FACULTY OF
BUSINESS &
ECONOMICS

HILDA PROJECT TECHNICAL PAPER SERIES
No. 2/14, November 2014

Outcomes from matching the HILDA Survey sample to the death register

Nicole Watson and Michelle Summerfield

The HILDA Project was initiated, and is funded, by the Australian Government Department of Social Services

Contents

INTRODUCTION **1**

DEATH REGISTER MATCHING PROCESS **1**

ETHICS APPROVAL 1
TIMEFRAME 1
SAMPLE MATCHED 1
MATCH WEIGHTS 2
REVIEW OF MATCHES 3
IMPACT ON THE CONDUCT OF WAVE 13 FIELDWORK 3

MATCH RESULTS **3**

NUMBER OF DEATHS 3
DATE OF DEATH 7
CAUSE OF DEATH 8
REVISION TO RESPONSE RATES 10
IMPACT ON WEIGHTS 13

INFORMATION PROVIDED IN RELEASE 13 ..... **13**

GENERAL RELEASE 13
UNCONFIDENTIALISED RELEASE 13

FUTURE PLANS **13**

REFERENCES **14**

APPENDIX 1: TIMING OF DEATHS BETWEEN WAVES **15**

Acknowledgements: The authors are grateful for the assistance of Michael Bradford and the HILDA Sample Team at Roy Morgan Research in resolving various sample queries associated with this work.

Introduction

Beginning in 2001, the Household, Income and Labour Dynamics in Australia (HILDA) Survey has become a major resource in the social sciences for understanding the dynamics of the economic, social and subjective well-being of Australians. The health-related questions facilitate the examination of changes in self-reported health status, long-term disabilities, impairments and health conditions, behavioural risk factors (such as smoking, exercise, and diet), and use of health care services. In 2014, the HILDA Survey sample was matched to the National Death Index (NDI) so that details of the date and cause of death could be added to the data files. This offers researchers a valuable opportunity to study the changes in life circumstances leading up to death. It also helps us to more accurately identify the eligibility of sample members over the course of the panel, which will assist in the production of response rates and population weights.

This report details the matching process, the outcomes, and the variables available.

Death register matching process

Ethics approval

Prior to the HILDA Survey sample being matched to the NDI, the process was approved by both the University of Melbourne's Human Research Ethics Committee and the Australian Institute of Health and Welfare's Ethics Committee.

Timeframe

The HILDA Survey sample was matched to the National Death Index in June 2013 and again January 2014. Roy Morgan Research (RMR), the fieldwork provider for the HILDA Survey, provided AIHW with sample information (first name, second name, third name, surname, date of birth, sex, and state of last contact).[1] A file of matched pairs was returned to RMR and a de-identified match file was provided to the Melbourne Institute.

Sample matched

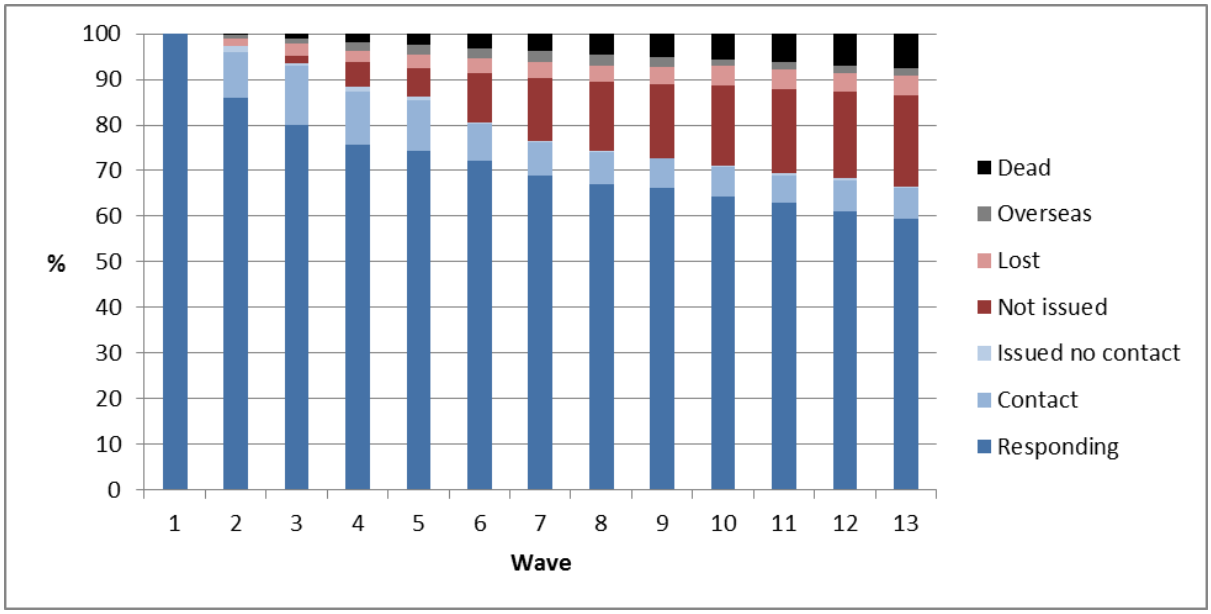
The entire HILDA Survey sample was matched to the NDI. The sample comprised all members of responding households in any wave up to and including wave 12. Matching the entire sample to the death register (rather than just matching known deaths and non-contacts) helps us understand the quality of name and date of birth information held in the respondent database and the quality of the matches obtained.

We expect that most of the newly identified deaths would come from the part of the sample that we no longer have contact with (and hence are no longer issued to field), have become untraceable following a move, or were not able to be contacted in the latest wave. As shown in Figure 1, 25 per cent of the wave 1 respondents were not contacted in wave 12: 4 per cent were untraceable (or 'lost'), 19 per cent were not issued to the interviewers for fieldwork, and 0.4 per cent were issued in wave 12 but were not able to be contacted.[2] This has grown from only 3 per cent in wave 2 and will continue to grow over the life of the panel.

[1] While provided, state of last contact was not used in the matching process. Second and third names are rare in the sample information.

[2] When we consider the larger group of wave 1 enumerated persons (i.e., people belonging to households where at least one person responded), there were 26% that were not contacted in wave 12.

Figure 1: Outcome of wave 1 respondents (n=13,969)



Match weights

The Australian Institute of Health and Welfare matched the sample details (name, date of birth, and sex) to the NDI. They provided a match weight for each potential matched pair of records that indicates how close the sample details were to the death record details (AIHW, undated). For the HILDA Survey sample, this match weight ranged from 0.8 to 68.0, with the higher the weight, the better the quality of the match. The main contribution to the weight is made by the name, in terms of both how unusual the name is and how exact the match is. The match of day, month and year of birth contribute next to the weight, with the smallest contribution coming from the match on sex.[3] Where a sample member matched to more than one death record, the match pair with the highest weight was chosen.

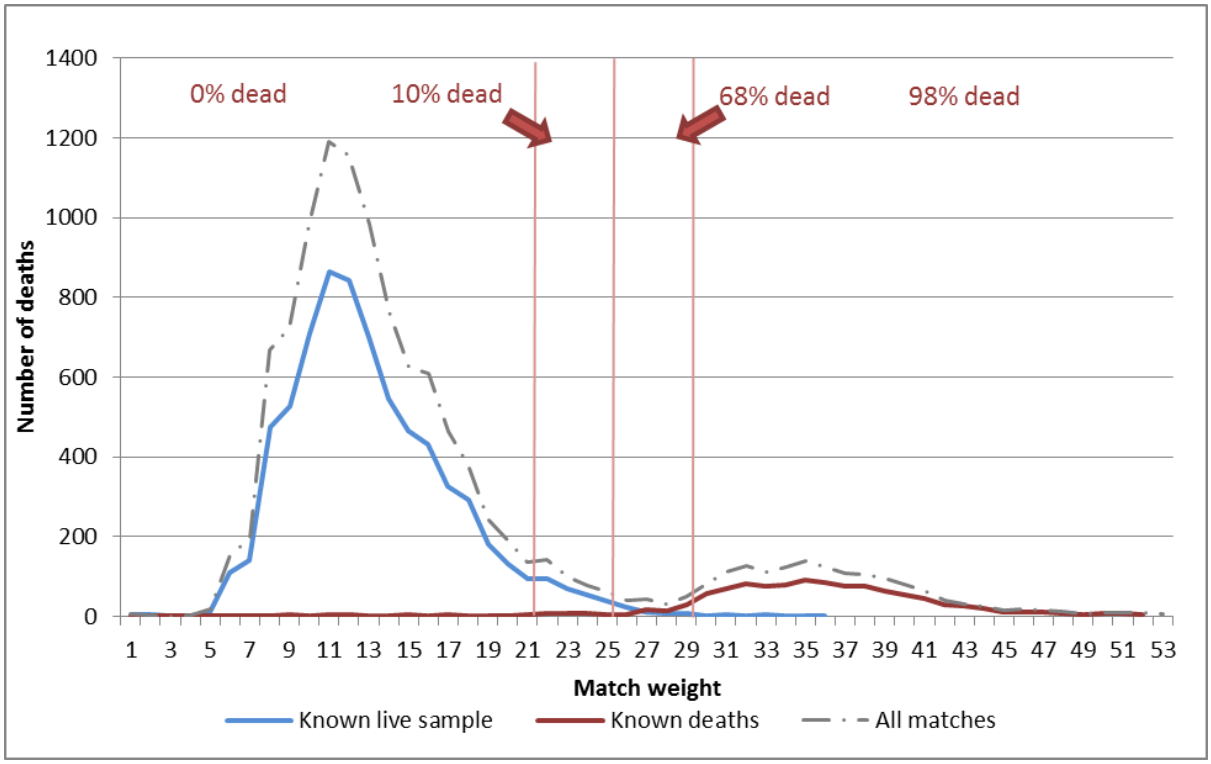
To understand the implications of the weights, the match weights of the known live sample (those who belonged to responding households in wave 12) and the known dead sample (identified by HILDA interviewers) were plotted against the match weights of all primary match pairs in Figure 2. We found the following:

- 98 per cent of all (known live and known dead) cases with a match weight of 29.5 or higher were dead.
- 68 per cent of all cases with a match weight between 25.5 and 29.5 were dead.
- 10 per cent of all cases with a match weight between 21.5 and 25.5 were dead.
- 0.5 per cent of all cases with a match weight less than 21.5 were dead.

As a result, all matches to known deaths were clerically reviewed by RMR along with matches to sample members that have a weight above 21.5.[4] The HILDA team at the Melbourne Institute also reviewed these cases, though to a more limited extent given the data restrictions.

[3] Disagreement on each name component drops the weight by 10 points whereas disagreement on each year of birth drops the weight by 1 point.
[4] There were 9 known deaths with a match weight less than 21.5 that were considered a match.

Figure 2: National Death Index match weights



Review of matches

The matches were reviewed in terms of the closeness of the name, date of birth and sex matches, along with consistency of the date of death with fieldwork outcomes at that time. For the known deaths, the NDI date of death was compared with interviewer collected date of death for consistency. For the rest of the matches, the NDI date of death was compared against the date the sample member was last recorded as part of a responding household (where the household roster had been completed to determine who was or was not a member of the household). RMR also reviewed fieldwork notes at the time of the death for cases where the match weight was low or the date of death was inconsistent to help establish the plausibility of the match.

Impact on the conduct of wave 13 fieldwork

The initial match of the sample to the NDI occurred in June 2013, 4 weeks prior to the start of the wave 13 fieldwork. While the NDI matching did not affect which cases were ultimately issued for fieldwork, RMR added a note to the Household Form for issued cases with a match weight of 26 or higher to indicate that the sample member 'may have passed away'. The interviewers sought to confirm this outcome.

Match results

Number of deaths

A total of 1238 deaths have already been identified via the HILDA Survey fieldwork. The matching to the death register resulted in a further 304 deaths being identified, or a 25 per cent increase in the number of deaths recorded in the data files. Table 1 shows the distribution of these deaths across the various types of sample members. Disproportionately

more deaths were identified among wave 1 non-responding adults (who belonged to a wave 1 responding household) than among wave 1 respondents. This is because we were more likely to lose touch with these non-responding adults over the course of the study as they move out of responding households without further engaging with the study themselves. Deaths of new entrants not followed after they leave the household of a sample member (i.e., 'inactive' sample members) were not incorporated into the data files.

Table 1: Number of deaths

	Number in sample	Deaths identified in field	Additional deaths from death register	Total number of deaths
Wave 1 sample members				
Respondents	13,969	1038	265	1303
Non-responding adults[1]	1,158	56	28	84
Children	4,787	12	2	14
Wave 11 top-up sample members				
Respondents	4,009	36	2	38
Non-responding adults[1]	273	5	0	5
Children	1,179	0	0	0
New entrants				
Added in wave 13[2]	1,290	-	-	-
Added earlier and active in wave 13	6,384	91	7	98
Added earlier and inactive in wave 13	4,378	-	-	-
Total	37,427	1,238	304	1,542

Notes: 1. The non-responding adults belong to households where at least one person responded in wave 1 (or wave 11 for the top-up sample).

2. Any new entrants added in wave 13 were not matched to the NDI.

There were a further 29 deaths identified from the death register matching that have been held over to wave 14, as they occurred after the structure of the household had been set in wave 13. Of these 29 people, 19 were interviewed in wave 13 and the remaining 10 were unable to be interviewed due to ill health. These deaths will be incorporated into Release 14.

A greater number of deaths were missing from the sample in later waves due to the increasing portion of the sample that is not issued to field over time. Figure 3 shows how the gap between the cumulative number of known deaths compares to the combination of known and matched deaths at each wave.[5] The gap is about 10 per cent of known deaths in wave 5 and 25 per cent in wave 13.

[5] Deaths are allocated to waves according to the fit between household structure dates. A household structure date is the date the household composition is set with a sample member, or if the household is not responding then it is the date the final outcome is coded for the wave. If a household is not issued to field, the deaths are assigned to waves according to the calendar year for the wave. See Appendix 1 for more detail on how the deaths identified by interviewers are spread across the calendar months.

Figure 4 shows how the deaths are distributed by age (calculated at 30th June 2001) and sex. The rate of missing deaths differs with both of these characteristics. Male deaths are more likely to be known than female deaths, especially among those aged 85 and older. The deaths of 20 per cent of the males aged 85 and older were missed, whereas 28 per cent of the female deaths in the same age group were missed. The rate of missing deaths in the younger male cohorts is 15 per cent compared to 17 per cent in the younger female cohorts.

Figure 3: Cumulative number of deaths by wave

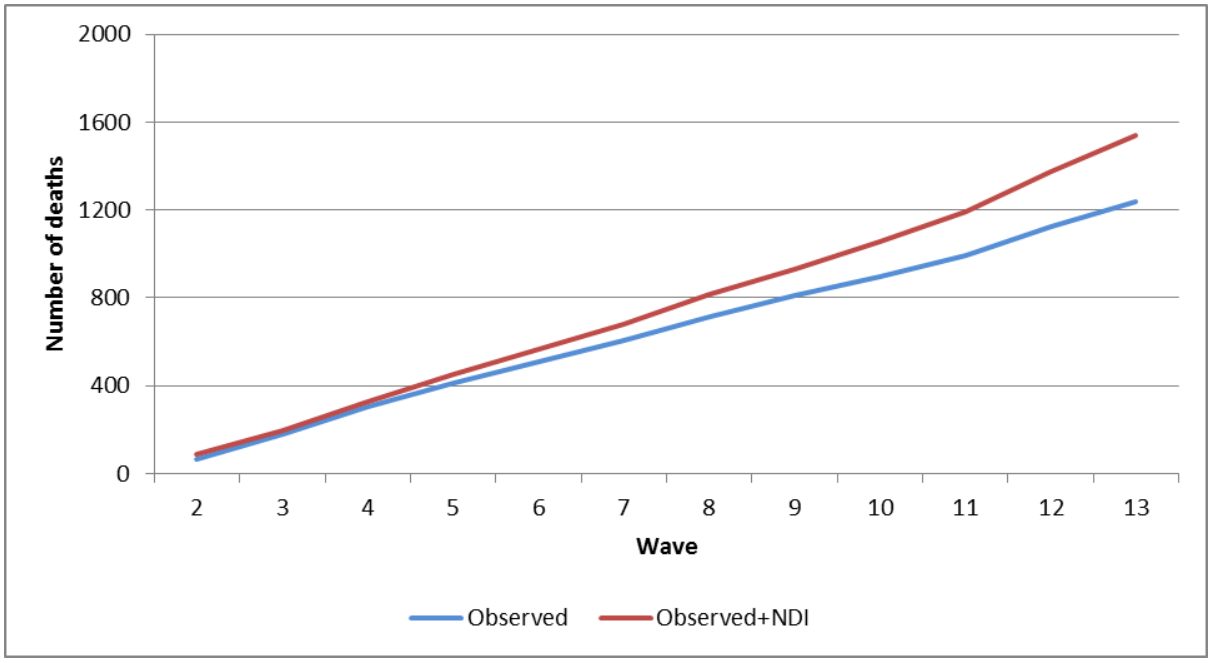


Figure 4: Number of deaths by age at 30 June 2001

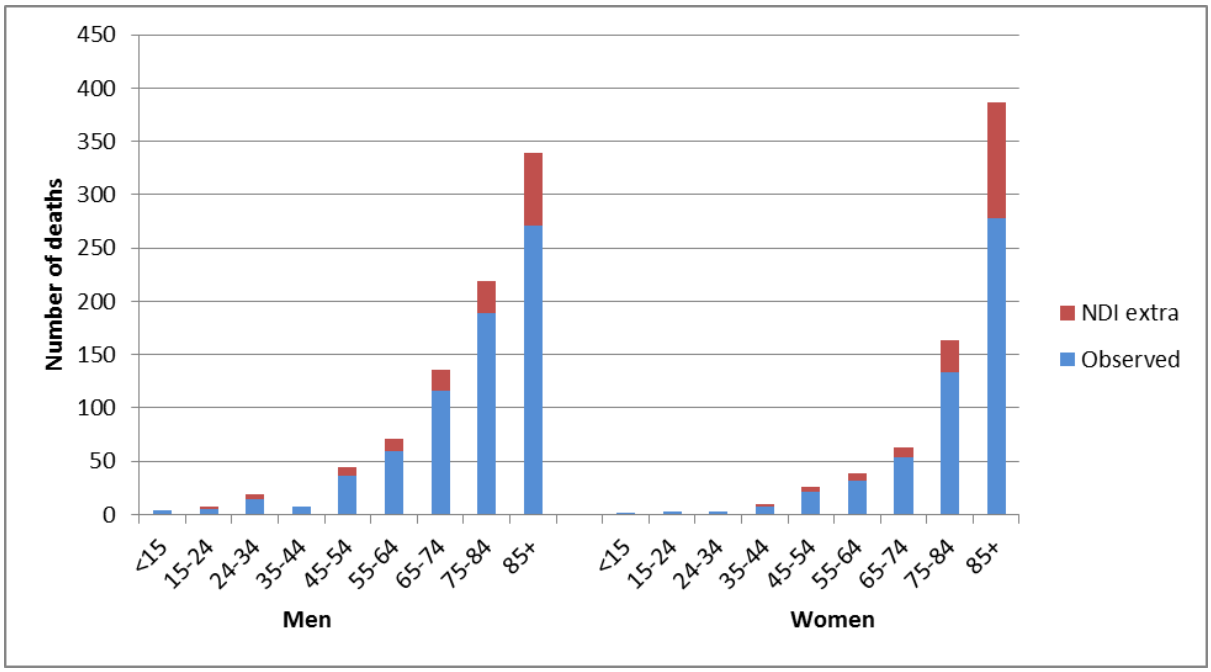


Figure 5 shows the number of deaths each calendar year for the known deaths (in light blue) and the known and matched deaths (in dark red). When counting just the known deaths, we often need to infer the year of death as it is unknown. Often we only know that the death occurred between two waves and we assign it to the calendar year relating to the wave. However, when the death is matched to the death register, the actual year of death is used to produce the counts shown by the red (dark) line. This produces the seemingly incorrect result of the number of known deaths being higher than the known and matched deaths in 2003: some of the interviewer recorded deaths in 2003 with an unknown year of death (so inferred to be in 2003) actually occurred in 2002.

There are five points to note regarding Figure 5. First, the number of deaths in the first couple of years is low, which reflects the sample being selected from private dwellings and so excluding people living in institutions (such as nursing homes). Second, from 2004, the number of deaths identified from the fieldwork generally declines as a greater portion of the sample is not issued to field, thus precluding us from knowing about any deaths in that part of the sample. Third, the increase in deaths in 2012 is simply a function of the expansion of the sample in that year. Fourth, the apparent decline in the number of deaths in 2013 is due to the fact that any deaths that occur after the date the household structure date was set are held over to the next data release (25 of the 29 deaths held over to wave 14 occurred in 2013). Lastly, from about 2004, the average number of deaths per year (when including both known and matched deaths) is relatively steady at 120 deaths per year (until 2012 when deaths in the top-up sample are included).

Figure 6 shows the source of the deaths by year for the known and matched deaths. It includes the deaths that we know about from the death register matches that have occurred after the wave 13 fieldwork, which brings the count for 2013 closer to the 2012 figure.

Figure 5: Number of deaths by year

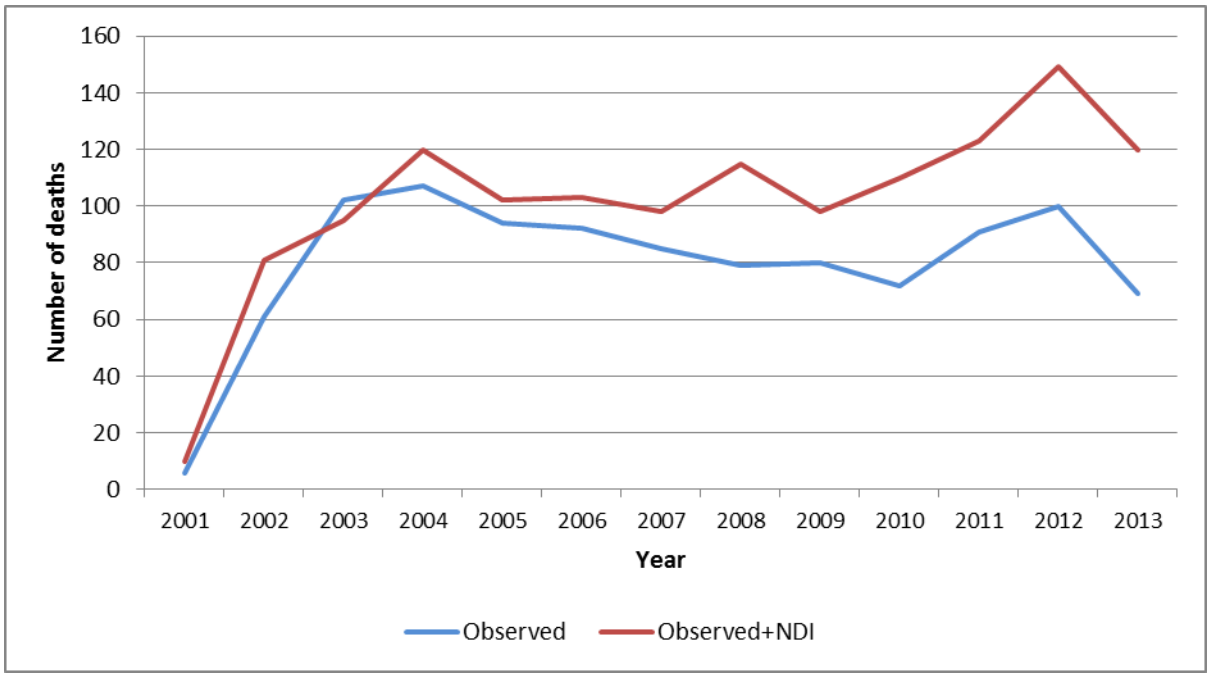
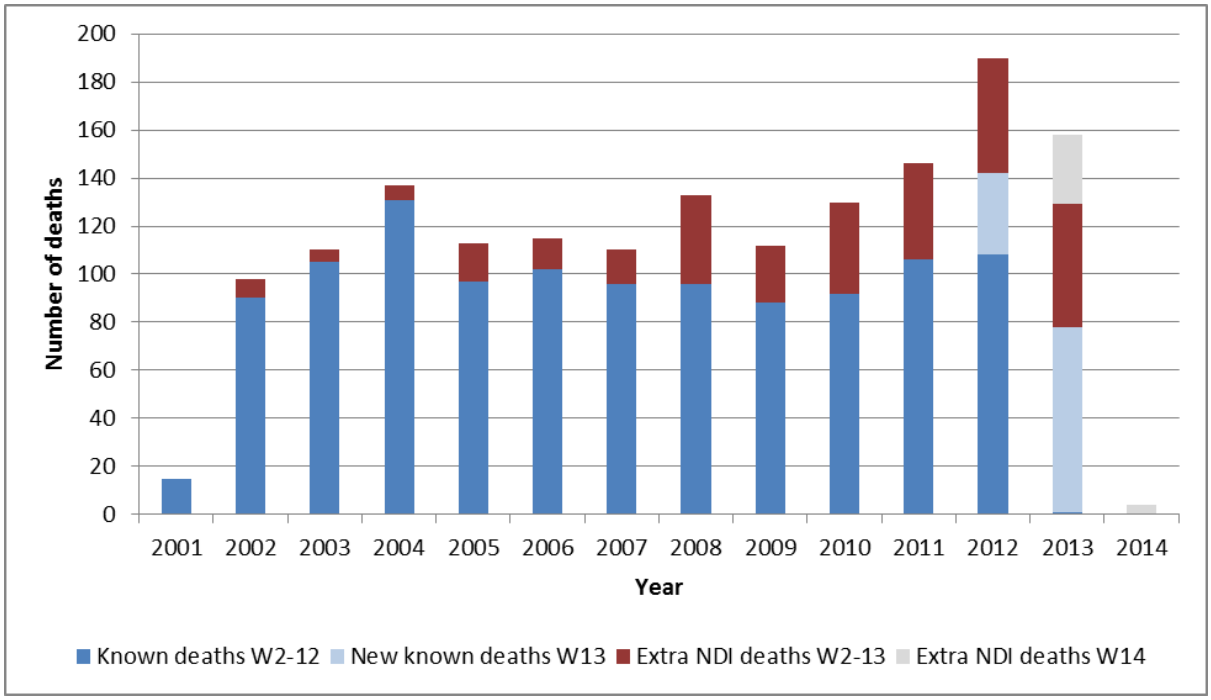


Figure 6: Source of deaths by year



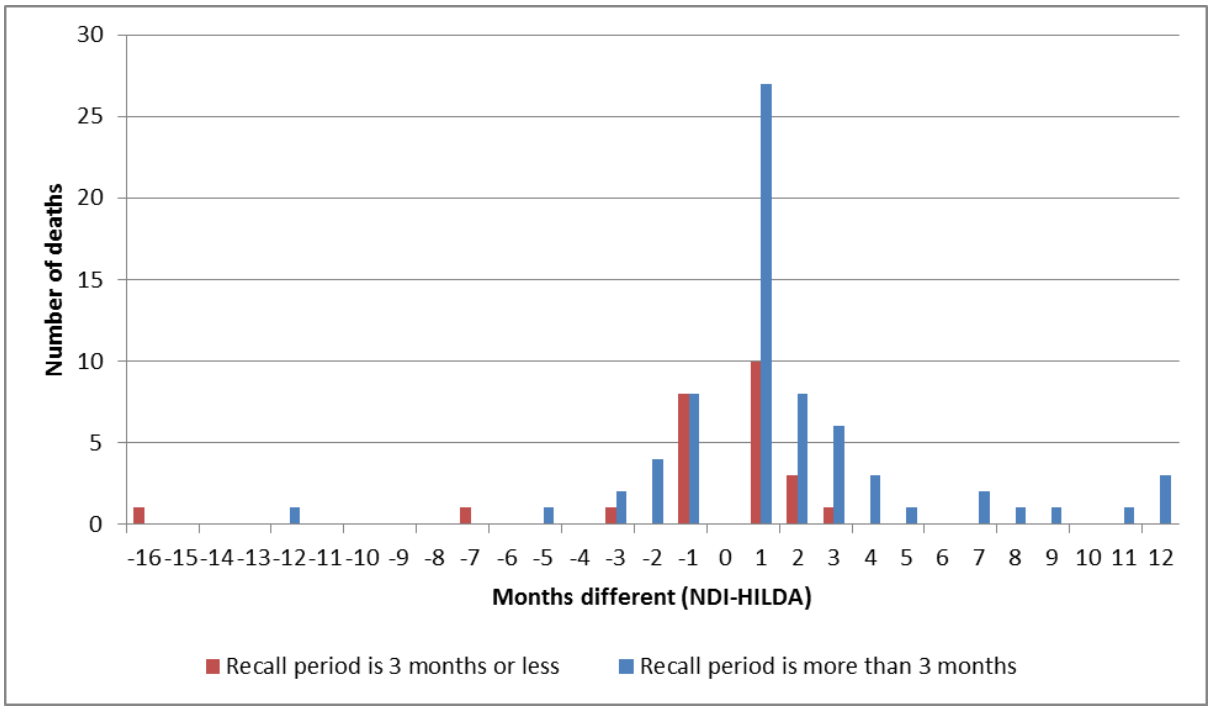
Date of death

The date of death of sample members reported to interviewers is frequently incomplete. Only 62 per cent of cases have both month and year of death. A further 3 per cent have year only. For the remaining 35 per cent of cases, we are left to infer that the death occurred between two waves. That is, we did not know about the death in the last wave, but we do in the current wave, so the death is assumed to have occurred between those two waves.[6]

Obtaining date of death information from the death register improves the quality of this information. For the cases where both month and year of death was provided to the interviewer, 87 per cent match exactly.[7] Figure 7 shows how the months differ for the cases that do not match exactly. Overall, two thirds of those that did not match have a date of death according to the death register that is earlier than the date of death reported to the interviewer. One possible explanation for this is forward telescoping bias, where the respondent is more likely to report the event occurring more recently than it actually happened. This is more likely to occur for events that are more distant in the past than for very recent events, which are more likely to suffer the opposite problem (Tourangeau et al. 2000, pp. 120-121). Our results are consistent with this explanation as there is little evidence of reporting bias in the deaths reported to interviewers within 3 months of the date of death recorded on the death register and much more bias in the reported dates of deaths that were more distant in the past.

[6] This assumption may not be accurate if the sample member could not be contacted in the previous wave.
[7] For the cases where we only have year or where we assume the death occurred between two waves, 90 per cent match exactly to the death register information. In 8 per cent of cases the death register information places the death earlier by 1-2 years and in 2 per cent of cases it is later.

Figure 7: Difference in month of death reported to interviewer and on death register (where reported both month and year and a difference exists)



Cause of death

The death register provides cause of death information classified to the International Classification of Diseases, Version 10 (ICD-10).[8] Table 2 shows the proportion of deaths accumulated in the HILDA Survey sample in the top 10 leading cause of death categories in Australia, based on all deaths registered in 2007 (ABS 2014, Table 2.1).[9] The (unweighted) distribution of HILDA deaths align reasonably well to the ABS figures even though the HILDA deaths are observed over a 13-year period and the ABS figures are for deaths in a single year.

Unfortunately the cause of death information is not very complete. Thirty one per cent of deaths are missing cause of death information. The reasons are as follows:

- 4 per cent of deaths could not be matched to the death register;
- 6 per cent of deaths have the cause of death withheld as the death had been before a coroner (such information is withheld for deaths that occur in 2006 or later); and
- 21 per cent of deaths do not have any cause of death information at the present time (almost all of these cases are in 2012 to 2014 and are a result of processing delays and should become available in a couple of years).

[8] More details on ICD-10 are provided on the World Health Organisation website (http://www.who.int/classifications/icd/en/) and the ICD-10 coding instruction manual (WHO, 2010).
[9] The method the ABS uses to define the leading causes of death follow the World Health Organisation recommendations provided by Becker et al (2006). Also, the year 2007 is used as the approximate midpoint of the range over which we observe the deaths in the HILDA sample.

Table 2: Leading cause of death

Leading cause of death (ICD-10 codes)	All registered deaths in 2007 (ABS)		HILDA sample deaths 2001-2013		
	Ranking	%	Number of cases	% (of non-missing)	% (of all deaths)
Ischaemic heart diseases (I20-I25)	1	16.7	182	17.0	
Cerebrovascular diseases (I60-I69)	2	8.3	80	7.5	
Trachea, bronchus and lung cancer (C33-C34)	3	5.5	59	5.5	
Dementia and Alzheimer disease (F01, F03, G30)	4	5.3	50	4.7	
Chronic lower respiratory diseases (J40-J47)	5	4.2	53	5.0	
Colon, sigmoid, rectum and anus cancer (C18-C21)	6	3.0	33	3.1	
Diabetes (E10-E14)	7	2.8	25	2.3	
Blood and lymph cancer (including leukaemia) (C81-C96)	8	2.6	43	4.0	
Heart failure (I50-I51)	9	2.5	21	2.0	
Diseases of the urinary system (N00-N39)	10	2.3	23	2.2	
Other causes of death		46.7	500	46.8	
Total with cause of death information		100.0	1069	100.0	69.3
Cause withheld			87		5.6
No cause provided			322		20.9
Not matched to death register			64		4.2
Total missing cause of death information			473		30.7
Total deaths			1542		100.0

Another way researchers typically analyse cause of death information is by ICD-10 Chapters. Table 3 shows the proportion of accumulated deaths in the sample by selected ICD-10 Chapters and compares this to all deaths registered in 2007 (ABS 2014, Table 1.2). Again, the (unweighted) distribution of deaths in the HILDA Survey sample align reasonably well to the ABS figures. There does, however, seem to be a greater portion of deaths from Neoplasms in the sample compared with the rate observed in the Australian population.

Table 3: Underlying cause of death (by selected ICD-10 Chapter level)

Underlying cause of death (ICD-10 codes)	All registered deaths in 2007 (ABS)	HILDA sample deaths 2001-2013		
	%	Number of cases	% (of non-missing)	% (of all deaths)
Neoplasms (C00-D48)	29.3	392	36.7	
Endocrine, nutritional and metabolic diseases (E00-E90)	3.9	34	3.2	
Mental and behavioural disorders (F00-F99)	4.1	37	3.5	
Diseases of the nervous system (G00-G99)	4.0	49	4.6	
Diseases of the circulatory system (I00-I99)	34.1	338	31.6	
Diseases of the respiratory system (J00-J99)	8.4	89	8.3	
Diseases of the digestive system (K00-K93)	3.5	30	2.8	
Diseases of the genitourinary system (N00-N99)	2.4	24	2.2	
External causes of morbidity and mortality (V01-Y98)	6.0	26	2.4	
Other causes of death	4.4	50	4.7	
Total with cause of death information	100.0	1069	100.0	69.3
Total missing cause of death information		473		30.7
Total deaths		1542		100.0

Revision to response rates

Incorporating additional deaths into the sample outcomes affects response rates. Rather than treating these individuals as in-scope non-respondents, they are now counted as out-of-scope and excluded from the calculation of response rates. Figure 8 shows how the outcomes of wave 1 respondents have changed as a result of the death register matching (thus providing an update to Figure 1 before the matching). The proportion of wave 1 respondents excluded from the denominator for the wave 13 re-interview rate increases from 9.2 per cent (being 1.7 per cent overseas and 7.4 per cent dead) to 11.1 per cent (i.e., 1.7 per cent overseas and 9.3 per cent dead). As a result, the re-interview rates for wave 1 respondents differ by up to 1.4 percentage points – the wave 13 rate increases from 66.0 per cent before the death register matching to 67.4 per cent after the death register matching (see Table 4).

Far less is the effect on the wave-on-wave response rates. Table 5 shows that the response rate series affected the most is that of previous wave non-respondents, and even for this group the response rates differ by 0.2 percentage points or less.

Figure 8: Outcome of wave 1 respondents (n=13,969), after death register matching

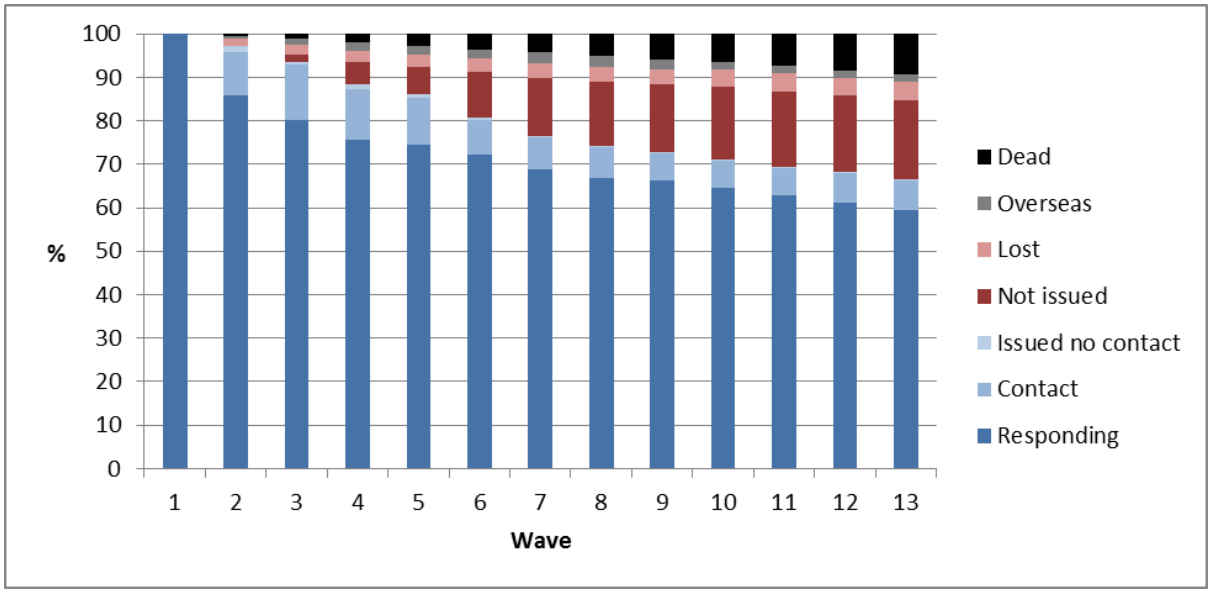


Table 4: Re-interview rates of initial wave respondents (excluding out-of-scope)

Wave	Wave 1 respondents		Wave 11 top-up respondents	
	Before death register matching	*After* death register matching	*Before* death register matching	*After* death register matching
2	86.8	86.9	-	-
3	82.0	82.1	-	-
4	78.5	78.6	-	-
5	77.9	78.1	-	-
6	76.3	76.6	-	-
7	73.6	74.0	-	-
8	72.0	72.5	-	-
9	71.5	72.1	-	-
10	70.0	70.8	-	-
11	68.8	69.8	-	-
12	67.5	68.7	92.3	92.3
13	66.0	67.4	87.4	87.4

Table 5: Wave-on-wave individual level response rates, before and after the death register matching

	Main sample												Top-up sample	
	W2	W3	W4	W5	W6	W7	W8	W9	W10	W11	W12	W13	W12	W13
***Before* death register matching**														
All people														
Previous wave respondent	86.8	90.4	91.6	94.4	94.9	94.7	95.2	96.3	96.3	96.5	96.2	96.4	92.3	92.7
Previous wave non-respondent	19.7	17.6	12.7	14.7	8.4	5.6	5.7	8.5	4.5	3.8	3.5	3.8	16.5	13.9
Previous wave child	80.4	71.5	70.7	76.0	76.0	70.8	73.5	73.7	72.1	70.0	65.9	65.0	87.8	86.1
New entrant this wave	73.3	76.1	70.4	81.8	81.4	79.5	79.5	81.3	82.8	80.7	78.8	74.8	80.5	80.4
People attached to responding household in previous wave														
Previous wave respondent	86.8	90.4	91.6	94.4	94.9	94.7	95.2	96.3	96.3	96.5	96.2	96.4	92.3	92.7
Previous wave non-respondent	19.7	19.8	17.7	25.3	18.3	13.3	15.0	25.9	16.2	15.4	11.0	11.3	16.5	11.0
Previous wave child	80.4	82.0	81.2	88.6	89.8	90.5	90.9	93.0	92.2	93.0	89.0	89.5	87.8	92.9
New entrant this wave	73.3	78.4	71.6	85.2	81.5	80.0	81.0	81.3	83.5	82.0	79.2	77.1	80.5	80.4
***After* death register matching**														
All people														
Previous wave respondent	86.9	90.4	91.6	94.4	94.9	94.7	95.2	96.3	96.3	96.5	96.2	96.4	92.3	92.7
Previous wave non-respondent	19.8	17.7	12.8	14.9	8.5	5.7	5.8	8.7	4.7	3.9	3.6	3.9	16.5	13.9
Previous wave child	80.4	71.5	70.7	76.0	76.0	70.8	73.5	73.7	72.1	70.0	65.9	65.0	87.8	86.1
New entrant this wave	73.3	76.1	70.4	81.8	81.4	79.5	79.5	81.3	82.8	80.7	78.8	74.8	80.5	80.4
People attached to responding household in previous wave														
Previous wave respondent	86.9	90.4	91.6	94.4	94.9	94.7	95.2	96.3	96.3	96.5	96.2	96.4	92.3	92.7
Previous wave non-respondent	19.8	19.8	17.8	25.4	18.3	13.3	15.0	26.0	16.2	15.5	11.0	11.3	16.5	11.0
Previous wave child	80.4	82.1	81.2	88.6	89.8	90.5	90.9	93.0	92.2	93.0	89.0	89.5	87.8	92.9
New entrant this wave	73.3	78.4	71.6	85.2	81.5	80.0	81.0	81.3	83.5	82.0	79.2	77.1	80.5	80.4

Impact on weights

The inclusion of these additional deaths will have a small impact on how the weights are constructed from Release 13. For Release 12, unknown deaths in the non-responding part of the sample were allowed for in the attrition adjustment (Watson 2012, p. 14). With the death register matching, approximately 96 per cent of these deaths have now been identified. Just as there were 4 per cent of known deaths that could not be matched to the death register for various reasons, there will be a similar group of individuals among the unknown deaths. In the attrition adjustments in the weights, the known and matched deaths can now be earmarked to receive no attrition adjustment in waves 2 to 13. From wave 14, a new method of flagging potential deaths among the non-contacted part of the sample will need to be used (possibly using a life-table method which predicts the probability of death by age and sex).

Information provided in Release 13

General Release

On the General Release datasets, the following variables will be included:

- Year of death (on the master file);
- Age at death (on master file);
- Source of death (on the master file); and
- Revised individual and household outcome codes (on the master file, household file, enumerated person file, responding person file and combined file).

The source of death flag will indicate whether the death was identified only in field, identified only from the death register matching, or identified from both sources.

Unconfidentialised Release

On the Unconfidentialised Release datasets, the following variables will be included:

- Date of death (on the master file);
- Age at death (on master file);
- Source of death (on the master file);
- Leading cause of death and major ICD-10 categories (on the master file); and
- Revised individual and household outcome codes (on the master file, household file, enumerated person file, responding person file and combined file).

Future plans

In approximately two years, we will have the death register matching re-run to update the cause of death information for deaths occurring between 2012 and 2014. We hope to undertake another full matching of the sample to the death register in 5-10 years.

References

Australian Bureau of Statistics (2014), *Cause of Death, Australia, 2012* (ABS cat. no. 3303.0), ABS, Canberra.

Australian Institute of Health and Welfare (undated), 'Linking research data files to the National Death Index (NDI)', Canberra.
[http://www.aihw.gov.au/WorkArea/DownloadAsset.aspx?id=60129544301, accessed 29 April 2014]

Becker, R., Silvi, J., Ma Fat, D., L'Hours, A., and Laurenti, R. (2006), 'A method for deriving leading causes of death', *Bulletin of the World Health Organisation*, vol. 84, pp. 297-304.

Tourangeau, R., Rips, L.J., and Rasinski, K. (2000), *The Psychology of Survey Response*, Cambridge University Press, Cambridge.

Watson, N. (2012), 'Longitudinal and Cross-sectional Weighting Methodology for the HILDA Survey', HILDA Project Technical Paper Series, No. 2/12, Melbourne Institute of Applied Economic and Social Research, University of Melbourne.

World Health Organisation (2010), *International Statistical Classification of Disease and Related Health Problems: Volume 2 Instruction Manual* (10th revision), World Health Organisation, Malta.
[http://www.who.int/classifications/icd/ICD10Volume2_en_2010.pdf?ua=1, accessed 18 August 2014]

Appendix 1: Timing of deaths between waves

The fieldwork for a wave spans up to 8 months (from July to February). This can result in a gap between one wave and the next for any particular sample member of between 5 and 20 months. Interviewer reported deaths in a particular wave can therefore also span such a period.

Figure A1 shows the distribution of the month of death for the deaths reported each wave. Here we are using the month of death from the death register (as the reported month of death is subject to recall error, with the more recent months being recalled more easily and more distant months being less likely to be recalled). In the x-axis 'y-1' and 'y+1' are used to refer to months in the previous and subsequent calendar years respectively. Deaths in October to December are more likely to be reported in the following year, consistent with the fact that approximately half of the interviews each wave are completed by the end of September.

Figure A1: Timing of known deaths