement allows a link status to be assigned to record-pairs. Link status is defined as the status assigned via the data linking methodology; the possibilities being either a link or non-link. Link status is different to match status. Match status is defined as the true status of a record-pair; the possibilities being either a match or a non-match. A match means that the records belong to the same person. A non-match means that the records belong to different people. Match status is typically unobserved. Ideally, the data linking methodology will assign a link status that aligns with the match status. However, there will be cases in which the data linking methodology incorrectly assigns a record-pair as a link or non-link. The comparison of link and match status forms the basis of data linking quality measures that are further discussed in Section 6.4. The terms 'link' and 'match' are used throughout this report, so it is important to remember how they differ.

When assigning a link status using the data linking methodology, some records may have a high level of agreement with more than one record from the other file. To address this issue, an algorithm was applied to choose an optimal set of unique record-pairs (Christen and Churches, 2005). Using the algorithm, a record from one file could not be linked to more than one record from the other file. This is called one-to-one assignment of record-pairs.

In this paper, the linking methodology has been broken down into the following steps:

- blocking strategy;
- linking variables and comparison functions;
- input probabilities;
- choosing cut-off weights; and
- clerical review.

Each of these steps is discussed in detail in Section 5.

An important stage of the data linking process is the preparation of the two datasets. Preparation includes a number of steps such as verification, removing inconsistencies, and parsing text fields resulting in standardised files. This data preparation takes place against a background of determining which variables will be used as linking variables. Section 4 describes the data in this quality study and its preparation for use in linking.

For a detailed description of the theory underlying probabilistic data linking, see Conn and Bishop (2006).

3.2 Software and hardware

The linking software chosen for use in this project was Febrl (Christen and Churches, 2005). The release version used was Febrl 0.3, and it was released under an open source licence. The ABS made significant changes to Febrl 0.3 to improve the speed of access to records and to add provision for clerical review and acceptance sampling.

The hardware was designed to cater for the memory-intensive requirements of Febrl, and consisted of a server which had four 2.8GHz dual core AMD Opteron processors with 64 GB RAM, 250 GB hard disk, and running a 64-bit Windows 2003 Server operating system.

Operational arrangements for managing data flows within the ABS included restricting access to personal information through functional separation of roles. In particular the process of bringing together two datasets and analysis of linked datasets were separated. The unit responsible for linking the datasets did not have access to the full original sources, nor the resulting combined dataset. They were responsible for determining the best linkages and providing a key that would enable the selected datasets to be brought together.

3.3 Impact of quality study objectives on linking process

Before setting some of the linking parameters for this quality study, it was important to consider how the linked file would be used. It was to be used for understanding the differences in reported Indigenous status between the death registrations and Census data. Once a link had been established, there was a two-way matrix of possible Indigenous reporting permutations, as presented in table 3.1.

3.1 Matrix of possible Indigenous reporting permutations

	Census 2006		
	Indigenous	*Non-Indigenous*	*Not stated*
Deaths			
Indigenous	—	—	—
Non-Indigenous	—	—	—
Not stated	—	—	—

The Indigenous reporting permutations from the linked file would then be used to investigate a number of analytical questions regarding the reporting of Indigenous status. Thus two important issues to consider when performing the linking for this quality study were:

- The two datasets needed to be linked in a way that was independent of reported Indigenous status, so that any future analysis would not be affected by bias introduced in the linking process; and
- It was important to ensure that a high level of agreement between records was necessary before a link was assigned, because it was the differences in Indigenous status for individual record-pairs that was of primary analytical interest.

4. THE DATA

4.1 Timing issues

2006 Census records were linked to death registration records for deaths that occurred after the 2006 Census, which was conducted on 8 August 2006.

To enable the use of name and address as linking variables, linking had to be performed during the Census processing period, which ended in late October 2007. This was the only time name and address would be available on the Census dataset. Furthermore, the number of death records to be linked needed to be maximised to ensure that the linked dataset had enough observations so that meaningful analysis could be performed. This meant waiting for the latest update of death registrations while still allowing enough time for the linkage to be performed before the Census processing period finished.

In the end, the scope of deaths [1] used in the linking process were deaths that had occurred during the period 9 August 2006 to 30 June 2007. Note that some deaths are not registered and processed immediately for various reasons, so the data extracted for linking would not have full coverage of deaths within scope. The most serious case of undercoverage was for deaths registered in Victoria, because only deaths that had been registered by the end of April were available for this quality study.

4.2 Variables common to both datasets

The full lists of variables available on each dataset were compared and the variables in table 4.1 were identified as potentially useful for linking.

Mesh blocks are a new building block of statistical and administrative geography that was introduced with the 2006 Census. There are in excess of 340,000 Mesh blocks covering the whole of Australia, and they may contain a residential area, an administrative area such as Parliament House, or a geographic feature such as a national park. A residential Mesh block typically contains 30 to 60 dwellings. Mesh blocks serve as a useful geographical indicator to be used in linking. Mesh blocks for the 2006 Census were experimental, as indicated in an ABS information paper (2008c). The more complete and robust introduction of rural addressing standards in non-metroplitan areas, principally for emergency management purposes, will improve the degree of mesh block coding in 2011.

1 The scope of deaths in this paper has been defined based on the stated aims of this quality study, as listed in Section 2. Analyses that use the results of this quality study, such as ABS publications (2008a, 2008b, 2009a), may define the scope of deaths differently to what has been defined in this paper, depending on how the linked data is being used.

4.1 Variables common to both datasets

Name	*Personal characteristics*	*Geographic*
First name	Day of birth (DD)	Mesh block (MB)
Surname	Month of birth (MM)	Statistical Local Area (SLA)
	Full Date of birth (DDMMYYYY)	Street number
	Age	Street name
	Sex	Suburb
	Place of birth	Postcode
	Year of arrival	
	Marital status	
	Number of children (females only)	

Note – *Indigenous status* was common to both datasets, but was not used as a linking variable.

Statistical local areas (SLA) are a broader statistical geography than Mesh blocks. In the 2006 *Australian Standard Geographical Classification* (ASGC), there were 1,426 SLAs in Australia. In regards to the usefulness of SLAs for linking purposes, they provide less identifying power compared to Mesh blocks. However, SLAs were still considered useful in cases where a record did not have enough address information to be coded to a Mesh block.

4.3 Census data issues

The 2006 Census file used for this study consisted of 19,046,302 records, excluding overseas visitors and imputed persons. The latter are people known to exist but for whom no Census form was returned and so a statistical method was used to impute their demographic information.

Name repair

Names on the Census dataset were of poorer quality than names on the deaths dataset. Names that are hand-written on forms and then read using optical character recognition often contain errors (as in the case of the Census). Census names are deleted following the Census processing period, and thus Census names normally undergo little processing to improve their quality. Therefore, special procedures had to be developed for the repair of names so that their usefulness as linking variables could be improved.

Census names were subjected to automatic name repair using name repair software and a standard name dictionary. Approximately 80% of Census names passed through this automated repair with a satisfactory result; this left about four million Census names unrepaired and requiring manual repair by clerical staff. There were not

enough resources to manually repair all four million records, so specific groups of interest were targeted for manual repair. These groups were of interest in either this quality study or one of the others conducted using the 2006 Census. Of interest to this quality study, Indigenous Australians were targeted for manual name repair. Even after automatic and selected manual repair, there were still some unrepaired names remaining in the dataset. In order to retain as much information as possible (even if poorer quality), these unrepaired names were still retained for linking.

Census undercount and persons temporarily overseas

Whilst the deaths data is a very comprehensive administrative list of deaths in Australia, each record will not always have an equivalent Census record to be linked to.

There will be some people who were in scope of the Census but were missed (undercount), and there will also be some people who were out of scope of the Census (persons temporarily overseas). Estimates from the Census Post Enumeration Survey also indicate that the Indigenous population has a larger undercount rate.

These issues are further discussed in Section 6.2, which includes an estimate of the number of death records with no equivalent Census record.

4.4 Deaths data issues

The death registrations data used for linking contained 106,945 records.

Two processing systems

A slight complication for this project was that the ABS changed its mortality data processing system at 1 January 2007. That is, the old data processing system was used until the end of 2006, and the new system was implemented from 1 January 2007. Therefore, the deaths data for this project (9 August 2006 to 30 June 2007) came from both the old and new processing systems.

The main issue that arose from this change was that the old processing system did not output Mesh block of usual residence, but the new system did. To overcome this problem, the death records from the old processing system underwent Mesh block coding using the ABS Address Autocoder. Records that could not be coded by the Autocoder were manually Mesh block coded by clerical staff where enough address information was present. In the end, 21.28% of the 106,945 death records did not have a Mesh block code; they could not be coded automatically or manually. It should be noted that virtually all of the records with no Mesh block had Statistical Local Area (SLA) codes, and many still had some components of address, but not enough to assign a Mesh block code.

Apart from the Mesh block coding issue, other differences between the two processing systems that impacted on this quality study were relatively minor and only required some re-formatting of variables.

Place of usual residence

Another data issue was that a person could have potentially changed their place of usual residence between the Census date and the Date of death. This could have led to records belonging to the same person having different addresses.

4.5 Quality of responses and missingness in both datasets

The deaths data are routinely processed by the ABS to resolve incomplete and invalid answers to standardised fields. The majority of records had very good quality reporting for the key linking variables of *Name*, *Address* (although not all had fine level detail), *Date of birth* and *Age*.

Compared to deaths data, the Census data contained more incomplete and invalid responses. This was even more evident for records identified as Indigenous. Table 4.2 shows the percentage of records missing key linking variables. Note that invalid and incomplete responses are not counted as missing in table 4.2, and that the Census data used for linking had more of these types of responses. Consequently, it is hard to make direct comparisons of missing rates between Census and deaths data.

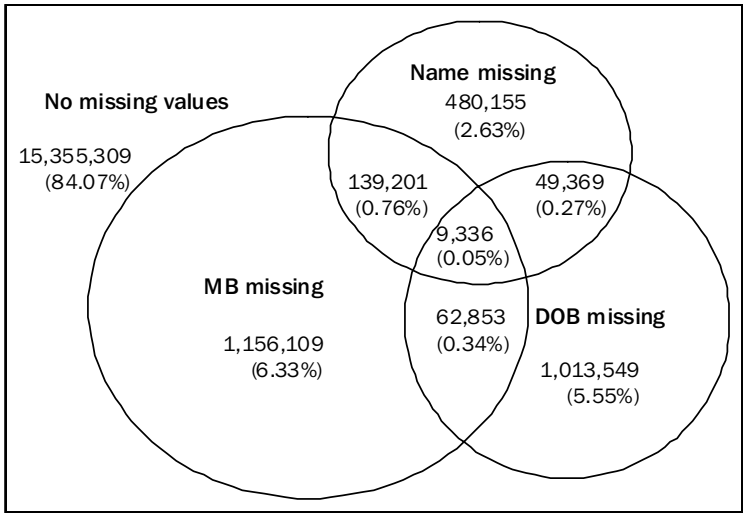
4.2 Percentage of Census and Death records with missing responses for key linking variables

	Census (19,046,302 records)			*Deaths (106,945 records)*		
	Non-Indigenous	*Indigenous*	*Not stated*	*Non-Indigenous*	*Indigenous*	*Not stated*
First name	3.18%	4.68%	7.51%	0.00%	0.06%	0.00%
Surname	3.67%	5.22%	8.79%	0.00%	0.06%	0.00%
DOB	6.21%	11.66%	13.24%	0.02%	0.78%	0.00%
Street number	5.09%	11.55%	13.03%	19.39%	32.72%	20.38%
Street name	3.27%	8.20%	10.06%	11.32%	24.50%	11.05%
Suburb	3.18%	13.42%	10.15%	11.57%	24.83%	9.76%
Postcode	3.26%	6.20%	10.13%	0.93%	4.61%	2.68%
Mesh block	7.49%	20.67%	15.74%	21.03%	36.22%	20.64%
No. of records	18,265,881	454,993	325,428	103,987	1,800	1,158

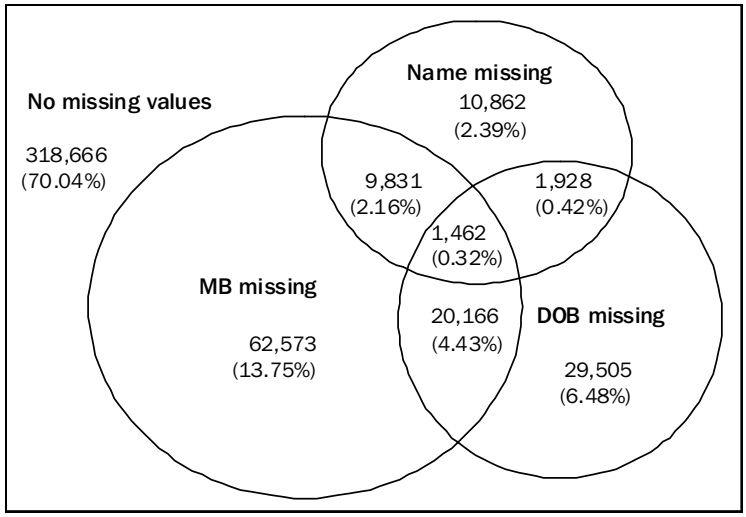
Note: In this table, *Date of birth* is considered missing if any component of *Date of birth* (*Day*, *Month* or *Year*) is missing. Also, respondents to the Census had the option of completing *Date of birth* or *Age at last birthday*. Many of the respondents who did not report *Date of birth* did report *Age at last birthday*.

4.3 Venn diagram of the number of Census records with missing responses for key linking variables, by Indigenous status

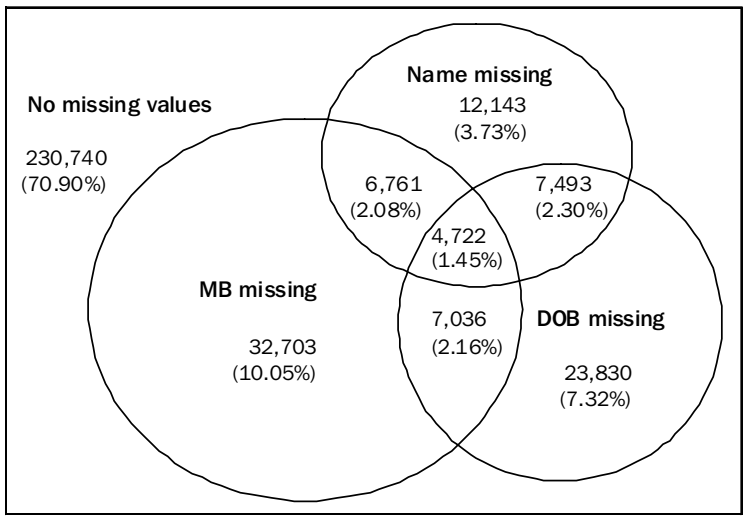
(a) Non-Indigenous



(b) Indigenous



(c) Not stated



'Name missing' refers to records that are either missing *First name* or *Surname* or both.

'DOB missing' refers to records that are missing any component of *Date of birth* (*Day*, *Month* or *Year*).

From table 4.2 it is clear that Indigenous records consistently have higher missing rates than non-Indigenous records for all key linking variables on both Census and deaths datasets. This indicates that it will be harder to establish agreement between Indigenous Census and death records, and thus they will be harder to link.

It also appears that the deaths data has lower missing rates for *Name* and *Date of birth*. While broader geography (*Postcode*) is less likely to be missing in deaths data, finer level geography items have higher missing rates in deaths, than in Census. This is reflective of two reasons:

- some death records specified a usual residence without *Street number* and *Street name* information, such as only reporting the name of a nursing home and its postcode; and
- address strings on death records had to be parsed from one string into the several address components using a parsing method contained within the ABS Address Autocoder, and some information might have been lost in this process.

It should be noted that although the Census missing rates for *Street number*, *Street name* and *Suburb* are lower than for deaths in table 4.2, the amount of processing the deaths data undergoes ensures that it is better quality and more informative if present; the Census data contains some invalid and incomplete responses.

Apart from the missing rates for individual variables, as presented in table 4.2, it is also important to consider if records are missing multiple variables. The Venn diagrams in figure 4.3 provide further insight into missing information on Census records by Indigenous status.

Figures 4.3 (a), (b), and (c) show that for some Census records, much of the key linking information was missing, and thus it would be difficult to establish agreement with death records. Specifically, 7.33% of Indigenous Census records had missing values for two or more of the important linking fields displayed in figure 4.3. This is compared to 1.42% for non-Indigenous Census records, and 7.99% for Census records with Indigenous status not stated.

4.6 Standardisation between datasets

This step of the process ensured that variables from the two datasets could be compared in a meaningful way. This meant formatting and coding the data in a common way and ensuring the concepts being measured by equivalent variables lined up as closely as possible. This step takes place against a background of determining which variables will be used for linking.

Below is a description of the standardisation required for each linking variable:

- *First name*

  On Census, respondents sometimes would report more than one name in this field. Rather than trying to guess which names were valid first names or middle names, the whole string was left in this field and compared with the *First name* string from the deaths data. This was done because the approximate string comparator would still return some agreement, even if there was a second name present. This would also prevent valid two word first names being deleted.

- *Surname*

  No standardisation was required.

- *Address*

  Address strings from both the Census and the deaths data were run through the ABS Address Autocoder and the resulting parsed address information (*Street number*, *Street name*, *Suburb*, *Postcode*) were used as the linking variables. This ensured that the same process was used for parsing and standardising address information for each dataset.

- *Date of birth*

  No standardisation was required.

- *Age*

  For deaths data, *Age at census* was calculated using *Date of birth* from the deaths data. For Census data, reported *Age* was used, but if it was missing, *Age* was derived from *Year of birth* (if reported). If reported and derived *Age* were missing, *Age* was set to missing (no imputed values were used).

- *Sex*

  No standardisation was required.

- *Place of birth*

  Deaths data had Australian state codes recoded to born in Australia to align with Census data. Both Census and deaths data were recoded to pool particular countries and provinces into broader categories. For example, the various parts of the United Kingdom were combined to form one broad category.

- *Marital status*

  Deaths data was recoded to align with Census data coding. 'Separated' was recoded as 'Registered married' on both datasets. On deaths data, 'married in de facto', 'tribally married', and 'tribally married now widowed' were all set to missing.

- *Year of arrival in Australia* (applicable for people born overseas)

  Duration of residence on deaths data was converted to *Year of arrival* so that it was equivalent to the Census variable. For duration of residence in category '97 or more', this was converted to *Year of arrival* of 'pre1911'. The Census variable *Year of arrival* was coded to have the category 'pre1911'.

- *Number of children ever born* (this variable records the number of children ever born (live births) to each female)

  Deaths data was recoded to align with Census data coding by creating the '6 or more' category.

Generally speaking, if variables had categories of inadequately described or not stated, these values were set to missing.

5. METHODOLOGICAL DETAIL IN THIS QUALITY STUDY

As discussed in Section 3.1, this section has broken down the linking methodology into the following steps:

- blocking strategy;
- linking variables and comparison functions;
- input probabilities;
- choosing cut-off weights; and
- clerical review.

This section discusses each of these steps, and describes the details and decisions made that were specific to the linking of Census records to death registrations.

Information on how the ABS applied probabilistic data linking in other CDE quality studies can be found in Solon and Bishop (2009) and Wright *et al.* (2009).

5.1 Blocking strategy

Once the data files have been prepared, records from each file can be compared to see whether they are likely to be a match. However, if the files are large, there may be too many record-pairs to conduct the comparison with the resources available. The 'blocking' stage reduces the number of comparisons needed by only comparing record-pairs where matches are more likely to be found (such as records with the same *Date of birth*).

A feature of the Census to deaths linkage was that four linking passes were conducted; each pass being defined by the blocking variables used. The reason for using multiple linking passes is that records that belong to the same person might disagree on one blocking variable, but agree on others. Using multiple passes with different blocking variables increases the chances that records belonging to the same person are compared at least once.

The available linking and blocking variables (as listed in table 4.1) were classified into three general categories: name information, personal characteristics, geographic information. For a record-pair to be linked, it was deemed that records would have to have a high level of agreement on at least one of these categories with supporting evidence from one or both of the other categories. This allowed for the fact that there may have been some errors in some categories for matches, but not all categories. Thus for record-pairs that did not have any of the three categories up to the standard required for getting the two records into corresponding blocks, it was deemed that there was not enough information to assign a link and thus it did not matter that they were not compared. Table 5.1 outlines the blocking strategy that was used.

5.1 Blocking strategy

Blocking Pass 1	**Fine geography**
	Mesh block (MB)
Blocking Pass 2	**Name information**
	First two initials in each of first name and surname (e.g. John Smith: 'JOSM')
	Sex
Blocking Pass 3	**Personal characteristics**
	Day of birth
	Month of birth
	Age
Blocking Pass 4	**Broader geography**
	Statistical Local Area (SLA)
	Sex

The first pass aimed to consider those records that had good quality geographic information. *Mesh block* was used as the blocking variable because the process of obtaining a *Mesh block* code was a form of standardisation that could be consistently applied to both datasets. *Mesh block* also produced blocks that reduced the number of comparisons to a computationally feasible level.

The second pass considered those records that had good quality name information, but may have been missed in pass one because they either provided poor quality geographic information or moved usual residence between Census night and date of death. *First two initials of first name and surname* was used as a blocking variable because it produced a feasible number of comparisons whilst still maintaining some allowance for differences in the rest of the names. *Sex* was also used as a blocking variable because it was a good quality variable on both datasets and it contributed to reducing the number of comparisons to a feasible level. Using a phonetic algorithm (e.g. double metaphone) to group names instead of using initials was also considered, however the phonetic concept was not as applicable to data captured in written form. Double metaphone also produced block sizes that were smaller than necessary, thus allowing fewer comparisons.

The third pass considered those records that had good quality personal characteristics information, but may have been missed in pass one because of disagreement on geographic information, and missed in pass two because they did not agree on name initial information. *Day of birth*, *Month of birth*, and *Age* were used as the blocking variables. Using *Age* instead of *Year of birth* enabled those records that did not have *Date of birth* information to still be included in a block. However, the marginal gain of capturing these records was probably not much because virtually all death records had *Date of birth* information.

The fourth pass was the final pass and was very broad in an attempt to compare any remaining matches that had not been compared in the first three passes. This very broad pass could be processed using the available computer resources because the first three passes had already linked the majority of the death records, and thus there were only about 10,000 unlinked death records left for comparison. *Statistical Local Area* and *Sex* were used as the blocking variables because they were very well reported and produced a feasible number of comparisons.

It was assumed that any true matches that could not be compared at least once in the four passes described above, did not have enough information to be assigned a link, and thus the four passes were deemed adequate for the purposes of linking.

Some information on the number of record-pair comparisons at each pass and the associated computing time is shown in table 5.2.

5.2 Number of record-pair comparisons and computing time for each pass

Pass	*Number of comparisons*	*Time taken for linking software to run*
Pass 1	9,588,669	1 hr 32 mins
Pass 2	91,052,466	12 hrs 42 mins
Pass 3	4,309,542	1 hr 20 mins
Pass 4	164,267,699	37 hrs 44 mins

5.2 Linking variables and comparison functions

After the 'blocking' stage has reduced the number of record-pair comparisons down to a computationally feasible level, records from the two files are compared using a full suite of linking variables. The choice of linking variables is different for each pass, and is related to the blocking variables for that pass. For each linking variable, a comparison function is used to determine the amount of agreement between values from the two files. There are different comparison functions that allow various types of comparisons of strings and numbers, and allow for scenarios in which there is only partial agreement between linking variables. The comparison functions used in this quality study included exact string comparison, approximate string comparison, and numeric comparison with tolerances. Christen and Churches (2005) provide thorough descriptions of these comparison functions.

Table 5.3 below presents the blocking and linking variables for each pass, and also the types of comparison functions used for the linking variables.

5.3 Blocking and linking variables for each pass (B = blocking, L = linking)

Variable	*Comparison function when used as linking variable*	*Pass 1*	*Pass 2*	*Pass 3*	*Pass 4*
Name information					
First name (e.g. John)	Approximate string	L	L	L	L
Surname (e.g. Smith)	Approximate string	L	L	L	L
Initials_4 (e.g. JOSM)			B		
Personal characteristics					
Sex	Exact string	L	B	L	B
Day of birth (DD)	Numeric with absolute tolerance (±2 days)	L	L	B	L
Month of birth (MM)	Exact String	L	L	B	L
Age	Numeric with absolute tolerance (±2 years)	L	L	B	L
Place of birth	Exact string	L	L	L	L
Year of arrival	Numeric with absolute tolerance (±2 years)	L	L	L	L
Marital status	Exact string	L	L	L	L
Address Information					
Mesh block	Exact string	B			
Statistical Local Area	Exact string				B
Street number	Exact string	L	L	L	L
Street name	Approximate string	L	L	L	L
Suburb	Approximate string		L	L	
Clerical review information					
Number of children					
Postcode					

In Passes 1,2, and 4, the individual components of *Date of birth* (*Day of birth*, *Month of birth*, *Age*) were used as linking variables because from observations of the data, it was evident that reporting error rates differed between the individual components. For example, a record may have had *Day of birth* out by a few days, but *Month of birth* was reported correctly. By using the components of *Date of birth*, as opposed to one combined field, the subtle errors and data qualities of the components could be accounted for in the linking process, thus producing a more accurate distribution of agreement weights. In Pass 3, *Date of birth* variables were used in blocking, and thus they were not used as linking variables.

For Passes 1 and 4, in which *Mesh block* and *Statistical Local Area* were used as blocking variables respectively, *Street number* and *Street name* still provided some distinguishing power within the blocks and thus were used as linking variables. Typically, all records within a *Mesh block* or *SLA* have the same *Suburb* and *Postcode* and thus they were not used as linking variables because they did not add any distinguishing power within the blocks.

For Passes 2 and 3, in which there was no geographic information in the blocking variables, *Street number*, *Street name* and *Suburb* were used as linking variables instead of *Mesh block* or *SLA*, because overall it provided address information that had more distinguishing power and also could take into account subtle errors in individual fields. *Postcode* was not used as a linking variable because in most cases it was very close in definition to *Suburb*, however it was still used as clerical review information.

The quality of the variable for *Number of children ever born* (to each female aged 15 and over) was deemed questionable so it was decided not to use it as a linking variable but to still include it in the clerical review process so that it could add some information. It was particularly useful for reviewing a record-pair where a woman may have had many children (e.g. five children).

The justification for choosing particular comparison functions for the linking variables is outlined below:

- *Approximate string comparison* was used for linking variables that contained strings of text such as names and addresses. It allowed for partial agreement when strings were almost the same but had some differences, possibly due to misspellings and other errors.
- *Numeric comparison with absolute tolerance* was used for those linking variables that had numeric values that could differ by one or two but still belong to the same person. Based on observations of the data and linking performed in other quality studies, it was deemed appropriate to use a numeric comparator with $\pm$ 2 absolute tolerance for the linking variables *Day of birth*, *Age* and *Year of arrival*.
- An *exact string comparator* was used for variables where it was deemed that only exact agreement should produce a positive agreement weight.

5.3 Input probabilities

For every record-pair comparison, each linking field is compared and the level of agreement is measured by calculation of a field weight. Field weights are then summed to form an overall record-pair comparison weight. Calculation of the field weights required two probabilities:

- The probability of a comparison outcome given that the record-pair belongs to the same person (m-probability).
- The probability of a comparison outcome given that the record-pair belongs to two different people (u-probability).

Three comparison outcomes were used in this quality study: 'agree', 'disagree' or 'missing from one or both records'. For each linking field, m- and u-probabilities were calculated for each of these comparison outcomes. These probabilities were calculated separately for each pass because they were dependent on agreement between the blocking fields.

For details about how input probabilities are used to calculate field weights, see Conn and Bishop (2006).

5.3.1 Calculation of the input probabilities

The m- and u-probabilities were derived from a similar linking project, in which registered deaths that occurred between 9 August 2005 and 8 August 2006 were linked to the 2005 Census Dress Rehearsal (CDR). In regards to the validity of using this method, the main issue was that the CDR was only conducted in purposively selected regions. This means there was a smaller number of linked records to calculate the input probabilities, and also potential for bias due to the purposive selection of the CDR. Another potential issue was that the amount of data processing and cleaning may have differed between the CDR and Census. Despite these issues, it was deemed that using the CDR_Deaths linkage was the most appropriate method to calculate the m- and u-probabilities for the Census_Deaths linkage.

Detailed descriptions of how the m- and u-probabilities were calculated are given below:

- m-probability ('agree')

  The estimation of the m-probabilities for the Census to deaths linkage assumed that the Census and deaths matched pairs had similar levels of agreement to the CDR and deaths linked pairs. The m-probability for a linking field was estimated by first counting the number of CDR_Deaths record-pairs which agreed on the blocking fields. Then, the proportion of these pairs which agreed further on the linking field was taken as the estimate of the linking field's block-specific m-probability, m('agree').

- u-probability ('agree')

  Calculating the u-probabilities for the 'agree' outcome involved a number of steps:

  1. From the set of all Census_Deaths record-pairs, the number of pairs which agreed on the blocking fields were counted.

  2. The CDR_Deaths links were used to estimate the number of Census_Deaths matches which agreed on the blocking fields. The balance was the estimated number of Census_Deaths non-matches which agree on the blocking fields.

3. The next step was to count the number of Census_Deaths record-pairs which agree on the blocking fields and the linking field.

4. The CDR_Deaths links were then used to estimate the number of Census_Deaths matches which agreed on the blocking fields and the linking field. The balance was the estimated number of Census_Deaths non-matches which agree on the blocking fields and the linking field.

5. The ratio of the numbers estimated in steps 4 and 2 above was the estimate of the block-specific u-probability for each field, u('agree').

- m-probability ('missing') and u-probability ('missing')

  In this quality study, a field weight of zero was assigned whenever the field was missing from one or both records. This assignment was based on the assumption that matched pairs and non-matched pairs were equally likely to have missing data. All candidate pairs were examined (those record-pairs which agreed on the blocking fields) and the proportion which had the linking field missing on either record was found. Consistent with the above assumption, this proportion was assigned to each of m('missing') and u('missing').

- m-probability ('disagree') and u-probability ('disagree')

  Given that the list of comparison outcomes is exhaustive, the m-probability for 'disagree' was set equal to 1 – m('agree') – m('missing'), and the u-probability for 'disagree' was set equal to 1 – u('agree') – u('missing').

5.3.2 Example input probabilities and interpretation

Appendix A contains a comprehensive list of the m- and u-probabilities for the three comparison outcomes by each field and blocking pass.

As an example, table 5.4 lists the m- and u-probabilities for the variable *First name* in Pass 1.

5.4 m- and u-probabilities for variable 'First name' in Pass 1 (Mesh block)

	First name – Pass 1 (Mesh block)	
Comparison outcome	*m-probability*	*u-probability*
Agree	0.714568	0.001258
Disagree	0.281995	0.995305
Missing from one or both records	0.003437	0.003437

To assist in the understanding of what the probabilities in table 5.4 represent, consider the following interpretations of the m- and u-probabilities for the variable *First name* in Pass 1:

- m('agree') – Given that the record-pair belongs to the same person, and the records agree exactly on *Mesh block* (the blocking variable), there is a 0.714568 probability that the records will *agree* exactly on *First name*.
- m('disagree') – Given that the record-pair belongs to the same person, and the records agree exactly on *Mesh block*, there is a 0.281995 probability that the records will *disagree* on *First name*.
- m('missing') – Given that the record-pair belongs to the same person, and the records agree exactly on Mesh block, there is a 0.003437 probability that First name will be missing on one or both records.
- u('agree') – Given that the record-pair belongs to two different people, and the records agree exactly on *Mesh block*, there is a 0.001258 probability that the records will *agree* exactly on *First name*.
- u('disagree') – Given that the record-pair belongs to two different people, and the records agree exactly on *Mesh block*, there is a 0.995305 probability that the records will *disagree* on *First name*.
- u('missing') – Given that the record-pair belongs to two different people, and the records agree exactly on *Mesh block*, there is a 0.003437 probability that *First name* will be missing on one or both records.

5.4 Choosing cut-off weights

After record-pair comparison, the record-pairs were sorted by total comparison weight. A decision rule was then used to determine whether a record-pair is linked, not linked, or considered as a possible link. This was done by comparing the total comparison weights with cut-off weights. The decision rule in this quality study had upper and lower cut-offs. Record-pairs with a weight above the upper cut-off were declared links, while those with a weight below the lower cut-off were declared non-links. The record-pairs with weights between the upper and lower cut-offs were not automatically assigned a status and were designated for clerical review. Clerical review is the process in which a person manually reviews record-pairs in order to assess whether the record-pair is a match or non-match. Not only is clerical review used to assign a link status for record-pairs between the upper and lower cut-offs, it is also used in setting the cut-offs. The purpose of this section is to explain how clerical review was used to set the level of the cut-off weights. Section 5.5 gives more information about how clerical review decisions were made, and how many record-pairs between the upper and lower cut-off weights were assigned as links through clerical review.

The cut-off weights are generally set by clerically inspecting record-pairs at different points along the total comparison weight distribution in order to identify where cut-offs should be positioned to trade off missing true links against including false links. Clerical review is a time and labour intensive task, and given the resources available for this quality study, it was important to plan how the clerical review would be managed. For this reason, the amount of clerical review was reduced by using a sampling scheme. The record-pairs were ordered by total comparison weight and divided into batches of equal weight ranges. From each batch, a random sample of record-pairs was selected and clerically reviewed to estimate the proportions of matches and non-matches. These estimates were then used to determine at which weight range the cut-offs should be set. This decision was made by considering the number of non-matches that would be accepted and the number of matches that would be missed. Based on the estimated distributions of matches and non-matches, and the objectives of this quality study, the upper cut-off weight was set when the estimated proportion of non-matches reached 20%. The lower cut-off weight was set when the estimated proportion of non-matches reached 60%. The two threshold values of 20% and 60% take into consideration the risks of incorrectly assigning matches as non-links, and non-matches as links, while also considering the available resources (staff and time) to perform clerical review. Note that the chosen threshold values are based on the unique circumstances of this quality study, and thus for any future linkages it would be prudent to review the threshold values. Guiver (2010) provides further details about sampling-based clerical review methods in probabilistic linking.

An assessment of the effect of non-matches accepted as links (false links) on the analysis of the linked file is included in Section 6.5.

Figure 5.5 shows the results of the sampling scheme used for Pass 1.

As can be seen in figure 5.5, not all samples from the different batches were clerically reviewed, but specific batches were targeted to try and identify the points in the weight distribution at which the estimated proportion of non-matches were at the threshold values for deciding where to set the cut-offs. For example, in figure 5.5, the results of batches 23 to 29 were used to assume that batches 0 to 22 had a tolerable proportion of non-matches. The results of batches 34 and 35 were used to assume that batches 36 to 39 would have a low proportion of matches.

Using this method, the upper cut-off used in Pass 1 was set at 13.44; the lower cut-off was set at 9.45. Table 5.6 shows the cut-off weights for the four passes.

Results from the sampling schemes for Passes 2,3, and 4 can be found in Appendix B.

5.5 Sampling scheme results for Pass 1

No. of record pairs: 64171 No. of Batches: 40 Sample Size: 65

Batch#	Status	Min weight	Max weight	Size	Sample Size	Est proportion of non-matches
0	N/A	42.45	43.44	4019	65	
1	N/A	41.44	42.40	1410	65	
2	N/A	40.45	41.43	375	65	
3	N/A	39.45	40.42	1013	65	
4	N/A	38.44	39.44	17071	65	
5	N/A	37.44	38.43	2217	65	
6	N/A	36.44	37.43	2052	65	
7	N/A	35.44	36.43	2953	65	
8	N/A	34.44	35.44	7377	65	
9	N/A	33.44	34.44	3606	65	
10	N/A	32.44	33.44	2079	65	
11	N/A	31.44	32.44	1563	65	
12	N/A	30.44	31.44	2592	65	
13	N/A	29.44	30.44	2436	65	
14	N/A	28.44	29.43	1816	65	
15	N/A	27.44	28.44	1798	65	
16	N/A	26.44	27.44	1047	65	
17	N/A	25.44	26.44	989	65	
18	N/A	24.44	25.44	1017	65	
19	confirmed	23.44	24.44	1124	65	0.0%
20	N/A	22.44	23.44	712	65	
21	N/A	21.45	22.43	481	65	
22	N/A	20.44	21.44	447	65	
23	confirmed	19.44	20.44	514	65	1.5%
24	confirmed	18.44	19.43	375	65	7.7%
25	confirmed	17.45	18.43	406	65	4.6%
26	confirmed	16.44	17.43	215	65	18.5%
27	confirmed	15.44	16.43	186	65	4.6%
28	confirmed	14.44	15.43	192	65	12.3%
29	confirmed	13.44	14.44	187	65	12.3%
30	clerical	12.45	13.43	185	65	30.8%
31	clerical	11.44	12.42	115	65	32.3%
32	clerical	10.44	11.44	113	65	43.1%
33	clerical	9.45	10.44	116	65	41.5%
34	rejected	8.44	9.44	151	65	86.7%
35	rejected	7.45	8.44	164	65	67.2%
36	N/A	6.46	7.43	162	65	
37	N/A	5.45	6.44	127	65	
38	N/A	4.44	5.40	154	65	
39	N/A	3.44	4.43	340	65	

Status is 'N/A' when the batch sample was not assessed (clerically inspected).
Status is 'confirmed' when the estimated proportion of non-matches is less than 20%.
Status is 'clerical' when the estimated proportion of non-matches is between 20% and 60%.
Status is 'rejected' when the estimated proportion of non-matches is greater than 60%.

Table 5.6 shows that different cut-offs were used for each pass. This is because different blocking and linking variables were used for each pass, not because different levels of agreement were considered acceptable. The cut-offs are not comparable between passes.

5.6 Upper and lower cut-offs for each pass

Pass	*Lower cut-off*	*Upper cut-off*
Pass 1	9.45	13.44
Pass 2	8.62	12.61
Pass 3	7.86	11.86
Pass 4	12.02	21.02

5.5 Clerical review

In clerical review, record-pairs are assessed by inspection to decide if the record-pair is a match or non-match, and hence whether or not it should be assigned as a link or non-link. Typically, the clerical reviewer is able to identify variations in names and common transcription errors (e.g. 1 and 7) that were not picked up using the comparison options described in Section 5.2. In addition to the variables used in the blocking and linking, the clerical reviewer can also inspect other variables to assist in the decision-making. In this quality study, postcode and number of children were not used in the automated linking (see Section 5.2 for details), but were used for clerical review.

To help inform the person performing clerical review, some simple frequency tables from the Census were produced for variable combinations such as *Name* and *Age*. This helped to provide some empirical evidence on the uniqueness of particular variable combinations, rather than just relying solely on a reviewer's intuition or personal knowledge. It should be noted that this was done to obtain a general impression to inform decision making, specific clerical review rules were not made up and the reviewer still had a lot of freedom to make decisions on whether a record-pair should be linked or not; the assessment of record-pairs was a subjective process.

Some typical types of record-pairs accepted as links through clerical review included:

- Pairs in which the name variables from the Census were very poor quality (e.g. ZF##RFY). Quite often, the general structure of a name could still be identified, despite the approximate string comparator giving a negative weight. For example, a clerical reviewer could still identify a seemingly invalid name such as ZF##RFY as potentially JEFFREY.
- Record-pairs in which fields in one file (usually Census because the data had been processed less) differed by a character or number that can often be mistaken because of hand writing style or scanning errors (e.g. 1 and 7, J and S, 5 and 6).

- Older people with missing fields on the Census. It was noted from the frequency tables that a name was much more likely to uniquely identify an individual whose *Age* was greater than 85 years.
- Record-pairs where *First names* have alternative forms (e.g. Bob and Robert).

For each linking pass, clerical review was performed in the sampling scheme to choose cut-offs, and then to assign links between the upper and lower cut-offs, before running the next pass. Clerical review results for choosing the cut-offs are discussed in Section 5.4. Results for the clerical review of record-pairs between the upper and lower cut-offs are shown in table 5.7. Record-pairs could either be accepted or rejected as a link.

5.7 Number of record-pairs for clerical review

		Accepted		*Rejected*	
Pass	*Number of pairs*	*Number*	*%*	*Number*	*%*
Pass 1	529	375	70.9%	154	29.1%
Pass 2	1,464	1,001	68.4%	463	31.6%
Pass 3	1,237	1,008	81.5%	229	18.5%
Pass 4	1,538	999	65.0%	539	35.0%
Total	4,768	3,383	71.0%	1,385	29.0%

5.6 Linking summary and results

In summary, the linked file was formed by conducting four linking passes; multiple passes combat the problem of missed links caused by blocking. Each pass consisted of the following steps:

- determining a set of blocking variables;
- determining a set of linking variables and setting appropriate comparison functions;
- calculating input probabilities for the linking variables;
- conducting the linking by running Febrl software;
- using a sampling method implemented in the ABS version of Febrl to determine upper and lower cut-offs; and
- clerically review record-pairs between the upper and lower cut-offs using the ABS version of Febrl.

At the end of the four passes, the output was a final links file consisting of two variables: a person identifier from the deaths file, and a person identifier from the Census file. This file could then be used to construct a linked file with analysis variables from each dataset.

Some basic results from the linking process are given below.

Records on Census file = 19,046,302

Records on deaths file = 106,945

Number of linked death records = 98,898

Number of unlinked death records = 8,047

Table 5.8 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.

5.8 Records linked at each stage of the linking process

	Number of links		
Pass	*Automatically assigned*	*Confirmed through clerical review*	*Total*
Pass 1	62,269	375	62,644
Pass 2	29,796	1,001	30,797
Pass 3	2,972	1,008	3,980
Pass 4	478	999	1,477
Total	95,515	3,383	98,898

It is clear from table 5.8 that the majority of links were automatically assigned in Passes 1 and 2. This indicates that many of the links had a high degree of agreement between linking variables.

For records linked in Pass 4, a smaller proportion of the links were automatically assigned compared to the other passes. This is probably because most of the matches with lots of agreement had been picked up in the earlier passes, and any potential matches compared in Pass 4 were those that had data quality issues that prevented them from being linked in earlier passes, but were compared in Pass 4 because of the broader blocking strategy (SLA and sex). Record-pairs with data quality issues are often the types of record-pairs in the clerical review range, in which clerical inspection can identify and discern data issues that the automatic comparison functions cannot.

6. EVALUATION OF THE LINKAGE

It is important to evaluate the quality of the linked data to help inform and set the context for any conclusions drawn from uses of the linked data. There are several ways to evaluate the quality of linked datasets. For the Census to deaths linkage, the following were considered:

- population characteristics of linked and unlinked death records;
- reasons for unlinked death records;
- estimating the number of false links;
- estimates of match-link rate and link accuracy; and
- effect of false links on analysis.

This section provides a summary of the investigations performed.

6.1 Population characteristics of linked and unlinked death records

Some investigations were performed to obtain an idea of the population characteristics for the linked and unlinked records in terms of *Indigenous status*, *Gender* and *Age*. The results are presented in tables 6.1 to 6.4.

6.1 Reported Indigenous status for linked Death records (98,898 records)

	Census records			
	Indigenous	*Non-Indigenous*	*Not stated*	*Total*
	Number			
Death records				
Indigenous	1,056	231	40	1,327
Non-Indigenous	302	91,076	5,153	96,531
Not stated	21	970	49	1,040
Total	1,379	92,277	5,242	98,898
	Proportion (%)			
Death records				
Indigenous	1.07%	0.23%	0.04%	1.34%
Non-Indigenous	0.31%	92.09%	5.21%	97.61%
Not stated	0.02%	0.98%	0.05%	1.05%
Total	1.39%	93.31%	5.30%	100.00%

6.2 Other information on Indigenous status for Census and Death data

	Census records (Census Indigenous status)		*All Death records (Deaths Indigenous status)*		*Unlinked Death records (Deaths Indigenous status)*	
	No.	*%*	*No.*	*%*	*No.*	*%*
Indigenous	454,993	2.39%	1,800	1.68%	473	5.88%
Non-Indigenous	18,265,881	95.90%	103,987	97.23%	7,456	92.66%
Not stated	325,428	1.71%	1,158	1.08%	118	1.47%
Total	19,046,302		106,945		8,047	

Some key observations from tables 6.1 and 6.2 are:

- 302 individuals that were identified as Indigenous on the Census were identified as non-Indigenous on the death record;
- 231 individuals that were identified as non-Indigenous on the Census were identified as Indigenous on the death record; and
- Indigenous people (as identified on the death records) were over-represented in the unlinked death records. Records identified as Indigenous made up 5.88% of the unlinked death records, compared to 1.68% of all death records available for linking.

Table 6.3 shows that males are over-represented in the unlinked death records. Males make up 60.01% of the unlinked death records, compared to 51.34% of all death records available for linking.

6.3 Linkage status by Gender

	Linked Death records (Deaths gender)		*Unlinked Death records (Deaths gender)*		*All Death records (Deaths gender)*	
Gender	*No.*	*%*	*No.*	*%*	*No.*	*%*
Males	50,075	50.63%	4,829	60.01%	54,904	51.34%
Females	48,823	49.37%	3,218	39.99%	52,041	48.66%
Total	98,898		8,047		106,945	

Table 6.4 shows that people under the age of 60 are over-represented in the unlinked death records. Records with ages in the range of 15 to 59 years make up 31.97% of the unlinked death records, compared to 14.54% of all death records available for linking.

6.4 Linkage status by Age group

	Linked Death records (Deaths age)		*Unlinked Death records (Deaths age)*		*All Death records (Deaths age)*	
Age group	*No.*	*%*	*No.*	*%*	*No.*	*%*
0–14 years	477	0.48%	91	1.13%	568	0.53%
15–59 years	12,975	13.13%	2,572	31.97%	15,547	14.54%
60 and over	85,446	86.40%	5,384	66.90%	90,830	84.93%
Total	98,898		8,047		106,945	

It is clear from tables 6.1 to 6.4 that deaths in the linked file are not completely representative of all death records. This is important to remember when considering potential uses of the linked data, for the purposes of making appropriate adjustments or performing error assessments. This issue is further discussed in the ABS publications in which results from the Indigenous Mortality Quality Study have already been released (2008a, 2008b, 2009a).

6.2 Reasons for unlinked death records

Ideally, all death records (106,945) would have been linked to their equivalent Census record. However, this did not occur. Reasons for unlinked death records included:

- The person was in-scope of the Census, but was missed, thus contributing to Census undercount;
- The person was temporarily out of the country on Census night (out of scope of the Census);
- A person immigrated to Australia after Census night, and then died;
- The Census and death records were missing information (or they disagreed) in two or more of the important linking fields; these being *Name*, *Address* and *Date of birth*.

Census undercount rates obtained from the Post Enumeration Survey were used to obtain an estimate of the number of death records that did not have an equivalent Census record because they were missed in the Census. The undercount rates were applied to the deaths population at the 5-year age by sex level. These groups were used because the age structure of the deaths population (predominance of older people) was the characteristic that was most different from the general population. Undercount adjustment factors by 5-year age by sex level by Indigenous status were not available. However, it is known that the overall Census undercount rate for Indigenous Australians is much higher than for non-Indigenous Australians. For more details, see *Census of Population and Housing: Details of Undercount* (ABS, 2006b). Using the method described above, it was estimated that 3,747 Death records did not have an equivalent Census record because they were missed in the Census. This explains part of the 8,047 unlinked death records.

When applying the 2006 Census undercount rates to the deaths data, it was implicitly assumed that the undercount rates (by 5-year age by sex level) for the general population were the same as the population who were going to die in the period 9 August 2006 to 30 June 2007. This may not necessarily be true.

Another reason that death records may not have had an equivalent Census record was that the person was temporarily overseas on the night of the Census. These type of people were out of scope of the Census and were not accounted for in the undercount adjustment. An estimated figure of around 345,000 Australian people were temporarily overseas on the 2006 Census night (ABS, 2007a). This was approximately 1.5% of the Australian population. It is not known if this rate is applicable to the population of people with death records within scope of this linking project, but it nonetheless gives an indication of the likely magnitude of this issue.

Persons with a death record but no Census record because they immigrated to Australia after Census night, would possibly account for some of the unlinked death records. However, when considering the total number of unlinked death records, this would only be a very minor reason.

The amount of missing or disagreeing information between the two datasets most likely explains the remaining number of unlinked death records. Table 4.2 and the associated discussion in Section 4 give an indication of the extent of this issue.

In summary, it appears that approximately half of the unlinked death records could not be linked because no equivalent Census record existed, and the other half could not be linked because although there was an equivalent Census record, not enough agreement could be established between records because of data quality issues.

6.3 Estimating the number of false links

A point of interest was how many of the record-pairs assigned as links were false links (records that do not belong to the same person). The presence of false links could potentially affect conclusions from analysis performed on the linked data. It was known that some false links were present in the linked file, simply because of the linking procedure used (see Section 5.4 'Choosing cut-off weights').

An estimate of false links was obtained by using the results of the sampling scheme for choosing cut-off weights. For the record-pairs that were automatically linked (above the upper cut-off), the estimated proportion of non-matches was used as the estimated proportion of false links. This method assumed that the person performing clerical review would make the correct decision on whether a record-pair was a true or false link.

The sample available for this estimation was a by-product of the sampling scheme for choosing cut-off weights, and thus was not specifically designed for this purpose. It was a probability based sample but had incomplete coverage, remembering that the sampling scheme targeted particular regions of the comparison weight distribution and not all batches were investigated (see figure 5.5). Therefore, some interpolations based on assumptions had to be made to estimate the complete distribution of non-matches for record-pairs above the upper cut-off weight. Using the estimated proportion of non-matches from the batch samples as reference points, it was assumed that the distribution of non-matches declined linearly as the weight range of the batches increased. This linear assumption was considered conservative because the distribution of non-matches is expected to decline more rapidly. Hence the method used is likely to introduce some upward bias into the estimates of false links.

It is also important to remember that the estimated percentage of false links in each batch is subject to sampling error. Thus the overall estimate of false links should be considered in light of the sampling error.

After interpolation of the non-match distribution above the upper cut-off for each pass, each batch accepted as links had an estimated proportion of false links. Treating each batch as a stratum, the overall sample design was considered as stratified simple random sampling without replacement. The following estimate of the number of false links in the linked file was obtained:

$$\textit{Estimated number of false links} = 1{,}202 \text{ (1.2\% of total linked records)}$$

$$\textit{Upper bound of 95\% confidence interval} = 1{,}617$$

6.4 Estimates of match-link rate and link accuracy

Using the estimates obtained in Sections 6.2 and 6.3, estimates of two data linking quality measures (link accuracy and match-link rate) could be calculated. These measures are defined using the terms 'match' and 'link'; these terms were defined in Section 3.1.

By comparing the match status to the link status, record-pairs can be classified into the following four groups shown in table 6.5.

6.5 Classification of matches and links

		Match status (true status)		
		Matches	Non-matches	
Link status (assigned in linking process)	Links	*True links* (matches that are linked)	*False links* (non-matches that are linked)	Total links
	Non-links	*Missed links* (matches that are not linked)	*True non-links* (non-matches that are not linked)	
		Total matches		

The link accuracy is defined as the proportion of links in the linked dataset that are matches. Using the terms defined in table 6.5, it is defined as:

$$Link\ accuracy = \frac{True\ links}{Total\ links}$$

Match-link rate is defined as the proportion of possible matches that are actually linked in the linked dataset. Using the terms defined in table 6.5, it is defined as:

$$Match\text{-}link\ rate = \frac{True\ links}{Total\ matches}$$

The terms in these equations can be calculated using the estimates obtained in Sections 6.2 and 6.3.

$$\begin{aligned} True\ links &= Total\ links - False\ links \\ &= 98,898 - 1,202 \\ &= 97,696 \end{aligned}$$

$$\begin{aligned} \textit{Total matches} &= \textit{number of death records} - \textit{records without a Census record} \\ &= 106{,}945 - 3{,}747 \\ &= 103{,}198 \end{aligned}$$

Note that for *records without a Census record*, only the estimate of Census undercount has been used. This was considered conservative, because other reasons for records not having a Census record have not been included in this estimate. Also, from the above equation, it is implicit that the number of total matches equals the number of death records that have an equivalent Census record.

Using the values calculated above, link accuracy and match-link rate can be calculated:

$$\textit{link accuracy} = 97{,}696 / 98{,}898 = 0.9878$$

$$\textit{match-link rate} = 97{,}696 / 103{,}198 = 0.9467$$

It is estimated that 98.78% of the links in the linked dataset are matches. It is estimated that 94.67% of possible matches were actually linked in the linked dataset. Individuals who did not have records in both datasets were not considered as possible matches.

6.5 Effect of false links on analysis

The primary use of the linked data was to investigate the differences in reported Indigenous status between the Census and deaths data. Therefore, it was important to know how many of these differences could be explained simply because the linked records were from two different people (a false link).

An indication of the distribution of false links across the possible Indigenous status reporting permutations was obtained by making a conservative assumption that the false links were randomly linked records. This is a conservative assumption because when using probabilistic linking, there is an expectation that even records in false links will agree on some characteristics, thus reducing the impact of false links on the analysis of these characteristics. Although Indigenous status was not used as a linking variable, it is expected to be correlated with other linking variables such as mesh block.

Table 6.6 shows the nine possible combinations of reported Indigenous status for a linked pair. It contains the actual count observed from the linked file for each combination, and the conservative apportioning of the estimated 1,202 false links based on the assumption discussed above.

6.6 Indigenous status counts from linked file, including the estimated number of false links

	Census records							
	Indigenous		*Non-Indigenous*		*Not stated*		*Total*	
	Count	*False*	*Count*	*False*	*Count*	*False*	*Count*	*False*
Death records								
Indigenous	1,056	***0***	231	***19***	40	***0***	1,327	***19***
Non-Indigenous	302	***28***	91,076	***1,121***	5,153	***20***	96,531	***1,169***
Not stated	21	***0***	970	***12***	49	***0***	1,040	***12***
Total	1,379	***28***	92,277	***1,152***	5,242	***20***	98,898	***1,202***

Table 6.6 indicates that some of the differences in reported Indigenous status could have been explained by false links, although the majority of differences appeared to be true links.

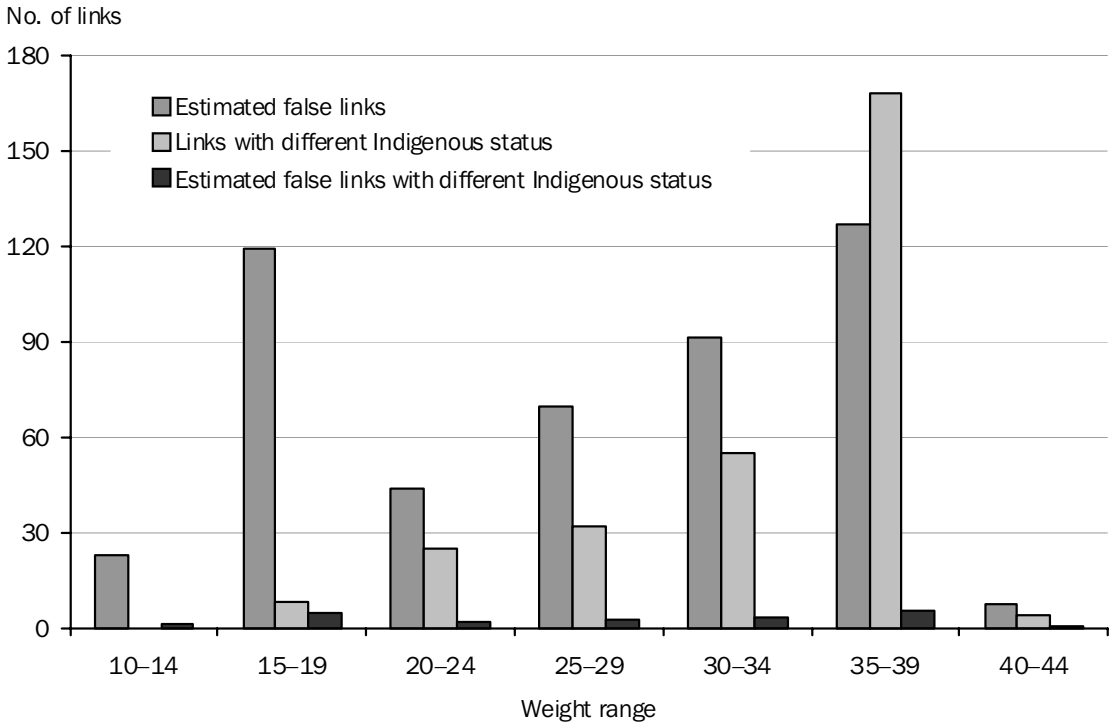
Although the apportioning assumption may not be very realistic, it at least demonstrates that even in a scenario where no evidence was used to link records (randomly linked), the expected number of linked records differing on Indigenous status would still be far less than what was actually observed in this probabilistic data linking quality study. This indicates the majority of linked records that differ on Indigenous status are in fact true links.

Some analysis was also performed on the distribution of record-pair comparison weights for linked record-pairs that had reported different Indigenous status from the two data sources. Linked records that differed on Indigenous status had roughly the same weight distribution as the rest of the linked record population for each pass. Many of the links with different Indigenous status had high comparison weights, well above the upper cut-off.

Figure 6.7 shows the weight distribution for Pass 1 of linked records that differ on Indigenous status, and the estimated number of false links at each weight interval, and also the estimated number of false links that differ on Indigenous status.

It can be seen in figure 6.7 that the number of links with different Indigenous status far outnumber the estimated number of false links that differ on Indigenous status, particularly for record-pairs with higher agreement weights. Similar patterns were observed in the other passes.

6.7 Number of links, by weight – Pass 1



Note:

Estimated false links are calculated as described in Section 6.3.

Estimated false links with different Indigenous status are calculated based on the same assumptions used to apportion false links in table 6.6. Note that record-pairs where one record has Indigenous status 'not stated', are not counted as links with different Indigenous status in figure 6.7.

Looking at all the evidence discussed above in this section, it appears that any major trends or patterns in linked records that have different reported Indigenous status would not have been masked by false links. However, any hypotheses that depended on only a few linked record-pairs could have been heavily influenced by the presence of some false links.

7. CONCLUSIONS

Overall, the application of probabilistic data linking methodology to the linking of Census data and death registration data has proved satisfactory. Any outstanding quality issues with the linked data are primarily due to the quality of the data being fed into the data linking methodology, not the methodology itself.

In regards to the usefulness of the linked file, the evaluation of the linkage in Section 6 identified two issues: the number of false links in the linked file, and the number of unlinked death registrations.

Regarding the number of false links, there are false links in any large scale data linking exercise, and the amount of them can be controlled by the chosen cut-off weights. In this quality study, the number of false links and their effect on analysis was found to be a minor issue. It is not thought that false links influenced any conclusions from analysis of the linked file.

The more pertinent issue, in terms of potential influence on analysis and scope to improve the linkage, is the number of unlinked records. Section 6.1 showed that the linked records are not perfectly representative of the entire death record population, and hence inferences made from the linked dataset may be biased if appropriate adjustments are not made. Of particular relevance to this quality study, Indigenous Australians were under-represented in the linked file.

As discussed in Section 6.2, the two main reasons for unlinked death records were that the person was not recorded in the Census, or that the person had too many missing values (or disagreeing values) for important linking variables. Of the 8,047 unlinked records, these two reasons contributed approximately half each.

It is assumed that the number of people temporarily overseas on Census night will remain approximately the same each Census. For those people in Australia on Census night, there will probably always be some Census undercount, but it may decrease in the future due to improved Census enumeration procedures. In particular, the ABS's Indigenous Community Engagement Strategy (ABS, 2009b) aims to enhance engagement with Indigenous communities in data collection. New and improved Indigenous enumeration procedures are being planned for the 2011 Population Census.

In terms of reducing the amount of missing or disagreeing information on the two datasets, some improvements may be made for the next Census. Deaths data from 2007 onwards will be Mesh block coded as part of the regular mortality data processing system; remembering that 21.28% of the death records used for linking this time did not have a Mesh block code. Furthermore, Mesh blocks for the 2006 Census were experimental, as indicated in an ABS information paper (2008c), and

there are likely to be improvements in the degree of Mesh block coding in 2011, particularly for non-metropolitan areas. This will positively impact the Census and deaths data and hopefully mean more records can be linked.

The number of Indigenous unlinked death records could be potentially decreased by improving the repair and standardisation of Indigenous names, through creating a more comprehensive name dictionary and corresponding list of alternative names (nicknames). For this quality study, the name dictionary was obtained from the Australian Electoral Commission. For future studies, a knowledge of Indigenous naming conventions and alternative name forms could help improve the name dictionary, and hence the linking.

In conclusion, the largest improvements in linking Census records to death registrations are to be found in improving the quality of the data being fed into the linking process. Therefore, the outstanding data linking issues are really just the outstanding issues with the datasets. That is, enumerating the entire in-scope population, and also ensuring the individual variables are reported accurately.

ACKNOWLEDGEMENTS

The author wishes to acknowledge the large team effort in many different aspects of the project that enabled the data linking to be conducted. Particular mention goes to Glenys Bishop, Tenniel Guiver, and Matthew Hardy for their technical advice and valuable comments on this paper. Thanks also to Peter Rossiter for his assistance in the preparation of this paper.

REFERENCES

Australian Bureau of Statistics (2005a) *Census Data Enhancement – Statement of Intention*, (last viewed on 19 January 2010) <http://www.abs.gov.au/websitedbs/D3110124.NSF/f5c7b8fb229cf017ca256973001fecec/5812a287d6a2e78fca2571ee001a7a49!OpenDocument>

—— (2005b) *Enhancing the Population Census: Developing a Longitudinal View*, Discussion Paper, cat. no. 2060.0, ABS, Canberra.

—— (2006a) *Census Data Enhancement Project: An Update*, Information Paper, cat. no. 2062.0, ABS, Canberra.

—— (2006b) *Census of Population and Housing – Details of Undercount*, August 2006, cat. no. 2940.0, ABS, Canberra.

—— (2007a) *Australian Demographic Statistics, December Quarter 2006*, cat. no. 3101.0, ABS, Canberra.

—— (2008a) *Census Data Enhancement – Indigenous Mortality Quality Study*, Information Paper, cat. no. 4723.0, ABS, Canberra.

—— (2008b) *Assessment of Methods for Developing Life Tables for Aboriginal and Torres Strait Islander Australians*, Discussion Paper, cat. no. 3302.0.55.002, ABS, Canberra.

—— (2008c) *Outcomes from the Review of the Australian Standard Geographical Classification*, Information Paper, cat. no. 1216.0.55.002, ABS, Canberra.

—— (2009a) *Experimental Life Tables for Aboriginal and Torres Strait Islander Australians, 2005–2007*, cat. no. 3302.0.55.003, ABS, Canberra.

—— (2009b) *ABS Indigenous Community Engagement Strategy*, ABS, Canberra, (last viewed 19 January 2010) <http://www.abs.gov.au/websitedbs/corporate.NSF/4a256353001af3ed4b2562bb00121564/e0ad7ea6a36195e1ca2574b3007f7281!OpenDocument>

Christen, P. and Churches, T. (2005) *Febrl 0.3 Documentation*, (last viewed on 19 January 2010) <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.

Guiver, T. (2010, forthcoming) "Sampling-Based Clerical Review Methods in Probabilistic Linking", *Methodology Research Papers*, Australian Bureau of Statistics, Canberra.

Solon, R. and Bishop, G. (2009) "A Linkage Method for the Formation of the Statistical Longitudinal Census Dataset", *Methodology Research Papers*, cat. no. 1351.0.55.025, Australian Bureau of Statistics, Canberra.

Wright, J.; Bishop, G. and Ayre, T. (2009) "Assessing the Quality of Linking Migrant Settlement Records to Census Data", *Methodology Research Papers*, cat. no. 1351.0.55.027, Australian Bureau of Statistics, Canberra.

APPENDIXES

A. *m*- AND *u*-PROBABILITIES FOR EACH PASS

	m-probabilities				*u-probabilities*			
Variable	*Pass 1*	*Pass 2*	*Pass 3*	*Pass 4*	*Pass 1*	*Pass 2*	*Pass 3*	*Pass 4*
				'agree'				
First name	0.714568	0.784379	0.735174	0.721426	0.001258	0.189025	0.003783	0.003486
Surname	0.891885	0.910773	0.848072	0.848850	0.004756	0.043309	0.000585	0.000561
DOB-Day	0.906163	0.900569	—	0.892441	0.029325	0.030587	—	0.030547
DOB-Month	0.933711	0.932408	—	0.928108	0.076983	0.078664	—	0.078172
Age	0.959016	0.946962	—	0.945304	0.012597	0.008504	—	0.005857
Sex	0.996554	—	0.998092	—	0.503129	—	0.500564	—
Place of birth	0.907973	0.916534	0.886287	0.912695	0.531722	0.563998	0.468692	0.493517
Year of arrival	0.049244	0.035439	0.049396	0.054116	0.002996	0.001019	0.002399	0.001342
Marital status	0.753941	0.735166	0.855998	0.716550	0.278356	0.266832	0.444803	0.237630
Street number	0.718954	0.579425	0.605374	0.640895	0.074736	0.009437	0.009855	0.010575
Street name	0.895910	0.765310	0.773423	0.831554	0.330954	0.000114	0.000188	0.004368
Suburb	—	0.731220	0.739621	—	—	0.000358	0.000462	—
				'disagree'				
First name	0.281995	0.215611	0.261607	0.274740	0.995305	0.810965	0.992998	0.992680
Surname	0.099098	0.089218	0.141345	0.142098	0.986227	0.956682	0.988832	0.990387
DOB-Day	0.035589	0.041020	—	0.044809	0.912427	0.911002	—	0.906703
DOB-Month	0.008214	0.009324	—	0.009337	0.864942	0.863068	—	0.859273
Age	0.040984	0.048088	—	0.048287	0.987403	0.986546	—	0.987734
Sex	0.003446	—	0.001908	—	0.496871	—	0.499436	—
Place of birth	0.052721	0.050139	0.060292	0.053923	0.428972	0.402675	0.477887	0.473101
Year of arrival	0.036508	0.022817	0.033395	0.038332	0.082756	0.057237	0.080392	0.091106
Marital status	0.047978	0.076320	0.089516	0.068315	0.523563	0.544654	0.500711	0.547235
Street number	0.148020	0.189812	0.205444	0.168657	0.792238	0.759800	0.800963	0.798977
Street name	0.018536	0.097981	0.103309	0.056331	0.583492	0.863177	0.876544	0.883517
Suburb	—	0.129487	0.122383	—	—	0.860349	0.861542	—
				'missing'				
First name	0.003437	0.000010	0.003219	0.003834	0.003437	0.000010	0.003219	0.003834
Surname	0.009017	0.000009	0.010583	0.009052	0.009017	0.000009	0.010583	0.009052
DOB-Day	0.058248	0.058411	—	0.062750	0.058248	0.058411	—	0.062750
DOB-Month	0.058075	0.058268	—	0.062555	0.058075	0.058268	—	0.062555
Age	0.000000	0.004950	—	0.006409	0.000000	0.004950	—	0.006409
Sex	0.000000	—	0.000000	—	0.000000	—	0.000000	—
Place of birth	0.039306	0.033327	0.053421	0.033382	0.039306	0.033327	0.053421	0.033382
Year of arrival	0.914248	0.941744	0.917209	0.907552	0.914248	0.941744	0.917209	0.907552
Marital status	0.198081	0.188514	0.054486	0.215135	0.198081	0.188514	0.054486	0.215135
Street number	0.133026	0.230763	0.189182	0.190448	0.133026	0.230763	0.189182	0.190448
Street name	0.085554	0.136709	0.123268	0.112115	0.085554	0.136709	0.123268	0.112115
Suburb	—	0.139293	0.137996	—	—	0.139293	0.137996	—

B. SAMPLING SCHEME RESULTS

B.1 Sampling scheme results for Pass 2

No. of record pairs: 35061 No. of Batches: 55 Sample Size: 65

Batch#	Status	Min weight	Max weight	Size	Sample Size	Est proportion of non-matches
0	N/A	57.63	58.63	457	65	
1	N/A	56.64	57.62	128	65	
2	N/A	55.64	56.61	116	65	
3	N/A	54.63	55.62	78	65	
4	N/A	53.63	54.59	74	65	
5	N/A	52.64	53.52	1821	65	
6	N/A	51.63	52.62	598	65	
7	N/A	50.63	51.62	333	65	
8	N/A	49.63	50.62	171	65	
9	N/A	48.64	49.62	201	65	
10	N/A	47.63	48.62	167	65	
11	N/A	46.65	47.62	299	65	
12	N/A	45.63	46.62	142	65	
13	N/A	44.65	45.62	610	65	
14	N/A	43.64	44.60	281	65	
15	N/A	42.67	43.61	173	65	
16	N/A	41.63	42.59	171	65	
17	N/A	40.63	41.62	152	65	
18	N/A	39.65	40.61	252	65	
19	N/A	38.64	39.62	108	65	
20	N/A	37.63	38.62	197	65	
21	N/A	36.63	37.61	158	65	
22	N/A	35.63	36.62	169	65	
23	N/A	34.63	35.62	287	65	
24	N/A	33.63	34.60	217	65	
25	N/A	32.65	33.62	177	65	
26	N/A	31.63	32.61	264	65	
27	N/A	30.63	31.62	152	65	
28	N/A	29.63	30.62	440	65	
29	N/A	28.63	29.61	660	65	
30	N/A	27.63	28.62	253	65	
31	N/A	26.63	27.60	394	65	
32	N/A	25.63	26.62	204	65	
33	N/A	24.63	25.62	295	65	
34	N/A	23.63	24.61	7145	65	
35	N/A	22.63	23.62	2361	65	
36	N/A	21.63	22.62	2056	65	
37	N/A	20.63	21.62	1174	65	
38	confirmed	19.63	20.61	806	65	1.5%
39	confirmed	18.63	19.62	1723	65	1.5%
40	confirmed	17.63	18.62	664	65	3.1%
41	confirmed	16.63	17.63	405	65	3.1%
42	confirmed	15.63	16.61	1337	65	1.5%
43	confirmed	14.63	15.62	623	65	7.7%
44	confirmed	13.63	14.62	895	65	13.8%
45	confirmed	12.63	13.62	408	65	15.4%
46	clerical	11.64	12.60	358	65	20.0%
47	clerical	10.63	11.61	320	65	30.8%
48	clerical	9.63	10.62	366	65	35.4%
49	clerical	8.63	9.62	418	65	58.5%
50	rejected	7.63	8.62	543	65	64.6%
51	rejected	6.63	7.62	613	65	69.2%
52	rejected	5.63	6.62	573	65	83.1%
53	N/A	4.63	5.62	667	65	
54	N/A	3.63	4.62	849	65	

See footnotes to table 5.5.

B.2 Sampling scheme results for Pass 3

No. of record pairs: 6039 No. of Batches: 49 Sample Size: 65

Batch#	Status	Min weight	Max weight	Size	Sample Size	Est proportion of non-matches
0	confirmed	52.86	52.86	1	1	0.0%
1	N/A	50.89	51.27	3	3	
2	N/A	50.11	50.39	5	5	
3	N/A	48.87	49.69	12	12	
4	N/A	47.87	48.63	10	10	
5	N/A	46.86	47.81	12	12	
6	N/A	46.03	46.86	30	30	
7	N/A	44.89	45.86	43	43	
8	N/A	43.87	44.83	68	65	
9	N/A	42.90	43.85	43	43	
10	N/A	41.92	42.69	35	35	
11	N/A	40.93	41.82	32	32	
12	N/A	39.87	40.83	135	65	
13	N/A	38.92	39.84	57	57	
14	N/A	37.90	38.85	42	42	
15	N/A	36.88	37.85	45	45	
16	N/A	35.88	36.86	65	65	
17	N/A	34.87	35.81	45	45	
18	N/A	33.87	34.85	77	65	
19	N/A	32.93	33.85	49	49	
20	N/A	31.87	32.85	72	65	
21	N/A	30.89	31.86	66	65	
22	N/A	29.96	30.86	42	42	
23	N/A	28.88	29.83	40	40	
24	N/A	27.88	28.86	61	61	
25	N/A	26.92	27.81	33	33	
26	N/A	25.86	26.85	63	63	
27	N/A	24.91	25.84	41	41	
28	N/A	23.87	24.86	46	46	
29	N/A	22.87	23.85	47	47	
30	N/A	21.86	22.78	35	35	
31	N/A	20.89	21.84	66	65	
32	N/A	19.87	20.83	46	46	
33	N/A	18.87	19.85	100	65	
34	N/A	17.90	18.85	150	65	
35	N/A	16.87	17.79	193	65	
36	N/A	15.86	16.86	223	65	
37	confirmed	14.87	15.84	247	65	4.6%
38	confirmed	13.87	14.86	199	65	4.6%
39	confirmed	12.86	13.86	184	65	6.2%
40	confirmed	11.87	12.85	209	65	6.2%
41	clerical	10.87	11.85	545	65	30.8%
42	N/A	9.86	10.85	261	65	
43	clerical	8.87	9.85	157	65	23.1%
44	clerical	7.87	8.85	274	65	47.7%
45	rejected	6.87	7.86	422	65	80.0%
46	rejected	5.87	6.86	256	65	69.2%
47	N/A	4.86	5.86	226	65	
48	N/A	3.87	4.85	490	65	

See footnotes to table 5.5.

B.3 Sampling scheme results for Pass 4

No. of record pairs: 8301 No. of Batches: 46 Sample Size: 65

Batch#	Status	Min weight	Max weight	Size	Sample Size	Est proportion of non-matches
0	N/A	46.02	46.02	1	1	
1	N/A	42.12	42.12	1	1	
2	N/A	40.53	40.53	1	1	
3	N/A	40.13	40.13	1	1	
4	N/A	38.64	38.64	1	1	
5	N/A	38.38	38.38	1	1	
6	N/A	37.58	37.58	1	1	
7	N/A	37.35	37.35	1	1	
8	N/A	35.32	35.32	1	1	
9	N/A	35.23	35.23	1	1	
10	N/A	34.77	34.77	1	1	
11	N/A	34.25	34.64	6	6	
12	N/A	33.04	33.91	7	7	
13	N/A	32.02	32.81	5	5	
14	N/A	31.07	31.49	5	5	
15	N/A	30.11	31.00	8	8	
16	N/A	29.05	29.99	12	12	
17	N/A	28.02	28.97	10	10	
18	N/A	27.03	27.99	11	11	
19	N/A	26.07	26.96	37	37	
20	N/A	25.04	25.97	27	27	
21	N/A	24.02	24.96	34	34	
22	N/A	23.07	24.01	88	65	
23	confirmed	22.03	23.00	105	65	10.8%
24	confirmed	21.02	22.01	112	65	13.8%
25	clerical	20.02	21.01	163	65	20.0%
26	clerical	19.03	20.00	131	65	24.6%
27	N/A	18.02	19.01	157	65	
28	N/A	17.04	18.01	145	65	
29	N/A	16.02	17.00	149	65	
30	N/A	15.02	16.01	179	65	
31	N/A	14.03	15.00	149	65	
32	rejected	13.02	14.02	267	65	61.5%
33	rejected	12.03	13.02	198	65	69.2%
34	rejected	11.02	12.02	239	65	89.2%
35	N/A	10.02	11.01	376	65	
36	rejected	9.02	10.01	508	65	95.4%
37	N/A	8.02	9.01	486	65	
38	rejected	7.02	8.02	663	65	100.0%
39	N/A	6.02	6.99	639	65	
40	N/A	5.02	6.01	756	65	
41	N/A	4.03	5.01	626	65	
42	N/A	3.02	4.01	641	65	
43	N/A	2.02	3.01	501	65	
44	N/A	1.02	2.01	553	65	
45	N/A	0.04	1.01	260	65	

See footnotes to table 5.5.

FOR MORE INFORMATION . . .

INTERNET **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.

PHONE 1300 135 070

EMAIL client.services@abs.gov.au

FAX 1300 135 211

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

FREE ACCESS TO STATISTICS

All statistics on the ABS website can be downloaded free of charge.

WEB ADDRESS www.abs.gov.au