Linking death registration and survey data: Procedures and cohort profile for The Irish Longitudinal Study on Ageing [version 1; peer review: 2 approved, 1 approved with reservations]

Mark Ward¹, Peter May², Robert Briggs¹,³, Triona McNicholas¹,³, Charles Normand², Rose Anne Kenny¹,³, Anne Nolan¹,⁴

¹The Irish Longitudinal Study on Ageing, Trinity College Dublin, Dublin, Ireland
²Centre for Health Policy and Management, Trinity College Dublin, Dublin, Ireland
³Department of Medical Gerontology, St James's Hospital, Dublin, Ireland
⁴The Economic and Social Research Institute, Dublin, Ireland

Abstract

Background: Research on mortality at the population level has been severely restricted by an absence of linked death registration and survey data in Ireland. We describe the steps taken to link death registration information with survey data from a nationally representative prospective study of community-dwelling older adults. We also provide a profile of decedents among this cohort and compare mortality rates to population-level mortality data. Finally, we compare the utility of analysing underlying versus contributory causes of death.

Methods: Death records were obtained for 779 (90.3% of all confirmed deaths at that time) and linked to individual level survey data from The Irish Longitudinal Study on Ageing (TILDA).

Results: Overall, 9.1% of participants died during the nine-year follow-up period and the average age at death was 75.3 years. Neoplasms were identified as the underlying cause of death for 37.0%; 32.9% of deaths were attributable to diseases of the circulatory system; 14.4% due to diseases of the respiratory system; while the remaining 15.8% of deaths occurred due to all other causes. Mortality rates among younger TILDA participants closely aligned with those observed in the population but TILDA mortality rates were slightly lower in the older age groups. Contributory cause of death provides similar estimates as underlying cause when we examined the association between smoking and all-cause and cause-specific mortality.

Conclusions: This new data infrastructure provides many opportunities to contribute to our understanding of the social, behavioural, economic, and health antecedents to mortality and to
inform public policies aimed at addressing inequalities in mortality and end-of-life care.

**Keywords**
mortality, ageing, death certification, TILDA, data linkage

This article is included in the TILDA gateway.

This article is included in the Ageing Populations collection.

**Corresponding author:** Mark Ward (wardm8@tcd.ie)

**Author roles:**
- **Ward M:** Conceptualization, Data Curation, Formal Analysis, Methodology, Project Administration, Writing – Original Draft Preparation, Writing – Review & Editing
- **May P:** Conceptualization, Methodology, Writing – Review & Editing
- **Briggs R:** Formal Analysis, Validation, Writing – Review & Editing
- **McNicholas T:** Formal Analysis, Validation, Writing – Review & Editing
- **Normand C:** Conceptualization, Funding Acquisition, Supervision, Writing – Review & Editing
- **Kenny RA:** Conceptualization, Data Curation, Funding Acquisition, Resources, Supervision, Writing – Review & Editing
- **Nolan A:** Conceptualization, Funding Acquisition, Methodology, Supervision, Writing – Original Draft Preparation, Writing – Review & Editing

**Competing interests:** No competing interests were disclosed.

**Grant information:** Health Research Board, Ireland [ILP-PHR-2017-022], Investigator-led projects scheme. TILDA is co-funded by the Government of Ireland through the Department of Health, by Atlantic Philanthropies, and by Irish Life PLC. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

**Copyright:** © 2020 Ward M et al. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

**How to cite this article:** Ward M, May P, Briggs R et al. Linking death registration and survey data: Procedures and cohort profile for The Irish Longitudinal Study on Ageing [version 1; peer review: 2 approved, 1 approved with reservations] HRB Open Research 2020, 3:43 https://doi.org/10.12688/hrbopenres.13083.1

**First published:** 08 Jul 2020, 3:43 https://doi.org/10.12688/hrbopenres.13083.1
Introduction

Linking data from death registers with survey and other individual-level data is commonplace in many countries. This practice has enabled a number of prospective cohort studies collecting rich individual-level data, such as the English Longitudinal Study on Ageing (ELSA) and the Health and Retirement Study (HRS), to examine associations between mortality and a wide range of factors (for example see: Lewer et al., 2017; Wu et al., 2016). The Republic of Ireland has lacked an equivalent data infrastructure and analyses of Irish mortality have therefore been largely limited to unlinked Census data (Layte & Banks, 2016). Consequently, researchers’ ability to identify the determinants of mortality at the population level has been severely restricted.

In 2007, The Central Statistics Office (CSO) conducted a limited linkage exercise, linking all deaths that occurred in the year after the 2006 Census of Population to their Census record. However, this linked dataset was of limited utility due to the short one-year follow-up period and the very limited information collected as part of the census. Furthermore, both the census and mortality data files have limited socioeconomic status (SES) information, and no information on disease risk factors or antecedents (CSO, 2010). Our linking of longitudinal survey and death register data enables us to supplement the rich data available from longitudinal surveys with detailed data on cause of death available from official mortality registers.

Previous research has highlighted the numerous limitations inherent in using unlinked Census data, including the large gap between Census observation periods, and the dependence on unlinked numerators (count of deaths) and dominators (population grouping variable) (Layte et al., 2015; Layte & Banks, 2016; Layte & Nolan, 2016). Furthermore, Census data on SES variables in Ireland and elsewhere is particularly problematic due to the large amount of missing data. This missing data is often systematic being higher among younger age groups, women, and those not in paid employment at the time of Census data collection (Layte & Nolan, 2016). Importantly, individuals with missing SES information have also been shown to have higher mortality rates, which means that previous research on the association between SES and mortality in Ireland will likely have underestimated the true strength of this association (Layte & Banks, 2016). Beyond the issue of missing data inherent in analysing unlinked census data, even in cases were SES data is available, there is a large question mark over its validity. For example, White et al. (White et al., 2008) compared individual level social class from death records with that from the previous census in England and Wales and found that almost half of the records did not match. This incongruence is therefore another source of error. In light of the above, the necessity of linked survey-mortality data to properly identify the determinants of mortality rates is clear (Mackenbach et al., 2015).

As well as these problems with denominators, there may also be issues with the death counts themselves, particularly when interested in specific cause(s) of death rather than simply the event. For example, Daking and Dodds (Daking & Dodds, 2007) found differences in ICD-10 coding between Australian Census and coroners’ data and inconsistencies between population-based cancer registry data and death certificate data for cancer mortality have also been identified (German et al., 2011). A further complicating factor is that, in many cases, more than one condition may be compatible with the manner of death and indeed variability in the ascertainment of underlying cause of death has been well documented (Danilova, 2016).

All of the above is not to say that death certificate records are themselves necessarily free of error. Indeed, coding and other errors have been widely documented (Danilova et al., 2016; Danilova, 2016; Harteloh, 2018; Harteloh et al., 2010; McGivern et al., 2017). These studies have highlighted numerous inconsistencies in both the recording of information on death certificates by physicians (Myers & Farquhar, 1998) and coding practices across space and time (Danilova, 2016), particularly at the most detailed level of ICD-10 codes. These inconsistencies are exacerbated when the goal is to identify an underlying cause of death when more than one condition is recorded on the death certificate (Harteloh, 2018).

Here, we describe the steps taken to link death registration information with survey data derived from a large nationally representative prospective study of community-dwelling older adults. We also provide a profile of decedents among this cohort and compare mortality rates in this cohort to population-level data. Finally, we consider the utility of analysing underlying and contributory causes of death. This new data infrastructure provides many opportunities to contribute to our understanding of the social, behavioural, economic, and health antecedents to mortality and to inform public policies aimed at addressing inequalities in mortality and end-of-life care.

Methods

Death register data

Every death in the Republic of Ireland must be registered with the General Register Office (GRO). Registration is legally required, and non-registration is rare because of the necessity of a death certificate for many legal purposes. Firstly, the attending physician completes the medical certificate of the primary and contributory causes of death. This information, together with socioeconomic and demographic information provided by the next of kin or other qualified informant, is entered electronically at one of the 25 civil registration offices around the country and forwarded to the GRO. The GRO provides these records to the CSO on a weekly basis where it is collated for statistical reports on mortality. The CSO also administer a research micro-data file which includes individual-level data on date of death, residential address of decedent, place of death, primary and contributory causes of death, occupation of deceased, age of deceased, sex of deceased and marital status of deceased. All deaths registered on or after 1st January 2007 are coded according to ICD-10 rules. The CSO use Iris software to automatically assign ICD-10 to all diagnostic conditions and underlying cause of death from death certificates (CSO, 2018).

Survey data

The Irish Longitudinal Study on Ageing (TILDA) is a prospective nationally representative study of community dwelling adults aged ≥ 50 years resident in the Republic of Ireland. Details
of the methodology employed by TILDA are fully described elsewhere (Donoghue et al., 2018; Kearney et al., 2011; Kenny et al., 2010; Whelan & Savva, 2013). Briefly, TILDA participants were selected using multi-stage stratified random sampling method whereby 640 geographical areas, stratified by socio-economic characteristics, were selected, followed by 40 households within each area. The Irish GeoDirectory listing of all residential addresses provided the sampling frame. The first Wave of data collection was conducted between 2009 and 2011, with subsequent Waves collected at two-year intervals. Details of the sample maintenance strategies used by TILDA are also available elsewhere (Donoghue et al., 2017). TILDA collects information on a broad range of topics including health, economic, social, and family circumstances. Data collection consists of a number of components. Computer-assisted personal interviews (CAPI) and self-completion questionnaires (SCQ) were completed at each Wave of data collection and a comprehensive health assessment, conducted by trained nurses, was carried out at Waves 1 and 3, and will be repeated at Wave 6 in 2020. From Wave 2 onwards, End-of-Life (EOL) interviews have been completed with a spouse, relative or friend in cases where a participant had passed away (May et al., 2017). TILDA is a member of the HRS family of studies and is therefore harmonised with a number of large prospective cohort studies on ageing including ELSA, HRS, and The Survey of Health, Ageing and Retirement in Europe (SHARE).

Data linkage
TILDA was granted approval from the GRO to link TILDA respondents to their corresponding death certificate information. As there is no unique personal identifier in Ireland that could be used to match TILDA decedents to their death certificate record, matching was performed on the basis of name, address and month/year of birth (and age, to account for possible misreporting of age and/or month/year of birth on either file). Where records could not be linked based on this information, additional information such as marital status was used. The first round of data matching was conducted by the CSO in 2013 and a second was undertaken by the GRO in early 2018. Matching was performed for all individuals who died between Wave 1 (2009/2011) and March 2018. This procedure will be repeated at Wave 6 in 2020. From Wave 2 onwards, End-of-Life (EOL) interviews have been completed with a spouse, relative or friend in cases where a participant had passed away (May et al., 2017). TILDA is a member of the HRS family of studies and is therefore harmonised with a number of large prospective cohort studies on ageing including ELSA, HRS, and The Survey of Health, Ageing and Retirement in Europe (SHARE).

Matched death records were provided to TILDA in excel format. Each record consisted of a unique identifier, an immediate or proximal cause of death, and contributory factors. Of a total of 863 confirmed deaths among the TILDA sample, matching death records were obtained for 779 (90.3% of all known deaths at that time). Table 1 shows the timing of all deaths among TILDA participants, including those for whom it was not possible to match to death records. The smaller number of deaths identified after Wave 4 are due to the fact that data linkage was carried out at the beginning of Wave 5 data collection.

Coding of cause of death
Iris is a software tool for coding multiple causes of death and for the selection of the underlying cause of death. It is the preferred mortality coding tool of Eurostat. While early versions of Iris used the Centre of Disease Control-developed Medical Mortality Data System (MMDS), since version 5 it uses the Multicausal and Unicausal Selection Engine (MUSE). MUSE operates based on internationally agreed decision tables which are based on the most recent version of ICD-10. We used Iris version 5.4.0. The Iris software is free to use and can be downloaded, along with supporting materials, from the Iris institute.

Firstly, Iris attempts to code all diagnostic expressions included in each death certificate according to the World Health Organisation (WHO) ICD-10 classification system. Once all diagnostic expressions have been assigned an ICD-10 code, Iris then selects an underlying cause according to the MUSE decision tables which are regularly reviewed by the Iris consortium. Iris also provides a text format explanation on how the WHO mortality coding guidelines were applied when assigning underlying cause from the list of diagnostic conditions. Where possible, each condition reported in the death records were coded at the four-digit ICD-10 level. In cases where this automated coding system fails to assign an ICD-10 code or an underlying cause, manual coding was required. In our case, Iris successfully coded 18% of the 1,605 diagnostic expressions and assigned an underlying cause to 5.3% of the cases.

Underlying cause of death
We have operationalised underlying cause of death according to the WHO definition as “the disease or injury which initiated the train of morbid events leading directly to death, or the circumstances of the accident or violence which produced the fatal injury” (United Nations, 1991). We grouped underlying causes of death to ICD-10 chapters in order to adhere to TILDA data protection policies regarding minimum cell sizes for reporting purposes and also to ensure that groupings were large enough to enable statistically robust analyses. Of the 779 deaths, cancer was identified as the underlying cause of death for 37.0%; 32.9% of deaths were attributable to diseases of the circulatory system; 14.4% due to diseases of the respiratory system; while the remaining 15.8% of deaths occurred due to all other causes (Table 2).

Statistical analysis
Descriptive statistics included counts, percentages, and 95% confidence intervals. We used Cox proportional hazards regression models to estimate sex-adjusted hazard ratios for smoking as a risk factor for cause-specific mortality. Respondents lost to follow up were right-censored at the end of the

Table 1. Timing of deaths in TILDA.

<table>
<thead>
<tr>
<th>Timing of deaths</th>
<th>N</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Deceased between Wave 1 &amp; Wave 2</td>
<td>243</td>
<td>28.2</td>
</tr>
<tr>
<td>Deceased between Wave 2 &amp; Wave 3</td>
<td>329</td>
<td>38.1</td>
</tr>
<tr>
<td>Deceased between Wave 3 &amp; Wave 4</td>
<td>226</td>
<td>26.2</td>
</tr>
<tr>
<td>Deceased between Wave 4 &amp; Wave 5</td>
<td>65</td>
<td>7.5</td>
</tr>
<tr>
<td>Total</td>
<td>863</td>
<td>100</td>
</tr>
</tbody>
</table>

Page 4 of 16
follow-up period (March 2018). The results of this analysis are presented in Figure 3. All analyses were conducted using Stata/MP 14.2 (StataCorp, 2015).

**Results**

**Description of sample**

As shown in Table 3, the mean age of TILDA participants at baseline was 64 years (95% CI: 63.6, 64.3); 51.8% were women (95% CI: 50.7, 52.8). Almost one-third (31.5%, 95% CI: 30.0, 33.1) had primary level education while 22.2% had completed tertiary education (95% CI: 21.0, 23.5). A similar proportion of participants were employed (36.0%, 95% CI: 34.5, 37.4) or retired (36.6%, 95% CI: 35.1, 38.1) with the remainder unemployed, in full-time education or training, permanently sick or disabled, or looking after the family home on a full-time basis. In terms of household social class, 25.1% (95% CI: 23.8, 26.5) of participants were in the professional, managerial or technical social class while 21.0% (95% CI: 19.7, 22.4) in the semi- or un-skilled class. The remaining unclassified group included participants for whom there was not enough information to assign to a social class and those who were never economically active. The mean annual household income was €34,285.

Overall 9.1% of TILDA participants died during the nine-year follow-up period and the average age at death was 75.3 years (95% CI: 74.3, 76.3). The average age at death from cancers was 72.2 years (95% CI: 70.8, 73.7); diseases of the circulatory system 77.4 years (95% CI: 75.8, 79.0); and diseases of the respiratory system 77.8 years (95% CI: 75.3, 83.0). Mortality

### Table 2. Distribution of main ICD-10 Chapters classification among TILDA decedents.

<table>
<thead>
<tr>
<th>ICD-10 chapters</th>
<th>ICD-10 code</th>
<th>% (n)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Neoplasms</td>
<td>C00-D49</td>
<td>37.0 (288)</td>
</tr>
<tr>
<td>Diseases of the circulatory system</td>
<td>I00-I99</td>
<td>32.9 (256)</td>
</tr>
<tr>
<td>Diseases of the respiratory system</td>
<td>J00-J99</td>
<td>14.4 (112)</td>
</tr>
<tr>
<td>All others</td>
<td></td>
<td>15.8 (123)</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>100 (779)</td>
</tr>
</tbody>
</table>

### Table 3. Distribution of key sample characteristics for baseline sample, all-cause and cause-specific mortality.

<table>
<thead>
<tr>
<th></th>
<th>Baseline sample (n=8,174) % (95% CI)</th>
<th>All-cause mortality (n=779) % (95% CI)</th>
<th>Cancers (n=288) % (95% CI)</th>
<th>Circulatory (n=256) % (95% CI)</th>
<th>Respiratory (n=112) % (95% CI)</th>
<th>Other causes (n=123) % (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Mean age</strong></td>
<td>64.0 (63.6,64.3)</td>
<td>75.3 (74.3,76.3)</td>
<td>72.2 (70.8,73.7)</td>
<td>77.4 (75.8,79.0)</td>
<td>77.8 (75.3,80.3)</td>
<td>74.6 (72.1,77.1)</td>
</tr>
<tr>
<td><strong>Men</strong></td>
<td>48.2 (47.2,49.3)</td>
<td>53.1 (48.4,57.8)</td>
<td>53.7 (46.3,61.0)</td>
<td>56.5 (48.7,63.9)</td>
<td>40.9 (29.8,52.9)</td>
<td>55.4 (43.7,66.6)</td>
</tr>
<tr>
<td><strong>Women</strong></td>
<td>51.8 (50.7,52.8)</td>
<td>46.9 (42.2,51.6)</td>
<td>46.3 (39.0,53.7)</td>
<td>43.5 (36.1,51.3)</td>
<td>59.1 (47.1,70.2)</td>
<td>44.6 (33.4,56.3)</td>
</tr>
<tr>
<td><strong>Education</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Primary</td>
<td>31.5 (30.0,33.1)</td>
<td>53.1 (48.6,57.6)</td>
<td>45.2 (38.1,52.4)</td>
<td>59.9 (52.1,67.1)</td>
<td>56.0 (43.9,67.5)</td>
<td>52.3 (41.3,63.1)</td>
</tr>
<tr>
<td>Secondary</td>
<td>46.3 (44.9,47.7)</td>
<td>34.8 (30.6,39.3)</td>
<td>40.1 (33.3,47.3)</td>
<td>31.4 (24.7,38.9)</td>
<td>33.5 22.7,46.4)</td>
<td>32.4 (23.2,43.2)</td>
</tr>
<tr>
<td>3rd level</td>
<td>22.2 (21.0,23.5)</td>
<td>12.1 (9.7,14.9)</td>
<td>14.8 (11.1,19.4)</td>
<td>8.8 (5.7,13.4)</td>
<td>10.5 (5.1,20.2)</td>
<td>15.3 (9.4,23.9)</td>
</tr>
<tr>
<td><strong>Principal economic status</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Employed</td>
<td>36.0 (34.5,37.4)</td>
<td>11.4 (9.0,14.3)</td>
<td>14.4 (10.0,20.3)</td>
<td>10.5 (6.9,15.6)</td>
<td>8.1 3.5,17.4)</td>
<td>9.8 (5.2,17.7)</td>
</tr>
<tr>
<td>Retired</td>
<td>36.6 (35.1,38.1)</td>
<td>64.6 (60.3,68.7)</td>
<td>64.1 (56.9,70.6)</td>
<td>64.6 (57.3,71.4)</td>
<td>68.7 (56.0,79.1)</td>
<td>61.8 (49.8,72.5)</td>
</tr>
<tr>
<td>Other*</td>
<td>27.5 (26.2,28.8)</td>
<td>24.0 (20.4,28.0)</td>
<td>21.5 (16.3,27.8)</td>
<td>24.8 (18.7,32.1)</td>
<td>23.2 (14.1,35.8)</td>
<td>28.4 (19.1,40.0)</td>
</tr>
<tr>
<td><strong>Occupation social class</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Professionals</td>
<td>25.1 (23.8,26.5)</td>
<td>16.9 (13.4,21.0)</td>
<td>25.0 (18.3,33.3)</td>
<td>12.2 (7.9,18.5)</td>
<td>9.9 (3.9,23.1)</td>
<td>18.0 (10.1,30.0)</td>
</tr>
<tr>
<td>Non-manual</td>
<td>28.7 (27.4,29.9)</td>
<td>29.0 (24.3,34.2)</td>
<td>27.6 (20.2,36.3)</td>
<td>33.0 (25.1,42.1)</td>
<td>23.9 (13.7,38.2)</td>
<td>26.8 (16.2,40.9)</td>
</tr>
<tr>
<td>Skilled manual</td>
<td>17.5 (16.4,18.7)</td>
<td>17.1 (13.3,21.8)</td>
<td>21.7 (14.9,30.4)</td>
<td>14.8 (9.5,22.4)</td>
<td>12.1 (5.8,23.4)</td>
<td>18.0 (9.9,30.6)</td>
</tr>
<tr>
<td>Semi- &amp; un-skilled</td>
<td>21.0 (19.7,22.4)</td>
<td>21.8 (17.7,26.7)</td>
<td>16.0 (10.0,24.5)</td>
<td>24.7 (17.5,33.7)</td>
<td>29.0 (17.8,43.4)</td>
<td>20.2 (11.3,33.4)</td>
</tr>
<tr>
<td>Not classified</td>
<td>7.7 (6.9,8.6)</td>
<td>15.1 (11.4,19.9)</td>
<td>9.7 (5.5,16.5)</td>
<td>15.2 (9.5,23.6)</td>
<td>25.1 (13.7,41.5)</td>
<td>17.0 (8.3,31.6)</td>
</tr>
<tr>
<td><strong>Mean household income</strong></td>
<td>€34,285 (32526,36043)</td>
<td>€21,184 (18762,23606)</td>
<td>€23,547 (17595,29499)</td>
<td>€19,024 (16154,21894)</td>
<td>€20,344 (17055,23632)</td>
<td>€22,074 (18337,25812)</td>
</tr>
</tbody>
</table>

* The Other occupational group includes: unemployed, in full-time education or training, permanently sick or disabled, or looking after family home on a full-time basis.
rates were higher among less educated participants, manual occupation social class groups, and those with lower average annual household incomes.

Comparison of mortality rates to CSO life tables
In order to assess the representativeness of the TILDA mortality data in the Irish population, we compared our data to the Census of Population life tables. For this exercise, we used un-weighted data so that every death was counted equally. Figure 1a, b show the mortality rate for men and women, respectively, with CSO life tables for 2010–2012. The mortality rate on the y-axis was based on the hazard function which was calculated as the number of deaths at age x / the number of persons surviving to exact age x out of the original 100,000 aged 0. The x-axis was truncated at 94 years due to the small number of deaths that occurred after that age. Overall, mortality rates among younger TILDA participants aligned closely with those observed in the population. We did however observe some important differences with higher mortality rates observed among older decedents in our sample compared to the wider population.

Figure 2 shows the cause-specific failure curves for the major disease groups which highlight important differences. There were fewer deaths due to diseases of the respiratory system, particularly before the 70 years of age. Most of the deaths before this age occurred due to neoplasms and other causes including accidental deaths. After 70 years, a similar pattern was observed for diseases of the circulatory and respiratory system while neoplasms accounted for the greatest number of deaths.

Figure 1. TILDA mortality by age compared to CSO life tables 2010–2012. (a) Male mortality; (b) female mortality.
As well as the underlying cause of death described above, the death certificates also contained information on other diseases, injuries, or events that contributed to death. Among the 779 death records, up to seven contributory causes were also recorded and 67.5% of records had at least one contributory cause listed. One of the key advantages of our approach to data linkage is that we were able to assign an ICD-10 code to every contributory cause of death, thus enabling us to consider these contributory factors as well as the underlying cause of death. Through this procedure we identified neoplasms as being a contributory factor in 40.8% of deaths, while diseases of the circulatory system and diseases of the respiratory system were mentioned in 52.6% and 34.4% respectively (Table 4).

To assess the utility of contributory cause of death versus underlying cause, Figure 3 shows the sex-adjusted hazard ratios for smoking as a risk factor for cause-specific mortality according
to both underlying and contributory (any mention) cause of death. In each instance, we observed similar estimates whether we assigned death due to an underlying or contributory cause.

**Discussion**

We have described the procedures employed to link death registration information to survey data among a large sample derived from a nationally representative cohort of community-dwelling older adults. From the first round of data collection in TILDA to early 2018 (nine-year follow-up), there were 863 confirmed deaths and it was possible to link to death registration data of 779 (90.3%) of these. This compares favourably to a similar exercise conducted by the CSO whereby all deaths occurring in the year after the 2006 Census of Population were matched to their corresponding Census record which resulted in a matching rate of 79.8%. The Northern Ireland Mortality Study, which links death certificate information with the 1991, 2001 and 2011 UK Census of Population, obtained a matching rate of 94% using names and addresses (CSO, 2010).

Comparison with life tables from the CSO showed that mortality rates among younger participants closely aligned with those in the wider population. While TILDA mortality rates were lower in the older age groups, this divergence is unsurprising given that the TILDA sample was drawn from adults living in the community which means that they were on average healthier than the total population of older adults. Furthermore, this pattern is similar to that reported from the Health and Retirement Study (Weir, 2016).

There are a number of important advantages to the approach to data coding and linkage described here. Having access to detailed death registration information provides us the opportunity to operationalise the causes of mortality in a number of different ways: underlying all-cause and cause-specific, contributory, and multiple cause of death. The richness and breadth of information collected by TILDA over multiple waves provides us with a unique opportunity to contribute to the study of mortality.

Having complete death registration data is particularly important when concerned with assessing multiple causes of death. For example, recent studies demonstrated how a multiple-cause-of-death approach is useful to characterise the contribution of diabetes (Rodriguez et al., 2019) and falls (Kiadaliri et al., 2019) to mortality. Here, we assessed the utility of contributory cause of death versus underlying cause of death using the example of smoking as a risk factor for cause-specific mortality. We observed similar estimates whether we assigned death due to an underlying or contributory cause, which suggests the use of either contributory or underlying cause may not greatly impact on estimates of the association between risk factors and mortality. Indeed, one potential benefit of using contributory causes is increased statistical power due to larger numbers and a reduction in the associated error. More broadly, the utility of contributory cause of death in epidemiological research has also been shown to be similar to that of underlying cause while reducing the risk of measurement error due to the potential identification of an underlying cause.

The application of standardised coding dictionaries and decision tables in the Iris software can aid harmonisation across data sources and jurisdictions. This harmonisation is critical to enable researchers better understand differences in the mortality rates and the mechanisms that explain differences between populations. However, our initial application of IRIS software for assigning ICD-10 codes to all conditions contained in the death registration data and subsequently identifying an underlying cause of death required substantial manual input. The failure to automatically assign codes was due mostly to syntax and semantic differences between the terms included on death certificates and the Iris dictionary. For example, Iris failed to automatically code cases of “ischaemic heart disease” as it searched for “ischemic”. When such failures occurred, researchers had to manually enter the appropriate ICD-10 code. The Iris dictionary was then amended so that subsequent incidences of ischaemic heart disease were automatically coded. This procedure will greatly improve the automation of the coding process in future Waves of TILDA.

**Limitations**

While every effort has been made to ensure that an appropriate underlying cause of death was assigned to each decedent, we cannot account for potential errors in the recording of individual death certificates. For example, a comparison of death certificate data with associated medical records showed high error rates on death certificates, including ICD-10 coding (McGivern et al., 2017). However, our application of broader diagnostic categories in the form of ICD-10 Chapters and our ability to include contributory conditions and multiple-cause-of-death in our analyses should minimise the impact of these potential errors. For example, consistency in coding of mortality has been shown to improve when cause of death is grouped into broad diagnostic categories (Danilova, 2016).

There is necessarily a time lag whereby, unbeknownst to us, participants may have died since the last round of data collection. This is inevitable as we do not have an automated linkage system with the GRO. The practical effect of this is that we have likely underestimated the rates of mortality for the most recent

---

**Table 4. Distribution of main ICD-10 Chapters classification of underlying cause of death and contributory diseases among TILDA decedents.**

<table>
<thead>
<tr>
<th>ICD-10 chapters</th>
<th>Underlying cause % (n)</th>
<th>Contributory % (n)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Neoplasms</td>
<td>37.0 (288)</td>
<td>40.8 (318)</td>
</tr>
<tr>
<td>Diseases of the circulatory system</td>
<td>32.9 (256)</td>
<td>52.6 (410)</td>
</tr>
<tr>
<td>Diseases of the respiratory system</td>
<td>14.4 (112)</td>
<td>34.4 (268)</td>
</tr>
<tr>
<td>All others</td>
<td>15.8 (123)</td>
<td>44.5 (347)</td>
</tr>
<tr>
<td>Total</td>
<td>100 (779)</td>
<td></td>
</tr>
</tbody>
</table>
period. The potential impact of this on our current analyses will be assessed during subsequent rounds of data linkage.

**Conclusion**
This is the first time that death registration data has been linked to survey data in the Republic of Ireland. This work therefore provides an important data infrastructure for research on mortality in Ireland. The rich and wide-ranging data collected by TILDA, including objective health assessment data, means that we have a unique opportunity to contribute to our understanding of the social, behavioural, economic, and health antecedents to mortality and to inform public policies aimed at addressing inequalities in mortality and end-of-life care. Finally, because TILDA is harmonised with other large prospective cohort studies within the HRS family of studies, this new data infrastructure also provides opportunities for researchers and policy makers interested in examining difference in the nature of mortality and its antecedents between populations.

**Data availability**

**Underlying data**
The first four waves of TILDA data are available from the Irish Social Science Data Archive (ISSDA) at www.ucd.ie/issda/data/tilda/. Due to the sensitive nature of death registration data, the cause of death data reported here are not publicly accessible at this time. Requests to access this data can be made directly to TILDA (tilda@ucd.ie) and will be considered on a case-by-case basis.

To access the TILDA survey data, please complete an ISSDA Data Request Form for Research Purposes, sign it, and send it to ISSDA by email (issda@ucd.ie).

For teaching purposes, please complete the ISSDA Data Request Form for Teaching Purposes, and follow the procedures, as above. Teaching requests are approved on a once-off module/workshop basis. Subsequent occurrences of the module/workshop require a new teaching request form.

**References**


ISSDA Data Request Form for Research Purposes. www.ucd.ie/issda/data/tilda/.

ISSDA Data Request Form for Teaching Purposes. www.ucd.ie/issda/data/tilda/.

TILDA is harmonised with other large prospective cohort studies within the HRS family of studies, this new data infrastructure also provides opportunities for researchers and policy makers interested in examining difference in the nature of mortality and its antecedents between populations.
This is an important piece of linkage study that is relevant to the Irish context when such data linkages are available elsewhere. It is also important to note that linkage studies are methodologically challenging in Ireland because of the lack of a unique identifier. The CSO did make attempts earlier to undertake such linkage research but was insufficient and was both labour and resource intensive. The current study builds on earlier linkage studies undertaken both by CSO and GRO in 2013 and 2018, respectively.

My main concern is the lack of explicit description of the linkage methodology in the current paper, which will not be very helpful for a researcher towards reproducibility. There are currently no standardized quality appraisal tools available to assess quality and bias of any linkage studies. However, it is essential that a linkage study must meet the following characteristics:

- Completeness of source databases
- Accuracy of data sources
- Linkage methodology and technology
- Ethical and data security considerations.

In the context of the current study - the first two criteria are broadly met. However, my main concern is with the linkage methodology and technology. My understanding is that the TILDA researchers were not primarily involved in the linkage methodology given that matching of records were undertaken separately by CSO in 2013 and by GRO in 2018. The TILDA team had a role to get an approval and forward their data to these two data sources team who in fact undertook the matching process - the details of which are not available to us. It also appears that the technology (software) used is IRIS, which is a broadly validated accepted tool for coding purposes employed by EUROSTAT and CSO in the past. However, this software also had limitations in capturing and coding all the diagnostic expressions - only 18% and 5% of all the cases. The rest of the matching was done manually - by whom and how is unclear. This is a crucial step for which sufficient information and clarity is lacking. Second, the matching was not 100% accurate - around 10% of records were unmatched - and further analyses of these unmatched records are essential to rule out systematic bias - measurement error, and such sensitivity analyses (false positives and
false negatives) have not been provided. Third, the matching variables employed were only three - name, address, and age (and marital status for some, but not sure for how many?). Names, especially for females can change once married; addresses are not always permanent - and age is also variable. Therefore, further details on how these methodological limitations during the process of matching were handled are unclear. There is also limited information on ethical and data security considerations for this linkage study when personal data have been used, especially from a GDPR perspective.

Furthermore, the coding practices of causes of death are crucial for any linkage studies. The authors have undertaken a separate analysis of exploring contributory versus underlying causes of deaths for the participants, and I believe that this piece of research is the sole contribution of the TILDA team to this paper. However, this could have been explained further and there is lack of clarity on how the unclassified causes of deaths within each of the three main types of causes of deaths (cancer, cardiovascular and respiratory) were handled. The CSO website clearly indicates ‘unclassified’ causes of cancer deaths and likewise for other conditions - and the Global Burden of Disease (GBD) Study team call these as ‘garbage’ codes. The GBD studies on causes of death have shown that there is a good proportion of ‘garbage’ codes for any death registry, and they have also developed a statistical technique on how to ‘redistribute’ these garbage codes. No such information is available to us in the current study.

In short, I approve the study but has methodological limitations and caveats which could have been addressed.

**Is the rationale for developing the new method (or application) clearly explained?**
Yes

**Is the description of the method technically sound?**
Yes

**Are sufficient details provided to allow replication of the method development and its use by others?**
Partly

**If any results are presented, are all the source data underlying the results available to ensure full reproducibility?**
Partly

**Are the conclusions about the method and its performance adequately supported by the findings presented in the article?**
Yes

**Competing Interests:** No competing interests were disclosed.

**Reviewer Expertise:** Tobacco Control; Non-communicable epidemiology; Global Burden of Disease (GBD) methodology.

I confirm that I have read this submission and believe that I have an appropriate level of
expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Reviewer Report 22 July 2020

https://doi.org/10.21956/hrbopenres.14183.r27634

© 2020 Harteloh P. This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Peter Harteloh
Statistics Netherlands (CBS), The Hague, The Netherlands

Linkage studies are important for enhancing the analytical power of cause-of-death registrations. They provide insight in associations between causes of death and their determinants. Linkage studies improve the utility of cause-of-death registrations for health policy or research. The study of Ward et al. is a fine example of such a linkage study. It is clear and well written. It shows associations between social economic status and causes of death both from a traditional approach by selecting one underlying cause of death per deceased and by a multiple cause coding approach. I would surely recommend its indexing, but ask for some minor revisions and answers to some questions.

Abstract: “Death records were obtained for 779 (90.3% of all confirmed deaths at that time) and linked to individual level survey data from The Irish Longitudinal Study on Ageing (TILDA).” Typo: Close brackets after 90.3% in stead of after “time”.

Methods. Coding of cause of death: “In our case, Iris successfully coded 18% of the 1,605 diagnostic expressions and assigned an underlying cause to 5.3% of the cases.” Usually about 60-70% of the records are coded automatically: see Harteloh, 20181. Can the authors explain this poor performance? If the performance of Iris is really that bad, I would not recommend using the software. I would consider the records coded manually. Could the authors say something about the instructions for manual coding i.e. processing the records not being coded automatically by Iris. Are all medical expressions on the death certificate coded and do the coders use volume 2 of the ICD-10? Are there any instructions deviating from volume 2 of the ICD-10 used? (as local certifying practice sometimes requires).

Also, if a record was rejected by Iris and then handled manually by coding all the expressions on a death certificate, Iris can select the underlying cause of death automatically in most of the cases (about 95%). I wonder why this function of Iris has not been used by the authors?

In short, I would like to have some more information about the use of Iris in the coding process in order to understand the multiple cause coding approach of the authors.

Methods. Data linkage. Can the authors say something about the ethics of linking survey data with cause of death registrations? They seem to suggest (“We grouped underlying causes of death to ICD-10 chapters in order to adhere to TILDA data protection policies regarding minimum cell sizes for reporting purposes”) some ethical restrictions. I wonder if the participant of the survey study gave permission for linkage to other data sources such as a cause of death registration.
Methods. A definition (explanation) of “contributory cause of death” is missing. It is commonly defined as a cause of death, not being selected as underlying cause of death (and mentioned in part 2 of the death certificate). However, the authors seem to use it for causes of death being mentioned on a death certificate. Otherwise, I cannot understand so many malignancies not being underlying cause of death (see table 4). So please explain the use of this concept (or replace it by “being mentioned”, regardless of being underlying cause of death).

Methods. Why did the authors (specifically) focus on the relationship between smoking and causes of death? What about other SES determinants? In order to avoid fishing expeditions, the selection of determinants to be studied should be clearly motivated.

Results. “while diseases of the circulatory system and diseases of the respiratory system were mentioned in 52.6% and 34.4% respectively”. Did the authors count records mentioning at least one cause of death of the group under consideration?

Results. Table 4. I think mentioned (of a death record) instead of contributory cause of death is meant here. Also in the column counting contributory causes of death: is this a count of records mentioning at least one malignancy etc... Otherwise, the numbers seem very low to me.

Results. Figure 3. Very interesting approach. Could the authors explain the fact that smoking is not a statistically significant determinant of cancer death? I assume lung cancer is the most prevalent cancer as cause of death.

Results. “In each instance, we observed similar estimates whether we assigned death due to an underlying or contributory cause.” Not clear. Please explain or show these estimates.

Results. “We observed similar estimates whether we assigned death due to an underlying or contributory cause, which suggests the use of either contributory or underlying cause may not greatly impact on estimates of the association between risk factors and mortality.” A bit far fetched for such an important conclusion when the estimates are not shown. In addition, could the negative result be explained by the grouping of causes of death? I would like to see the result of associations between risk factors and major causes of death such as dementia, lung cancer or cerebrovascular accidents if the privacy rules are not violated.

Discussion. “For example, Iris failed to automatically code cases of “ischaemic heart disease” as it searched for “ischemic”. This example is not clear to me. When you put “ischaemic heart disease” in your dictionary Iris will be able to code the expression automatically. Please explain.

Conclusion. “This is the first time that death registration data has been linked to survey data in the Republic of Ireland. This work therefore provides an important data infrastructure for research on mortality in Ireland.” I agree! This is a very important aspect of this study. It deserves to be indexed.

Outcome of my review: approved. Some minor issues to be addressed. Most important: clear up the use of the term “contributory cause of death”. Finally, I would like to compliment the authors on their research and encourage further analysis.
References

Is the rationale for developing the new method (or application) clearly explained?
Yes

Is the description of the method technically sound?
Yes

Are sufficient details provided to allow replication of the method development and its use by others?
Yes

If any results are presented, are all the source data underlying the results available to ensure full reproducibility?
No source data required

Are the conclusions about the method and its performance adequately supported by the findings presented in the article?
Yes

**Competing Interests:** No competing interests were disclosed.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Reviewer Report 13 July 2020
https://doi.org/10.21956/hrbopenres.14183.r27635

© 2020 Lewer D. This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Dan Lewer
Department of Epidemiology and Public Health, University College London, London, UK

Thank you for inviting me to review this article. It provides a clear summary of a linkage exercise conducted between a community health survey of older people and national mortality data in Ireland.

The data is a valuable resource and researchers will find this technical article useful.

To my knowledge this type of data is not commonplace (as per first line of introduction), which strengthens the international importance of this data.
I think a central use of this data is analyses of the association between longitudinal information on exposures and mortality (e.g. what is the effect of weight loss, quitting smoking, or cognitive decline?). This is not discussed in the article, and I think it might be worth mentioning this as a potential use of the dataset. In general, I would find it useful to know some of the key research questions that the authors think the dataset might address (though of course it's not possible to anticipate all the different research uses).

A few questions/comments:
1. What is a confirmed death? If not from the linked mortality records, how do you find out that a participant has died (i.e. how do you know that 863 participants died?). Apologies if I missed an explanation of this in the text.

2. Is it worth adding some information on the associations with successful linkage? (i.e. were certain types of participant less likely to be linked?)

3. For participants who are linked, what is the probability of correct linkage? Did the linkage process use an existing method, and is there any validation that the linkages are correct?

4. I like the analysis of smoking. It might be worth adding a brief justification for this analysis to the introduction (e.g. that the relationship between smoking and different causes of death is well-researched in other sources, so it acts as a kind of validation - you would expect a stronger association between smoking and respiratory causes of death than between smoking and all-cause mortality; or because it allows you to evaluate the difference between the derived 'underlying cause' of deaths and contributing causes?). Would it be possible to add the association between ever-smoking and all-cause mortality to figure 3 for comparison?

5. In the results, you mention that "mortality rates were higher among less educated participants, manual occupation social class groups, and those with lower average annual household incomes." I can see in Table 3 that (for example) 53% of deaths were among people with only primary education, while 32% of the baseline sample had only primary education. This does suggest higher mortality rates in this group, but does not explicitly show the rates or the association between education and mortality. I'd suggest either omitting this from the results, or adding specific results that support this association.

6. I like the age-specific comparison to the general population provided in Figure 1. The results say that "Overall, mortality rates among younger TILDA participants aligned closely with those observed in the population. We did however observe some important differences with higher mortality rates observed among older decedents in our sample compared to the wider population." However, in the figure, mortality rates look lower for the TILDA participants at both younger and older ages. It may help to (a) plot these charts with a log y-axis, and (b) use a model to plot a smooth curve with confidence limits that can be more easily compared to the general population. It looks like a simple exponential model would work, (c) report the age-standardised mortality rate for both the cohort and the general population. Also note that the mortality rate is not among decedents but among the population/participants.
7. In the limitations, you note that “There is necessarily a time lag whereby, unbeknownst to us, participants may have died since the last round of data collection. This is inevitable as we do not have an automated linkage system with the GRO. The practical effect of this is that we have likely underestimated the rates of mortality for the most recent period.” It may be possible to address this by ending follow-up at an earlier date, e.g. 6 months before the final linkage date, to increase the likelihood that your study includes all deaths for the follow-up period.

Is the rationale for developing the new method (or application) clearly explained?
Yes

Is the description of the method technically sound?
Yes

Are sufficient details provided to allow replication of the method development and its use by others?
Yes

If any results are presented, are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions about the method and its performance adequately supported by the findings presented in the article?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Research using electronic health records; public health; health and social exclusion; health inequalities.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.