METHOD ARTICLE

REVISED Linking death registration and survey data: Procedures and cohort profile for The Irish Longitudinal Study on Ageing (TILDA) [version 2; peer review: 3 approved]

Mark Ward[1], Peter May[2], Robert Briggs[1,3], Triona McNicholas[1,3], Charles Normand[2], Rose Anne Kenny[1,3], Anne Nolan[1,4]

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

v2 **First published:** 08 Jul 2020, **3**:43
https://doi.org/10.12688/hrbopenres.13083.1
Latest published: 19 Nov 2020, **3**:43
https://doi.org/10.12688/hrbopenres.13083.2

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 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.

Open Peer Review

Reviewer Status ✓ ✓ ✓

	Invited Reviewers		
	1	**2**	**3**
version 2 (revision) 19 Nov 2020	✓ report		✓ report
version 1 08 Jul 2020	✓ report	✓ report	? report

1. **Dan Lewer**, University College London, London, UK
2. **Peter Harteloh**, Statistics Netherlands (CBS), The Hague, The Netherlands
3. **Zubair Kabir**, University College Cork, Cork, Ireland

Any reports and responses or comments on the article can be found at the end of the article.

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.

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 (TILDA) [version 2; peer review: 3 approved]** HRB Open Research 2020, **3**:43 https://doi.org/10.12688/hrbopenres.13083.2

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

REVISED **Amendments from Version 1**

The acronym of The Irish Longitudinal Study on Ageing, 'TILDA', has been appended to the title of the manuscript. Text referring to the % of confirmed deaths has been removed from the abstract.

A new reference to a data linkage exercise carried out by the Central Statistics Office has been added to the second paragraph of the introduction (CSO, 2019).

A new paragraph has been added at the beginning of the section on 'Data linkage'. This text describes how decedents among TILDA participants were identified. In the second paragraph of this section, we have now clarified that the data linkage described in the manuscript was undertaken with the General Registers Office in 2018 and not with the Central Statistics Office in 2013. The third paragraph in this section also clarifies that the 84 deaths that were known to the authors but not included in the data linkage, occurred after matching had taken place. These deaths, and more recent ones, will be linked to their official death records when the data linkage is repeated in 2021.

In our discussion of Figure 1, we have included a new reference to the results of a similar data linkage exercise carried out by the Health and Retirement Study (Weir , 2016).

Figure 2 now includes estimates for the association between smoking and all-cause mortality.

A justification for the analysis of the association between smoking and mortality to compare the estimates derived from including underlying cause versus contributory cause of death has been included. We have also provided a reference to a similar analysis (Batty *et al.,* 2019). We have now explained our use of the term 'contributory' cause of death

The numbering of the bibliography has been changed to incorporated the new citations described above, that we have included in this revision.

Any further responses from the reviewers can be found at the end of the article

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, and again in 2017, the Central Statistics Office (CSO) conducted a limited data linkage exercise, linking all deaths that occurred in the year after the 2006 Census of Population to their Census record. However, these linked datasets are 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; CSO, 2019). 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 long time 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. 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 assignation 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 whereby 640 geographical areas, stratified by socioeconomic 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 2021. 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

Deaths among TILDA participants were identified by a number of methods. In many cases, spouses or other relatives of decedents contacted TILDA to inform them of the death of the participant. Other deaths were identified when interviewers visited the home of decedents to conduct subsequent waves of data collection. Also, where it was not possible to contact a participant, the TILDA data management team identified deaths through searches of the obituary website RIP.ie which is dedicated to publishing death notices in Ireland and deathevents.gov.ie, an online service that reports information on death events to public sector bodies. Finally, for a number of the remaining cases where the status of participants was not known, GRO records were interrogated in order to identify those who had died.

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. Data matching was conducted with 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 as subsequent waves of TILDA data become available.

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. The 84 deaths not captured in this data linkage occurred after we completed the exercise and will be captured when we repeat data linkage in 2021. 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

Table 1. Timing of deaths in TILDA.

Timing of deaths	N	%
Deceased between Wave 1 & Wave 2	243	28.2
Deceased between Wave 2 & Wave 3	329	38.1
Deceased between Wave 3 & Wave 4	226	26.2
Deceased between Wave 4 & Wave 5	65	7.5
Total	863	100

number of deaths identified after Wave 4 is 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 in the first iteration 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).

Table 2. Distribution of main ICD-1O Chapters classification among TILDA decedents.

ICD-10 chapters	ICD-10 code	% (n)
Neoplasms	C00-D49	37.0 (288)
Diseases of the circulatory system	I00-I99	32.9 (256)
Diseases of the respiratory system	J00-J99	14.4 (112)
All others		15.8 (123)
Total		100 (779)

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 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

For reference, the distribution of important socio-demographic characteristics of the full TILDA sample and those who died over the course of the study are presented 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 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

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

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

* 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.

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 more closely than those among older decedents, with those observed in the population. This pattern is similar to that reported by the Health and Retirement Study (Weir, 2016).

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 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.

Underlying versus contributory cause of death

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. A contributory cause of death is a disease or condition that contributed to the death but was not directly implicated and recorded in part two of death certificates. While this information has been rarely used in epidemiological research, recent evidence suggests that it may have some methodological utility (Batty *et al.,* 2019). For present purposes, contributory causes include diseases and conditions listed anywhere on the death certificate.

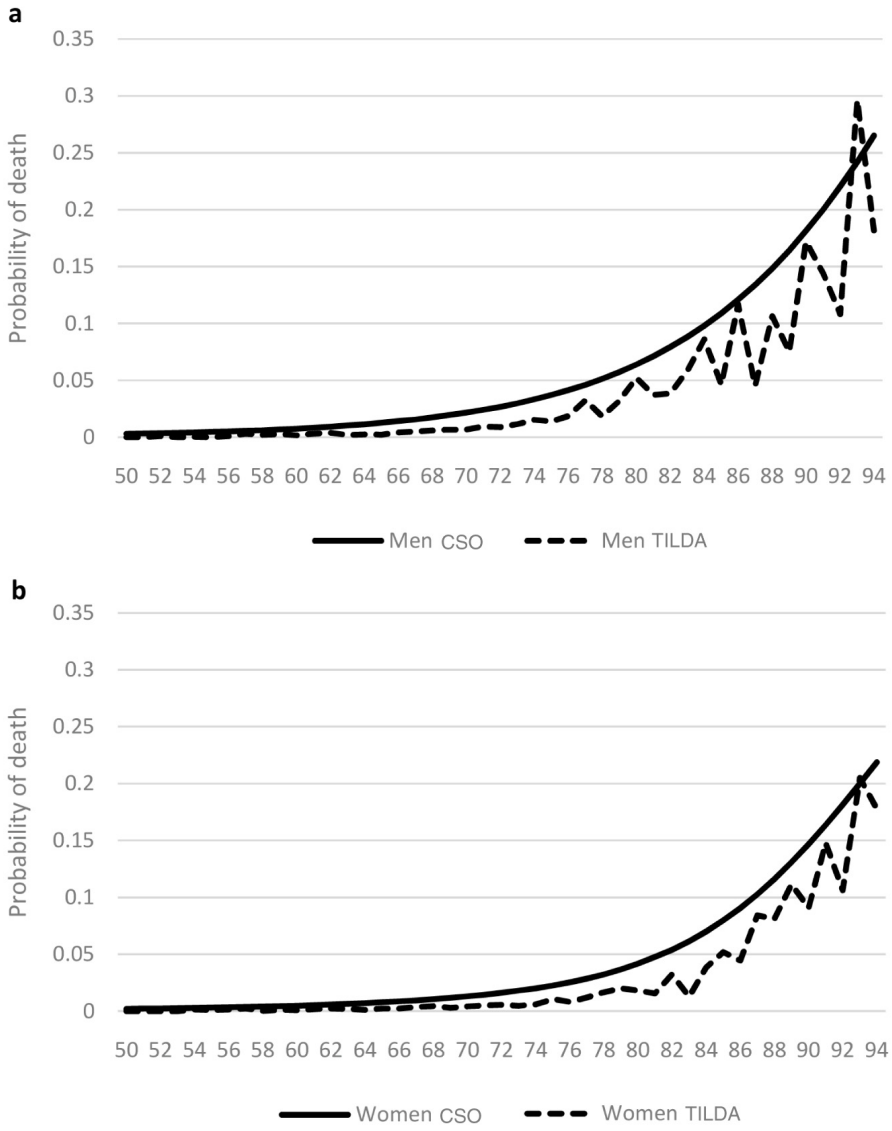

Figure 1. TILDA mortality by age compared to CSO life tables 2010–2012. (**a**) Male mortality; (**b**) female mortality.

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 all-cause, and cause-specific mortality according to both underlying and contributory (any mention) cause of death. We chose smoking to test our hypothesis that similar estimates would be derived from both underlying and contributory conditions as smoking is an established risk factor for mortality and it has been used for a similar purpose previously (Batty *et al.,* 2019). In each instance, we observed similar estimates whether we assigned death due to an underlying or contributory cause. Smokers, including those who had quit, had an increased all-cause mortality risk (HR= 1.38, 95% CI:1.16-1.62) compared to participants who never smoked. The estimates for both cardiovascular and respiratory, contributory (any mention) and underlying cause of death were similar. The precision of the estimates was better when including the contributory conditions due to the increased number of cases included in these groups.

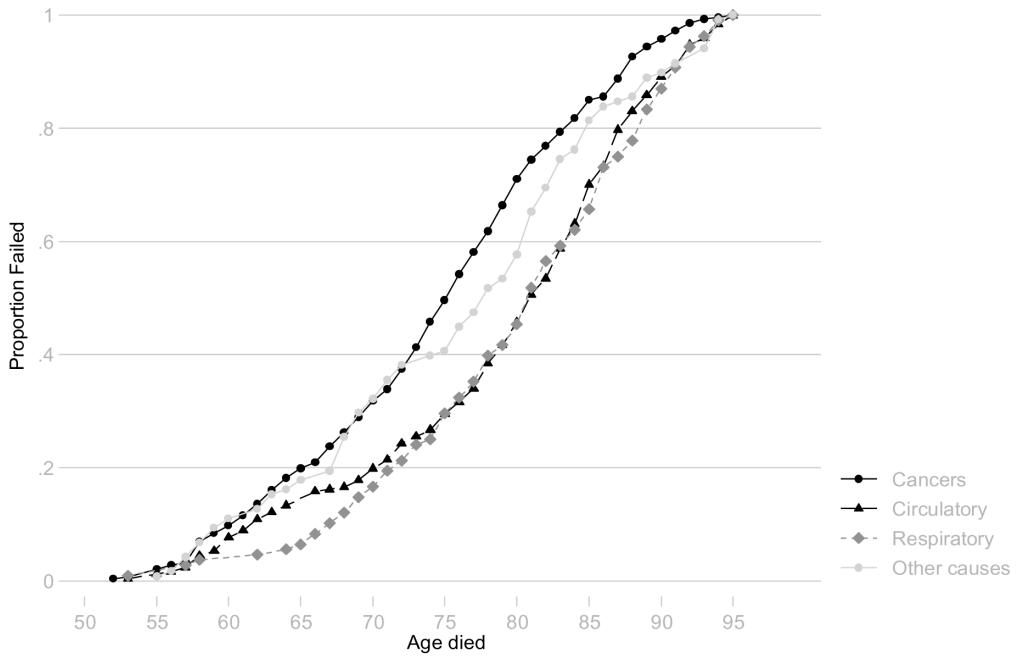

Figure 2. Cause-specific failure curves.

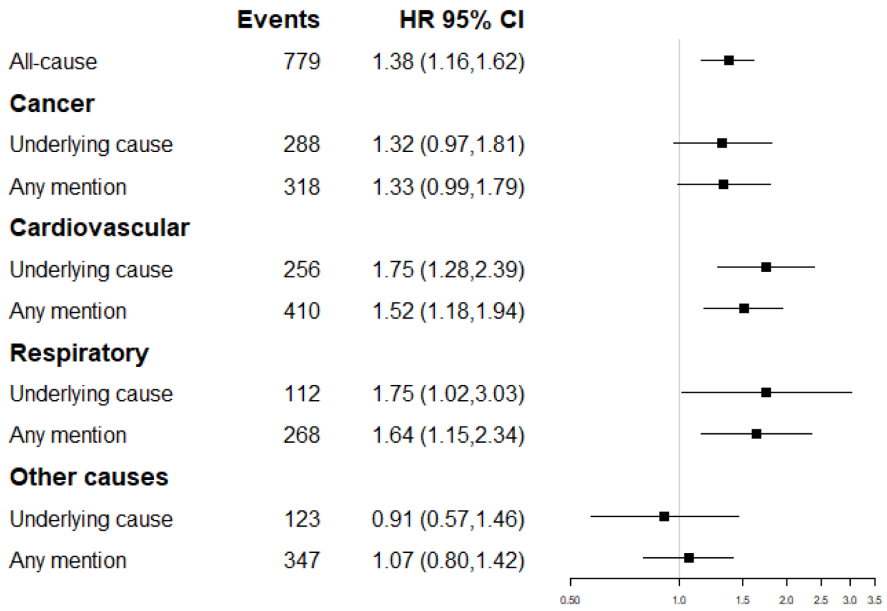

Figure 3. Sex-adjusted hazard ratios for ever smoking in relation to underlying and contributory cause-specific mortality.

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), it was possible to link to death registration data of 779 confirmed deaths. 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

Table 4. Distribution of main ICD-1O Chapters classification of underlying cause of death and contributory diseases among TILDA decedents.

ICD-10 chapters	Underlying cause % (n)	Contributory % (n)
Neoplasms	37.0 (288)	40.8 (318)
Diseases of the circulatory system	32.9 (256)	52.6 (410)
Diseases of the respiratory system	14.4 (112)	34.4 (268)
All others	15.8 (123)	44.5 (347)
Total	100 (779)	

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. This finding is similar to that reported previously by Batty *et al.* (2019) and an earlier study by Crews *et al.* (1991). 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 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@tcd.ie) and will be considered on a case-by-case basis.

To access the TLDA 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

Batty GD, Gale CR, Kivimäki M: **Assessment of Relative Utility of Underlying vs Contributory Causes of Death.** *JAMA Netw Open.* 2019; **2**(7): e198024.
PubMed Abstract | Publisher Full Text | Free Full Text

Crews DE, Stamler J, Dyer A: **Conditions other than underlying cause of death listed on death certificates provide additional useful information for epidemiologic research.** *Epidemiology.* 1991; **2**(4): 271–275.
PubMed Abstract | Publisher Full Text

CSO: **Mortality Differentials in Ireland. An Analysis Based on the Census Characteristics of Persons Who Died in the Twelve Month Period after Census Day 23 April 2006**. Dublin. **2010.**
Reference Source

CSO: **Mortality Differentials in Ireland. An Analysis Based on the Census Characteristics of Persons Who Died in the Twelve Month Period after Census Day 24 April 2016.** Dublin. 2019. [Accessed: October 2020].
Reference Source

CSO: **Implementation of IRIS Software for the Automated Coding of Deaths in Ireland**. Dublin. **2018.**
Reference Source

Daking L, Dodds L: **ICD - 10 Mortality Coding and the NCIS: A Comparative Study.** *Health Inf Manag.* 2007; **36**(2): 11–23; discussion 23–5.
PubMed Abstract | Publisher Full Text

Danilova I, Shkolnikov VM, Jdanov DA, *et al.*: **Identifying Potential Differences in Cause-of-Death Coding Practices across Russian Regions.** *Popul Health Metr.* 2016; **14**: 8.
PubMed Abstract | Publisher Full Text | Free Full Text

Danilova IA: **Problems of the Quality of Cause-Specific Mortality Statistics at Old Ages.** *Advances in Gerontology.* 2016; **6**(1): 1–5.
Publisher Full Text

Donoghue O, Foley M, Kenny RA: **Cohort Maintenance Strategies Used by The Irish Longitudinal Study on Ageing (TILDA)**. Dublin. **2017.**
Reference Source

Donoghue OA, McGarrigle CA, Foley M, *et al.*: **Cohort Profile Update: The Irish Longitudinal Study on Ageing (TILDA).** *Int J Epidemiol.* 2018; **47**(5): 1398–1398l.
PubMed Abstract | Publisher Full Text

German RR, Fink AK, Heron M, *et al.*: **The Accuracy of Cancer Mortality Statistics Based on Death Certificates in the United States.** *Cancer Epidemiol.* 2011; **35**(2): 126–31.
PubMed Abstract | Publisher Full Text

Harteloh P: **The Implementation of an Automated Coding System for Cause-of-Death Statistics.** *Inform Health Soc Care.* 2018; **45**(1): 1–14.
PubMed Abstract | Publisher Full Text

Harteloh P, De Bruin K, Kardaun J: **The Reliability of Cause-of-Death Coding in The Netherlands.** *Eur J Epidemiol.* 2010; **25**(8): 531–38.
PubMed Abstract | Publisher Full Text

Kearney PM, Cronin H, O'Regan C, *et al.*: **Cohort Profile: The Irish Longitudinal Study on Ageing.** *Int J Epidemiol.* 2011; **40**(4): 877–84.
PubMed Abstract | Publisher Full Text

Kenny RA, Whelan B, Cronin H, *et al.*: **The Design of the Irish Longitudinal Study on Ageing**.
Reference Source

Kiadaliri AA, Rosengren BE, Englund M: **Fall-Related Mortality in Southern Sweden: A Multiple Cause of Death Analysis, 1998-2014.** *Inj Prev.* 2019; **25**(2): 129–35.
PubMed Abstract | Publisher Full Text

Layte R, Banks J: **Socioeconomic Differentials in Mortality by Cause of Death in the Republic of Ireland , 1984 - 2008.** *Eur J Public Health.* 2016; **26**(3): 451–58.
PubMed Abstract | Publisher Full Text

Layte R, Banks J, Walsh C, *et al.*: **Trends in Socio-Economic Inequalities in Mortality by Sex in Ireland from the 1980s to the 2000s.** *Ir J Med Sci.* 2015; **184**(3): 613–21.
PubMed Abstract | Publisher Full Text

Layte R, Nolan A: **Socio-Economic Differentials in Male Mortality in Ireland 1984-2008.** *The Economic and Social Review.* 2016; **47**(3): 361–90.
Reference Source

Lewer D, McKee M, Gasparrini A, *et al.*: **Socioeconomic Position and Mortality Risk of Smoking: Evidence from the English Longitudinal Study of Ageing (ELSA).** *Eur J Public Health.* 2017; **27**(6): 1068–73.
PubMed Abstract | Publisher Full Text | Free Full Text

Mackenbach JP, Menvielle G, Jasilionis D, *et al.*: **Measuring Educational Inequalities in Mortality Statistics.** Paris. 2015.
Reference Source

May P, Mcgarrigle C, Normand C: **The End of Life Experience of Older Adults in Ireland.** Dublin. 2017.
Reference Source

McGivern L, Shulman L, Carney JK, *et al.*: **Death Certification Errors and the Effect on Mortality Statistics.** *Public Health Rep.* 2017; **132**(6): 669–75.
PubMed Abstract | Publisher Full Text | Free Full Text

Myers KA, Farquhar DR: **Improving the Accuracy of Death Certification.** *CMAJ.* 1998; **158**(10): 1317–23.
PubMed Abstract | Free Full Text

Rodriguez F, Blum MR, Falasinnu T, *et al.*: **Diabetes-Attributable Mortality in the United States from 2003 to 2016 Using a Multiple-Cause-of-Death Approach.** *Diabetes Res Clin Pract.* 2019; **148**: 169–78.
PubMed Abstract | Publisher Full Text

StataCorp: **Stata Statistical Software: Release 14**. College Station, TX: StataCorp LP. 2015.

United Nations: **Handbook of Vital Statistics Systems and Methods.** Volume 1: Legal, Organizational and Technical Aspects, United Nations Studies in Methods, Glossary. 1991.
Reference Source

Weir DR: **Validating Mortality Ascertainment in the Health and Retirement Study**. 2016.
Reference Source

Whelan BJ, Savva GM: **Design and Methodology of the Irish Longitudinal Study on Ageing.** *J Am Geriatr Soc.* 2013; **61**(Suppl 2): S265–68.
PubMed Abstract | Publisher Full Text

White C, Edgar G, Siegler V: **Social Inequalities in Male Mortality for Selected Causes of Death by the National Statistics Socio-economic Classification, England and Wales, 2001-03.** *Health Stat Q.* 2008; **38**: 19–32.
PubMed Abstract

Wu C, Odden MC, Fisher GG, *et al.*: **Association of Retirement Age with Mortality: A Population-Based Longitudinal Study among Older Adults in the USA.** *J Epidemiol Community Health.* 2016; **70**(9): 917–23.
PubMed Abstract | Publisher Full Text | Free Full Text

Open Peer Review

Current Peer Review Status: ✔ ✔ ✔

Version 2

Reviewer Report 23 November 2020

https://doi.org/10.21956/hrbopenres.14315.r28409

✔ **Zubair Kabir**
School of Public Health, University College Cork, Cork, Ireland

The authors have adequately addressed all my previous concerns. Happy to approve this version. Well done!

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.

Reviewer Report 19 November 2020

https://doi.org/10.21956/hrbopenres.14315.r28410

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

No further comments.

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.

Version 1

Reviewer Report 31 July 2020

https://doi.org/10.21956/hrbopenres.14183.r27633

?

Zubair Kabir
School of Public Health, University College Cork, Cork, Ireland

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.

Author Response 06 Nov 2020

Mark Ward, Trinity College Dublin, Dublin, Ireland

Response to Reviewer 1 comments – Dan Lewer

Comment 1. 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.

Response 1. Thank you for taking the time to review our manuscript and providing insightful comments. Indeed, this is the first time that this data linkage exercise has been conducted in the Republic of Ireland and as such we hope that it will be a valuable resource for researchers who wish to better understand the antecedents of mortality among older adults.

Comment 2. 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).

Response 2. This data linkage exercise was the first step in a wider programme of research being conducted within TILDA. This research is funded by the Health Research Board (ILP-PHR-2017-022)
The project is titled “Do we die as we live? Age, socioeconomic status, healthcare utilisation and pathways to death in Ireland” and is led by Professor Rose Anne Kenny (PI, TCD) and Dr Anne Nolan (Lead applicant, ESRI).
Three broad research questions are being examined in this project:
1) How do patterns of all-cause, cause-specific and amenable mortality in the over 50s in Ireland vary across groups defined by socioeconomic status, co-existing conditions, and cause of death?
2) What are the possible mechanisms (e.g., underlying health conditions, differential health behaviours, accessibility of healthcare services, etc.) that underlie these patterns?
3) What are the determinants of healthcare utilisation and costs at the end of life among the over 50s in Ireland?

Comment 3. 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.

Response 3. Deaths among TILDA participants were identified through a number of sources. In many cases, spouses or other relatives of decedents contacted TILDA to inform the research team of the death. Other deaths were identified when interviewers visited the home of decedents to conduct subsequent waves of data collection. Also, where it was not possible to contact a participant, the TILDA data management team identified some deaths through searches of the obituary website dedicated to publishing death notices in Ireland, RIP.ie. Finally, in the remaining cases where the status of participants were not known, GRO records were interrogated in order to identify those who had died. We have now included text to reflect this in the 'data linkage' section on page 4.

Comment 4. Is it worth adding some information on the associations with successful linkage? (i.e. were certain types of participant less likely to be linked?).

Response 4. On reflection our referring to the 863 total deaths among TILDA participants has led to some confusion. The 779 death records that we successfully matched were all the deaths confirmed by us at the time we carried out data linkage. The remaining 84 (863 - 779) deaths occurred after we had requested the death records from the GRO. These included the 65 deaths noted in Table 1 that occurred between waves 4 and 5 of TILDA data collection. We fully expect that the death records of these individuals will be included in the next round of data linkage in 2021.
We have now included the following text where we describe Table 1: *"The 84 deaths not captured in this data linkage occurred after we completed the exercise and will be captured when we repeat data linkage in 2021".*

Comment 5. 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?

Response 5. Unfortunately we have no way of checking this. However, we are confident that the participants we have linked were correct. As described in the text we used a number of participant characteristics to ensure that we correctly identified individuals – "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." Furthermore, as discussed in response to comment 3, in many cases this information was confirmed by a family member prior to the linkage exercise. Of course, every care was taken to ensure the accuracy of the characteristics used to identify death records in the GRO files.
As also noted in the manuscript, Ireland does not have a unique health identifier which could have been used for the purpose of matching participant records, nor is there an automated notification of death available to use. The latter is the method used by a number of similar cohort studies to identify deaths among their participants.

Comment 6. 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?

Response 6. This is an excellent suggestion. Thank you.
This particular analysis was informed by similar work carried out using UK Biobank data by Batty et al. The aim of this research, and our aim also, was to assess the utility of cause of death data extracted from the underlying cause field versus any location on the death certificate. The estimates do also confirm a stronger association between smoking and respiratory causes of death compared to all-cause mortality which is re-assuring but was not our main aim in this analysis.
Our choice of smoking as a risk factor was, as you identify, because it is so well established. Smoking was also one of three risk factors included in the Batty et al. analysis. We have now included the following text in the manuscript to justify this analysis: *"We chose smoking to test our hypothesis that similar estimates would be derived from both underlying and contributory conditions as smoking is an established risk factor for mortality and it has been used for a similar purpose previously (Batty et al. 2019)."*
As suggested, we have also now included the estimates for all-cause mortality in Figure 3 and described these results more fully in the text describing that graph.
Batty GD, Gale CR, Kivimäki M, Bell S. Assessment of Relative Utility of Underlying vs Contributory Causes of Death. JAMA Netw Open. 2019 Jul 3;2(7):e198024. doi: 10.1001/jamanetworkopen.2019.8024. PMID: 31365105; PMCID: PMC6669894.

Comment 7. 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.

Response 7. An important purpose of this paper is to provide an overview of the linked mortality data available in TILDA. Indeed, an important deliverable of the mortality project discussed above is the development of a data infrastructure of linked mortality / survey data. We hope that this manuscript will be an important reference for researchers using this new data resource.
With this in mind, our intention in including the information in Table 3 was to provide a brief description of decedents within the TILDA sample. We did not intend to suggest associations as such. Indeed, as also described above, explicitly and rigorously testing these associations is a central aim of the project and a number of manuscripts are currently in development that do just that.
In an effort to make this clearer to readers we have now included the following text: *"For reference, the distribution of important socio-demographic characteristics of the full TILDA sample and those who have died over the course of the study are presented in Table 3."*

Comment 8. 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.

Response 8. Our understanding is that the y-axis hazard rates are in effect standardised as described in the text *"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."*
That said, we did try to find an alternative means of presenting this comparison as suggested by you. Unfortunately we were unable to create an informative and easily interpreted solution. One difficulty is the small number of deaths observed within years, or indeed age bands. For example, for suggestion b, this leads to massive CIs among older ages in particular.
Also, the approach we have taken is similar to that of Weir (2016) when validating mortality data for the TILDA sister study, the Health and Retirement Study. Our representation therefore aids comparability of the two studies. We do however appreciate these suggestions and hope to have greater success in our efforts to incorporate them when we repeat this exercise in 2021.
We have replaced 'older decedents' with 'older ages' in the offending sentence.

Comment 9. 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.
Response 9. This is an interesting suggestion. Thank you. TILDA intends to collect its 6th wave of data in 2021 and during that time we will repeat this data linkage exercise. We know that there have been a quite a number of deaths since we carried out this exercise and given the large numerator (count of deaths) this will result in, we will consider, as you suggest, trimming our survival time.

Response to Reviewer 2 comments – Peter Harteloh
Comment 1. 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.

Response 1. We wish to thank Dr Harteloh for his positive review of our manuscript. This is the first time that this data linkage exercise has been conducted in the Republic of Ireland and as such we hope that it will be a valuable resource for researchers who wish to better understand the antecedents of mortality among older adults.
As also discussed in response to Reviewer 1, this data linkage exercise was the first step in a wider programme of research being conducted within TILDA. This research is funded by the Health Research Board (ILP-PHR-2017-022)
The project is titled “Do we die as we live? Age, socioeconomic status, healthcare utilisation and pathways to death in Ireland”.

Comment 2. 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% instead of after “time”.

Response 2. This has been corrected.

Comment 3. 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, 2018 . 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.

Response 3. The poor performance of Iris in assigning an ICD-10 code to the conditions mentioned in the individual death records was largely due to the fact that the death records had not been cleaned prior to our receiving them. As these records were provided as strings, their quality / consistency was variable. As this was the first time we had used the Iris software, the generic data dictionary included with the software, failed to identify conditions with different spellings, random spaces, and other typographical errors.
One recurring example which we believe exemplifies this was the case of “ischemic” / “ischaemic”. The of-the-shelf dictionary in the software correctly identified the former but not the latter, which was in fact the more common spelling in the death certificates. As part

of our data processing we appended the in-built data dictionary with common variations of spellings and descriptors we encountered and as a result, Iris performed this task increasingly well as we progressed. We plan to conduct data matching again in 2021 and are confident that we will have a higher success rate in our next attempt to assign ICD-10 codes automatically to individual death records.
We confirm that we coded the string expressions on the death records according to volume 2 of the ICD-10 with no local deviations.
Once ICD-10 codes were inputted (either automatically or manually), Iris performed excellently when selecting an underlying cause of death using the decision tables described in the manuscript. Indeed, this is the function that attracted us to using software for this purpose as it removed the possibility of subjective, or coder variation in the assignation of underlying cause.

Comment 4. 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.

Response 4. TILDA has full ethical approval in place for all data collection waves and further gains informed consent from all participants prior to data collection. Ethical approval is approved by the Faculty of Health Sciences Research Ethics Committee, Trinity College Dublin. Participants are informed through the Participant Information Leaflet that their data is shared in a confidential manner as part of the TILDA study.
The TILDA Privacy Policy gives more detailed information about data linkage with the GRO. It is important to note also that GDPR and the Irish Health Research Regulations do not apply to the personal data of deceased individuals. For the situation where a participant may be lost to follow up and their status unknown, TILDA have been granted a consent declaration by the Health Research Consent Declaration Committee (HRCDC) to process their data for GRO Linkage. A HRCDC declaration is granted in a case where the public interest of doing the research significantly outweighs the need for explicit consent.
A data transfer agreement is signed between TCD and GRO which commits to protecting the confidentiality of data. Physical and technical safeguards are also in place.

Comment 5. 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)

Response 5. Our use of the term 'contributory' was informed by a study by Batty et al. who use the term to refer to "Other diseases or injuries that contributed to the death but were not directly implicated" (p.2).
We have now explained our use of the term 'contributory' and provided a reference to the

Batty et al. paper. *"A contributory cause of death is a condition that contributed to the death but were not directly implicated and are recorded in part two of death certificates. While this information has been rarely used in epidemiological research, recent evidence suggest it may have some methodological utility (Batty et al. 2019). For present purposes, contributory causes include diseases and conditions listed anywhere on the death certificate."*
Batty GD, Gale CR, Kivimäki M, Bell S. Assessment of Relative Utility of Underlying vs Contributory Causes of Death. JAMA Netw Open. 2019 Jul 3;2(7):e198024. doi: 10.1001/jamanetworkopen.2019.8024. PMID: 31365105; PMCID: PMC6669894.

Comment 6. 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.

Response 6. This valid point was also raised by another reviewer. In response, this particular analysis was informed by similar work carried out using UK Biobank data by Batty et al. The aim of this research, and our aim also, was to assess the utility of cause of death data extracted from the underlying cause field versus any location on the death certificate. Our choice of smoking as a risk factor was, as you identify, because it is so well established. Smoking was also one of three risk factors included in the Batty et al. analysis. We have now included the following text in the manuscript to justify this analysis: *"We chose smoking to test our hypothesis that similar estimates would be derived from both underlying and contributory conditions as smoking is an established risk factor for mortality and it has been used for a similar purpose previously (Batty et al. 2019)."* Batty GD, Gale CR, Kivimäki M, Bell S. Assessment of Relative Utility of Underlying vs Contributory Causes of Death. JAMA Netw Open. 2019 Jul 3;2(7):e198024. doi: 10.1001/jamanetworkopen.2019.8024. PMID: 31365105; PMCID: PMC6669894.

Comment 7. 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?

Response 7. We hope we have interpreted this question correctly, but we confirm that the figures refer to the proportion of death certificates that included any cause from the ICD-10 chapter of diseases of the circulatory system as a contributory cause of death (52.6 %) and any cause from the ICD-10 chapter of diseases of the circulatory system (34.4%).

Comment 8. 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.

Response 8. Yes. This is a count of records that included at least one malignancy per record. We hope that the additional text we have included in response to your comment 5 in defining our use of 'contributory' has made this clearer to readers.

Comment 9. 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.

Response 9. Lung cancer was indeed the most common type accounting for 19% of cancers. We note that the association between smoking and cancer death is positive, but non-significant due to wide 95% confidence bands. We also note that our smoking variable identifies 'ever' as well as 'current' smokers, so some of the smokers may have quit some time ago.

Comment 10. 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.

Response 10. These estimates (HRs with 95% CIs) are presented in Figure 3. In responses to another reviewers suggestion, we have now also included the estimates for all-cause mortality. We also now more fully describe the results presented in this figure. We hope that this fuller description also provides clearer support for our contention that choice of contributory or underlying cause may not make much difference to these estimates. This final point is more fully discussed in response to comment 11 below and comment 6 from Reviewer 1.

Comment 11. 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.

Response 11. As in our response to the previous comment, these estimates are presented in Figure 3 and the text describing these results has been extended.
Our contention that it appears that underlying and contributory cause of death may have similar utility for studies examining mortality risk factors is supported by the work discussed above by Batty et al. (2019) and a smaller scale study by Crews et al. (1991). We have now referenced both of these studies in support of the contention we made here.
We are also going to repeat the data linkage exercise in 2021 when TILDA will conduct its 6 th wave of data collection. The increased number of deaths will provide us with an appropriately large sample size to examine the association of major risk factors and specific causes of death. Initial results from this work are anticipated in late 2021.

Comment 12. 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.

Response 12. We have again checked this and can confirm that the Iris data dictionary does

not identify “ischaemic heart disease”, only “ischemic heart disease”. The reason we chose to refer to this example was because it occurred so often.
As part of our data processing we appended the in-built data dictionary with common variations of spellings and descriptors we encountered and as a result, Iris performed this task increasingly well as we progressed. We plan to conduct data matching again in 2021 and are confident that we will have a higher success rate in our next attempt to assign ICD-10 codes automatically to individual death records.

Comment 13. 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.

Response 13. Thank you. We are glad that you agree with the importance of this exercise. As described above, we hope that project that this work stems from will make an important contribution to research on mortality in Ireland. We also hope that this particular data linkage demonstrates the great potential of combining rich individual level survey data with administrative data sources. Unfortunately, to date Ireland somewhat lags behind other jurisdictions who have well developed data linkage infrastructures.

Comment 14. 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.

Response 14. Again, we wish to thank Dr Harteloh for his constructive feedback. We believe that the revisions have greatly improved the manuscript and provided clarification as to the meaning of contributory cause in this context. As discussed above, this is the first of many publications from this work. If interested, we have recently published another methodological paper using this data which compares the utility of cause of death data from official records and reports from end-of-life interviews. Ward, M, May, P, Normand, C, Kenny, RA, and Nolan, A. Comparing Underlying and Contributory Cause of Death in Registry Data With End-of-Life Proxy Interviews: Findings From The Irish Longitudinal Study on Ageing (TILDA). Journal of Applied Gerontology. [In Press].
https://doi.org/10.1177/0733464820935295

Response to Reviewer 3 comments – Dr Zubair Kabir
Comment 1. 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.

Response 1. Thank you Dr Kabir for taking the time to review our manuscript and for your helpful observations. As you rightly say, this type of exercise is challenging within the Irish data infrastructure and we do hope that our efforts contribute to improving this situation.

Comment 2. 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.

Response 2. We have done our best to describe as fully as possible the steps we took to achieve this data linkage. We hope that our responses to yours' and other reviewers suggestions have further improved this.
Naturally, many of our decisions and subsequent actions are specific to the data environment in which the work was conducted. By this, we mean that we were confined to the data that was available to use in TILDA, for example, the individual identifiers and so on. As such, it may well not be possible to replicate our procedures with other studies in Ireland. However, we feel strongly that we have been fully transparent and as specific as possible in our description of the steps we have taken to link the individual-level survey data available in TILDA to official death records. Indeed, given the richness of the data available to us in TILDA, we have many advantages not necessarily available to other studies.
As you correctly state, there are no standardised quality assurance tools available to use to assess the validity of our data linkage procedures and it was partly due to the absence of such a tool that we felt compelled to describe our methods as fully as possible and importantly to make this manuscript freely available to all.

Also importantly, our intention with this manuscript was not to suggest a one-size fits all method but rather to describe a new data infrastructure within TILDA that researchers interested in studying mortality in Ireland might avail of. How a similar task might be approached using a different study sample will be study dependent. That said, we do believe that our use of the Iris software tool for coding and identifying underlying cause of death is one way in which our work might be replicated and could help ensure standardisation in at least this aspect of the linkage across studies.

Completeness of source databases

As TILDA is prospective cohort study we are confident of the accuracy of the participant contact information and status as participants are contacted regularly and the status of non-responders is followed up via the participants or their proxies. The contact database is regularly updated so that participants can be contacted for future rounds of data collection. The GRO is the official register of all deaths in Ireland and provides information on deaths to the CSO for use in official statistics. As such, we are confident that it is a reliable and comprehensive source of data on deaths in Ireland.

Ethical and data security considerations

TILDA has full ethical approval in place for all data collection waves and further gains informed consent from all participants prior to data collection. Ethical approval is approved by the Faculty of Health Sciences REC, Trinity College Dublin. Participants are informed through the Participant Information Leaflet that their data is shared in a confidential manner as part of the TILDA study.

It is important to note also that GDPR and the Irish Health Research Regulations do not apply to the personal data of deceased individuals. For the situation where a participant may be lost to follow up and their status unknown, TILDA have been granted a consent declaration by the Health Research Consent Declaration Committee to process their data for GRO Linkage. A HRCDC declaration is granted in a case where the public interest of doing the research significantly outweighs the need for explicit consent.

A data transfer agreement is signed between TCD and GRO which commits to protecting the confidentiality of data. Physical and technical safeguards are also in place.

Linkage methodology and technology

Our stating that data matching was conducted by the CSO in 2013 was in error and has now been removed from the manuscript. The only time data matching took place was in 2018 with the GRO.

The TILDA data team did undertake the data matching through the GRO search room facility. Once the TILDA team member identified the decedent within these records, the GRO then provided the detailed death certificate information for this person.

We have provided further clarification to these points in response to earlier comments. We have also appended our description of these measures within the manuscript and hope that they adequately address each of the points raised here.

You may also be interested to know that the CSO have repeated their 2013 data linkage using 2016 census data. You will find the results here: CSO: Mortality Differentials in Ireland. An Analysis Based on the Census Characteristics of Persons Who Died in the Twelve Month Period after Census Day 24 April 2016. 2019. Dublin. Source: https://www.cso.ie/en/releasesandpublications/in/mdi/mortalitydifferentialsinireland2016-2017/ [Accessed: October 2020].

Comment 3. 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.

Response 3. We hope we have clarified that the full data linkage exercise was conducted by the TILDA team. In practice the GROs sole involvement was to provide the team with the death certificate information of decedents identified among TILDA participants.
In light of these, we believe we have made three contributions here. (1) We performed the data linkage, (2) provided an overview of a new data infrastructure and, (3) provided an assessment of the utility of contributory versus underlying cause in estimating the association between risk factors and mortality risk.
As also detailed in response to Reviewers 1 and 2 above, in this amended version of the manuscript we have better described our use of the term 'contributory' as: *"A contributory cause of death is a condition that contributed to the death but were not directly implicated and are recorded in part two of death certificates. While this information has been rarely used in epidemiological research, recent evidence suggest it may have some methodological utility (Batty et al. 2019). For present purposes, contributory causes include diseases and conditions listed anywhere on the death certificate."*
Batty GD, Gale CR, Kivimäki M, Bell S. Assessment of Relative Utility of Underlying vs Contributory Causes of Death. JAMA Netw Open. 2019 Jul 3;2(7):e198024. doi: 10.1001/jamanetworkopen.2019.8024. PMID: 31365105; PMCID: PMC6669894.
Also to re-state an earlier response, this particular analysis was informed by similar work carried out using UK Biobank data by Batty et al. The aim of this research, and our aim also, was to assess the utility of cause of death data extracted from the underlying cause field versus any location on the death certificate. The estimates do also confirm a stronger association between smoking and respiratory causes of death compared to all-cause mortality which is re-assuring but was not our main aim in this analysis.
Our choice of smoking as a risk factor was, as you identify, because it is so well established. Smoking was also one of three risk factors included in the Batty et al. analysis. We have now included the following text in the manuscript to justify this analysis: *"We chose smoking to test our hypothesis that similar estimates would be derived from both underlying and contributory conditions as smoking is an established risk factor for mortality and it has been used for a similar purpose previously (Batty et al. 2019)."* Again, we sincerely thank Dr Kabir for his insightful comments and appreciate his sharing his vast experience in this area with us.

Competing Interests: No competing interests were disclosed.

Reviewer Report 22 July 2020

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

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, 2018[1]. 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

1. Harteloh P: The implementation of an automated coding system for cause-of-death statistics. *Inform Health Soc Care*. 2020; **45** (1): 1-14 PubMed Abstract | Publisher Full Text

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.

Author Response 06 Nov 2020

Mark Ward, Trinity College Dublin, Dublin, Ireland

Response to Reviewer 1 comments – Dan Lewer

Comment 1. 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.

Response 1. Thank you for taking the time to review our manuscript and providing insightful comments. Indeed, this is the first time that this data linkage exercise has been conducted in the Republic of Ireland and as such we hope that it will be a valuable resource for researchers who wish to better understand the antecedents of mortality among older adults.

Comment 2. 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).

Response 2. This data linkage exercise was the first step in a wider programme of research being conducted within TILDA. This research is funded by the Health Research Board (ILP-PHR-2017-022)
The project is titled “Do we die as we live? Age, socioeconomic status, healthcare utilisation and pathways to death in Ireland” and is led by Professor Rose Anne Kenny (PI, TCD) and Dr Anne Nolan (Lead applicant, ESRI).
Three broad research questions are being examined in this project:
1) How do patterns of all-cause, cause-specific and amenable mortality in the over 50s in Ireland vary across groups defined by socioeconomic status, co-existing conditions, and cause of death?
2) What are the possible mechanisms (e.g., underlying health conditions, differential health behaviours, accessibility of healthcare services, etc.) that underlie these patterns?
3) What are the determinants of healthcare utilisation and costs at the end of life among the over 50s in Ireland?

Comment 3. 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.

Response 3. Deaths among TILDA participants were identified through a number of sources. In many cases, spouses or other relatives of decedents contacted TILDA to inform the research team of the death. Other deaths were identified when interviewers visited the home of decedents to conduct subsequent waves of data collection. Also, where it was not possible to contact a participant, the TILDA data management team identified some deaths through searches of the obituary website dedicated to publishing death notices in Ireland, RIP.ie. Finally, in the remaining cases where the status of participants were not known, GRO records were interrogated in order to identify those who had died. We have now included text to reflect this in the ‘data linkage’ section on page 4.

Comment 4. Is it worth adding some information on the associations with successful linkage? (i.e. were certain types of participant less likely to be linked?).

Response 4. On reflection our referring to the 863 total deaths among TILDA participants has led to some confusion. The 779 death records that we successfully matched were all the deaths confirmed by us at the time we carried out data linkage. The remaining 84 (863 - 779) deaths occurred after we had requested the death records from the GRO. These included the 65 deaths noted in Table 1 that occurred between waves 4 and 5 of TILDA data collection. We fully expect that the death records of these individuals will be included in the next round of data linkage in 2021.

We have now included the following text where we describe Table 1: *"The 84 deaths not captured in this data linkage occurred after we completed the exercise and will be captured when we repeat data linkage in 2021".*

Comment 5. 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?

Response 5. Unfortunately we have no way of checking this. However, we are confident that the participants we have linked were correct. As described in the text we used a number of participant characteristics to ensure that we correctly identified individuals – "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." Furthermore, as discussed in response to comment 3, in many cases this information was confirmed by a family member prior to the linkage exercise. Of course, every care was taken to ensure the accuracy of the characteristics used to identify death records in the GRO files.
As also noted in the manuscript, Ireland does not have a unique health identifier which could have been used for the purpose of matching participant records, nor is there an automated notification of death available to use. The latter is the method used by a number of similar cohort studies to identify deaths among their participants.

Comment 6. 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?

Response 6. This is an excellent suggestion. Thank you.
This particular analysis was informed by similar work carried out using UK Biobank data by Batty et al. The aim of this research, and our aim also, was to assess the utility of cause of death data extracted from the underlying cause field versus any location on the death certificate. The estimates do also confirm a stronger association between smoking and respiratory causes of death compared to all-cause mortality which is re-assuring but was not our main aim in this analysis.
Our choice of smoking as a risk factor was, as you identify, because it is so well established. Smoking was also one of three risk factors included in the Batty et al. analysis. We have now included the following text in the manuscript to justify this analysis: *"We chose smoking to test our hypothesis that similar estimates would be derived from both underlying and contributory conditions as smoking is an established risk factor for mortality and it has been used for a similar purpose previously (Batty et al. 2019)."*
As suggested, we have also now included the estimates for all-cause mortality in Figure 3 and described these results more fully in the text describing that graph.
Batty GD, Gale CR, Kivimäki M, Bell S. Assessment of Relative Utility of Underlying vs

Contributory Causes of Death. JAMA Netw Open. 2019 Jul 3;2(7):e198024. doi: 10.1001/jamanetworkopen.2019.8024. PMID: 31365105; PMCID: PMC6669894.

Comment 7. 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.

Response 7. An important purpose of this paper is to provide an overview of the linked mortality data available in TILDA. Indeed, an important deliverable of the mortality project discussed above is the development of a data infrastructure of linked mortality / survey data. We hope that this manuscript will be an important reference for researchers using this new data resource.
With this in mind, our intention in including the information in Table 3 was to provide a brief description of decedents within the TILDA sample. We did not intend to suggest associations as such. Indeed, as also described above, explicitly and rigorously testing these associations is a central aim of the project and a number of manuscripts are currently in development that do just that.
In an effort to make this clearer to readers we have now included the following text: *"For reference, the distribution of important socio-demographic characteristics of the full TILDA sample and those who have died over the course of the study are presented in Table 3."*

Comment 8. 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.

Response 8. Our understanding is that the y-axis hazard rates are in effect standardised as described in the text *"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."*
That said, we did try to find an alternative means of presenting this comparison as suggested by you. Unfortunately we were unable to create an informative and easily interpreted solution. One difficulty is the small number of deaths observed within years, or indeed age bands. For example, for suggestion b, this leads to massive CIs among older ages in particular.

Also, the approach we have taken is similar to that of Weir (2016) when validating mortality data for the TILDA sister study, the Health and Retirement Study. Our representation therefore aids comparability of the two studies. We do however appreciate these suggestions and hope to have greater success in our efforts to incorporate them when we repeat this exercise in 2021.
We have replaced 'older decedents' with 'older ages' in the offending sentence.

Comment 9. 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.
Response 9. This is an interesting suggestion. Thank you. TILDA intends to collect its 6th wave of data in 2021 and during that time we will repeat this data linkage exercise. We know that there have been a quite a number of deaths since we carried out this exercise and given the large numerator (count of deaths) this will result in, we will consider, as you suggest, trimming our survival time.

Response to Reviewer 2 comments – Peter Harteloh
Comment 1. 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.

Response 1. We wish to thank Dr Harteloh for his positive review of our manuscript. This is the first time that this data linkage exercise has been conducted in the Republic of Ireland and as such we hope that it will be a valuable resource for researchers who wish to better understand the antecedents of mortality among older adults.
As also discussed in response to Reviewer 1, this data linkage exercise was the first step in a wider programme of research being conducted within TILDA. This research is funded by the Health Research Board (ILP-PHR-2017-022)
The project is titled "Do we die as we live? Age, socioeconomic status, healthcare utilisation and pathways to death in Ireland".

Comment 2. 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% instead of after "time".

Response 2. This has been corrected.

Comment 3. 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, 2018 . 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.

Response 3. The poor performance of Iris in assigning an ICD-10 code to the conditions mentioned in the individual death records was largely due to the fact that the death records had not been cleaned prior to our receiving them. As these records were provided as strings, their quality / consistency was variable. As this was the first time we had used the Iris software, the generic data dictionary included with the software, failed to identify conditions with different spellings, random spaces, and other typographical errors.
One recurring example which we believe exemplifies this was the case of "ischemic" / "ischaemic". The of-the-shelf dictionary in the software correctly identified the former but not the latter, which was in fact the more common spelling in the death certificates. As part of our data processing we appended the in-built data dictionary with common variations of spellings and descriptors we encountered and as a result, Iris performed this task increasingly well as we progressed. We plan to conduct data matching again in 2021 and are confident that we will have a higher success rate in our next attempt to assign ICD-10 codes automatically to individual death records.
We confirm that we coded the string expressions on the death records according to volume 2 of the ICD-10 with no local deviations.
Once ICD-10 codes were inputted (either automatically or manually), Iris performed excellently when selecting an underlying cause of death using the decision tables described in the manuscript. Indeed, this is the function that attracted us to using software for this purpose as it removed the possibility of subjective, or coder variation in the assignation of underlying cause.

Comment 4. 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.

Response 4. TILDA has full ethical approval in place for all data collection waves and further gains informed consent from all participants prior to data collection. Ethical approval is

approved by the Faculty of Health Sciences Research Ethics Committee, Trinity College Dublin. Participants are informed through the Participant Information Leaflet that their data is shared in a confidential manner as part of the TILDA study.
The TILDA Privacy Policy gives more detailed information about data linkage with the GRO. It is important to note also that GDPR and the Irish Health Research Regulations do not apply to the personal data of deceased individuals. For the situation where a participant may be lost to follow up and their status unknown, TILDA have been granted a consent declaration by the Health Research Consent Declaration Committee (HRCDC) to process their data for GRO Linkage. A HRCDC declaration is granted in a case where the public interest of doing the research significantly outweighs the need for explicit consent.
A data transfer agreement is signed between TCD and GRO which commits to protecting the confidentiality of data. Physical and technical safeguards are also in place.

Comment 5. 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)

Response 5. Our use of the term 'contributory' was informed by a study by Batty et al. who use the term to refer to "Other diseases or injuries that contributed to the death but were not directly implicated" (p.2).
We have now explained our use of the term 'contributory' and provided a reference to the Batty et al. paper. *"A contributory cause of death is a condition that contributed to the death but were not directly implicated and are recorded in part two of death certificates. While this information has been rarely used in epidemiological research, recent evidence suggest it may have some methodological utility (Batty et al. 2019). For present purposes, contributory causes include diseases and conditions listed anywhere on the death certificate."*
Batty GD, Gale CR, Kivimäki M, Bell S. Assessment of Relative Utility of Underlying vs Contributory Causes of Death. JAMA Netw Open. 2019 Jul 3;2(7):e198024. doi: 10.1001/jamanetworkopen.2019.8024. PMID: 31365105; PMCID: PMC6669894.

Comment 6. 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.

Response 6. This valid point was also raised by another reviewer. In response, this particular analysis was informed by similar work carried out using UK Biobank data by Batty et al. The aim of this research, and our aim also, was to assess the utility of cause of death data extracted from the underlying cause field versus any location on the death certificate. Our choice of smoking as a risk factor was, as you identify, because it is so well established. Smoking was also one of three risk factors included in the Batty et al. analysis. We have now included the following text in the manuscript to justify this analysis: *"We chose smoking to test our hypothesis that similar estimates would be derived from both underlying and contributory conditions as smoking is an established risk factor for mortality and it has been*

used for a similar purpose previously (Batty et al. 2019).” Batty GD, Gale CR, Kivimäki M, Bell S. Assessment of Relative Utility of Underlying vs Contributory Causes of Death. JAMA Netw Open. 2019 Jul 3;2(7):e198024. doi: 10.1001/jamanetworkopen.2019.8024. PMID: 31365105; PMCID: PMC6669894.

Comment 7. 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?

Response 7. We hope we have interpreted this question correctly, but we confirm that the figures refer to the proportion of death certificates that included any cause from the ICD-10 chapter of diseases of the circulatory system as a contributory cause of death (52.6 %) and any cause from the ICD-10 chapter of diseases of the circulatory system (34.4%).

Comment 8. 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.

Response 8. Yes. This is a count of records that included at least one malignancy per record. We hope that the additional text we have included in response to your comment 5 in defining our use of ‘contributory’ has made this clearer to readers.

Comment 9. 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.

Response 9. Lung cancer was indeed the most common type accounting for 19% of cancers. We note that the association between smoking and cancer death is positive, but non-significant due to wide 95% confidence bands. We also note that our smoking variable identifies ‘ever’ as well as ‘current’ smokers, so some of the smokers may have quit some time ago.

Comment 10. 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.

Response 10. These estimates (HRs with 95% CIs) are presented in Figure 3. In responses to another reviewers suggestion, we have now also included the estimates for all-cause mortality. We also now more fully describe the results presented in this figure. We hope that this fuller description also provides clearer support for our contention that choice of contributory or underlying cause may not make much difference to these estimates. This final point is more fully discussed in response to comment 11 below and comment 6 from Reviewer 1.

Comment 11. 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.

Response 11. As in our response to the previous comment, these estimates are presented in Figure 3 and the text describing these results has been extended.
Our contention that it appears that underlying and contributory cause of death may have similar utility for studies examining mortality risk factors is supported by the work discussed above by Batty et al. (2019) and a smaller scale study by Crews et al. (1991). We have now referenced both of these studies in support of the contention we made here.
We are also going to repeat the data linkage exercise in 2021 when TILDA will conduct its 6 th wave of data collection. The increased number of deaths will provide us with an appropriately large sample size to examine the association of major risk factors and specific causes of death. Initial results from this work are anticipated in late 2021.

Comment 12. 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.

Response 12. We have again checked this and can confirm that the Iris data dictionary does not identify "ischaemic heart disease", only "ischemic heart disease". The reason we chose to refer to this example was because it occurred so often.
As part of our data processing we appended the in-built data dictionary with common variations of spellings and descriptors we encountered and as a result, Iris performed this task increasingly well as we progressed. We plan to conduct data matching again in 2021 and are confident that we will have a higher success rate in our next attempt to assign ICD-10 codes automatically to individual death records.

Comment 13. 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.

Response 13. Thank you. We are glad that you agree with the importance of this exercise. As described above, we hope that project that this work stems from will make an important contribution to research on mortality in Ireland. We also hope that this particular data linkage demonstrates the great potential of combining rich individual level survey data with administrative data sources. Unfortunately, to date Ireland somewhat lags behind other jurisdictions who have well developed data linkage infrastructures.

Comment 14. 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.

Response 14. Again, we wish to thank Dr Harteloh for his constructive feedback. We believe that the revisions have greatly improved the manuscript and provided clarification as to the meaning of contributory cause in this context. As discussed above, this is the first of many publications from this work. If interested, we have recently published another methodological paper using this data which compares the utility of cause of death data from official records and reports from end-of-life interviews. Ward, M, May, P, Normand, C, Kenny, RA, and Nolan, A. Comparing Underlying and Contributory Cause of Death in Registry Data With End-of-Life Proxy Interviews: Findings From The Irish Longitudinal Study on Ageing (TILDA). Journal of Applied Gerontology. [In Press].
https://doi.org/10.1177/0733464820935295

Response to Reviewer 3 comments – Dr Zubair Kabir
Comment 1. 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.

Response 1. Thank you Dr Kabir for taking the time to review our manuscript and for your helpful observations. As you rightly say, this type of exercise is challenging within the Irish data infrastructure and we do hope that our efforts contribute to improving this situation.

Comment 2. 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.

Response 2. We have done our best to describe as fully as possible the steps we took to achieve this data linkage. We hope that our responses to yours' and other reviewers suggestions have further improved this.
Naturally, many of our decisions and subsequent actions are specific to the data environment in which the work was conducted. By this, we mean that we were confined to the data that was available to use in TILDA, for example, the individual identifiers and so on. As such, it may well not be possible to replicate our procedures with other studies in Ireland. However, we feel strongly that we have been fully transparent and as specific as possible in our description of the steps we have taken to link the individual-level survey data available in TILDA to official death records. Indeed, given the richness of the data available to us in TILDA, we have many advantages not necessarily available to other studies.
As you correctly state, there are no standardised quality assurance tools available to use to assess the validity of our data linkage procedures and it was partly due to the absence of such a tool that we felt compelled to describe our methods as fully as possible and importantly to make this manuscript freely available to all.
Also importantly, our intention with this manuscript was not to suggest a one-size fits all method but rather to describe a new data infrastructure within TILDA that researchers interested in studying mortality in Ireland might avail of. How a similar task might be approached using a different study sample will be study dependent. That said, we do believe that our use of the Iris software tool for coding and identifying underlying cause of death is one way in which our work might be replicated and could help ensure standardisation in at least this aspect of the linkage across studies.
Completeness of source databases
As TILDA is prospective cohort study we are confident of the accuracy of the participant contact information and status as participants are contacted regularly and the status of non-responders is followed up via the participants or their proxies. The contact database is regularly updated so that participants can be contacted for future rounds of data collection. The GRO is the official register of all deaths in Ireland and provides information on deaths to the CSO for use in official statistics. As such, we are confident that it is a reliable and comprehensive source of data on deaths in Ireland.
Ethical and data security considerations
TILDA has full ethical approval in place for all data collection waves and further gains informed consent from all participants prior to data collection. Ethical approval is approved by the Faculty of Health Sciences REC, Trinity College Dublin. Participants are informed through the Participant Information Leaflet that their data is shared in a confidential manner as part of the TILDA study.
It is important to note also that GDPR and the Irish Health Research Regulations do not

apply to the personal data of deceased individuals. For the situation where a participant may be lost to follow up and their status unknown, TILDA have been granted a consent declaration by the Health Research Consent Declaration Committee to process their data for GRO Linkage. A HRCDC declaration is granted in a case where the public interest of doing the research significantly outweighs the need for explicit consent.
A data transfer agreement is signed between TCD and GRO which commits to protecting the confidentiality of data. Physical and technical safeguards are also in place.
Linkage methodology and technology
Our stating that data matching was conducted by the CSO in 2013 was in error and has now been removed from the manuscript. The only time data matching took place was in 2018 with the GRO.
The TILDA data team did undertake the data matching through the GRO search room facility. Once the TILDA team member identified the decedent within these records, the GRO then provided the detailed death certificate information for this person.
We have provided further clarification to these points in response to earlier comments. We have also appended our description of these measures within the manuscript and hope that they adequately address each of the points raised here.
You may also be interested to know that the CSO have repeated their 2013 data linkage using 2016 census data. You will find the results here: CSO: Mortality Differentials in Ireland. An Analysis Based on the Census Characteristics of Persons Who Died in the Twelve Month Period after Census Day 24 April 2016. 2019. Dublin. Source: https://www.cso.ie/en/releasesandpublications/in/mdi/mortalitydifferentialsinireland2016-2017/ [Accessed: October 2020].

Comment 3. 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.

Response 3. We hope we have clarified that the full data linkage exercise was conducted by the TILDA team. In practice the GROs sole involvement was to provide the team with the death certificate information of decedents identified among TILDA participants.
In light of these, we believe we have made three contributions here. (1) We performed the data linkage, (2) provided an overview of a new data infrastructure and, (3) provided an assessment of the utility of contributory versus underlying cause in estimating the association between risk factors and mortality risk.

As also detailed in response to Reviewers 1 and 2 above, in this amended version of the manuscript we have better described our use of the term 'contributory' as: *"A contributory cause of death is a condition that contributed to the death but were not directly implicated and are recorded in part two of death certificates. While this information has been rarely used in epidemiological research, recent evidence suggest it may have some methodological utility (Batty et al. 2019). For present purposes, contributory causes include diseases and conditions listed anywhere on the death certificate."*
Batty GD, Gale CR, Kivimäki M, Bell S. Assessment of Relative Utility of Underlying vs Contributory Causes of Death. JAMA Netw Open. 2019 Jul 3;2(7):e198024. doi: 10.1001/jamanetworkopen.2019.8024. PMID: 31365105; PMCID: PMC6669894.
Also to re-state an earlier response, this particular analysis was informed by similar work carried out using UK Biobank data by Batty et al. The aim of this research, and our aim also, was to assess the utility of cause of death data extracted from the underlying cause field versus any location on the death certificate. The estimates do also confirm a stronger association between smoking and respiratory causes of death compared to all-cause mortality which is re-assuring but was not our main aim in this analysis.
Our choice of smoking as a risk factor was, as you identify, because it is so well established. Smoking was also one of three risk factors included in the Batty et al. analysis. We have now included the following text in the manuscript to justify this analysis: *"We chose smoking to test our hypothesis that similar estimates would be derived from both underlying and contributory conditions as smoking is an established risk factor for mortality and it has been used for a similar purpose previously (Batty et al. 2019)."* Again, we sincerely thank Dr Kabir for his insightful comments and appreciate his sharing his vast experience in this area with us.

Competing Interests: No competing interests were disclosed.

Reviewer Report 13 July 2020

https://doi.org/10.21956/hrbopenres.14183.r27635

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.

Author Response 06 Nov 2020

Mark Ward, Trinity College Dublin, Dublin, Ireland

Response to Reviewer 1 comments – Dan Lewer
Comment 1. 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.

Response 1. Thank you for taking the time to review our manuscript and providing

insightful comments. Indeed, this is the first time that this data linkage exercise has been conducted in the Republic of Ireland and as such we hope that it will be a valuable resource for researchers who wish to better understand the antecedents of mortality among older adults.

Comment 2. 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).

Response 2. This data linkage exercise was the first step in a wider programme of research being conducted within TILDA. This research is funded by the Health Research Board (ILP-PHR-2017-022)
The project is titled “Do we die as we live? Age, socioeconomic status, healthcare utilisation and pathways to death in Ireland” and is led by Professor Rose Anne Kenny (PI, TCD) and Dr Anne Nolan (Lead applicant, ESRI).
Three broad research questions are being examined in this project:
1) How do patterns of all-cause, cause-specific and amenable mortality in the over 50s in Ireland vary across groups defined by socioeconomic status, co-existing conditions, and cause of death?
2) What are the possible mechanisms (e.g., underlying health conditions, differential health behaviours, accessibility of healthcare services, etc.) that underlie these patterns?
3) What are the determinants of healthcare utilisation and costs at the end of life among the over 50s in Ireland?

Comment 3. 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.

Response 3. Deaths among TILDA participants were identified through a number of sources. In many cases, spouses or other relatives of decedents contacted TILDA to inform the research team of the death. Other deaths were identified when interviewers visited the home of decedents to conduct subsequent waves of data collection. Also, where it was not possible to contact a participant, the TILDA data management team identified some deaths through searches of the obituary website dedicated to publishing death notices in Ireland, RIP.ie. Finally, in the remaining cases where the status of participants were not known, GRO records were interrogated in order to identify those who had died. We have now included text to reflect this in the ‘data linkage’ section on page 4.

Comment 4. Is it worth adding some information on the associations with successful linkage? (i.e. were certain types of participant less likely to be linked?).

Response 4. On reflection our referring to the 863 total deaths among TILDA participants has led to some confusion. The 779 death records that we successfully matched were all the

deaths confirmed by us at the time we carried out data linkage. The remaining 84 (863 - 779) deaths occurred after we had requested the death records from the GRO. These included the 65 deaths noted in Table 1 that occurred between waves 4 and 5 of TILDA data collection. We fully expect that the death records of these individuals will be included in the next round of data linkage in 2021.
We have now included the following text where we describe Table 1: *"The 84 deaths not captured in this data linkage occurred after we completed the exercise and will be captured when we repeat data linkage in 2021".*

Comment 5. 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?

Response 5. Unfortunately we have no way of checking this. However, we are confident that the participants we have linked were correct. As described in the text we used a number of participant characteristics to ensure that we correctly identified individuals – "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." Furthermore, as discussed in response to comment 3, in many cases this information was confirmed by a family member prior to the linkage exercise. Of course, every care was taken to ensure the accuracy of the characteristics used to identify death records in the GRO files.
As also noted in the manuscript, Ireland does not have a unique health identifier which could have been used for the purpose of matching participant records, nor is there an automated notification of death available to use. The latter is the method used by a number of similar cohort studies to identify deaths among their participants.

Comment 6. 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?

Response 6. This is an excellent suggestion. Thank you.
This particular analysis was informed by similar work carried out using UK Biobank data by Batty et al. The aim of this research, and our aim also, was to assess the utility of cause of death data extracted from the underlying cause field versus any location on the death certificate. The estimates do also confirm a stronger association between smoking and respiratory causes of death compared to all-cause mortality which is re-assuring but was not our main aim in this analysis.
Our choice of smoking as a risk factor was, as you identify, because it is so well established. Smoking was also one of three risk factors included in the Batty et al. analysis. We have now included the following text in the manuscript to justify this analysis: *"We chose smoking to test our hypothesis that similar estimates would be derived from both underlying and*

contributory conditions as smoking is an established risk factor for mortality and it has been used for a similar purpose previously (Batty et al. 2019).”
As suggested, we have also now included the estimates for all-cause mortality in Figure 3 and described these results more fully in the text describing that graph.
Batty GD, Gale CR, Kivimäki M, Bell S. Assessment of Relative Utility of Underlying vs Contributory Causes of Death. JAMA Netw Open. 2019 Jul 3;2(7):e198024. doi: 10.1001/jamanetworkopen.2019.8024. PMID: 31365105; PMCID: PMC6669894.

Comment 7. 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.

Response 7. An important purpose of this paper is to provide an overview of the linked mortality data available in TILDA. Indeed, an important deliverable of the mortality project discussed above is the development of a data infrastructure of linked mortality / survey data. We hope that this manuscript will be an important reference for researchers using this new data resource.
With this in mind, our intention in including the information in Table 3 was to provide a brief description of decedents within the TILDA sample. We did not intend to suggest associations as such. Indeed, as also described above, explicitly and rigorously testing these associations is a central aim of the project and a number of manuscripts are currently in development that do just that.
In an effort to make this clearer to readers we have now included the following text: *"For reference, the distribution of important socio-demographic characteristics of the full TILDA sample and those who have died over the course of the study are presented in Table 3."*

Comment 8. 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.

Response 8. Our understanding is that the y-axis hazard rates are in effect standardised as described in the text *“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.”*

That said, we did try to find an alternative means of presenting this comparison as suggested by you. Unfortunately we were unable to create an informative and easily interpreted solution. One difficulty is the small number of deaths observed within years, or indeed age bands. For example, for suggestion b, this leads to massive CIs among older ages in particular.
Also, the approach we have taken is similar to that of Weir (2016) when validating mortality data for the TILDA sister study, the Health and Retirement Study. Our representation therefore aids comparability of the two studies. We do however appreciate these suggestions and hope to have greater success in our efforts to incorporate them when we repeat this exercise in 2021.
We have replaced 'older decedents' with 'older ages' in the offending sentence.

Comment 9. 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.
Response 9. This is an interesting suggestion. Thank you. TILDA intends to collect its 6th wave of data in 2021 and during that time we will repeat this data linkage exercise. We know that there have been a quite a number of deaths since we carried out this exercise and given the large numerator (count of deaths) this will result in, we will consider, as you suggest, trimming our survival time.

Response to Reviewer 2 comments – Peter Harteloh
Comment 1. 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.

Response 1. We wish to thank Dr Harteloh for his positive review of our manuscript. This is the first time that this data linkage exercise has been conducted in the Republic of Ireland and as such we hope that it will be a valuable resource for researchers who wish to better understand the antecedents of mortality among older adults.
As also discussed in response to Reviewer 1, this data linkage exercise was the first step in a wider programme of research being conducted within TILDA. This research is funded by the Health Research Board (ILP-PHR-2017-022)
The project is titled "Do we die as we live? Age, socioeconomic status, healthcare utilisation and pathways to death in Ireland".

Comment 2. 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% instead of after “time”.

Response 2. This has been corrected.

Comment 3. 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, 2018 . 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.

Response 3. The poor performance of Iris in assigning an ICD-10 code to the conditions mentioned in the individual death records was largely due to the fact that the death records had not been cleaned prior to our receiving them. As these records were provided as strings, their quality / consistency was variable. As this was the first time we had used the Iris software, the generic data dictionary included with the software, failed to identify conditions with different spellings, random spaces, and other typographical errors.
One recurring example which we believe exemplifies this was the case of “ischemic” / “ischaemic”. The of-the-shelf dictionary in the software correctly identified the former but not the latter, which was in fact the more common spelling in the death certificates. As part of our data processing we appended the in-built data dictionary with common variations of spellings and descriptors we encountered and as a result, Iris performed this task increasingly well as we progressed. We plan to conduct data matching again in 2021 and are confident that we will have a higher success rate in our next attempt to assign ICD-10 codes automatically to individual death records.
We confirm that we coded the string expressions on the death records according to volume 2 of the ICD-10 with no local deviations.
Once ICD-10 codes were inputted (either automatically or manually), Iris performed excellently when selecting an underlying cause of death using the decision tables described in the manuscript. Indeed, this is the function that attracted us to using software for this purpose as it removed the possibility of subjective, or coder variation in the assignation of underlying cause.

Comment 4. 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.

Response 4. TILDA has full ethical approval in place for all data collection waves and further gains informed consent from all participants prior to data collection. Ethical approval is approved by the Faculty of Health Sciences Research Ethics Committee, Trinity College Dublin. Participants are informed through the Participant Information Leaflet that their data is shared in a confidential manner as part of the TILDA study.
The TILDA Privacy Policy gives more detailed information about data linkage with the GRO. It is important to note also that GDPR and the Irish Health Research Regulations do not apply to the personal data of deceased individuals. For the situation where a participant may be lost to follow up and their status unknown, TILDA have been granted a consent declaration by the Health Research Consent Declaration Committee (HRCDC) to process their data for GRO Linkage. A HRCDC declaration is granted in a case where the public interest of doing the research significantly outweighs the need for explicit consent.
A data transfer agreement is signed between TCD and GRO which commits to protecting the confidentiality of data. Physical and technical safeguards are also in place.

Comment 5. 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)

Response 5. Our use of the term 'contributory' was informed by a study by Batty et al. who use the term to refer to "Other diseases or injuries that contributed to the death but were not directly implicated" (p.2).
We have now explained our use of the term 'contributory' and provided a reference to the Batty et al. paper. *"A contributory cause of death is a condition that contributed to the death but were not directly implicated and are recorded in part two of death certificates. While this information has been rarely used in epidemiological research, recent evidence suggest it may have some methodological utility (Batty et al. 2019). For present purposes, contributory causes include diseases and conditions listed anywhere on the death certificate."*
Batty GD, Gale CR, Kivimäki M, Bell S. Assessment of Relative Utility of Underlying vs Contributory Causes of Death. JAMA Netw Open. 2019 Jul 3;2(7):e198024. doi: 10.1001/jamanetworkopen.2019.8024. PMID: 31365105; PMCID: PMC6669894.

Comment 6. 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.

Response 6. This valid point was also raised by another reviewer. In response, this particular analysis was informed by similar work carried out using UK Biobank data by Batty et al. The aim of this research, and our aim also, was to assess the utility of cause of death data extracted from the underlying cause field versus any location on the death certificate.

Our choice of smoking as a risk factor was, as you identify, because it is so well established. Smoking was also one of three risk factors included in the Batty et al. analysis. We have now included the following text in the manuscript to justify this analysis: *"We chose smoking to test our hypothesis that similar estimates would be derived from both underlying and contributory conditions as smoking is an established risk factor for mortality and it has been used for a similar purpose previously (Batty et al. 2019)."* Batty GD, Gale CR, Kivimäki M, Bell S. Assessment of Relative Utility of Underlying vs Contributory Causes of Death. JAMA Netw Open. 2019 Jul 3;2(7):e198024. doi: 10.1001/jamanetworkopen.2019.8024. PMID: 31365105; PMCID: PMC6669894.

Comment 7. 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?

Response 7. We hope we have interpreted this question correctly, but we confirm that the figures refer to the proportion of death certificates that included any cause from the ICD-10 chapter of diseases of the circulatory system as a contributory cause of death (52.6 %) and any cause from the ICD-10 chapter of diseases of the circulatory system (34.4%).

Comment 8. 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.

Response 8. Yes. This is a count of records that included at least one malignancy per record. We hope that the additional text we have included in response to your comment 5 in defining our use of 'contributory' has made this clearer to readers.

Comment 9. 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.

Response 9. Lung cancer was indeed the most common type accounting for 19% of cancers. We note that the association between smoking and cancer death is positive, but non-significant due to wide 95% confidence bands. We also note that our smoking variable identifies 'ever' as well as 'current' smokers, so some of the smokers may have quit some time ago.

Comment 10. 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.

Response 10. These estimates (HRs with 95% CIs) are presented in Figure 3. In responses to another reviewers suggestion, we have now also included the estimates for all-cause mortality. We also now more fully describe the results presented in this figure. We hope that this fuller description also provides clearer support for our contention that choice of contributory or underlying cause may not make much difference to these estimates. This

final point is more fully discussed in response to comment 11 below and comment 6 from Reviewer 1.

Comment 11. 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.

Response 11. As in our response to the previous comment, these estimates are presented in Figure 3 and the text describing these results has been extended.
Our contention that it appears that underlying and contributory cause of death may have similar utility for studies examining mortality risk factors is supported by the work discussed above by Batty et al. (2019) and a smaller scale study by Crews et al. (1991). We have now referenced both of these studies in support of the contention we made here.
We are also going to repeat the data linkage exercise in 2021 when TILDA will conduct its 6 th wave of data collection. The increased number of deaths will provide us with an appropriately large sample size to examine the association of major risk factors and specific causes of death. Initial results from this work are anticipated in late 2021.

Comment 12. 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.

Response 12. We have again checked this and can confirm that the Iris data dictionary does not identify “ischaemic heart disease”, only “ischemic heart disease”. The reason we chose to refer to this example was because it occurred so often.
As part of our data processing we appended the in-built data dictionary with common variations of spellings and descriptors we encountered and as a result, Iris performed this task increasingly well as we progressed. We plan to conduct data matching again in 2021 and are confident that we will have a higher success rate in our next attempt to assign ICD-10 codes automatically to individual death records.

Comment 13. 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.

Response 13. Thank you. We are glad that you agree with the importance of this exercise. As described above, we hope that project that this work stems from will make an important contribution to research on mortality in Ireland. We also hope that this particular data linkage demonstrates the great potential of combining rich individual level survey data with

administrative data sources. Unfortunately, to date Ireland somewhat lags behind other jurisdictions who have well developed data linkage infrastructures.

Comment 14. 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.

Response 14. Again, we wish to thank Dr Harteloh for his constructive feedback. We believe that the revisions have greatly improved the manuscript and provided clarification as to the meaning of contributory cause in this context. As discussed above, this is the first of many publications from this work. If interested, we have recently published another methodological paper using this data which compares the utility of cause of death data from official records and reports from end-of-life interviews. Ward, M, May, P, Normand, C, Kenny, RA, and Nolan, A. Comparing Underlying and Contributory Cause of Death in Registry Data With End-of-Life Proxy Interviews: Findings From The Irish Longitudinal Study on Ageing (TILDA). Journal of Applied Gerontology. [In Press].
https://doi.org/10.1177/0733464820935295

Response to Reviewer 3 comments – Dr Zubair Kabir
Comment 1. 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.

Response 1. Thank you Dr Kabir for taking the time to review our manuscript and for your helpful observations. As you rightly say, this type of exercise is challenging within the Irish data infrastructure and we do hope that our efforts contribute to improving this situation.

Comment 2. 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.

Response 2. We have done our best to describe as fully as possible the steps we took to achieve this data linkage. We hope that our responses to yours' and other reviewers suggestions have further improved this.
Naturally, many of our decisions and subsequent actions are specific to the data environment in which the work was conducted. By this, we mean that we were confined to the data that was available to use in TILDA, for example, the individual identifiers and so on. As such, it may well not be possible to replicate our procedures with other studies in Ireland. However, we feel strongly that we have been fully transparent and as specific as possible in our description of the steps we have taken to link the individual-level survey data available in TILDA to official death records. Indeed, given the richness of the data available to us in TILDA, we have many advantages not necessarily available to other studies.
As you correctly state, there are no standardised quality assurance tools available to use to assess the validity of our data linkage procedures and it was partly due to the absence of such a tool that we felt compelled to describe our methods as fully as possible and importantly to make this manuscript freely available to all.
Also importantly, our intention with this manuscript was not to suggest a one-size fits all method but rather to describe a new data infrastructure within TILDA that researchers interested in studying mortality in Ireland might avail of. How a similar task might be approached using a different study sample will be study dependent. That said, we do believe that our use of the Iris software tool for coding and identifying underlying cause of death is one way in which our work might be replicated and could help ensure standardisation in at least this aspect of the linkage across studies.
Completeness of source databases
As TILDA is prospective cohort study we are confident of the accuracy of the participant contact information and status as participants are contacted regularly and the status of non-responders is followed up via the participants or their proxies. The contact database is regularly updated so that participants can be contacted for future rounds of data collection. The GRO is the official register of all deaths in Ireland and provides information on deaths to the CSO for use in official statistics. As such, we are confident that it is a reliable and comprehensive source of data on deaths in Ireland.
Ethical and data security considerations
TILDA has full ethical approval in place for all data collection waves and further gains

informed consent from all participants prior to data collection. Ethical approval is approved by the Faculty of Health Sciences REC, Trinity College Dublin. Participants are informed through the Participant Information Leaflet that their data is shared in a confidential manner as part of the TILDA study.
It is important to note also that GDPR and the Irish Health Research Regulations do not apply to the personal data of deceased individuals. For the situation where a participant may be lost to follow up and their status unknown, TILDA have been granted a consent declaration by the Health Research Consent Declaration Committee to process their data for GRO Linkage. A HRCDC declaration is granted in a case where the public interest of doing the research significantly outweighs the need for explicit consent.
A data transfer agreement is signed between TCD and GRO which commits to protecting the confidentiality of data. Physical and technical safeguards are also in place.
Linkage methodology and technology
Our stating that data matching was conducted by the CSO in 2013 was in error and has now been removed from the manuscript. The only time data matching took place was in 2018 with the GRO.
The TILDA data team did undertake the data matching through the GRO search room facility. Once the TILDA team member identified the decedent within these records, the GRO then provided the detailed death certificate information for this person.
We have provided further clarification to these points in response to earlier comments. We have also appended our description of these measures within the manuscript and hope that they adequately address each of the points raised here.
You may also be interested to know that the CSO have repeated their 2013 data linkage using 2016 census data. You will find the results here: CSO: Mortality Differentials in Ireland. An Analysis Based on the Census Characteristics of Persons Who Died in the Twelve Month Period after Census Day 24 April 2016. 2019. Dublin. Source: https://www.cso.ie/en/releasesandpublications/in/mdi/mortalitydifferentialsinireland2016-2017/ [Accessed: October 2020].

Comment 3. 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.

Response 3. We hope we have clarified that the full data linkage exercise was conducted by the TILDA team. In practice the GROs sole involvement was to provide the team with the

death certificate information of decedents identified among TILDA participants.
In light of these, we believe we have made three contributions here. (1) We performed the data linkage, (2) provided an overview of a new data infrastructure and, (3) provided an assessment of the utility of contributory versus underlying cause in estimating the association between risk factors and mortality risk.
As also detailed in response to Reviewers 1 and 2 above, in this amended version of the manuscript we have better described our use of the term 'contributory' as: *"A contributory cause of death is a condition that contributed to the death but were not directly implicated and are recorded in part two of death certificates. While this information has been rarely used in epidemiological research, recent evidence suggest it may have some methodological utility (Batty et al. 2019). For present purposes, contributory causes include diseases and conditions listed anywhere on the death certificate."*
Batty GD, Gale CR, Kivimäki M, Bell S. Assessment of Relative Utility of Underlying vs Contributory Causes of Death. JAMA Netw Open. 2019 Jul 3;2(7):e198024. doi: 10.1001/jamanetworkopen.2019.8024. PMID: 31365105; PMCID: PMC6669894.
Also to re-state an earlier response, this particular analysis was informed by similar work carried out using UK Biobank data by Batty et al. The aim of this research, and our aim also, was to assess the utility of cause of death data extracted from the underlying cause field versus any location on the death certificate. The estimates do also confirm a stronger association between smoking and respiratory causes of death compared to all-cause mortality which is re-assuring but was not our main aim in this analysis.
Our choice of smoking as a risk factor was, as you identify, because it is so well established. Smoking was also one of three risk factors included in the Batty et al. analysis. We have now included the following text in the manuscript to justify this analysis: *"We chose smoking to test our hypothesis that similar estimates would be derived from both underlying and contributory conditions as smoking is an established risk factor for mortality and it has been used for a similar purpose previously (Batty et al. 2019)."* Again, we sincerely thank Dr Kabir for his insightful comments and appreciate his sharing his vast experience in this area with us.

Competing Interests: No competing interests were disclosed.