Australian Bureau of Statistics

1351.0.55.058

Research Paper

Death Registrations to Census Linkage Project – A Linked Dataset for Analysis

New Issue

Research Paper

Death Registrations to Census Linkage Project – A Linked Dataset for Analysis

Health & Disability Branch

Data Integration Microdata Futures Branch

Methodology Transformation Branch

AUSTRALIAN BUREAU OF STATISTICS

EMBARGO: 11.30 AM (CANBERRA TIME) WED 09 MAR 2016

ABS Catalogue no. 1351.0.55.058

Views expressed in this paper are those of the author(s), and do not necessarily represent those of the Australian Bureau of Statistics.
Where quoted, they should be attributed clearly to the author(s).

Produced by the Australian Bureau of Statistics

INQUIRIES

The ABS welcomes comments on the research presented in this paper.
For further information, please contact Mr James Eynstone-Hinkins, Director, Health & Vitals Statistics Unit, on Brisbane (07) 3222 6185.

DEATH REGISTRATIONS TO CENSUS LINKAGE PROJECT – A LINKED DATASET FOR ANALYSIS

EXECUTIVE SUMMARY

Mortality outcomes provide insights into the health and well-being of populations. The richer this information, the more opportunity it offers for researchers to understand the interplay between health, lifestyle and socioeconomic status and to identify groups in society who are at risk of premature death. Accordingly, governments can then base policy settings and resource allocation on better evidence.

The Australian Bureau of Statistics (ABS), with the support of State and Territory Registrars of Births Deaths and Marriages, has created a new information base for research into mortality and health outcomes in Australia by combining Death Registrations data with data from the 2011 Australian Census of Population and Housing.

The purpose of this report is to describe the record linkage process and demonstrate the potential of the linked dataset to provide valuable insights into contextual factors associated with mortality.

Since 2011, the ABS has used statistical data integration to combine a growing number of administrative datasets with the Census. The intent has been to enrich the value and maximise the use (and re-use) of publicly-collected information – thereby expanding the evidence base available to support research and policy development to enhance the well-being of society.

The Death Registrations data collected and maintained by State and Territory Registrars of Births, Deaths and Marriages are of very high quality. Mortality outcomes by cause of death are able to be accurately monitored because of generally high standards of death certification and access to coronial information through the National Coronial Information System. The Census is a regular snapshot of the Australian population with a rich set of sociodemographic information, including living arrangements, educational attainment and labour force status. By combining these data sources, the ABS has created a breakthrough information base for understanding contextual factors associated with mortality.

The ABS conducted a Death Registrations to Census record linkage in 2012. The primary aim of that project was to evaluate the consistency of Indigenous identification, as reported in the Death Registrations data and Census data, and thereby provide input into the compilation of life tables and life expectancy estimates for Aboriginal and Torres Strait Islander people.

The terms of the project permitted the use of name and address information, in conjunction with other personal characteristics reported on both datasets, to create a gold standard record linkage. This record linkage could be used to evaluate future record linkage activities, but dissemination of the gold standard linkage to external users was precluded. Details of all names and addresses were removed from the Census and Death Registrations datasets at the conclusion of the 2011 Census processing period.

The linked dataset created for this study will be available and suitable for much wider analytical applications. This report describes the record linkage methodology, assesses the properties of the new linked dataset relative to the gold standard linkage created for the earlier study, and proposes weighting strategies that can support appropriate statistical inference.

A particular focus of the report is to alert users to important, and possibly unfamiliar issues that arise in the analysis of linked data. Familiarity with these issues will assist users to frame suitable questions, select appropriate weighting options, conduct sound inference and recognise the magnitude and direction of potential biases in their results.

The report concludes that, with careful and informed use, the linked dataset can support a wide range of new analytical investigations. That is, it represents a significant advance in the information base available for understanding mortality in Australia.

There is, however, room for improvement. Future Death Registrations to Census record linkages will benefit from greater use of name information, either directly or in the form of anonymised name codes. Pertinent to this issue is the announcement, in December 2015, that the ABS will retain names and addresses collected in the 2016 Census of Population and Housing to support future data integration activities.

The use of name and address information may be expected to deliver more successful linkage outcomes for otherwise difficult-to-link populations – in particular, Aboriginal and Torres Strait Islander people living in remote communities. Indeed, while this report has focussed on the utility of the linked dataset as an information base for the general Australian population, further research is required to develop its capacity to provide insights into mortality outcomes for the Aboriginal and Torres Strait Islander population.

CONTENTS

DEATH REGISTRATIONS TO CENSUS LINKAGE PROJECT – A LINKED DATASET FOR ANALYSIS

ABSTRACT

Death registrations are provided to the Australian Bureau of Statistics by the State and Territory Registrars of Births, Deaths and Marriages. In this project, the records of individuals whose deaths were registered in 2011–12 have been augmented with personal information obtained from the 2011 Census. This has been achieved through a process of probabilistic record linkage. The resulting linked dataset promises to contribute significantly to a better understanding of health risk factors, leading to more informed targeting of health policy and support measures for the Australian community. This paper describes the creation of the linked dataset and provides necessary background material and advice for potential users.

1. INTRODUCTION

In 2012, the Australian Bureau of Statistics (ABS) conducted a Death Registrations to Census record linkage, as part of the 2011 Census Data Enhancement (CDE) program (ABS, 2010b). The primary aim of the project was to evaluate the consistency of Indigenous identification, as reported in the Death Registrations data and Census data, and thereby provide input into the compilation of life tables and life expectancy estimates for Aboriginal and Torres Strait Islander people.

The project sought to locate the Census records of 153,455 persons who were registered as deceased between 10 August 2011 and 27 September 2012 inclusive. Details have been reported in the Information Paper: *Death Registrations to Census Linkage Project – Methodology and Quality Assessment* (ABS, 2013d).

Under the terms of the CDE program, the record linkage was permitted to use name and address information, in conjunction with other personal characteristics reported on both datasets, to create a gold standard record linkage. This record linkage could then be used to evaluate future record linkage activities, but dissemination of the gold standard linkage to external users was precluded.

The terms of the CDE program also stipulated that all names and addresses were to be removed from the Census and Death Registrations data at the conclusion of the 2011 Census processing period.

The purpose of the current study is to reconstruct the Death Registrations to Census record linkage, without the benefit of having name and address information as linking fields, but using knowledge and insights obtained from the gold standard linkage to guide and validate the new record linkage process. It is proposed to make the resulting linked dataset available for research and analysis.

The ABS expects that the linked dataset will contribute to a better understanding of health risk factors and/or predictors of particular health conditions – allowing for more informed targeting of health policies, interventions and community support structures to improve health outcomes for Australians in general. The linked dataset will also enable Australia to participate in international studies that measure and compare health inequalities across countries (OECD, 2015).

This paper provides an overview of the record linkage methodology employed to construct the linked dataset (Section 2), an explanation of the weighting strategies designed to restore the representativeness of the linked records (Section 3), and a discussion of issues that may influence the appropriate use of the linked dataset for analysis (Section 4).

2. LINKAGE METHODOLOGY

As in the previous study, a probabilistic record linkage methodology, based upon the approach of Fellegi and Sunter (1969), has been employed. The key feature of this methodology is the ability to locate records on the two datasets that may refer to the same person by evaluating the degree of commonality exhibited by a range of personal identifying variables common to the two datasets. More specifically, patterns of agreement and disagreement on the common variables are converted to a single score that reflects the probability that the linked records do indeed refer to the same person (i.e. they constitute a 'match'). These scores permit the ranking of all potential matches, and facilitate a rational decision to assign probable match or non-match status to the linked records (Solon and Bishop, 2009). More detail on the probabilistic linking methodology employed in the construction of the gold standard linkage may be found in the original report (ABS, 2013d).

As already noted, detailed name and address information are now unavailable for linking. However, coded address information has been retained, in the form of Australian Statistical Geography Standard (ASGS) geographical areas (Meshblock, SA1, SA2, etc.)(ABS, 2010a). The description of the probabilistic linkage methodology used to construct the new linked dataset has been kept to a minimum, but the most pertinent details are reported in this section.

2.1 The linking strategy

The task of evaluating the status of all possible record pairs is often computationally infeasible, and usually also unnecessary. This is resolved by the process of 'blocking', which determines that only those records that agree on selected variables (the 'blocking' variables) will be further assessed on the remaining 'linking' variables.

Of course there will be instances where matching records will fail to agree on a selected blocking variable, due to missing, invalid or legitimately different responses on one or other record. To mitigate this, the broader linking strategy usually involves evaluating several combinations of blocking and linking variables.

Table 2.1 reports the variables available to support the current record linkage, and summarises the principal blocking and linking strategies employed (Runs 1–4). The different runs feature a trade-off between the level of agreement on geographical variables (Meshblock, SA1) versus personal characteristics (such as Date of birth, Country of Birth) required to establish credible links.

Within every run, the linking variables have associated *field weights* that reflect the probabilistic significance of agreement or disagreement on that variable. For every record pair considered under the blocking criteria, an aggregate *record pair weight* or *link weight* is computed by summing the applicable field weights.

2.1 Blocking and linking strategies

Variable	*RUN 1*	*RUN 2*	*RUN 3*	*RUN 4*
Geography	Meshblock	SA1	SA2	——
Birthday	BDAY	BDAY	BDAY	BDAY
Birth Year	——	——	BYEAR	BYEAR
Age	AGE (±1)	AGE	——	——
Sex	SEX	SEX	SEX	SEX
Marital Status	MST	MST	MST	MST
Country of Birth	COB	COB	COB	COB
Year of Arrival	YOA	YOA (±2)	YOA	YOA

Shading denotes blocking variables.

2.2 Implementation

Parameter estimates (specifically, the m- and u-probabilities[1] from which the field weights are calculated) were derived with reference to the matches and non-matches identified in the gold standard linkage.

It has been observed that u-probabilities in particular can vary significantly with respect to selected subpopulations in the two link files. By exploiting this information, and independently linking selected subpopulations, it is possible to extract superior information with which to classify probable matches and non-matches. Table 2.2 defines the broad population characteristics that have been used for this purpose.

2.2 Demographic categories used to define significant subpopulations

Age cohort	*Migrant*	*City*
A0= Born after 1945	M0= Born in Australia	C0= Does not live in a major city
A1= Born before 1946	M1= Born overseas	C1= Lives in a major city

For this study, all linking runs were conducted independently. This is a departure from the methodology employed in the gold standard linkage, which conducted sequential linking passes. The process of identifying the optimal links then proceeded as follows:

1. The links generated by each run were assigned adjusted aggregate link weights to permit comparability between runs, and all such links were then collated.
2. Where two records were linked within multiple runs, only the instance with the highest adjusted link weight was retained.

[1] The m-probabilities for a specified variable are the agreement/disagreement probabilities that apply for two matching records. u-probabilities are the probabilities that apply for two non-matching records.

3. For each record in the Death Registrations dataset, a search was conducted for a unique best link to a Census record. This required that both linked records were identified as the unique best link to the other. All unique best links were then extracted, and all remaining links involving the extracted records were deleted.
4. A linear assignment algorithm[2] was then employed to select the best one-to-one links from those remaining.

2.3 Diagnostics

As no constraints or cut-off weights were used in the linking runs or in the selection process, it was possible to link all Death Registration records that fell within the scope of the blocking strategies. However, the credibility of an assigned link drops rapidly as the adjusted aggregate link weight declines, and it is therefore necessary to determine an acceptable cut-off weight to control the number and quality of links selected for the final linked dataset.

In the gold standard record linkage, 142,697 of the 153,455 Death Registration records (93%) were successfully matched to Census records.

In this study, 3,151 Death Registration records (2%) remained unlinked under the specified blocking criteria. 2,133 of these missing records had been successfully linked in the gold standard linkage. Consequently, the best achievable outcome from this study would be to extract the remaining (142,697–2,133 =) 140,564 matches from the (153,455–3,151 =) 150,304 within-scope Death Registration records.

When all within-scope Death Registration records were linked, only 122,087 records were found to be correctly matched (87% of 140,564). 18,477 records that had previously been matched successfully were linked to different Census records, and 9,740 previously unmatched records were also incorrectly linked.

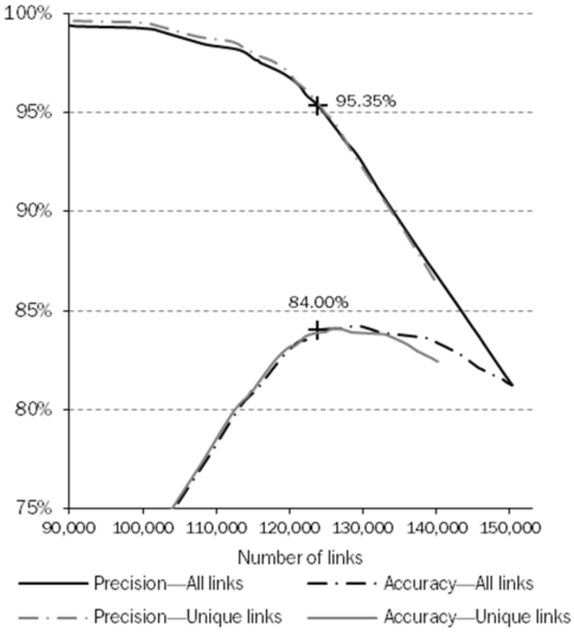
Table 2.3 and figure 2.4 illustrate how the diagnostic measures of *precision* and *accuracy* vary as successively lower cut-off weights are used to define the linked dataset. *Precision* measures the proportion of true positives, or matches, in the linked dataset. Maximum *accuracy* relates to the best attainable classification of both true positives and true negatives. (Formulae for the calculation of precision and accuracy are provided within table 2.3.)

[2] This refers to a numerical optimisation process designed to identify the one-to-one record linkage that delivers the maximum sum of record pair weights.

2.3 Diagnostics

	Cut-off weight				*Gold standard (in-scope)*
	29	***23.5***	*22*	*0*	
Linked					
True positives = TP	100,920	**118,145**	120,003	122,087	140,564
False positives = FP	815	**5,765**	9,227	28,217	0
Total = P	101,735	**123,910**	129,230	150,304	140,564
Not linked					
False negatives = FN	38,982	**18,309**	14,479	0	0
True negatives = TN	9,587	**8,085**	6,595	0	9,740
Total = N	48,569	**26,394**	21,074	0	9,740
Precision = TP/P (%)	99.2%	**95.3%**	92.9%	81.2%	100.0%
Accuracy = (TP+TN)/(P+N) (%)	73.5%	**84.0%**	84.2%	81.2%	100.0%

2.4 Accuracy and precision



2.4 Specification of the linked dataset

Selection of the linked record pairs that comprise the linked dataset was based upon the identification of a single appropriate cut-off weight. No clerical review or further refinement of the selection was attempted, and it is openly acknowledged that the linked dataset will therefore contain incorrect links.

The cut-off weight of 23.5 was selected as it produces a linked dataset with good *precision* (95.3%) and near maximum achievable *accuracy* (84%).

The resulting linked dataset contains 123,910 record pairs. The gold standard identifies 118,145 of these as true matches.

Under the higher cut-off weight of 29, almost 99% of the selected links involve persons whose Death Registration and Census records agree on MB or SA1 (i.e. they were linked in Runs 1 and 2). The chosen cut-off of 23.5 captures 7,000 persons whose residence moved further afield from the SA1. This subpopulation may be important to include, e.g. for studies that assess the impact of ready access to medical facilities.

At the lower cut-off weight of 22, maximum accuracy has been achieved, but the extra links added to the dataset include almost twice as many false positives as true positives – lowering the precision measure significantly for only a marginal improvement in accuracy. Below the cut-off of 22, false positives continue to be added to the dataset at a much faster rate than the remaining true positives.

2.5 Representativeness

The principal issue that must be recognised by all potential users of the linked dataset is that the record linkage is not perfect. As documented above, the precision realised in the linked dataset is 95%. However, while the information extracted from the 5% of links that are incorrect may be inaccurate for those individuals, it need not necessarily prove misleading for broader analysis.

Furthermore, the realised link rate of (123,910/153,455 =) 81% is not achieved uniformly over all possible subpopulations of interest within the Death Registrations dataset.

For example, link rates were lower than average for the overlapping subpopulations of

- Aboriginal and Torres Strait Islander people;
- People living in remote and very remote regions;
- Residents of the Northern Territory; and
- Persons aged under 50 years.

Among causes of death, the link rates for intentional self harm were lower than those of the other selected causes.

See table A.1 in the Appendix and Table S.5 in the accompanying Data Cube for more information on the link rates realised for selected subpopulations.

Section 3 explains how weighting schemes may be employed to mitigate proportional under- and over-representation of demographic subpopulations within the linked dataset.

Section 4 provides empirical examples that may assist users of the linked dataset to recognise the implications of precision and representativeness for their analyses.

3. WEIGHTING STRATEGY

To enable evaluation of the new linked dataset against the gold standard linkage, it is necessary to adjust or compensate for the differing population counts and characteristics observed in the two linked datasets. This is achieved by applying a consistent weighting approach to both datasets to replicate record counts on the complete Death Registrations dataset.

The use of weighting to restore representativeness presupposes that there are always persons in the linked dataset who are sufficiently similar to those persons who have not been matched. As discussed in Sections 2.4 and 2.5, this cannot be guaranteed for linked data since there may be specific characteristics that result in certain individuals not being linked. In addition, it may be prudent to question the wisdom of weighting unconvincing links. Users are encouraged to reflect on these caveats when evaluating weighting strategies and their impact on aggregate statistical analysis.

In this study, the proposed weighting strategy comprises two separate weighting processes. The crucial first stage weights identify and adjust for any relative under- or over-representation of selected subpopulations within the linked datasets. The second stage weights may be applied, as required, to fine-tune or calibrate selected weighted aggregates to corresponding totals from the complete dataset. Some caution should be exercised when using second-stage weights for analysis to ensure that they do not introduce unnecessary distortion to the resulting weighted estimates.

3.1 Stage 1 weights

The design of the first stage weighting strategy was closely aligned to the linkage strategy, in which specific subpopulations (see table 2.2) were sometimes linked independently. The selected subpopulations for Stage 1 weighting are defined by cross-tabulating the categories shown in table 3.1.

3.1 Demographic categories used to define subpopulations for Stage 1 weights

Sex	*Age cohort*	*Migrant*	*City*

M= Male
F= Female
	A0= Born after 1945
	A1= Born before 1946
		M0= Born in Australia
		M1= Born in Europe (incl. UK)
		M2= Born elsewhere overseas
			C0= Does not live in a major city
			C1= Lives in a major city

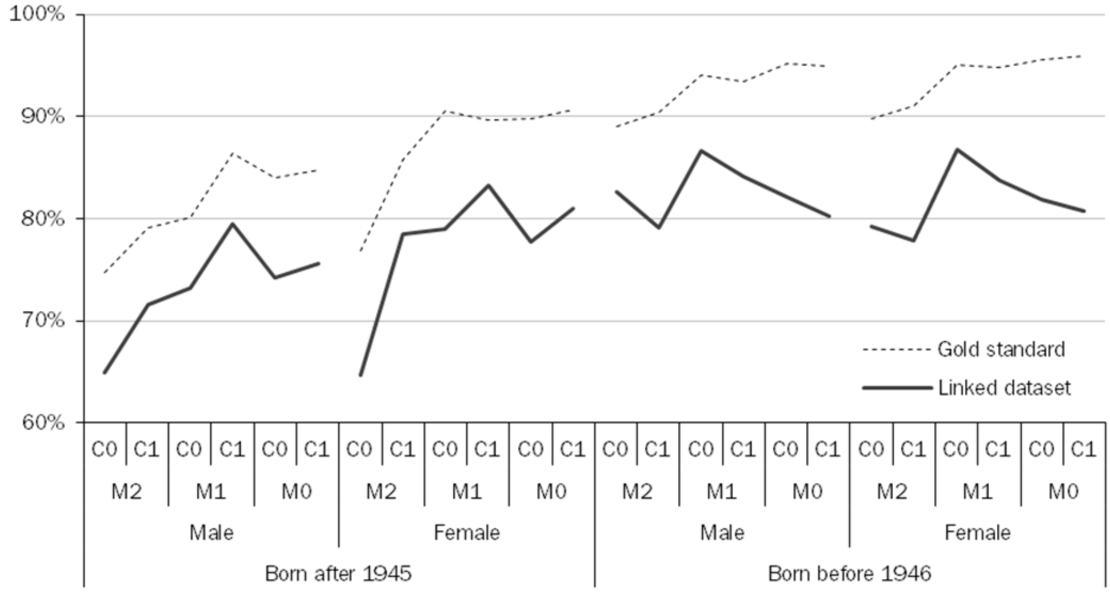
The categories in table 3.1 define (2 × 2 × 3 × 2 =) 24 subpopulations of interest. Selecting the categories and deciding how to partition them requires a degree of judgement. Too many categories with too many partitions may generate very small subpopulations for which the results may prove spurious in the sense that consistent results might not be observed if the linkage were conducted on another dataset of

Death Registration records. On the other hand, too few categories may fail to adequately capture the diversity of linking outcomes achieved.

An example of weighting follows. The Death Registrations dataset contains the records of 12,933 older European-born males who were living in a major city at the time of their death. The linked dataset contains 10,881 of these individuals, while the gold standard linkage contains 12,083 of them. Therefore the first stage weights required to restore the representativeness of this demographic subpopulation are 1.18859 for the linked dataset and 1.07035 for the gold standard.

Figure 3.2 reports the observed link rates from the linked dataset for the 24 reference subpopulations, with the link rates attained by the gold standard linkage provided for comparison.

3.2 Link rates for 24 subpopulations



It may be observed that there are significant differences in the link rates observed for the selected subpopulations, which confirms the suitability of the choice. Figure 3.2 also highlights some systematic patterns in the link rate outcomes. For example, the Death Registration records for older persons, females, European migrants and residents of major cities have been linked more successfully than those for other demographic groups.

The Stage 1 weights are derived from the reciprocals of the link rates. Thus, subpopulations with lower link rates are assigned higher weights, increasing their representation within the weighted linked datasets.

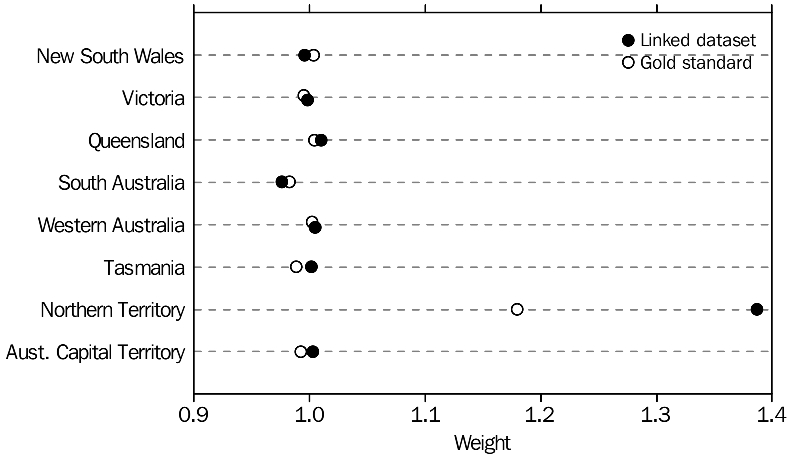
3.2 Calibration of Stage 1 weights

Stage 1 weights may be modified to derive Stage 2 weights when there exists a need to present weighted aggregates that correspond exactly to the record counts in selected categories in the original Death Registrations dataset. Alternatively, Stage 2 weights might be deemed necessary when there is some question regarding the ability of the Stage 1 weights to restore sufficient representativeness for important analytical dimensions of the linked dataset.

Examples of Stage 2 weights are presented to illustrate these considerations (see figures 3.3–3.5).

Figure 3.3 illustrates the Stage 2 weights required to align the weighted State totals derived from the linked dataset and the gold standard linkage with the State totals reported in the complete Death Registrations dataset.

3.3 Stage 2 weights required to align with State benchmarks



From figure 3.3 it can be observed that Stage 1 weights have satisfactorily reproduced the State totals for seven of the eight States, since the corresponding Stage 2 weights are almost identically equal to one.

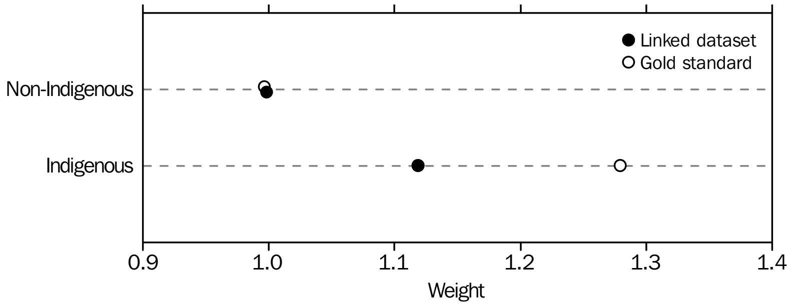
The large weights required to adjust the Northern Territory totals highlight the fact that low link rates have been identified for Northern Territory records in both the linked dataset and the gold standard.

Compared with other states, the Northern Territory reports disproportionately more deaths of younger people (born after 1945) in regional and remote areas. Three in five deaths in the Northern Territory pertain to persons born after 1945, compared with one in five for the remainder of Australia. Geographically, no part of the Northern Territory is classified as a major city, while 66% of all other Australian deaths are reported in the major cities.

While a high proportion of Northern Territory records relate to subpopulations that are inherently more difficult to link, it is still necessary to explain why the Stage 1 weights have failed to adequately compensate for this. In fact Northern Territory deaths are under-represented within all relevant demographic subpopulations.

Aboriginal deaths in remote and very remote regions comprise one-third of all Northern Territory deaths, and are the least likely to be linked. With link rates of 40% in the linked dataset and 58% in the gold standard linkage, this group remains significantly under-weighted as the broad Stage 1 weights do not adequately recognise the practical difficulties of enumerating remote and very remote communities.

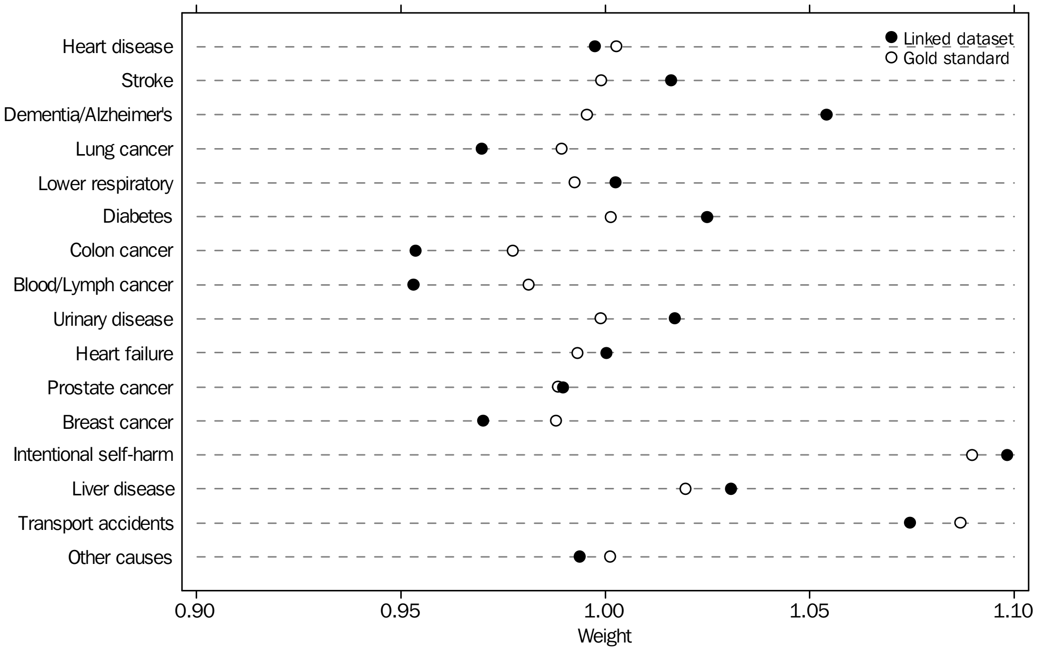
3.4 Stage 2 weights required to align with Indigenous status benchmarks



The wider issue of under-weighting the Aboriginal and Torres Strait Islander records in the linked dataset is highlighted in figure 3.4. Large Stage 2 weights are required to restore the counts of Aboriginal and Torres Strait Islander persons in both the linked dataset and the gold standard. This indicates the inadequacy of the weighting strategy to identify and adjust for factors leading to the under-representation of this subpopulation in both linkages. It will be necessary to develop a revised two-stage weighting strategy, employing Indigenous status benchmarks, to support future analyses that specifically focus on comparing the non-Indigenous and Aboriginal and Torres Strait Islander subpopulations.

Figure 3.5 graphs the Stage 2 weights that would be required to calibrate the weighted counts from the linked dataset and the gold standard linkage to the counts in the complete Death Registrations dataset for 15 reported causes of death. The selected causes of death include common causes reported for males, females and Aboriginal and Torres Strait Islander people, and are ranked (from top to bottom) from most to least prevalent within the complete Death Registrations dataset (except for the 'Other causes' category).

3.5 Stage 2 weights required to align with Cause of death benchmarks



Some variation is observable, and it should be noted that the sample sizes associated with many categories are quite small, implying that variation should be anticipated. The more notable weights correspond to Intentional self-harm, Transport accidents and Dementia/Alzheimer's. Differences in the demographic characteristics (e.g. age and sex) associated with specific causes of death, and varying proportions of difficult-to-link subpopulations, may explain much of the variation.

Certainly a high proportion of deaths by Intentional self-harm involve persons born after 1945 (75%, compared with 19% for all other causes). Young adult males are known to be under-enumerated in the Census, and this may help to explain the relatively poor link rates obtained in both the linked dataset and the gold standard.

In the case of Dementia/Alzheimer's, and perhaps other degenerative or terminal conditions requiring institutional care, it is reasonable to speculate that matching Census records may be relatively difficult to locate, if indeed they exist, due to incomplete or poor quality responses. Section 4.2.3 contains further discussion regarding the under-representation in the linked dataset of persons residing in non-private dwellings or institutions.

3.3 Implementation of Stage 2 weights

The analytical case studies presented in the next section employ Cause of death, State, Sex and Age cohort as cross-classifying variables, and this has determined the design of the second stage weighting process. After the application of Stage 2 weights, the linked dataset and the gold standard will both aggregate to the (8 × 2 × 2 =) 32 State by Sex by Age cohort totals and the 15 Cause of death totals from the Death Registrations dataset.

When Stage 2 weights are applied, counts for the 24 subpopulations identified in table 3.1 will no longer exactly align with the totals in the Death Registrations dataset as they would if only Stage 1 weights were applied.

Table 3.6 illustrates the impact of second stage weighting for a small subset of the data – Males born before 1946 whose deaths were registered in Victoria. It can be seen that the second stage weights restore the linked dataset and gold standard totals for all causes of death to that reported in the Death Registrations dataset (i.e. 14,649). This was only approximately true for the first stage weights.

Reproducing the State totals may be desirable for analyses that compare outcomes between States. However, scaling to State totals is irrelevant for intra-State analyses, and potentially distorting for analyses with a purely national focus.

3.6 Weighted estimates, by Cause of death – Males born before 1946, Victoria

				Weighted estimates			
	Unweighted record counts			*Stage 1*		*Stage 2*	
Cause of death	*Death registrations*	*Gold standard*	*Linked dataset*	*Gold standard*	*Linked dataset*	*Gold standard*	*Linked dataset*
Heart disease	2,211	2,097	1,848	2,225	2,253	2,225	2,251
Stroke	863	815	681	866	831	862	847
Dementia	899	849	642	902	783	895	828
Lung cancer	889	839	764	892	929	881	902
. . .	…	…	…	…	…	…	…
Transport accidents	39	37	36	40	44	43	47
Other causes	5,648	5,336	4,652	5,665	5,674	5,660	5,648
All causes	14,649	13,861	12,024	14,716	14,660	14,649	14,649

Note that the totals for individual causes of death are not exactly aligned by the second stage weighting. However, they will typically be better aligned than under the first stage weighting, and they are exactly aligned when summed across all 32 demographic subpopulations.

As observed for State totals, the desirability of calibrating to Cause of death benchmarks may vary depending whether the focus of the intended analysis is on a specific cause of death or involves comparisons between multiple causes of death.

4. ANALYSIS AND COMMENTARY

The purpose of this Section is to demonstrate the potential of the linked dataset to supplement the Death Registrations data with useful analytical variables from the 2011 Census, and to highlight some issues that may influence 'fitness for purpose'. These include demographic factors, weighting considerations and comparisons of output from the linked file and the gold standard.

By supplementing the Death Registrations data with 2011 Census data, it is possible to relate the characteristics of those who subsequently died within the following year to the general Australian population on Census Night.

The reasons for doing so may include:

- Descriptive analysis – learning more about persons who die from certain causes;
- Demographic analysis – calculation of mortality rates and projecting future population change;
- Policy analysis – identifying *at risk* populations, or the need to (re-)allocate resources to persons and regions of need.

Throughout this section, the gold standard linkage is presented as the best achievable quality benchmark, and indeed the results from the linked dataset generally compare very favourably against this standard.

There remain unavoidable differences in composition, however, between the gold standard linkage and the reference population (all deaths), and between the linked dataset and the gold standard. These differences generally reflect the fact that difficult-to-link subpopulations tend to be under-represented, or missing.

The analytical vignettes presented in Section 4.2 provide examples of the analytical opportunities provided by the new linked dataset, while also illustrating some ways in which linking outcomes and selective weighting processes can influence the analysis.

Section 4.3 provides a checklist of issues that careful users of the linked dataset (or indeed any linked dataset) might wish to consider.

4.1 Analysing mortality data

4.1.1 The role of demographic factors

Many characteristics measured by the Census vary according to the age and sex structure of the population. Particular causes of death may also relate to different age and sex cohorts (see Section 3) and because of this it is important to consider the interactions between these key demographics when cross-classifying Census variables by causes of death.

Deaths naturally occur mostly among the elderly. Some social and demographic characteristics of the elderly will differ from those of younger people. For instance, many older people are no longer part of the labour force, and levels of educational attainment are generally lower.

In addition to generational differences, the characteristics of an individual will also vary throughout that person's life. Characteristics such as income, family and living arrangements are all variable and are likely to change as a person ages. Snapshots such as the Census provide a point in time measure of a person's characteristics, which reflect their circumstances at that time only.

Analysis of the linked dataset needs to consider both of these issues. If deaths from a particular disease or condition generally occur at an older age, then factors associated with that group (such as educational attainment) should be compared with a similar demographic cohort. Analysis of factors which are more likely to change during a person's life (such as housing characteristics) should be presented in light of the person's likely circumstances at the time of the Census snapshot.

Particular diseases and conditions, such as cancers, heart conditions and dementia will impact on a person's circumstances in the year, or years, before their death. The Census snapshot therefore may capture a recent change in circumstances. This does not make the data any less informative, but the inferences that can be made from the data need to be contextualised appropriately.

It is possible that characteristics which may have influenced a person's health throughout their life are not able to be captured at the time close to death. This might be especially relevant to data on occupation, where either lifestyle or occupational hazards have impacted on long term health, but the occupation is no longer captured or has changed closer to the time of death.

In the linked dataset the cause listed for each individual is the *underlying cause of death* obtained from the mortality dataset. For most deaths a range of conditions are recorded on the death certificate. The underlying cause is "the disease or injury which initiated the train of morbid events leading directly to death". Among all people, but especially the elderly, multiple conditions may be present, and all can potentially impact on a person's life.

Causes of death are coded to the *International Statistical Classification of Diseases and Related Health Problems (Tenth Revision) (ICD-10)* (WHO, 2010). This classification enables conditions to be grouped for statistical analysis. The selected leading causes discussed in this section are all groupings from within the ICD-10, each of which contains multiple diseases or conditions – for example, "Dementia and Alzheimer's" includes: Vascular dementia (F01); Dementia unspecified (F03); and Alzheimer's disease (G30).

For further information on cause of death related information, see *Causes of Death, Australia* (ABS, 2013b).

4.1.2 The role of weights

The linked file has an almost endless set of possible cross-tabulations which could be created. However, the current two-stage weighting strategy, while designed to be a good general purpose approach, does not and indeed cannot, adjust for all possible analytical dimensions in the data. It is important that users evaluate the suitability of any weighting strategy on a case-by-case basis, and the discussions that follow aim to illustrate a range of weighting-related issues that may be encountered.

Section 3 described the way in which weights may be used to restore the relative contributions of selected demographic subpopulations within the linked dataset. In particular, as noted in Section 3.3, the second stage weighting process has been designed to facilitate analysis of the linked Census variables when cross-classified by Cause of death, State, Sex and Age cohort. An objective of the current section is to review the adequacy of this weighting strategy for the analytical vignettes that follow in Section 4.2.

As noted earlier (in Section 3.2), the weighting strategy adopted here for the general Australian population is not necessarily appropriate for conducting analysis of the Aboriginal and Torres Strait Islander population, nor for contrasting such findings with the non-Indigenous population.

In general, the conformably-weighted gold standard linkage will be used as the benchmark for assessing adequacy of the linked dataset, but it should be recognised that the gold standard may also have weaknesses.

For example, no matching Census record can be found if it does not exist, and varying rates of Census non-response from particular demographic subpopulations will affect both the gold standard and the linked dataset produced in this study. Analysis of particular Census variables may also be compromised within both linkages if there is systematic non-response to those questions (item non-response).

Difficult-to-link subpopulations are, of course, under-represented in both linkages. For example, persons who changed address were less likely to be (correctly) linked than those who didn't. Aboriginal and Torres Strait Islander persons in general, and especially those living in remote and very remote communities have proven difficult to link with confidence. Significant differences between the linked dataset and the gold standard can arise when difficult-to-link subpopulations are more successfully linked in the gold standard through the use of personal name information.

Whether difficult-to-link populations present a serious cause for concern depends upon whether they are represented at all in the linked datasets, and whether weighting strategies can satisfactorily compensate.

Where significant inconsistencies are observed between *weighted estimates* derived from the linked dataset and the gold standard linkage, they will generally be attributable to either

- failure of the weighting scheme to adequately compensate for the under-representation of relevant subpopulations in the linked dataset; and/or
- differences in the relative weights applied to components of the variable of interest in the linked dataset and the gold standard; and/or
- inconsistency arising from the contribution of incorrect links to the reported Census variables.

4.2 Analytical vignettes

In the sections that follow (Sections 4.2.1–4.2.4) , the interactions between social and demographic factors, causes of death, the linkage methodology and the timing of the Census snapshot are examined through several analytical vignettes. The analysis is not intended to provide comprehensive details of all interactions and does not seek to draw conclusions, but instead highlights how a user of the linked dataset must consider these wide ranging factors when drawing inferences from the data.

The Data Cube accompanying this document contains four main analytical tables. Each table features a selected analytical Census variable.[3]

4.1 Analytical tables available from the Data Cube

Table	*Description*
S.1	SEIFA Index of Relative Socio-Economic Disadvantage quintiles, by Cause of death
S.2	Highest level of educational attainment, by State and Age cohort (a) All causes of death (b) Breast cancer (c) Prostate cancer
S.3	Dwelling type and Household type, by State (a) All causes of death (b) Dementia/Alzheimer's (c) Intentional self-harm
S.4	Labour force status, by State and Sex, Persons born after 1945 (a) All causes of death (b) Intentional self-harm

[3] The *Census Dictionary* (ABS, 2011a) contains detailed descriptions of all Census variables, and may be a useful resource for obtaining a better understanding of the variables selected here, or for choosing variables for future investigation.

The tables report the estimated proportions of all Death Registration records falling into each subcategory of the analytical variable, cross-tabulated by further dimensions of the data (such as State, Age, Sex and Cause of death). The estimated proportions are calculated from weighted counts from the linked dataset and the gold standard linkage.

For some main tables, supplementary tables are provided that extract the data from the main table pertaining to selected causes of death. By this means it is possible to illustrate a diversity of interactions and analytical stories that are subsumed within the aggregate data.

Sections 4.2.1–4.2.4 comprise analytical vignettes that examine each of the four analytical variables in turn. Each section contains a rationale for selecting the variable of interest, and suggestions on the appropriate application of the variable in analysis. This is followed by illustrative results, contrasting the results from the linked dataset with the gold standard. Differences in the weighted estimates and the derived proportions are reviewed, and explanations proposed. Finally the supplementary tables are discussed, with the purpose of highlighting the context within which results need to be presented when analysing the linked dataset.

4.2.1 SEIFA Index of Relative Socio-Economic Disadvantage

Table S.1 presents the weighted and unweighted proportions of records from the Death Registrations dataset, the linked dataset and the gold standard linkage, classified to quintiles of the SEIFA Index of Relative Socio-Economic Disadvantage (ABS, 2013c), disaggregated by Sex and selected causes of death.

Context

The SEIFA Index of Relative Socio-Economic Disadvantage (IRSD) characterises the extent of relative socio-economic disadvantage within a geographical area (specifically at the ASGS Statistical Area 1 level) by combining a range of statistical indicators of well-being. These indicators include, *inter alia*, measures of household income, family composition, labour force participation and education.

Each IRSD quintile represents 20% of the total number of areas ranked using index characteristics. However, as populations of areas are not equal this does not guarantee that each quintile represents 20% of the total population. To minimise the impact of population differences across areas, the IRSD quintiles have been re-defined on a population basis, ensuring that the aggregate populations of each quintile are approximately equal on Census Night 2011.

Both the Census and the mortality datasets collect information on a person's 'usual residence'. However, the IRSD quintiles used in this study only relate to the addresses recorded on the Death Registration. Where the usual residence recorded at the time of the Census differs from that on the Death Registration, the IRSD quintile to which the person is allocated may differ from that which would have been obtained based on the usual residence from the Census record.

Changes to usual residence are potentially a significant issue for the linked dataset. Many deaths occur within hospitals or other care facilities (such as nursing homes or hospices). While short term stays are unlikely to affect usual residence, longer duration stays in care facilities may change both a person's usual residence and potentially their assigned IRSD quintile.

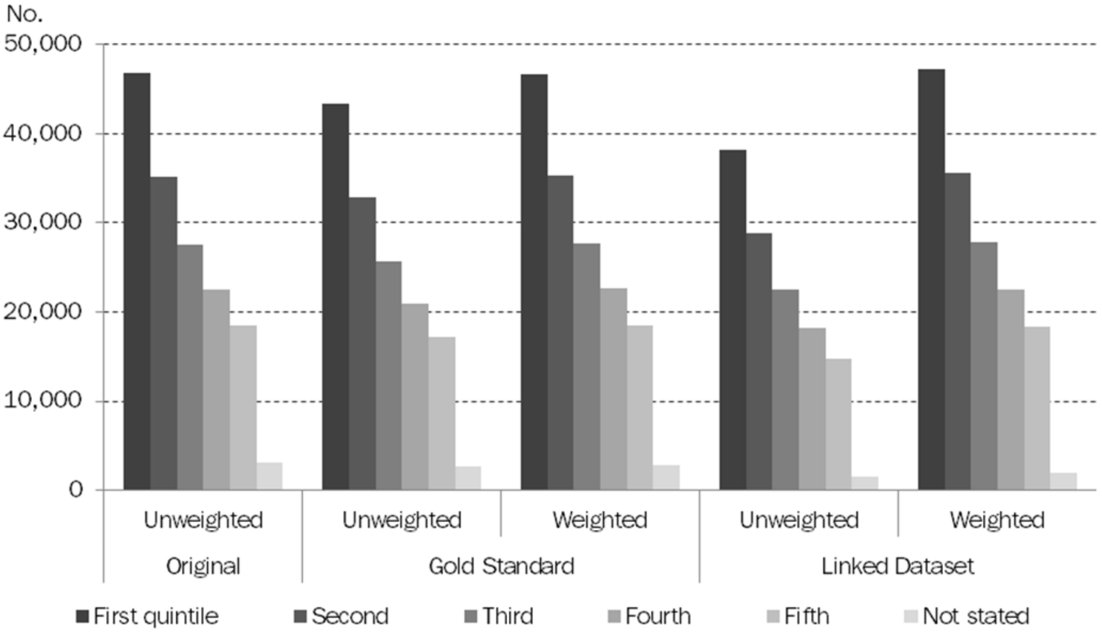
Despite this potential drawback, the clear advantage of using the addresses from the Death Registrations dataset for this study is that they are available for all records – allowing for an analysis of differential link rates.

Almost 80% of persons recorded in the Death Registrations dataset were aged over 65 years on Census Night. As previously discussed, age can impact on a person's labour force status, income and educational qualifications (see ABS, 2013c), and consequently their contribution to the IRSD quintile to which an area is assigned.

This highlights two potential issues for the linked dataset:

1. The age structure of an area may influence the IRSD quintile to which it is assigned.
2. The age ranges associated with a particular cause of death (i.e. median age at death) may influence or bias how the deaths are distributed across IRSD quintiles.

4.2 Number of deaths, by SEIFA IRSD quintile



Comparison

Figure 4.2 displays the weighted and unweighted record counts from the Death Registrations dataset, the linked dataset and the gold standard linkage, cross-classified by SEIFA IRSD quintile. It is notable that, although the quintiles represent equal shares of the Census population, the numbers of deaths reported in the most disadvantaged quintiles far exceed the numbers reported in the least disadvantaged quintiles. The following sections examine methodological and statistical issues that can inform the interpretation of this result.

The proportional distributions of the weighted counts for all causes of death across the quintiles were examined to ensure consistency of outcomes across methodologies (original, gold standard and the linked dataset). These are presented in table 4.3. All three datasets display strongly consistent outcomes, especially when allowance is made for differences in the "not stated" category.[4]

4.3 Distribution of deaths (%), by SEIFA IRSD quintile

	IRSD quintile						
	First	*Second*	*Third*	*Fourth*	*Fifth*	*Not stated*	*Total*
All causes of death				%			
Original	30.5	22.9	17.9	14.7	12.0	2.0	100.0
Gold standard	30.4	23.0	18.0	14.7	12.1	1.8	100.0
Linked dataset	30.8	23.2	18.2	14.7	12.0	1.2	100.0

Stage 2 weights calibrate the counts in the linked dataset and the gold standard linkage to the same Death Registrations benchmarks for Cause of death, but small discrepancies will arise when further cross-classification is introduced. Data users should consider this when interpreting results.

Table S.1 in the associated data cube provides counts for selected leading causes of death by IRSD quintile. Even where discrepancies occur in the weighted totals, the proportional distributions show strong consistency between the linked dataset and the gold standard at all disaggregated levels, especially when allowance is made for the "not stated" category. The largest differences, of up to two percentage points for some quintiles, are observed for Dementia/Alzheimer's and Intentional self-harm.

In terms of weighted total counts, the largest discrepancies are observed for male deaths from Dementia/Alzheimer's (4.8%) and female deaths from Intentional self-harm (4.0%). Significantly, it was observed in Section 3.2 that deaths from Dementia/Alzheimer's were relatively under-represented in the linked dataset, and deaths from Intentional self-harm were under-represented in both the linked dataset and the gold

[4] Death Registration records without ASGS address information were significantly harder to link – especially without detailed name information – and are therefore under-represented in the linked datasets.

standard linkage. The application of larger Stage 2 weights to these causes of death would naturally raise the implicit standard errors for the weighted counts.

Given the very high proportion of deaths that occurred in areas of greater relative socio-economic disadvantage, the link rates for each IRSD quintile were checked to eliminate weighting bias as a driver for this difference. Table 4.4 shows that uniform linkage rates have been realised over all quintiles.

4.4 Link rates (%), by SEIFA IRSD quintile

	IRSD quintile					
	First	*Second*	*Third*	*Fourth*	*Fifth*	*Total*
All causes of death			%			
Gold standard	92.7	93.4	93.6	93.4	93.3	93.0
Linked dataset	81.5	81.9	81.8	80.8	80.3	80.7

As noted earlier, a change in residence close to the time of death (e.g. to obtain nursing care) can affect assignment to a particular IRSD quintile. In addition, the age structure of the population within an area may influence the IRSD quintile to which the area is assigned.

Within this context, it is informative to look at differences between the proportional distributions for selected causes of death (see table 4.5 and Table S.1 in the data cube). It is apparent that for all of the selected leading causes, the highest proportions of deaths are in the most disadvantaged areas, and proportions progressively decrease across quintiles. However, the magnitude of this phenomenon can differ significantly for particular conditions.

4.5 Distribution of deaths (%), by SEIFA IRSD quintile and selected causes of death

	IRSD quintile						
	First	*Second*	*Third*	*Fourth*	*Fifth*	*Not stated*	*Total*
All causes of death				%			
Original	30.5	22.9	17.9	14.7	12.0	2.0	100.0
Gold standard	30.4	23.0	18.0	14.7	12.1	1.8	100.0
Linked dataset	30.8	23.2	18.2	14.7	12.0	1.2	100.0
Diabetes							
Original	36.9	23.8	16.6	12.2	8.4	2.1	100.0
Gold standard	37.1	23.9	16.6	12.2	8.3	2.0	100.0
Linked dataset	37.6	24.5	16.5	12.2	8.0	1.2	100.0
Breast cancer							
Original	23.1	21.3	19.5	17.8	16.9	1.4	100.0
Gold standard	22.7	21.2	19.6	18.2	17.0	1.2	100.0
Linked dataset	22.7	21.5	19.8	18.1	17.3	0.7	100.0

The greatest disparity is observed for Diabetes-related deaths (37.6% in the first quintile, decreasing to 8.0% in the fifth quintile). The median age at death for Diabetes is 81 years – approximately average for the whole population. Age at death is therefore likely to be an influencing factor, but the magnitude of the bias towards the lower IRSD quintiles may be greater than that which would be expected because of age factors alone. Therefore it is possible that there are IRSD related factors other than age which influence the likelihood of death from diabetes.

The least disparity is observed for Breast cancer (22.7% in the first quintile, decreasing to 17.3% in the fifth). The median age at death for Breast cancer is 69 years, much younger than for many other causes. It would therefore be expected that age-related IRSD bias would be less evident across those who died from breast cancer, and this is clearly shown in the dataset.

4.2.2 Educational attainment

Table S.2 presents estimates of the distribution of educational attainment levels for persons reported in the Deaths Registrations dataset, disaggregated by State and broad Age cohort.

Supplementary tables provide similar information, but pertaining specifically to females who died from Breast cancer and males who died from Prostate cancer.

Context

In the Australian population as a whole, educational attainment varies significantly over age cohorts and between males and females:

- Younger generations report higher levels of post-Year 12 qualifications than do older persons, while a considerable proportion of older persons report that they did not complete Year 12;
- Males are significantly more likely than females to have acquired certificate / diploma / trade qualifications, irrespective of age;
- Correspondingly, females are more likely than males to report that they have completed Year 12, without acquiring further qualifications.

Thus decomposition by age and sex is highly informative when examining educational attainment in the Deaths Registrations data. In general, the lower educational attainment levels observed for persons reported in the Deaths Registrations dataset are broadly consistent with the attainment levels of other Australians of corresponding age and sex.

Comparison

Figure 4.6(a) displays estimates of the educational attainment profile of persons who died from all causes within the 13 months following the 2011 Census. The proportions are derived from weighted counts from the linked dataset and the gold standard linkage. To illustrate the effects of age and sex on educational attainment among deaths, the charts show results by broad age group (0–65 years, 66 years and over) and examine separately deaths from Breast cancer and Prostate cancer. Table S.2 provides more detail, showing a State by broad Age cohort disaggregation.

The second stage weighting strategy (Section 3.3 and table 3.6) ensures that the weighted totals for all cross-tabulations in Table S.2 are identical with the Death Registrations benchmarks, for both the linked dataset and the gold standard linkage.

Consequently, any inconsistencies observed in the proportional distributions estimated from the linked dataset and the gold standard arise because

- different Death Registration records have been included in the two linkages;
- records common to both linkages may nevertheless be linked to different Census records; and
- different weights will be applied to the individual records within each linkage.

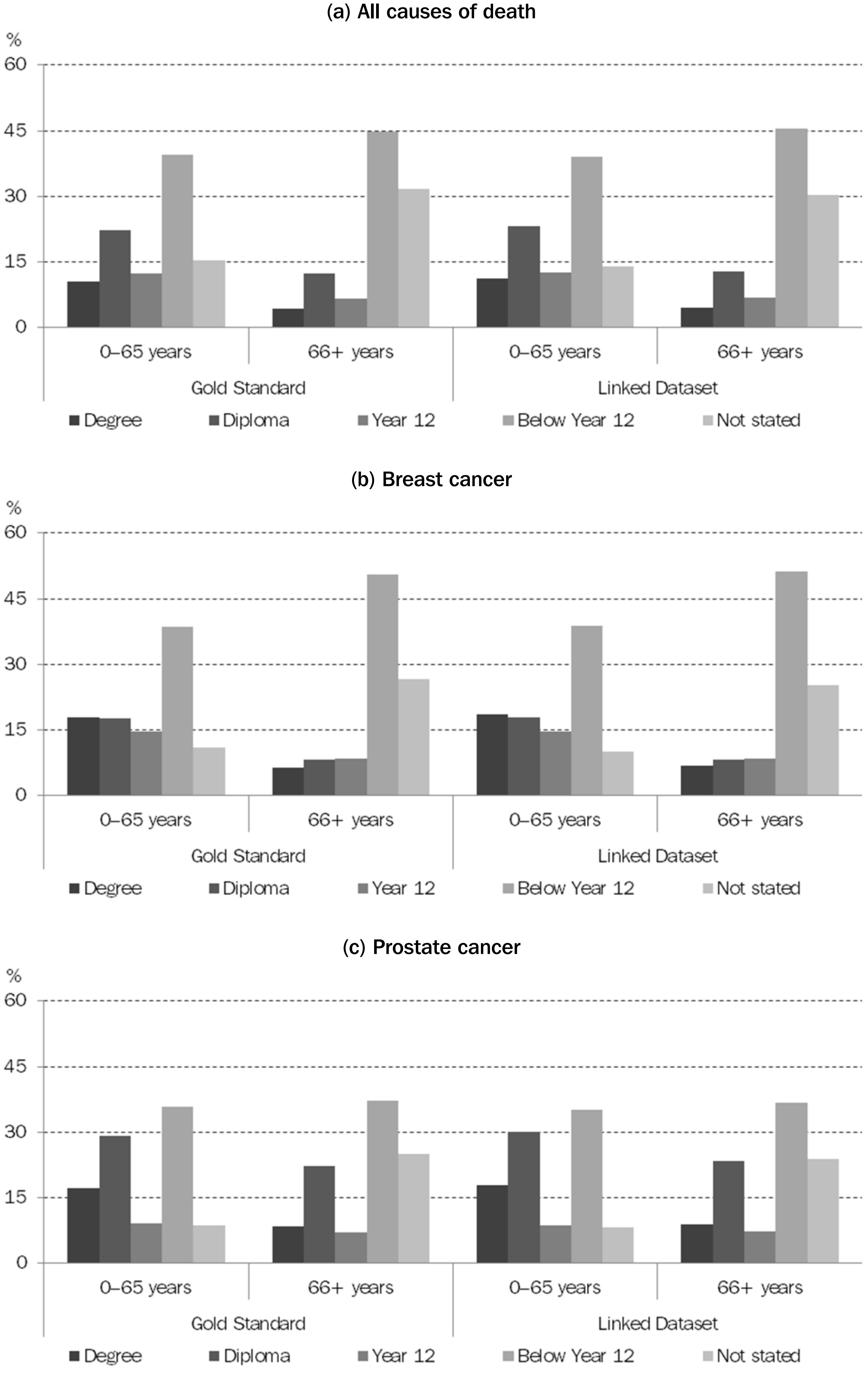
It is important to reiterate that the linked dataset contains incorrect links, which may report levels of educational attainment that are also incorrect for the individual – although, in this example, there is no reason to expect that these incorrect responses will necessarily bias the results.

A notable feature of figure 4.6(a) is the significant proportion of "Not stated/ Insufficient information" responses – 14% for younger persons (born after 1945) and 30% for older persons (born before 1946).

Non-response may be more prevalent for persons with particular medical conditions or Census Night circumstances, and so care is required when contrasting the distributions of valid responses for different causes of death. Persons residing in non-private dwellings (such as hospitals or nursing homes) on Census Night are usually required to complete a Personal Form (rather than the usual Household Form). For many people in this situation, this is an impractical task, and it is likely that many such forms will be returned with minimal information (possibly having been filled in by a carer or staff member). This is a possible explanation for the high non-response rates observed for the educational attainment question, and other questions requesting detailed personal information – especially for the older age cohort.

The "Not stated" proportion will generally be lower in the linked dataset than in the gold standard, because the factors leading to item non-response may also create difficulties for record linkage.

4.6 Educational attainment, by Age cohort



(a) All causes of death

(b) Breast cancer

(c) Prostate cancer

The supplementary tables provided for Table S.2 report the same cross-tabulations of the data, but restricted to two selected causes of death – Breast cancer and Prostate cancer. While these tables are not explicitly disaggregated by sex, they are each dominated by one sex only. Deaths from Breast cancer relate (almost) exclusively to females while deaths from Prostate cancer relate exclusively to males.

The educational attainment indicators for all deaths from Breast cancer are influenced by the relatively high proportion of deaths to persons aged 0–65 years (45%) whereas the corresponding indicators for all deaths from Prostate cancer are only marginally influenced by deaths in this age cohort (9%) (see Table S.2).

Figure 4.6(b) displays educational indicators for females who died from Breast cancer by age cohort and figure 4.6(c) does likewise for males who died from Prostate cancer. While both 4.6(b) and 4.6(c) are noticeably different from figure 4.6(a), they both align consistently with an age-sex decomposition of "all causes of death" and also with a similar decomposition of the broader Australian population.

The proportional distributions derived from the linked dataset and the gold standard linkage provide strongly consistent information for both of these causes of death.

Comparisons between the States and Territories show general uniformity across the six States, but some significant differences for the Territories:

- Reflecting educational attainment in the population overall, the average educational attainment levels reported for persons whose deaths were registered in the Australian Capital Territory are significantly higher in both the younger and older age cohorts.
- The results for the Northern Territory appear spurious and may be affected by small sample and poor link rate considerations.

Comparison of the estimated distributions obtained from the linked dataset and the gold standard linkage shows an average deviation[5] of less than 0.5 percentage points for all estimates pertaining to the largest five States. For Tasmania and the ACT, this rises to just under one percentage point.

The Northern Territory returns an average deviation of three percentage points, with the largest discrepancy reported for younger persons who did not complete Year 12 (6.1 percentage points). This result is most likely explained by the small sample sizes present, and the relatively poor link rates achieved in both record linkages (leading to higher second stage weights) (see Section 3.2). The higher educational attainment of the older age cohort suggests that older persons with tertiary qualifications may be easier to link, and are therefore over-represented in both datasets.

[5] Calculated as the root mean square deviation = $\sqrt{\sum_{i=1}^{n}\left(p_{1i}-p_{2i}\right)^2 / n}$.

4.2.3 Dwelling type and household composition

Table S.3 presents Census dwelling type and household composition information. Estimates are provided of the proportion of persons living in private and non-private dwellings, and within various household situations, disaggregated by State and Sex.

The supplementary tables provided for Table S.3 report the same Dwelling type and Household composition information, but cross-tabulated by State only, and restricted to two selected causes of death – Dementia/Alzheimer's and Intentional self-harm.

Context

The principal dwelling types classified in the Census are 'private dwellings' and 'non-private dwellings'. Self-contained accommodation in retirement villages is typically included with the former, while nursing home accommodation is generally classified under 'non-private dwellings'. Non-self-contained accommodation for the aged, hostels and refuges are also classified as 'non-private dwellings'. The household composition information elaborates on the circumstances of persons living in private dwelling accommodation – specifically whether the person lives alone, with a partner and/or children, or in a group household.

The living arrangements of individuals are influenced by life-cycle factors. For instance, younger persons in the Death Registrations dataset are more likely to be living with family in a private dwelling (as either a parent or child). Longer life expectancy for females compared to males make it more likely that elderly males will be living with a partner prior to death, while elderly females will more likely live alone or in nursing home accommodation.

The living arrangements of a person may also be influenced by health conditions from which they are suffering at the time of the Census snapshot. Some chronic, degenerative or terminal conditions may increase the person's need for institutional care prior to death. This would constitute a change from the person's 'normal' living arrangements but does reflect the living arrangements of people suffering those types of conditions.

On the other hand, people who die from sudden onset conditions or from many external causes (i.e. accidents, assaults or intentional self-harm) are more likely to be living in their 'normal' household type. Their living arrangements could therefore be more accurately compared with others in the population in a similar demographic group.

Comparison

The second stage weighting strategy (Section 3.3) ensures that the weighted totals for all cross-tabulations in Table S.3 are identical with the Death Registrations benchmarks, for both the linked dataset and the gold standard linkage.

The estimates shown in figure 4.7 indicate that, on Census Night, three out of four males who died within the following 13 months were living in private dwellings, and half of these were living in a "Couple only" household. About 15% of males were living in nursing home accommodation.

By contrast, it is estimated that 43% of females in the Death Registrations dataset were living in non-private dwellings (predominantly nursing homes) on Census Night, and 20% were living alone in a private dwelling. The proportion of females living in a "Couple only" household (18%) was half that observed for males.

Slightly more males (21%) than females (18.5%) were living in private dwellings with other family members or in a group housing situation.

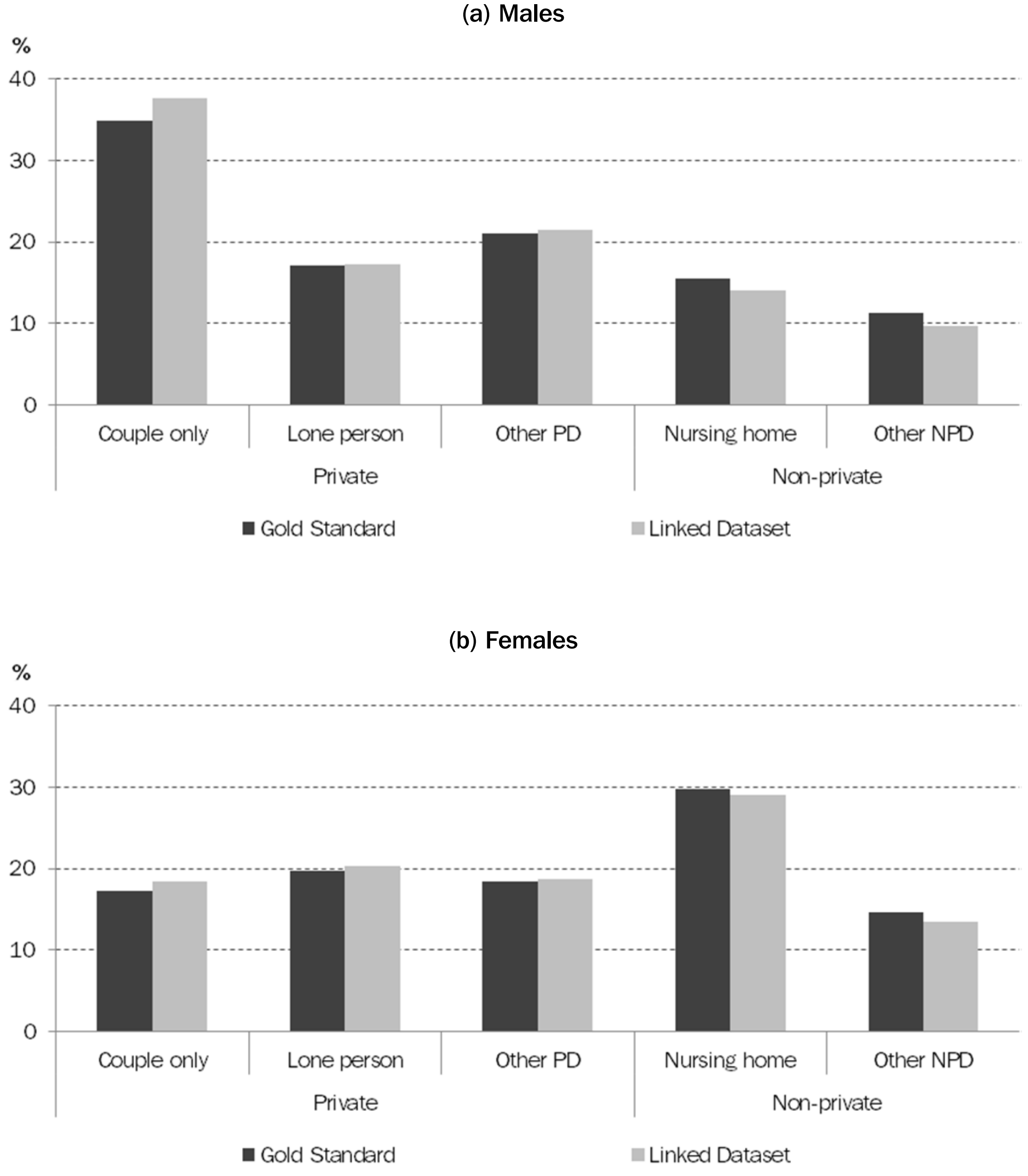
These housing characteristics are reasonably consistent across most States and Territories, but can be influenced significantly where the age and sex composition of deaths differs, or where the proportions of deaths from different conditions differ. This is the case in the Northern Territory, where a greater proportion of deaths occur in younger age groups as well as from external causes of death. Consequently Northern Territorians who died in the 13 months after the Census were more likely than other Australians to be living in a family ('Couple family with children' or 'One parent household') and less likely to be living in a non-private dwelling.

Throughout Table S.3 there is some evidence of systematic inconsistency between the proportions derived from the linked dataset and the gold standard linkage. In particular, the proportion of persons living in non-private dwellings is uniformly under-estimated in the linked dataset.

This is essentially due to the lower link rates realised for residents of non-private dwellings in the linked dataset (relative to the gold standard), and the fact that the weighting strategy cannot explicitly adjust for this under-representation.

The supplementary tables examine two selected causes of death – Dementia/ Alzheimer's and Intentional self-harm (see figures 4.8(a) and 4.8(b)). These causes of death were chosen to illustrate the underlying diversity of dwelling and household circumstances as they pertain to specific medical conditions and causes of death.

4.7 Dwelling type and Household type, by Sex, All causes of death



(a) Males

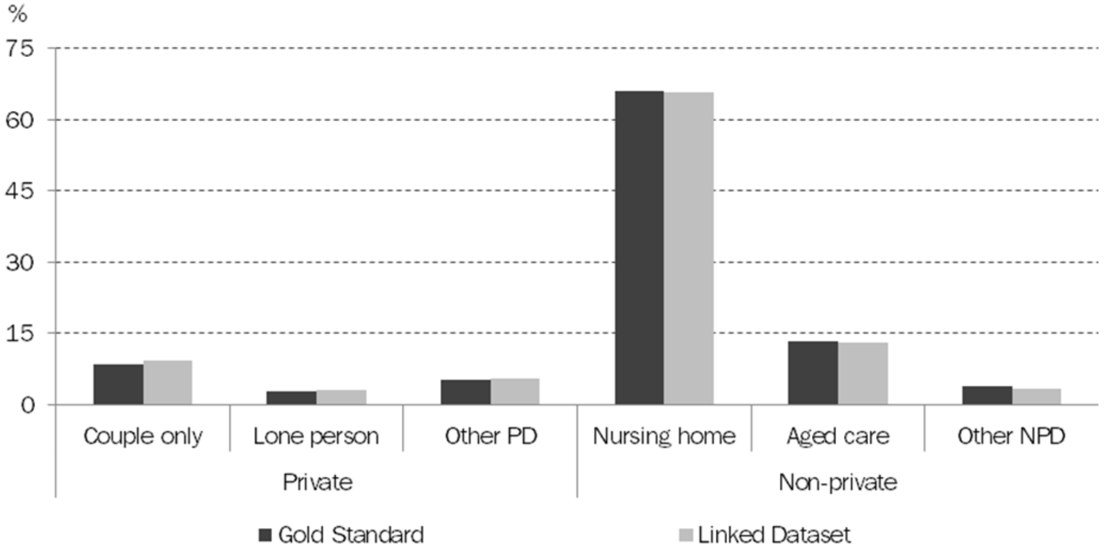
(b) Females

Dementia/Alzheimer's is a long-term degenerative condition, and it is unsurprising that more than 80% of those who died from this condition within the reference period were already living in non-private dwelling (mainly nursing home) accommodation. Most of those not already in care were living with their partner (9%).

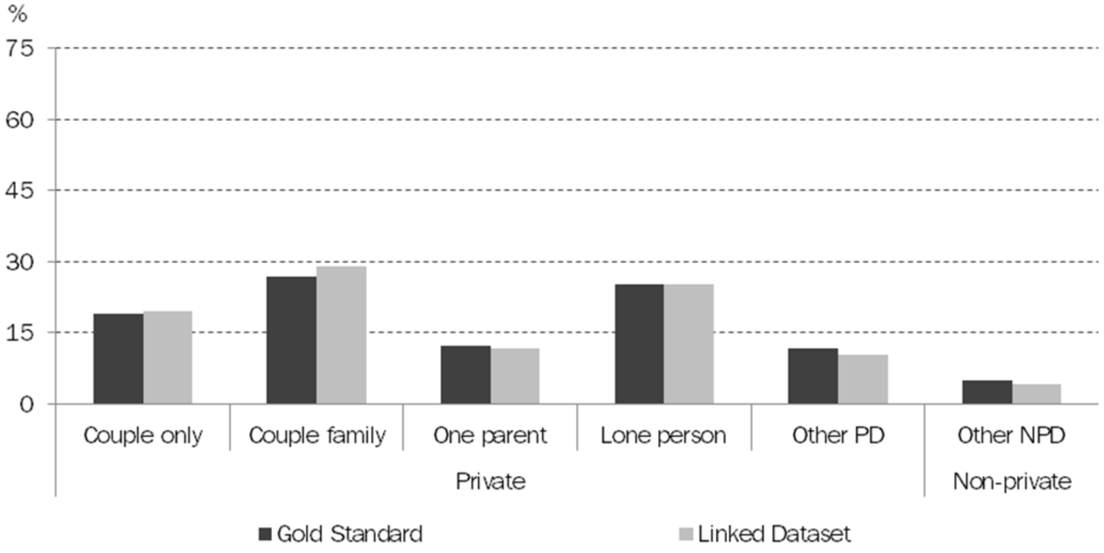
Almost 85% of deaths from Intentional self-harm involve persons in the younger Age cohort (born after 1945). Figure 4.8(b) shows that, on Census Night, 41% of persons in this category were living in a family (either as a parent or child), 20% were living with a partner and 25% were living alone.

4.8 Dwelling type and Household type, by selected causes of death

(a) Dementia/Alzheimer's



(b) Intentional self-harm



Despite some inconsistencies observed in the weighted totals, the proportional distributions derived from the linked dataset and the gold standard linkage provide strongly consistent information for both selected causes of death.

4.2.4 Labour force status

Table S.4 presents estimates of the degree of Labour Force engagement within the Death Registrations population, disaggregated by State and Sex, and restricted to the younger Age cohort (persons born after 1945) only.

The supplementary table provided for Table S.4 presents the Labour Force status on Census Night for persons who subsequently died from Intentional self-harm.

Context

Labour force status is unlikely to prove informative for the 80% of persons in the linked dataset who were born prior to 1946. Hence the table focuses on the younger age cohort, who were aged 66 years or under in 2011. This age cohort will align, to a reasonable approximation, with the 15–64 years age cohort presented in the *Labour Force, Australia* publication (ABS, 2011b).[6]

Two labour force statistics are likely to be of general analytical interest:

- the proportion of persons in the population who were in the labour force (i.e. either employed or unemployed) on Census Night (the participation rate), and
- the proportion of persons in the labour force who were unemployed (the unemployment rate).

Comparisons of participation rates and unemployment rates between the linked dataset and the broader Australian population will typically involve some hypothesis about the reasons for any observed differences. For example:

- Does a person's medical condition prevent them from seeking, finding or retaining satisfactory employment?
- Does an inability to find satisfactory employment limit the quality of care that a person can afford, or otherwise contribute to a deterioration in well-being?

That is, labour force status will often be used for more than descriptive analysis.

Comparison

The second stage weighting strategy (Section 3.3) ensures that the weighted totals for all cross-tabulations in Table S.4 are identical with the Death Registrations benchmarks, for both the linked dataset and the gold standard linkage.

[6] Of course, there remain discrepancies between the age-sex composition of the Death Registrations dataset and the Australian population aged 15–64 years. For example, the younger age cohort on the Death Registrations dataset includes a small number of persons aged under 15 years, but more significantly is dominated by persons aged 55 years and over.

The estimates supplied in table 4.9 show that in the linked dataset, among persons born after 1945, 42% of males and 29% of females were reported as being in the labour force on Census Night. Approximately 39% of males and 27% of females were identified as "Employed".

4.9 Labour Force status of persons born after 1945, by Sex and selected causes of death

	Labour Force Status					
	Labour Force			*Not in the Labour Force*	*Not stated*	*Total*
	Employed	*Unemployed*	*Total*			
All causes of death			%			
Males						
Gold standard	35.9	3.3	39.2	53.6	7.2	100.0
Linked dataset	38.5	3.2	41.7	52.2	6.1	100.0
Females						
Gold standard	25.4	1.8	27.2	65.4	7.4	100.0
Linked dataset	27.2	1.6	28.8	64.9	6.3	100.0

As discussed in earlier sections, the proportion of "Not stated" responses (6–7% in this case) may be attributable to factors related to the health or incapacity of the Census respondent, and will probably vary across the different causes of death.

Comparison of the proportional distributions computed from the weighted linked dataset and gold standard linkage reveals systematic over-estimation of the "Employed" category in the linked dataset. This is most likely the result of relatively poorer link rates associated with persons in the "Unemployed" and "Not in the labour force" categories, and the fact that the weighting strategy cannot adjust for under-representation in these categories.

An unresolved issue is the possibility that the gold standard estimates may themselves contain some systematic distortion arising from the relative difficulty of finding matches for unemployed persons.

The distortion to the estimates arising from relative under-representation of the "Unemployed" and "Not in the labour force" categories is amplified in the case of the Northern Territory, where poor overall link rates in both the linked dataset and the gold standard have already been identified (see Section 3.2) and the sample size is small. Relative to the gold standard, the linked dataset estimates over-state the proportion of employed males by 9 percentage points.

One of the stated advantages of appending Census variables to the Death Registrations dataset is the capacity to compare the characteristics of the persons who died with the characteristics of the broader Australian population at a specific point in time (Census Night).

Table 4.10 provides a simple example of such a comparison. Using the weighted estimates that underpin table 4.9, it is straightforward to compute estimates of the Participation Rate and the Unemployment Rate for the younger cohort of persons recorded on the Death Registrations dataset.[7] These may then be compared with the official Labour Force statistics for all Australians aged 15–64 years, released in August 2011 (ABS, 2011b).

4.10 Labour force derived indicators for persons born after 1945, by Sex and selected causes of death

	All causes of death		*Intentional self-harm*		*All Australians aged 15–64 years (ABS)*
	Gold standard	*Linked dataset*	*Gold standard*	*Linked dataset*	
Unemployment Rate (%)					
Males	8.4	7.8	12.1	11.7	5.5
Females	6.6	5.6	17.0	14.2	5.4
Participation Rate (%)					
Males	42.2	44.4	68.5	70.2	83.0
Females	29.4	30.7	52.8	52.5	70.7

The results in table 4.10 for "All causes of death" reveal significantly lower participation rates and higher unemployment rates for the Death Registrations population born after 1945. In line with the preceding discussion on the systematic over-estimation of "Employed" persons, it may be noted that the estimates obtained from the linked dataset again over-estimate the participation rate and under-estimate the unemployment rate, relative to the gold standard.

Table 4.10 highlights that those people in the linked dataset who died from Intentional self-harm had labour force participation rates that were higher than those for "All causes of death". However, the participation rates for this cohort are still appreciably lower than official participation rates for all Australian males and females at that time.

The unemployment rates for both males and females who died from Intentional self-harm are more than double the official rates for Australia at that time. They also significantly exceed the unemployment rates for all people in this age group who died in the 13 months after the Census.

Table S.4 in the accompanying data tables show additional State, Territory and age group breakdowns of employment status by selected causes of death.

[7] Weighted counts pertaining to the "Not stated" category were not used in the calculation.

4.3 A checklist for analysing linked data

Sections 4.2.1–4.2.4 provided a glimpse into the analytical possibilities provided by the new Death Registrations to Census record linkage. Comparisons with the gold standard linkage from the previous study have generally endorsed the validity of the distributional analyses presented, and this is a very positive indication of the 'fitness for purpose' of the new dataset.

These case studies were also selected to highlight some of the issues and weaknesses that will confront users of the linked dataset (and linked data generally). The intention has been to encourage such users to consider and recognise these issues, and form balanced judgements about the implications for statistical inference.

The remainder of this section recaps issues that users of linked data are advised to consider. Some issues will appear familiar to those who regularly analyse survey or administrative data, but the same issues can actually be significantly more complex in the context of linked data. Other issues are peculiar to linked data, and not all are necessarily capable of being recognised, quantified or corrected for.

Is the weighting strategy appropriate for the proposed analysis?

While the Death Registrations dataset is comprehensive of all deaths in the reference period, it is important to recognise that the linked dataset is *not* complete, nor does it constitute a representative sample. Difficulties encountered and decisions made in the record linkage process inevitably lead to some subpopulations being under-represented in the linked dataset – possibly even missing completely.

Section 3 discussed the potential to use weighting to compensate for the different link rates realised for certain key subpopulations. In particular it was noted that there is no universal weighting strategy that can anticipate the needs of all users of the data, and that the choice of weighting benchmarks can prove simultaneously beneficial and detrimental to answering different analytical questions.

The weighting strategies applied in this study were designed to be appropriate for the case studies presented – while also being sensible for general application. However, a critical user should consider the following questions when proposing to undertake further analyses:

- Is it better to use unweighted data, Stage 1 weights or Stage 2 weights?
- Are the Stage 2 weights as presented here suitable for within strata, across strata or population-wide comparisons?
- Do they restore or distort the key demographic dimensions relevant to the analysis?

As the weighting strategy devised for this linked dataset is required to support a very general research agenda, it is important to caution against the use of the weighted data for investigations that are too narrowly focussed on specific subpopulations.

For example, the use of this weighted dataset to analyse issues of Aboriginal and Torres Strait Islander mortality is not recommended, and the ABS is currently considering further work to develop an alternative two-stage weighting strategy that may be better suited to such investigations.

What is the extent of residual under-representation?

Weights are typically used to restore benchmark aggregates identified from the reference population (the Death Registrations dataset in this case). However, it is also usually the case that *only* benchmark variables from the reference population can be used to define the weights. These variables will typically be insufficient to address all instances of systematic under- and over-estimation in the linked data.

By examining the distributions of responses to Census analytical variables, Section 4 has identified under-representation for several demographic categories that cannot be resolved by weighting strategies – because suitable population benchmarks are not available. Examples include the residents of non-private dwellings (Section 4.2.3) and unemployed persons (Section 4.2.4). Such instances of unavoidable bias will typically be an artefact of the linkage process. Some demographic subpopulations have poor Census response rates, while others are simply difficult to link.

What is the impact of incorrect links

In devising the specification for the final linked dataset (Section 2.4), it was necessary to trade off 'representativeness' against 'precision'. Accepting only the most certain linked record pairs will have the potential to inadvertently eliminate characteristics of interest from the dataset. By choosing to retain diversity and representativeness, the user must contend with a known rate of linkage error.

The impact of incorrect links on analysis will typically be determined by the degree of correlation or causality that is expected between the variables under investigation.

This can be demonstrated by reference to two examples from Section 4.

1. Educational attainment is typically not highly correlated with cause of death, and incorrect links would not be expected to bias the overall profile of educational attainment. This assumes that a Census respondent who was incorrectly linked exhibits similar demographic characteristics (age, sex, geographical location) to the person on the Death Registrations dataset, and is likely to have had similar educational opportunities.

2. Conversely, Labour Force status may be directly related to a person's state of health, and an incorrect link is unlikely to return a satisfactory response on the basis of similar demographic characteristics alone.

Knowledge of the overall linkage error rate can provide a basis for estimating the bias contributed by incorrect links.

How does Sample size affect the analysis?

The case studies presented in this section identified a number of instances in which the analysis of small subpopulations produced spurious results, or results that differed markedly from the gold standard. In the absence of a gold standard, users must assess whether the quality of the results may have been adversely affected by poor link rates (and high compensating weights), poor quality or low precision links (producing incorrect responses to the variable of interest), or normal statistical variation.

What is the effect of missing records and item non-response

Users typically have no control over the degree of person-level or item non-response in the Census. However, an additional level of care is advisable for linked data. Subpopulations that have poor response rates may also have higher false link rates. Respondents who do not answer the question of interest may also have been linked on the basis of poorer quality or incomplete responses.

5. CONCLUSIONS

Following on from the *Death Registrations to Census Linkage Project* in 2012, the objective of the current study was to construct a new linked dataset that might provide a valuable ongoing resource for research, within the governance constraints of the earlier project. Effectively this required re-linking the Death Registration records to the 2011 Census, without the benefit of using detailed name and address information available to the earlier project. However, information from the earlier linkage was available to guide and validate the construction of the new linked dataset, and to provide a 'gold standard' by which to evaluate its 'fitness for purpose'.

Section 2 explained the probabilistic record linkage methodology used to construct the new linked dataset, and provided the diagnostic statistics that were used to define the quality and composition of the final product.

There is no doubt that the new linked dataset is inferior to the gold standard linkage in terms of the completeness, representativeness and precision of the assigned links. Nevertheless, the supplementary personal information attached to the Death Registration records in the linked dataset provides a unique opportunity to understand and illuminate many of the circumstances surrounding death. Also, the fact that the Death Registrations dataset documents the period immediately following Census Night 2011 provides opportunities to undertake richer demographic modelling of the broader Australian population.

However it is crucial that users understand the nature and limitations of probabilistically-linked datasets.

As already noted, specification of the linked dataset involved an informed trade-off between the 'representativeness' and the 'precision' of the final product. The informed user must acknowledge and accommodate the consequences of this decision.

Section 3 discussed the use of weights to restore the representativeness of selected subpopulations. Stage 1 weights were provided to adjust for differing link rates observed in the record linkage process, and Stage 2 weights were proposed to calibrate the dataset to benchmark totals that might be the focus of analysis. The message from Section 3 was that weighting is a useful tool for restoring the representativeness of selected subpopulations in the linked dataset. It cannot, however, completely overcome inadequacies in the linkage itself. Indeed, there may be occasions when analysis of the unweighted data is preferable.

Section 4 selected several analytical variables and examined the methodological issues that might arise in an empirical analysis based upon the linked dataset. The gold standard linkage from the earlier study was used to adjudge the value of the new linked dataset.

The linked dataset performed well in reproducing distributional analyses consistent with the gold standard, and should prove capable of delivering valuable and reliable insights into many other dimensions of the data.

Without detracting from the positive findings of the study that confirmed the analytical opportunities of the linked dataset, the extended discussion in Section 4 sought to provide insights into important issues and potential weaknesses that are inherent in linked data. Familiarity with these issues will assist users to frame suitable questions, select appropriate weighting options, conduct sound inference and recognise the magnitude and direction of potential biases in the results.

The combination of information on causes of death and contextual circumstances enables a range of public and research benefits for understanding mortality and developing health policy. The Australian Bureau of Statistics, together with the Registrars of Births, Deaths and Marriages, intends to reproduce similar linked datasets from subsequent Census and Death Registrations data. The greatest benefit to the utility of future linked datasets will be gained from improvements to linkage quality, and a primary way of achieving this will be through the expanded use of name information in linking. Pertinent to this issue is the announcement, in December 2015, that the ABS will “retain the names and addresses collected in the 2016 Census of Population and Housing to provide a richer and dynamic statistical picture of Australia through the combination of Census data with other survey and administrative data.”

ACKNOWLEDGEMENTS

The principal authors of this report were Peter Rossiter, Gwenda Thompson and Shamali Weeraratne. They would like to acknowledge the support and advice received from members of the multi-disciplinary ABS Working Group that had oversight of this important project, and in particular the direction and leadership provided by Andrew Webster.

REFERENCES

Australian Bureau of Statistics (2010a) *Australian Statistical Geography Standard (ASGS): Volume 1 – Main Structure and Greater Capital City Statistical Areas, July 2011*, cat. no. 1270.0.55.001, ABS, Canberra. < http://www.abs.gov.au/ausstats/abs@.nsf/mf/1270.0.55.001 >

—— (2010b) *Census Data Enhancement Project: An Update*, cat. no. 2062.0, ABS, Canberra. < http://www.abs.gov.au/ausstats/abs@.nsf/mf/2062.0 >

—— (2011a) *Census Dictionary, 2011*, cat. no. 2901.0, ABS, Canberra. < http://www.abs.gov.au/ausstats/abs@.nsf/mf/2901.0 >

—— (2011b) *Labour Force, Australia, August 2011*, cat. no. 6202.0, ABS, Canberra. < http://www.abs.gov.au/ausstats/abs@.nsf/mf/6202.0 >

—— (2013a) *Australian Statistical Geography Standard (ASGS): Volume 5 – Remoteness Structure, July 2011*, cat. no. 1270.0.55.005, ABS, Canberra. < http://www.abs.gov.au/ausstats/abs@.nsf/mf/1270.0.55.005 >

—— (2013b) *Causes of Death, Australia, 2013*, cat. no. 3303.0, ABS, Canberra. < http://www.abs.gov.au/ausstats/abs@.nsf/mf/3303.0 >

—— (2013c) *Census of Population and Housing: Socio-Economic Indexes for Areas (SEIFA), Australia, 2011*, cat. no. 2033.0.55.001, ABS, Canberra. < http://www.abs.gov.au/ausstats/abs@.nsf/mf/2033.0.55.001 >

—— (2013d) *Death Registrations to Census Linkage Project – Methodology and Quality Assessment*, cat. no. 3302.0.55.004, ABS, Canberra. < http://www.abs.gov.au/ausstats/abs@.nsf/mf/3302.0.55.004 >

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

Organisation for Economic Cooperation and Development (2015) *Health Inequalities*, OECD web site. < http://www.oecd.org/health/inequalities-in-health.htm >

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. < http://www.abs.gov.au/ausstats/abs@.nsf/mf/1351.0.55.025 >

World Health Organisation (2010) *International Statistical Classification of Diseases and Related Health Problems (Tenth Revision) (ICD-10)*, 2010 edition, WHO.

APPENDIX

A. LINK RATES FOR SELECTED SUBPOPULATIONS

A.1 Link rates, selected subpopulations

	Death Registrations (No.)	*Record counts (No.)*		*Link rates (%)*	
		Gold standard	*Linked dataset*	*Gold standard*	*Linked dataset*
All records	153,455	142,697	123,910	93.0	80.7
Sex					
Male	78,522	71,994	62,958	91.7	80.2
Female	74,933	70,703	60,952	94.4	81.3
Age					
0–9 years	464	384	328	82.8	70.7
10–19 years	687	568	508	82.7	73.9
20–29 years	1,591	1,163	1,049	73.1	65.9
30–39 years	2,511	1,889	1,636	75.2	65.2
40–49 years	4,971	4,085	3,665	82.2	73.7
50–59 years	10,190	8,944	7,972	87.8	78.2
60–69 years	19,103	17,516	15,792	91.7	82.7
70–79 years	31,719	29,831	26,558	94.0	83.7
80–89 years	55,370	52,698	45,002	95.2	81.3
90+ years	26,849	25,619	21,400	95.4	79.7
State					
New South Wales	53,358	49,422	43,193	92.6	80.9
Victoria	37,794	35,399	30,650	93.7	81.1
Queensland	28,065	25,954	22,360	92.5	79.7
South Australia	13,567	12,874	11,265	94.9	83.0
Western Australia	13,683	12,635	11,006	92.3	80.4
Tasmania	4,285	4,039	3,472	94.3	81.0
Northern Territory	869	658	490	75.7	56.4
Aust. Capital Territory	1,829	1,711	1,471	93.5	80.4
Remoteness					
Major cities	101,349	94,274	81,720	93.0	80.6
Inner regional	34,362	32,182	28,173	93.7	82.0
Outer regional	15,158	14,017	12,224	92.5	80.6
Remote	1,713	1,500	1,243	87.6	72.6
Very Remote	825	705	518	85.5	62.8
No usual address	48	19	32	39.6	66.7
Marital status					
Never married	15,081	12,828	10,837	85.1	71.9
Married	61,940	58,566	52,416	94.6	84.6
Widowed, divorced, separated	72,433	67,935	57,768	93.8	79.8
Not stated	4,001	3,368	2,889	84.2	72.2
Indigenous status					
Non-Indigenous	150,238	140,037	121,827	93.2	81.1
Aboriginal or Torres Strait Islander	2,345	1,884	1,435	80.3	61.2
Not stated	872	776	648	89.0	74.3

A.1 Link rates, selected subpopulations — continued

	Death Registrations (No.)	*Record counts (No.)*		*Link rates (%)*	
		Gold standard	*Linked dataset*	*Gold standard*	*Linked dataset*
All records	153,455	142,697	123,910	93.0	80.7
Country of birth					
Australia	105,827	99,019	84,863	93.6%	80.2%
New Zealand & Pacific	3,039	2,582	2,229	85.0%	73.3%
United Kingdom & Ireland	16,028	15,025	13,473	93.7%	84.1%
Europe	19,022	17,662	15,935	92.9%	83.8%
Africa & Middle East	2,922	2,604	2,289	89.1%	78.3%
Asia	5,296	4,668	4,133	88.1%	78.0%
Americas	936	826	735	88.2%	78.5%
Not stated	385	311	253	80.8%	65.7%
Cause of death					
Heart disease	21,541	20,077	17,485	93.2	81.2
Stroke	11,408	10,746	9,123	94.2	80.0
Dementia/Alzheimer's	10,422	9,930	8,077	95.3	77.5
Lung cancer	8,793	8,177	7,292	93.0	82.9
Diabetes	4,412	4,113	3,495	93.2	79.2
Colon cancer	4,405	4,155	3,713	94.3	84.3
Prostate cancer	3,396	3,211	2,787	94.6	82.1
Breast cancer	3,082	2,889	2,565	93.7	83.2
Intentional self-harm	2,521	2,003	1,766	79.5	70.1
Other causes	83,475	77,396	67,607	92.7	81.0

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.

LIBRARY A range of ABS publications are available from public and tertiary libraries Australia wide. Contact your nearest library to determine whether it has the ABS statistics you require, or visit our website for a list of libraries.

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, or purchase a hard copy publication. 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

1351.0.55.058 • RESEARCH PAPER: DEATH REGISTRATIONS TO CENSUS LINKAGE PROJECT – A LINKED DATASET FOR ANALYSIS

Produced by the Australian Bureau of Statistics