Mortality In The US By Education Level

January 13, 2019

Cristian Redondo Lourés[1] and Andrew J.G. Cairns[2]

Maxwell Institute for Mathematical Sciences and Department of Actuarial
Mathematics and Statistics, School of Mathematical and Computer Sciences,
Heriot-Watt University, Edinburgh, EH14 4AS, UK.

Abstract

Different mortality rates for different socioeconomic groups within a population have been consistently reported throughout the years. In this study, we aim to exploit data from multiple public sources, including highly-detailed cause-of-death data from the United States Centers for Disease Control and Prevention, to explore the mortality gap between the better and worse off in the US during the period 1989-2015, using education as a proxy.

Keywords

US mortality, education level, socioeconomic status, mortality inequalities.

Acknowledgements

This work forms part of the research programme "Modelling, Measurement and Management of Longevity and Morbidity Risk" funded by the Actuarial Research Centre of the Institute and Faculty of Actuaries, the Society of Actuaries and the Canadian Institute of Actuaries. The authors gratefully acknowledge feedback from the respective steering and oversight groups.

[1] E: C.Redondo_Loures@hw.ac.uk

[2] E: A.J.G.Cairns@hw.ac.uk   W: www.macs.hw.ac.uk/∼andrewc/ARCresources/

1 Introduction

It is a well established fact that different socioeconomic groups within a single population experience different levels of mortality, with the lower socioeconomic groups experiencing consistently higher mortality rates than their better off counterparts. Education has been used as a proxy for socioeconomic status in several mortality studies in the past because information on educational attainment of both decedents and the general population is more readily available than other socioeconomic indicators (e.g., income, wealth, occupation...). In particular, differences in mortality between different education categories in the United States have been reported in the past, see for example Jemal et al. (2008), Olshansky et al. (2012), or Case & Deaton (2015).

There are two reasons why we choose to analyse mortality by education level in the United States. First, the size of the population is such that, even when we focus on relatively rare causes of death, absolute death counts will be high enough that reliable rates can be computed separately for single years of age and calendar years, for males and females, and different education groups. The second reason is that the data needed for this kind of analysis is freely available online, and both a comprehensive list of all deaths in the US and certain characteristics of the resident population can be downloaded from the webpages of official US agencies.

To the best of our knowledge there are no other studies that analyse such a complete and detailed range of causes of death as the one we can construct with this data. This will help uncover which causes of death have influenced most heavily the recent trends in mortality for different education groups and which are the key drivers of the observed gap in all cause mortality. Having a comprehensive list of causes of deaths, each of them potentially caused by very different underlying factors (smoking, poor diet, genetic makeup...), will provide important insight on why mortality differences arise and what could drive their future evolution.

In this paper we will discuss how we construct the full dataset consisting of deaths and exposures by single year of age, single calendar year, and different education groups for males and females. We will also discuss briefly the resulting empirical all-cause mortality rates. A more thorough analysis of the data, including detailed cause of death analysis and considerations on the effect of grade inflation in our results, will be the subject of future research. The outline of the present work is as follows: Section 2 introduces the data sources and explains what they provide us with, as well as some of their shortcomings. Section 3 deals with how we can correct some of the issues that appear when separating our death counts and exposures by education level, explaining the procedure followed to build our final dataset. Sections 4 and 5 present some of the results obtained for both the proportion of people in different education groups and their death rates, as well as testing their reliability. Section 6 closes the work with a summary and some further discussion of the results.

2 Data

2.1 Death counts

Individual deaths data comes from the National Vital Statistics System (NVSS), a program managed by the National Centre for Health Statistics (NCHS) – part of the US Centers for Disease Control and Prevention (CDC). The NCHS collects vital data that has been acquired by the state authorities and makes it publicly available through their webpage, see NCHS (2016). Published data are transcribed directly from individual death certificates. These are anonymised to comply with data protection laws and published annually. The death certificate is partly filled in by the funeral director, see Rostron et al. (2010). This can cause problems with the reliability of certain items and will be discussed in more detail later. Although death certificates might not provide an exact count of all deaths it can be assumed that it is accurate for our purposes. The NCHS estimates that the coverage is above 99%, see US Department of Health and Human Services (1995).

The information that we are interested in is gender, age, calendar year, educational attainment, and cause of death. The first three are quite straightforward, and are recorded in almost every death (for example, only a negligible number, smaller than 0.02% over the whole period analysed, lacks information on age). We do not have any reasons to doubt the accuracy of any of these elements. The recording of education attainment on death certificates is less straightforward. This item started being recorded only in 1989, and even then not all states introduced it simultaneously, resulting in many entries with missing education in the earlier years. The death certificate was changed again in 2003, including a revised classification of educational attainment. Again, some states adopted this new certificate immediately, while some others used the old certificate for the whole period analysed. Due to this we have two different ways in which education is recorded that coexist during the 2003-2015 period. Some background on the treatment of educational attainment in death certificates can be found in Hoyert et al. (2006). We explain later how this is handled.

Regarding the quality of the data, there are several issues that are well known in the literature. For example, as already mentioned, some parts of the death certificate have to be filled by the funeral director. These might not have personally known the deceased, or have otherwise access to all the information required to fill all fields completely. This is especially true for the educational attainment item, which is known to be biased towards the "finished high school" level; i.e., people who attended (but never graduated) high school, as well as people who attended college but never got a degree, tend to be recorded as high school graduates who never attended university (see Rostron et al. (2010)). Some other times the funeral director, unable to find what the educational attainment of a person was, reports it as unknown, although this is relatively rare (the certificates with missing education are only about 3% in the last years included in our analysis).

However, many other entries are missing education because different states started recording it at different points in time. Entries with unknown education make almost 30% of the deaths in the earlier years, slowly declining to slightly below 10% in 1996. Fortunately, some states adopted the 1989 standard death certificate inmediately and have consistently reported education in more than 90% of their death certificates every year since 1989. These can be used to impute an education to all entries missing it using the ideas from Sasson (2016). In a later section we will explain how we implement this in our case.

Causes of death (CoD's) are recorded using the International Classification of Diseases (ICD) codes. Up to 1998 the ICD-9 standard was used, with later years using ICD-10. This change affects not only the way certain illnesses are coded, but also how the underlying cause of death is determined. After the chain of events that led to death by a practitioner has been listed, an automated system determines the underlying CoD following some certain rules. These rules change with each revision of the ICD, which means that discontinuities are introduced in the temporal trends of CoD mortality. The effect of this can be quite big for certain causes of death. As is common practice when ICD rules change, deaths for a certain year were classified using both standards, thus allowing the estimation of some comparability ratios (published in Anderson et al. (2001), which also contains more detail on how the coding and determination of underlying cause work). With this we can estimate the size of the discontinuity in the number of deaths attributed to certain causes when the ICD9-ICD10 change was implemented. In principle, applying these comparability ratios could make the death rates evolve smoothly, but changes in the trend can still occur that are caused by the change in practice rather than an actual change in causes of death. If we are, however, interested in the ratios between the mortalities experienced by different subpopulations, these would be unaffected by all these changes. In this paper, which deals exclusively with the construction of the database, cause of death information is only used in the education imputation procedure described in Section 3. A more thorough analysis of cause of death mortality in different education groups and its latest trends will be the subject of a further publication.

2.2 Exposures

In line with the deaths data, we seek to estimate exposures (average population) by education level as well as gender, year and age. To achieve this we exploit data from three sources: the Human Mortality Database (HMD), Current Population Survey (CPS) and American Community Survey (ACS).

- Exposures without subdivision by education have been extracted from the HMD. These have been then filtered using the methods proposed by Cairns et al. (2016) to adjust for potential anomalies in the data.

- For each group by gender, year and age, we need estimates of the proportion of the group that have attained different levels of education. This comes from two potential sources: the CPS and the ACS. The CPS data cover the full period 1989 to 2015, providing data by individual year of age up to age 79. The ACS data surveys a larger number of people, but only covers 2004 to the present. As a consequence, we use the CPS data for our main calculations, and use the larger ACS dataset for independent verification of aspects of our results.

The CPS is a monthly survey conducted by the Census Bureau. In their own words (U.S. Census Bureau (2000)):
"*The CPS is administered by the Census Bureau using a probability selected sample of about 60,000 occupied households. The fieldwork is conducted during the calendar week that includes the 19th of the month. The questions refer to activities during the prior week; that is, the week that includes the 12th of the month. Households from all 50 states and the District of Columbia are in the survey for 4 consecutive months, out for 8, and then return for another 4 months before leaving the sample permanently. This design ensures a high degree of continuity from one month to the next (as well as over the year). The 4-8-4 sampling scheme has the added benefit of allowing the constant replenishment of the sample without excessive burden to respondents.*"
"*To be eligible to participate in the CPS, individuals must be 15 years of age or over and not in the Armed Forces. People in institutions, such as prisons, long-term care hospitals, and nursing homes are ineligible to be interviewed in the CPS. [...] No upper age limit is used, and full-time students are treated the same as nonstudents. One person generally responds for all eligible members of the household. The person who responds is called the 'reference person' and usually is the person who either owns or rents the housing unit. If the reference person is not knowledgeable about the employment status of the others in the household, attempts are made to contact those individuals directly.*"
The CPS is designed to be a survey about the characteristics of the workforce, and therefore the sample is designed to obtain an error for the unemployment rate below a certain threshold. Once the participating households have been selected, their occupants are surveyed for 4 straight months, and then dropped for 8 months, before returning to the sample exactly one year after their initial inclusion. In this second round people are interviewed for another 4 months and then dropped permanently. Each month, 1/8 of the households are dropped permanently and another 1/8 enters the "freezing" period, with 1/8 of the sample for the following month being comprised of completely new households entering the survey, and another 1/8 of "defrozen" households. This way there is a 75% overlap between the households surveyed in two consecutive calendar months and a 50% overlap between the same month of two consecutive calendar years. More details on sample selection and rotation can be found in Chapter 3 of U.S. Census Bureau (2006), particularly in pages 1 and 10-14.
The rotation pattern means that the samples for the same calendar month of two consecutive calendar years are not independent. Half of the households interviewed in a certain year will be the same ones that were interviewed in the previous one, which links the two data points. This is done in order to smooth the yearly drift of the quantities of interest, but could be a problem for us if we want to treat the different points in our time series as independent random variables. However, because we are not interested in the total population estimates given by the CPS, but only in the ratio of respondents that reported to have a certain education, we can restrict the data we retrieve to be the one reported by households that had not taken part in the survey the previous year. We do this at the expense of losing approximately half of the counts for each of our groups, which will increase the sampling error by a $\sqrt{2}$ factor.

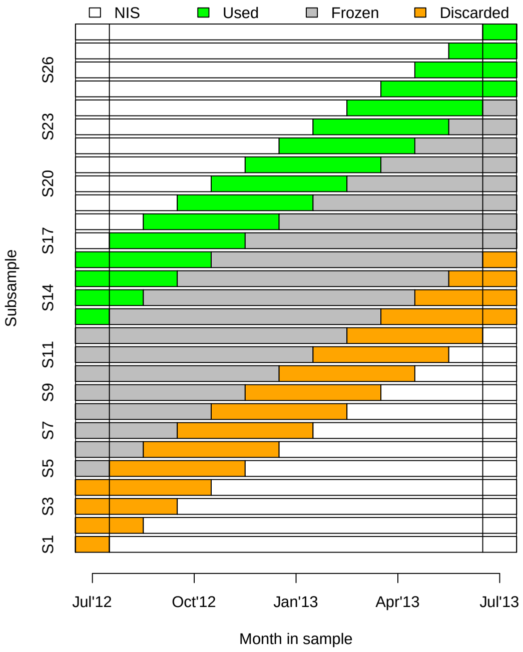


Figure 1: Plot showing the how 28 samples (labelled S1-S28) leave and enter the rotation group during a calendar year. Green: data is collected and not correlated with the one for the same month one year before. Orange: data is collected but correlated with the one collected the previous year, and is discarded. Grey: the subsample is in the waiting period. White: the subsample is not in sample (NIS).

The way the sample is designed still makes the not-in-sample-previous-year subsample of households representative of the whole population. Details on sampling can be found, as explained before, in Chapter 3 of U.S. Census Bureau (2006), but the basic idea is as follows. The whole territory of the US is divided into small, relatively homogeneous Primary Sampling Units (PSU's). These are grouped together in strata which are expected to have similar workforce characteristics, and one PSU is selected to be surveyed, with its results representing all other similar PSU's. In each PSU to be surveyed, a list of addresses is taken from the latest Census, and complemented with the addresses of recently built housing obtained from building permits. From these a sample of addresses is chosen to be surveyed, and split in several subsamples to be used in the rotation scheme. The way the rotations groups are set up is such that all subsamples are balanced at the levels that we are interested in, which are the state and national level.
Figure 1 shows how this rotation works. Each row represents a sample of the US population. In a given month, for example July 2012, 16 samples (labelled S1-S16 here) are part of the CPS sample. 4 of them (S1-S4) are being interviewed that month, but

they were also interviewed exactly one year ago. 8 others (S5-S12) are in the 8-month waiting period and are not interviewed. Finally, the last group of 4 (S13-S16) is being interviewed in July 2012, but was not in July 2011. We can see that as months pass some of the subsamples are being dropped and others take their place, and that in exactly one year's time (July 2013) the group S13-S16 is now being interviewed for the second time in a month of July, while the group S25-S28 is interviewed in July 2013 but was not in July 2012. Choosing only the uncorrelated data is equivalent to discarding the data coming from the orange part of the plot.

3 Educational attainment

The first thing we need to study educational attainment is a clear definition of the different levels that we are going to use. We need to find education groups that will produce informative results and that can be easily and consistently defined with the available data. Recent studies of mortality in the US that use education as a covariate (including Case & Deaton (2015)) classify people in three groups: people who never attended college (whether or not they graduated high school), people who attended university but never got a Bachelor's degree, and people who obtained a Bachelor's or higher qualification. A degree-based classification like this one is used in both the CPS and the 2003 death certificate (which, as already stated, some states had not yet adopted by 2015), but in the earlier years another classification based on years of schooling was used. This forces us to try to make an equivalence between the two systems to have consistent data throughout all the years in our analysis.

For the exposures data coming from the CPS we have in fact three different education reporting methods: Between 1989 and 1991 a "years of schooling model" was used, in which the number of years spent in elementary education (or college if applicable) was recorded; From 1994 on there is a "grade obtained" approach, in which people are classified according to the highest degree they have earned (with a special "some college, no degree" category); And in 1992-1993 a hybrid approach is used, in which years of schooling are used for people who did not graduate high school, and highest degree obtained is used as in the most modern standard for people with higher education. However, in death certificates there was no intermediate stage, with the 1989 model recording education as "years of schooling" and the 2003 revision recording "highest degree obtained". As already mentioned, not all states adopted any of the two death certificate models inmediately after they were introduced, which results in a high number of entries with missing education information in the early years after 1989, and a mixture of both standards in data for all years starting in 2003. Moreover, between 1992 and 2003 CPS population data and death counts data were recorded using completely different standards, which forces us to try and find a way to group these different education categories in a manner that is consistent throughout the years and regardless of the educational attainment recording standard used.

A complete, category-by-category basis correspondence between the two standards is extremely difficult, if not impossible, to create. However, it is possible to do some

		New Code		
Old Code	0 to HS diploma	Some college/associate	Bachelor's or higher	Total
0-12	39219	2832	147	42198
13-15	508	13683	350	14541
16+	70	881	16735	17686
Total	39797	17396	17232	74425

Table 1: Cross-tabulation of Old Code (highest grade completed) and New Code (highest grade completed or degree received). Source: Park (1999) – abbreviated.

grouping that will roughly correspond to the three education groups we want to work with. In the month of February 1990 people responding to the CPS survey question about their educational attainment were asked to provide two answers, one using the old (years based) and one using the new (degree based) standard. The results obtained are shown in Table 1 adapted from Park (1999). The following groups can be formed, which correspond to our definition of three educational attainment levels given before, and with memberships that are roughly similar in any of the standards:

- Lower educated group. This includes all people who have never been to college. It corresponds to categories 0 through to 12 years of schooling in the old standard (row labels in the attached table), and to categories from no formal schooling to HS Diploma in the new standard (0-HS Diploma in the attached table column headings).
- Medium educated group. These are people who have been to college, but have not achieved a Bachelor's degree. It corresponds to categories 13-15 years of schooling in the old standard and "Some College/Associate degree" in the new one.
- Higher educated group. This group is formed by people who hold a Bachelor's degree or higher. It is formed by people in the categories 16 or more years of schooling in the old standard, or in the categories Bachelor's or further in the new recording.

In principle, this grouping should solve most of the problems arising from the differences in education recording. The memberships of each group for February 1990 are as follows:

- Lower educated group: 42198 (old standard) and 39797 (new standard).
- Medium educated group: 14541 (old) and 17396 (new).
- Higher educated group: 17686 (old) and 17232 (new).

The main source of error is the misclassification of 2538 (out of the 2832) people who said they had been schooled for 12 years in the old standard (that is, until the end of high school) but claimed to have some college education when asked the degree-based question (probably because they attended college for less than a year). For the lower educated group, having a big membership, absorbing these extra members does not make a very big difference, but the medium educated group death rates could potentially be affected. Specifically, death rates for group 2 might be underestimated (death counts use old standard → underestimation of number of deaths compared to exposures using the new standard) in the period 1992-2003. The consistency of the membership of group 3 make any misclassification issues less of a problem for people in that group. Also note that we only studied the equivalence of the two educational attainment recording methods for a very specific case (CPS survey, February 1990), and aggregated on age for people aged 25-64. The results found there might not be representative of the equivalences between groups of education for all the length of our study, or for people outside that age range. Moreover, and even though our grouping should solve most of the problems coming from education being recorded in two different ways, any data quality problems (for example misreporting) still remain.

3.1 Education in the death records

Now that we have defined our education groups we can tackle the problem of missing information in the death counts. In general, missing education in a death certificate happens due to one of two reasons: individual entries for which education was not known (which is a small percentage every year), and state-wide non-reporting due to the late adoption of the 1989 death certificate by some states (which is the reason behind most "unknown education" entries in early years). For example, whereas in 2010 all states reported education in death certificates, and entries with missing data were less than 5% of the total, in 1989 (the first year in which educational attainment entered death certificates) only 21 states reported education with less than 10% missing data. We can nonetheless try to impute an educational attainment level to each entry without one using simply the death counts with education recorded and the exposures. We generalise the method explained in Sasson (2016), based on Bayes theorem:

$$p(e|C,X) = \frac{p(C|e,X)p(e|X)}{p(C|X)}, \tag{1}$$

where e represents education levels, C is cause of death, and X={Year, age, gender} captures the characteristics of the underlying live population as well as the matching characteristics at the time of death for those who die.

As long as we can provide an estimate for all three quantities on the right hand side we will be able to calculate the probability that an entry corresponding to a person who died from cause C, with characteristics X, had attained education level e.

The denominator on the right hand side is simple to compute. It is the aggregate mortality rate by cause of death across all education groups, and can be calculated using the death counts and total population data.

As the second term in the numerator we have the probability that a person with characteristics X has a certain level of education. This would be calculated as the exposures for education level e divided by total exposures for each X. As a first approximation, and because there is no "unknown education" category in the CPS, we can use the (unsmoothed) survey data to give an initial estimate of this quantity. Once we have death counts with complete education we will proceed to the smoothing of the exposures, described in the next subsection, and obtain a better estimate for $p(e|X)$, which we will use for a second round of imputation. After this, we will recalculate our exposures and repeat the process until we reach convergence, and the difference in the estimated death rates between two consecutive rounds of imputation-smoothing becomes negligible.

The first term in the numerator can be problematic, since we need the mortality rates by education level. In later years, and because the number of entries with missing information is relatively small, we can use the whole data set to give an initial estimate for this quantity and proceed to the imputation from there. But for the earlier years, which concentrate most of the entries with unknown education, the large proportion of missing data means that an estimate obtained using all of the death counts could be biased. In this case we have to use a subset of all states that have always reported educational attainment in at least 90% of death certificates to get a first estimate of these death rates, restricting both our death counts and exposures to them. This criterion has been used by the NCHS themselves when reporting educational-related data, see NCHS (1993). From year 1997 on, the missing education information is below the 10% threshold for the whole of the US, and therefore the first estimate for $p(C|e, X)$ can be calculated using all the available mortality data. From 1989 until 1996 we use data from the following states: Arizona, California, Colorado, Delaware, Florida, Hawaii, Idaho, Illinois, Iowa, Kansas, Michigan, Minnesota, Missouri, Montana, New Hampshire, Oregon, South Carolina, Utah, Vermont, Wisconsin, and Wyoming.

The entries without education are split evenly between all three education groups in the first iteration to initialise $p(C|e, X)$ for each X. This way we ensure that there will be an initial seed for all X and e for all causes of death. Otherwise, we could end up with deaths due to rare causes not being assigned to any group because there were no occurrences for which education was known. Comparing the value to which the death rates converge using different criteria for the initial assignment of these deaths shows that the effect of the specific choice is negligible. This is understandable since only in extremely rare cases the number of deaths in the three education groups will not be much bigger than in the unknown education group thanks to the 90% completeness criterion. Note that, with this method, we are assuming that the deaths with missing education are equally distributed in e given certain X and C as the deaths for which education was recorded.

3.2 Education in the exposures

The raw exposures obtained from the CPS are rather noisy, and even total population estimates for individual ages vary quite widely for consecutive calendar years (as does the estimate of the number of people of different ages in a single calendar year). Figure 2 clearly illustrates this: the HMD gives a smooth evolution of the total population, whereas the CPS (and even the more accurate ACS) estimates are noisier. As a consequence, we decided to take an approach in which we use the more reliable estimates for the total exposures from the HMD and then separate them by education level using the ratios of people in each education group as extracted from the CPS.

To smooth the exposures, we will calculate the evolution of the ratios of people in each education group for each of the cohorts that participate in the analysis. Crude estimates of these ratios based on raw CPS data will be quite erratic due to the sampling error. However, it is reasonable to expect that the true proportion of people with a certain education level *within a cohort* should evolve smoothly from year to year as the cohort ages.

We define the following quantities:

- Let c represent a cohort with a particular year of birth and gender.
- Let e represent the level of educational attainment.
- $E_C(i,e,c)$ = exposures for the cohort c with education level e, in year i of observation. $i = 0$ represents the first year that we observe this cohort. i then has a dual role as a proxy for both the calendar year and the age of the cohort.
- $E_C(i,c) = \sum_e E_C(i,e,c)$ = exposures for the cohort c across all levels of educational attainment, in year i of observation.

We now define

$$R_C(i,e,c) = \frac{E_C(i,e,c)}{E_C(i,c)}. \tag{2}$$

$R_C(i,e,c)$ is the true ratio of people in cohort c with education level e in year of observation i.

It is easy to derive the recurrence relation

$$\begin{aligned} R_C(i+1,e,c) &= \frac{E_C(i+1,e,c)}{E_C(i+1,c)} \\ &= \frac{E_C(i,e,c) - \Delta_C(i,e,c)}{E_C(i,c) - \Delta_C(i,c)} \\ &= \frac{R_C(i,e,c)E_C(i,c) - \Delta_C(i,e,c)}{E_C(i,c) - \Delta_C(i,c)}, \end{aligned} \tag{3}$$

where we have defined Δ_C as the decrease in the number of members of a certain cohort between years i and $i+1$. There are three possible contributions to this Δ_C: the members of the group that died between years i and $i+1$, net migratory flux, and people moving from a lower to a higher level of educational attainment (with the latter not affecting $\Delta_C(i,c)$, the decrease in the total cohort exposures). If we assume that at the ages analysed migration is small compared to the number of deaths, and the amount of people who choose to pursue further education is negligible, we can approximate Δ_i by the number of deaths:

$$\tilde{\Delta}_C(i,e,c) = D(t_0 + x_0 + i, x_0 + i, e, g), \tag{4}$$

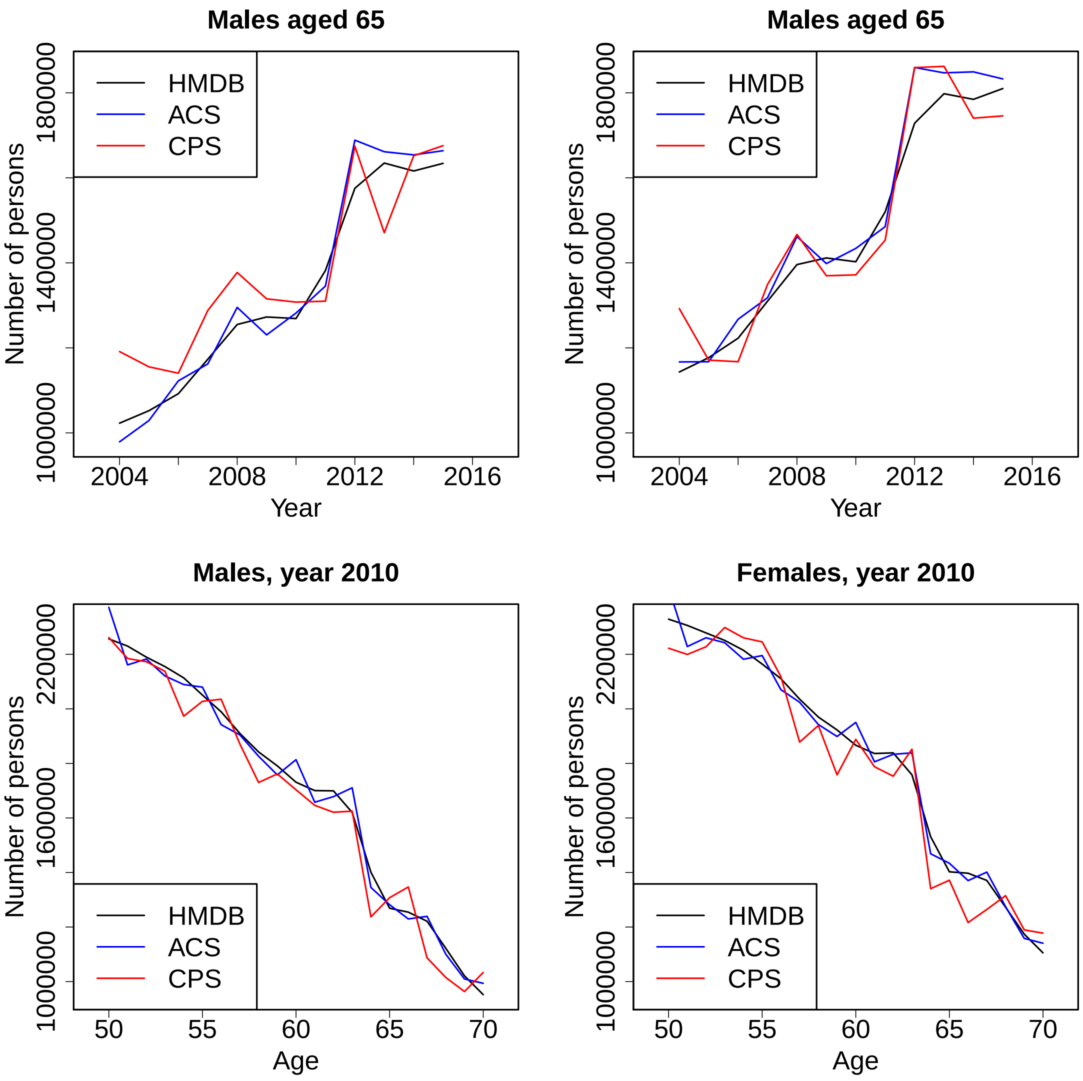


Figure 2: Comparison between the exposures coming from the CPS (red lines) and the ACS (blue lines) with the raw HMD exposures. The trends explained in the text are clearly visible in these examples, but they exist with more or less strength for most calendar years and ages.

where $D(t, x, e)$ is the number of deaths in calendar year t, age x, and education level e, t_0 is the year of birth of the cohort, x_0 the age at which the cohort enters our analysis, and g is the gender (note that all these three parameters are collectively represented by c on the left hand side). Additionally, we have

$$\tilde{\Delta}_C(i, c) = \sum_e D(t_0 + x_0 + i, x_0 + i, e, g). \tag{5}$$

This approximation is good as long as the original exposures data produces memberships for each cohort that decrease with age consistently with the number of deaths. This is precisely what Graphical Diagnostic 3 from Cairns et al. (2016) analyses.

Once we have specified an initial estimate for $\hat{R}_C(0, e, c)$, we can calculate estimates of the remaining ratios, $R_C(i, e, c)$ for $i = 1, 2, \ldots$ using the $\tilde{\Delta}_i$ (Equations (4) and (5)) and the recurrence relation (3). This produces a full series $\hat{R}_C(i, e, c)$, for that particular cohort and education level. Alongside this we also have the crude CPS ratios, $R_C^{CPS}(i, e, c)$.

To estimate this initial ratio it seems reasonable to minimise the quadratic deviation between our smoothed ratios using recurrence relation 3 and the crude CPS ratios:

$$\mathcal{O}(e, c) = \sum_i \omega(i, e, c) \left(\hat{R}_C(i, e, c) - R_C^{CPS}(i, e, c) \right)^2, \tag{6}$$

where the $\omega(i, e, c)$'s are weights that we will discuss later on. Once we have chosen to use this model there is one single parameter to be estimated for each cohort and education level: $\hat{R}_C(0, e, c)$.

The advantage of this method is that we make full use of all our data to estimate ratios that are compatible with the mortality observed. However, because we need the exact number of deaths for all education levels to estimate the ratios, and the ratios themselves are necessary for the education imputation discussed in the previous subsection, we need to proceed iteratively and repeat the imputation-smoothing procedure several times until convergence of the final death rates is obtained.

In the next section, when the detailed results are shown, we will see that this cohort-by-cohort based smoothing presents some issues. In particular, because we are smoothing the ratios for each cohort independently from its neighbours, we will see "ripples" in the estimated death rates (see Figure 6), as for some cohorts there will be a systematically over- or under-estimation of the ratio of people within a certain education group compared to its immediate neighbours. These type of oscillating changes in the ratios of educated people between consecutive neighbours are highly undesirable, and can potentially make genuine cohort effects more difficult to detect. Due to this, we will slightly modify the smoothing procedure so that we get rid of this problem making small adjustments to the ratios of educated people. Thanks to this we will obtain death rates that vary smoothly between consecutive cohorts.

In Cairns et al. (2016) one of the diagnostics of the quality of the exposures was concavity of the generated death rates. We will see that this is the most clear indicator that the ratios obtained by minimising function $\mathcal{O}(e, c)$ need to be corrected. The concavity function (suppressing reference to (e,g)) is defined as:

$$C(t,x) = \log\left(m(t,x)\right) - \frac{1}{2}\left(\log\left(m(t,x+1)\right) + \log\left(m(t,x-1)\right)\right), \tag{7}$$

where the $m(t,x)$ are empirical death rates. Assuming that $m(t,x)$ is smooth from age to age within year t, this should be close to zero. However, simply minimising the $\mathcal{O}(e,c)$ for each cohort in isolation, can lead to oscillations in the $m(t,x)$ and larger values for the concavity than are reasonable. We should, therefore, modify our target function to include a term that penalises ratios that would generate a large concavity. This new target function is:

$$\mathcal{O}_N = \sum_{i,e,c} \omega(i,e,c)\left[\left(\hat{R}_C(i,e,c) - R_C^{CPS}(i,e,c)\right)^2 + \alpha C(i,e,c)^{2\beta}\right] \tag{8}$$

for some constants $\alpha,\ \beta > 0$. The new term, which gives a measure of the concavity, is:

$$C(i,e,c) = \log\left(m(t,x,e,g)\right) - \frac{1}{2}\Big[\log\left(m(t,x+1,e,g)\right) + \log\left(m(t,x-1,e,g)\right)\Big], \tag{9}$$

where we define $t = t_0+x_0+i$ and $x = x_0+i$, and the $m(t,x,e,g) = D(i,e,c)/\hat{R}_C(i,e,c)E_C(i,c)$ are the empirical death rates for cohort c (c encapsulating gender g and year of birth $(t-x)$) with education level e.

Given that we know the number of deaths D for a cohort and the total exposures E at every age, and that we know how to estimate the ratios $\hat{R}_C(i)$ given the initial ratio $\hat{R}_C(0)$, the only thing we need to give this estimate of the concavity is the log of the death rates of the neighbouring cohorts (those that at year t, when the cohort born at year t_0 is age x, are ages $x+1$ and $x-1$). Because we are already using an iterative method to estimate our death rates (remember we did this so that we could repeat the imputation-smoothing several times until convergence was achieved), we can use for each iteration the mortality rates calculated in the previous one. As long as the method converges (the change in the death rates between consecutive iterations becomes close to zero), the death rates of the previous iteration will help us compute the concavity that we would observe at the end of the current one. Thanks to this, optimisation of equation (8) would be useful to calculate ratios that generate exposures from which we derive death rates with a small concavity.

After a thorough analysis it has been found that values of $\alpha = \beta = 1$ give results that do not greatly deviate from the ones obtained minimising the $\mathcal{O}(e,c)$, while reducing the problems we observed in that case. It has been checked that the choice of α and β does not heavily affect the resulting death rates as long as we restrict ourselves to values close to unity.

Lastly, it is worth mentioning that for both the unsmoothed ratios and the observed concavity a standard error can be estimated. For the ratios we use, the number of counts in each education group follows a conditional binomial distribution with $c(t,x,g)$ trials (i.e. the number of persons in the sample with characteristics (t,x,g)) and probability $R(t,x,e,g)$ (the "true" ratio or proportion of the group with education level e) of success.

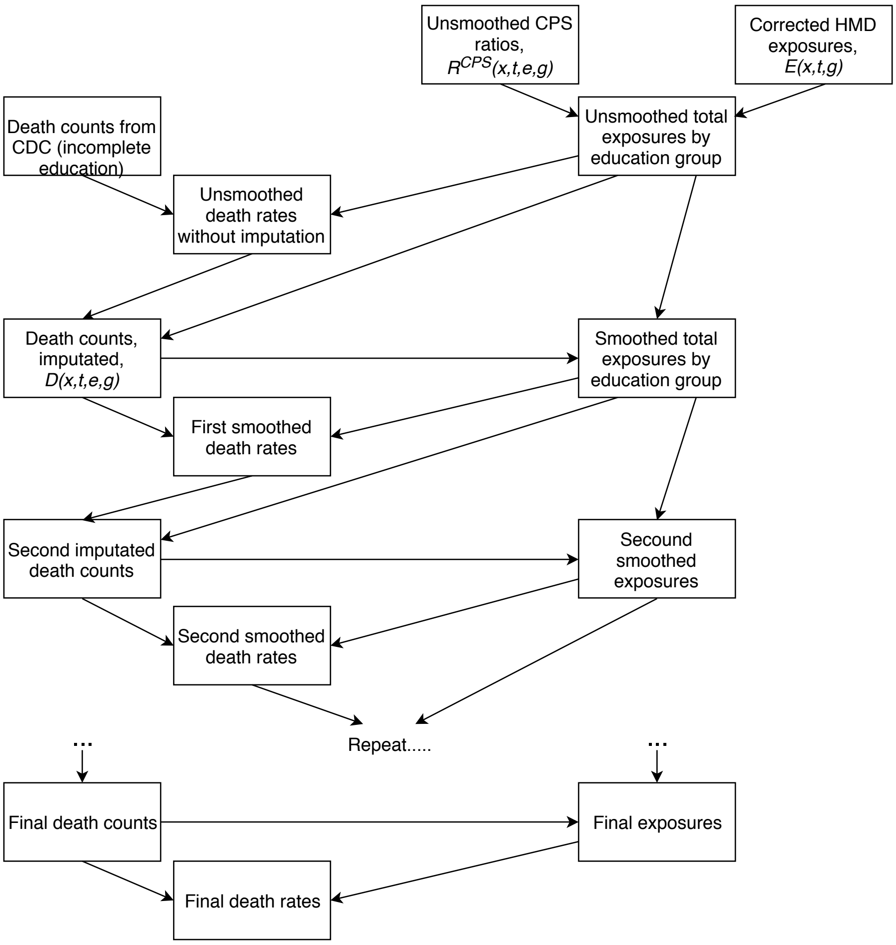

Unsmoothed CPS ratios, $R^{CPS}(x,t,e,g)$
Corrected HMD exposures, $E(x,t,g)$
Death counts from CDC (incomplete education)
Unsmoothed total exposures by education group
Unsmoothed death rates without imputation
Death counts, imputated, $D(x,t,e,g)$
Smoothed total exposures by education group
First smoothed death rates
Second imputated death counts
Secound smoothed exposures
Second smoothed death rates
Repeat.....
...
...
Final death counts
Final exposures
Final death rates

Figure 3: Flux diagram showing how the algorithm handles the initial data and obtains final death rates and ratios of educated people.

The $\hat{R}$ ratios we calculated before are an estimate of the true ratios R, and therefore their standard error, $\sigma(\hat{R}(t,x,e,g))$, is:

$$\sigma(\hat{R}(t,x,e,g)) = \frac{\sqrt{c(t,x,g)\hat{R}(t,x,e,g)(1-\hat{R}(t,x,e,g))}}{c(t,x,g)} = \sqrt{\frac{\hat{R}(t,x,e,g)(1-\hat{R}(t,x,e,g))}{c(t,x,g)}}. \tag{10}$$

For the concavity, the fact that the total number of deaths observed is a random variable creates a certain variability of the death rates that will generate an uncertainty in the value of the concavity. If we assume that the number of deaths follows a log-normal distribution, then the logarithm of the mortality would be normal. In particular, if we use the observed number of deaths D as the estimate of the mean of the distribution, and $\sqrt{D}$ as its standard deviation, then the mortality $m = D/E$ would be log-normally distributed with mean $n = D/E$ and standard deviation $d = \sqrt{D}/E$. Through the relation between the parameters of the log-normal distribution and the distribution of its natural logarithm, we know that $\log(m)$ would be normally distributed with mean μ and standard deviation σ, where:

$$\mu = \log\left(\frac{n}{\sqrt{1+\frac{d^2}{n^2}}}\right) \tag{11}$$

$$\sigma = \sqrt{\log\left(1+\frac{d^2}{n^2}\right)} \tag{12}$$

Since at the end of the iterative process that imputes education and estimates ratios of educated people we get a total number of deaths and exposures for each group, we can estimate the standard deviation of $\log(m)$ for each calendar year and age. Because three log mortalities (that are, additionally, conditionally independent of each other) enter the calculation of the concavity, the conditional standard deviation of the concavity is:

$$\sigma(C) = \sqrt{\mathrm{Var}\left(\log\left(m(t,x)\right)\right) + \frac{1}{4}\mathrm{Var}\left(\log(m(t,x+1))\right) + \frac{1}{4}\mathrm{Var}\left(\log(m(t,x-1))\right)} \tag{13}$$

This allows us to add confidence intervals to the plots of the concavity so that we can see if the dispersion of the points for the different education groups can be attributed to random fluctuations in the number of deaths or to issues with the data. This should be interpreted as a lower bound of the error, since here we are assuming that the only variable generating the error is the number of deaths, i.e., we assume that we know the exact exposures without any uncertainty. If we had added a certain error to E the value of $\sigma(C)$ would be bigger than the one obtained here.

4 Results: death rates and education ratios

This section will present the results obtained. Because the computation of the death rates and the ratios of people in each education group are intertwined in our model and neither can be calculated without the other, we will show the results for both the death rates and the ratios of educated people alongside each other.

Figure 4 shows, as an example, how the unsmoothed ratios look for certain cohorts. We can see the noisy nature of the data, with oscillations that are compatible with the sampling error, as shown by the dashed lines. The death rates extracted from this data will therefore be noisy as well, and thus the need to introduce the smoothing explained in the previous section.

In Figure 5 we show the smoothed version of these same ratios when the $\mathcal{O}$-smoothing and $\mathcal{O}_N$-smoothing procedures are used (with all weights ω_i being equal). We see that we obtain a smooth curve that falls really close to the centre of the 2σ intervals, and that the difference in the ratios due to the concavity penalisation is rather small. However, this small correction makes the smoothed death rates behave much better.

By way of example, resulting all cause death rates for medium-educated are shown in figure 6. We can clearly see the effect of each step of the smoothing procedure: when we go from the raw to the $\mathcal{O}$-smoothed exposures we see that the surface becomes smoother, but with some ripples (or ridges) due to the independent adjustment of neighbouring cohorts; when going from $\mathcal{O}$- to $\mathcal{O}_N$-smoothed ratios anomalous cohorts that give rise to these ripples are brought closer to their neighbours, resulting in a much smoother and more plausible mortality surface.

We now proceed to analyse the results by applying several diagnostic tools to our results. We will start with the concavity function. Figure 7 shows heat maps for the concavity function of medium educated males. The result is exactly as we would expect: for the unsmoothed ratios (panel A) the concavity is fairly random and large (top left, note the different scale), for the $\mathcal{O}$-smoothed ratios (panel B) we have a smaller concavity but patterns along some cohort diagonals, while for the $\mathcal{O}_N$-smoothed ratios (panel C) the concavity is random and fairly small everywhere.

Figure 8 shows the detailed process for a specific cohort, the medium educated males born in 1925. In the top left panel we can see that, when unsmoothed ratios are used, the concavity is large and positive, and that there is quite a lot of noise. Not a single point, or the average concavity, falls within the $0 \pm \sigma$ confidence interval marked by the blue lines. In the top right panel, where $\mathcal{O}$-smoothed ratios have been used, we see a decrease in the total concavity (note the change of scale), and all the dots come closer to the average marked by the red line. But even though we managed to reduce the noise, the average concavity is still higher than zero and outside the $\pm\sigma$ interval. Finally, in the bottom panel we see what happens with the $\mathcal{O}_N$-smoothed ratios. The pattern of the dots is extremely similar to the $\mathcal{O}$-smoothed case, but they have been brought down so that the average concavity nearly vanishes. The dispersion of the points is compatible with our estimation of the standard error coming from the expected random variations in the number of deaths.

The reason for this can be seen in figure 9. In this figure we see the death rate for the 1925 birth cohort of medium educated males as a solid black line, along with the death rates for the two previous (dashed black) and the two following (dashed red) cohorts. A systematically positive concavity would appear if the death rate for the central cohort

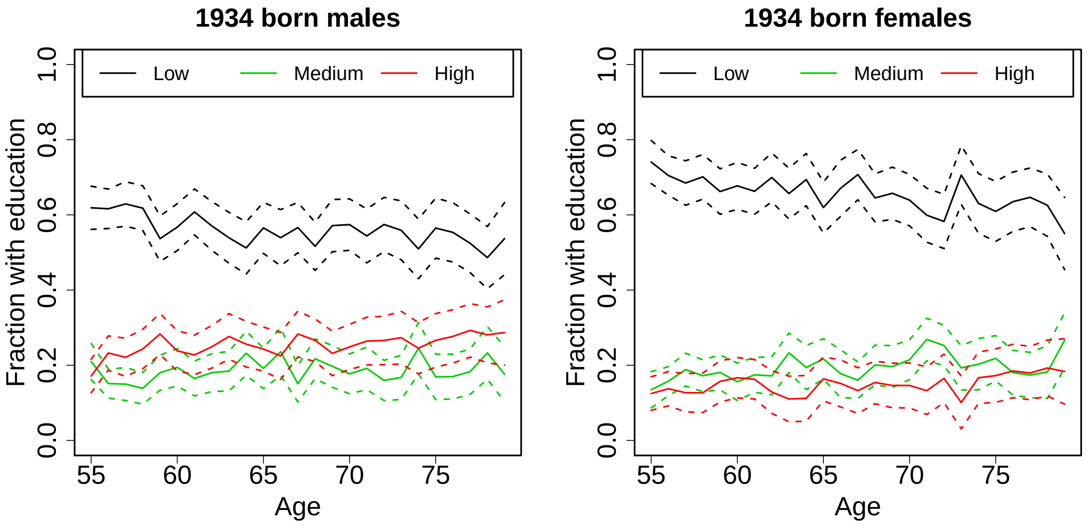


Figure 4: Unsmoothed ratios using the raw CPS data for male and female cohorts born in 1934, for the three groups low, medium and high education. The ratios (solid lines) and the $\pm 2\sigma$ confidence intervals (dashed lines) are shown.

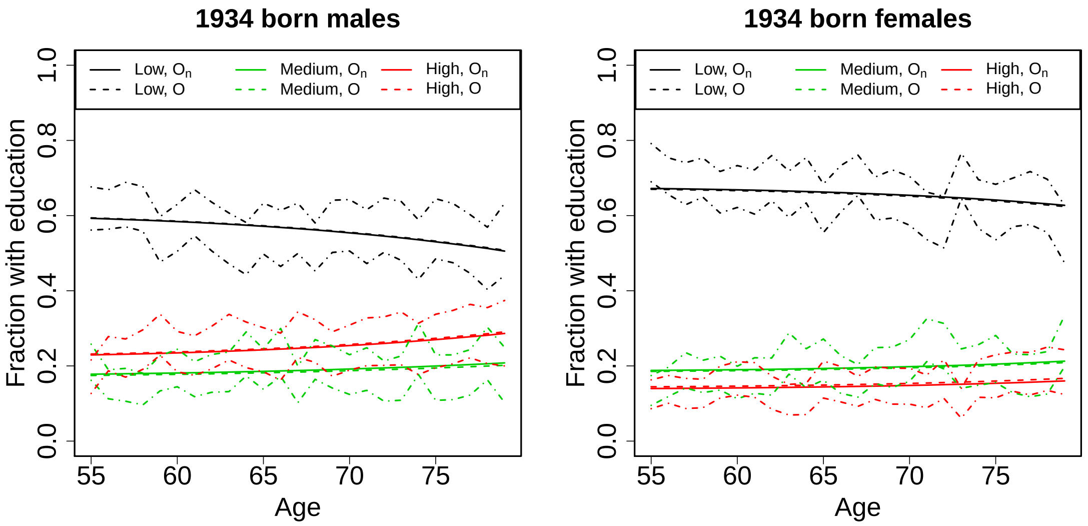


Figure 5: $\mathcal{O}$-smoothed (solid lines) and $\mathcal{O}_N$-smoothed (dashed lines) ratios for the same cohorts as in figure 4. The $\pm 2\sigma$ confidence intervals around the unsmoothed ratios from Figure 4 are shown as dot-dashed lines.

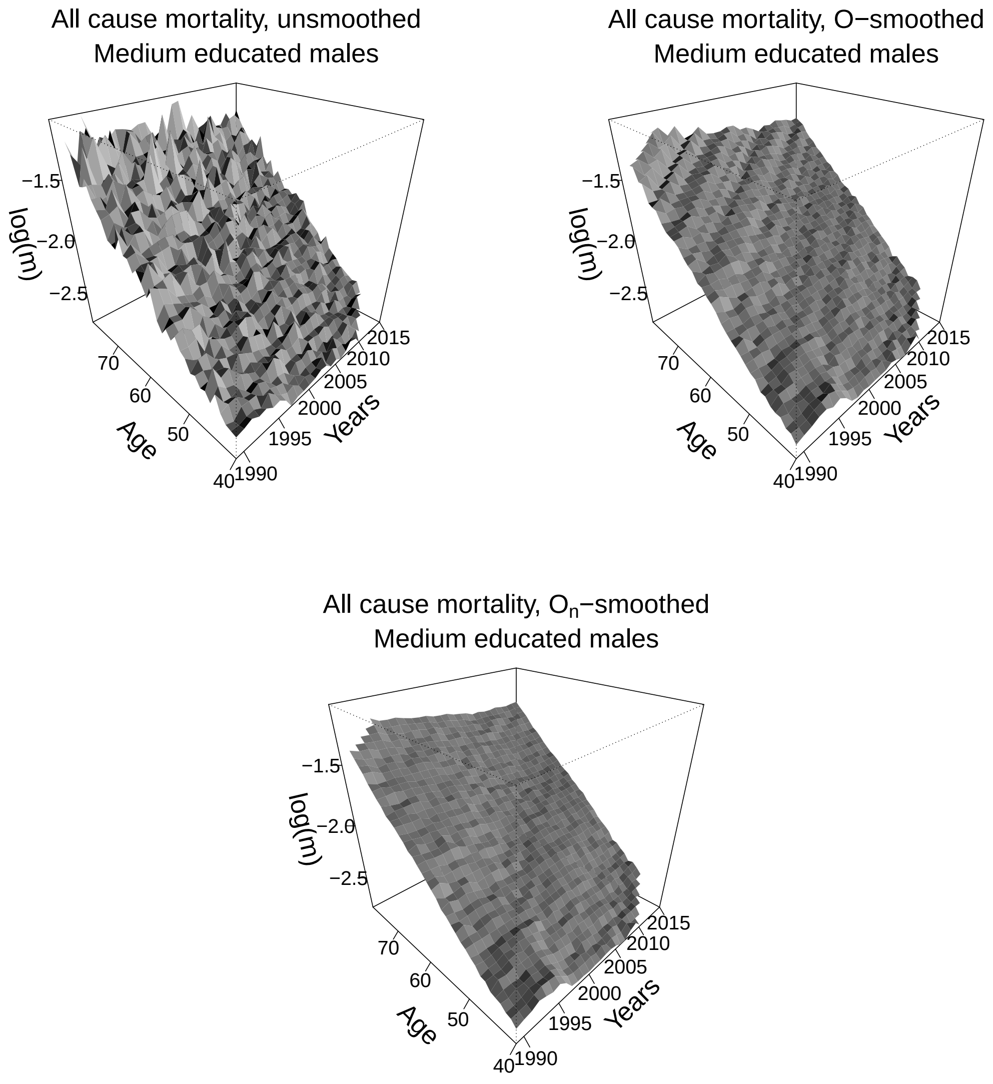


Figure 6: Death rates for medium educated males using the unsmoothed ratios (top left), smoothed ratios obtained minimising $\mathcal{O}$ (top right), and smoothed ratios obtained minimising $\mathcal{O}_N$ (bottom).

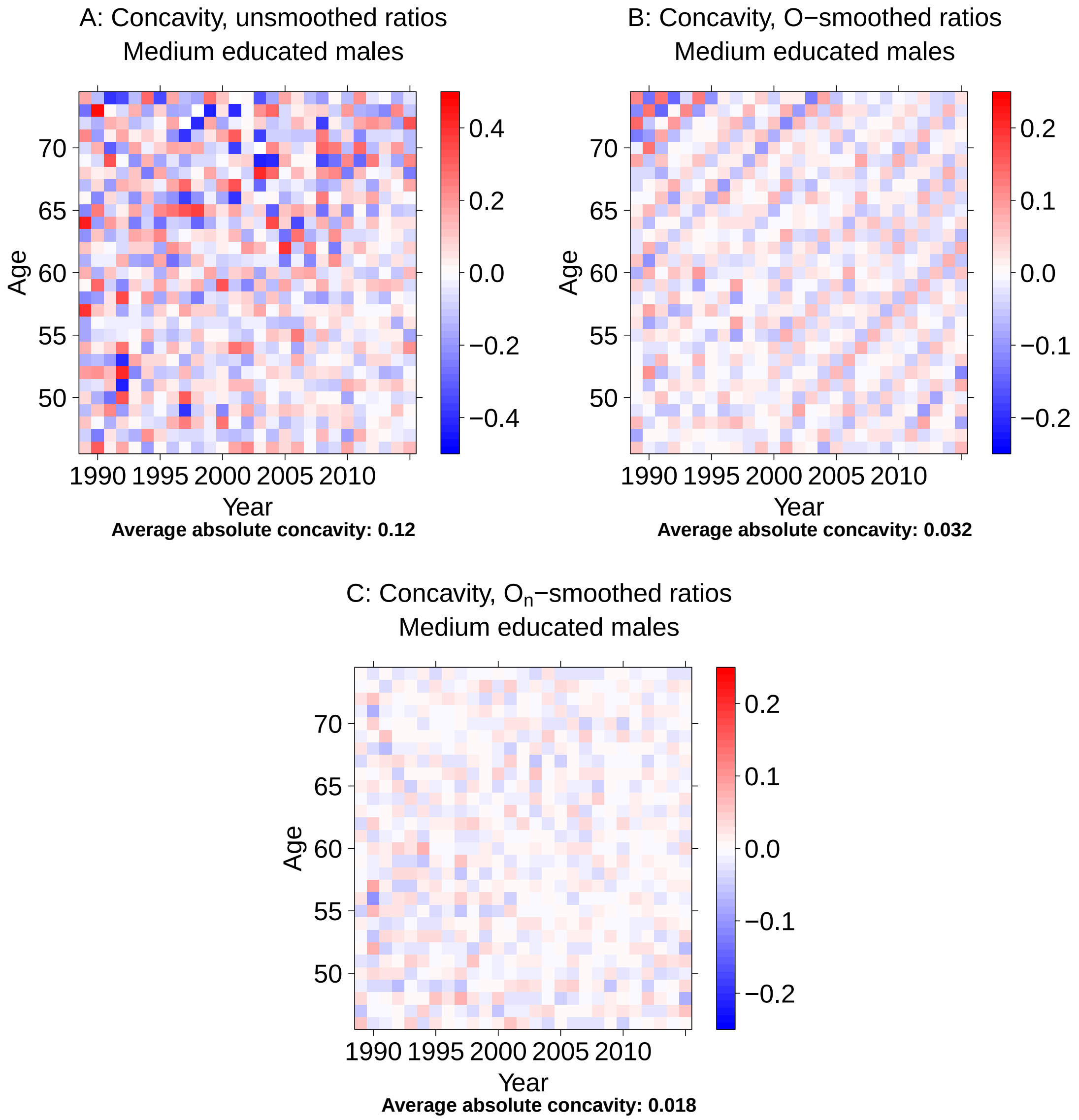


Figure 7: Heat maps of the concavity of the log death rates for medium educated males.

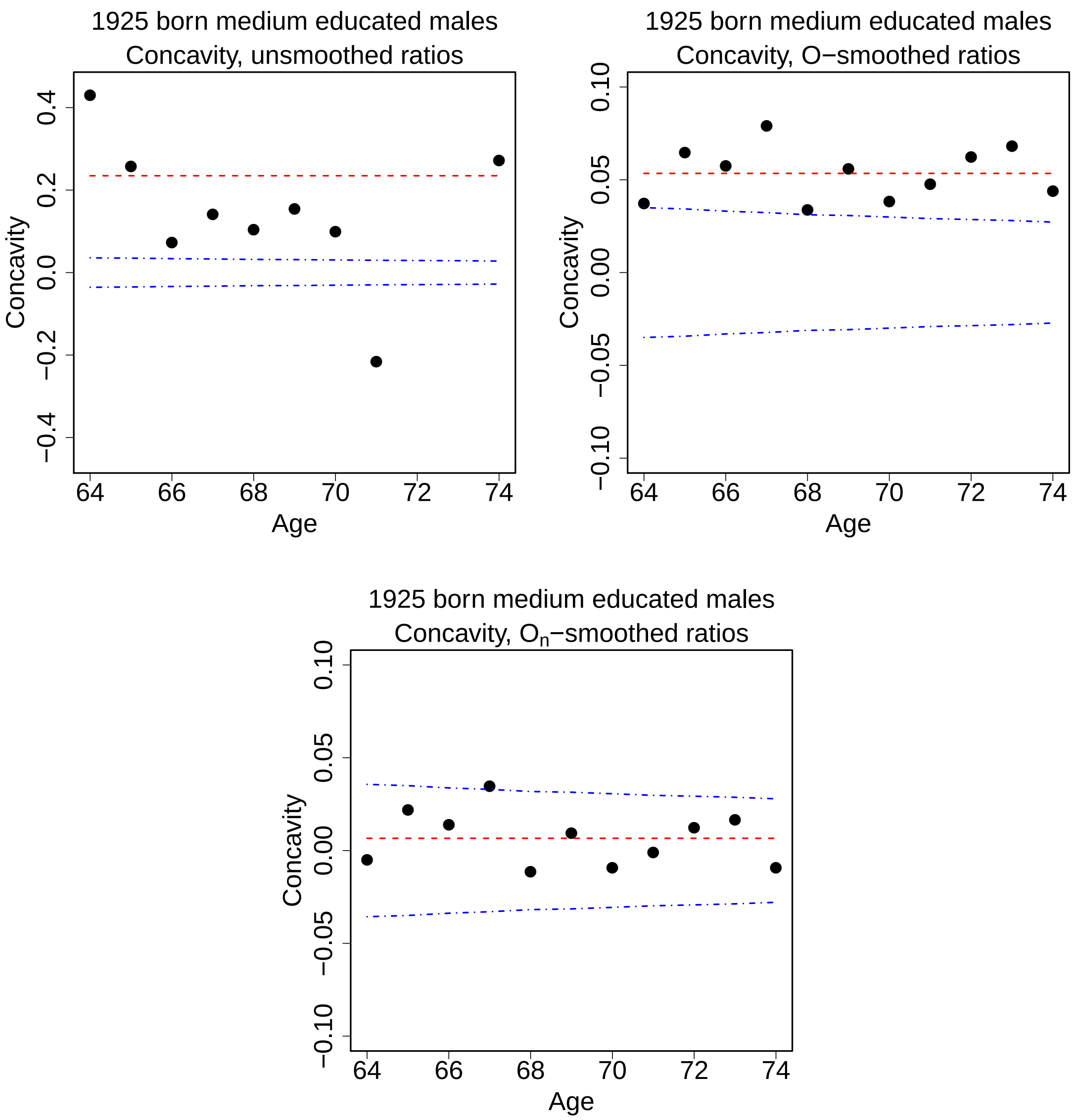


Figure 8: Concavity of the log death rates for the medium educated males born in 1925 (dots), along with its average (red line) and the $\pm\sigma$ interval around zero (blue lines), assuming underlying linearity.

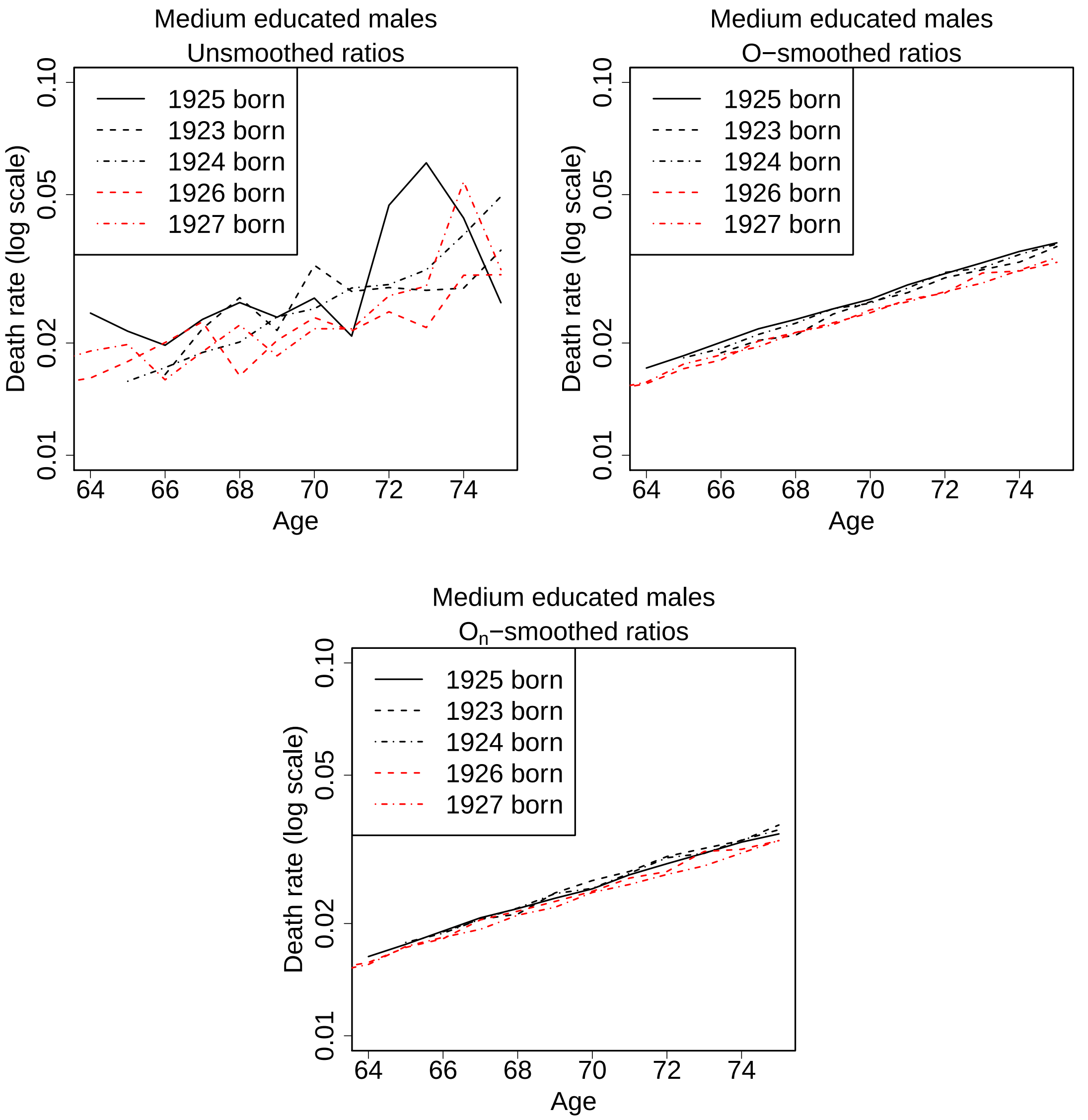


Figure 9: Death rates for the 1925 medium educated males cohort, along with the two previous (dashed black) and two next (dashed red) cohorts using unsmoothed, $\mathcal{O}$-smoothed and $\mathcal{O}_N$-smoothed ratios.

is higher than for its neighbours, which would be due to an underestimation of the exposures (underestimation of the initial ratio, $\hat{R}_0^{(e,c)}$). Indeed, when unsmoothed ratios are used the death rate for the central cohort is not only higher than for its neighbours, but has also some big jumps at higher ages, as expected from such noisy death rates as the ones seen in figure 9. When the $\mathcal{O}$-smoothing is applied to the ratios, death rates for each cohort become smooth independently. That eliminates all the big oscillations in both the death rates and the concavity, but the rate for some cohorts might be pushed "out of place". This could easily happen if just a few points in a ratio time series are systematically lower than their "true" value, either because of some systematic bias of the sample used in the CPS or due to pure random variation, with the latter probably playing a big role for cohorts with few observations. In particular, for the cohort we are studying, we see that for ages 73 and 74 the unsmoothed death rate is much higher than at any other age for any of the cohorts shown. It is probably due to this that $\mathcal{O}$-smoothing estimates a death rate that is higher at all ages for this specific cohort than those of its neighbours, since these "rogue" points are driving the whole curve up. However, and as expected, the older 1924 and 1923 cohorts have death rates higher than the 1926 and 1927 cohorts.

And here is when the "interaction" term in $\mathcal{O}_N$ becomes justifiable: we cannot let a few unusual data points affect our final results so heavily. By adjusting our ratios not only on a cohort by cohort basis, but comparing with their neighbours, strange oscillations like the ones seen at high ages for the medium educated 1925 males will be smoothed out to behave in a way closer to the one observed for similar cohorts. This is a reasonable expectation as long as we don't have any reasons to believe that the big jumps in the death rates are real and not a problem with the exposures (for example, a period effect that is also seen at the same calendar year in neighbouring cohorts; however, from figure 9, this does not seem to the case here). The bottom panel of figure 9 shows how, when $\mathcal{O}_N$-smoothed ratios are used, the death rate for the 1925 cohort has been "brought down" to be in a more reasonable position, quite close to the red and black dashed lines.

Now that we have produced smoothed education ratios we can test wether or not the observed, unsmoothed ratios from the CPS are compatible with them by means of statistical tests. We will briefly explain three different tests that can give us information about how good the smoothed ratios are and then discuss the results obtained from their application to our problem.

The goal of the tests is to check how the CPS results compare to a binomial distribution with number of trials equal to the number of people interviewed and probability of success the smoothed ratios we calculated. Tests will be implemented following the cohort dimension, i.e., for each cohort (people with the same gender, education level, and year of birth) we have an unsmoothed trajectory and a smoothed one (the CPS and $\mathcal{O}_N$-smoothed ratios of people of that gender born in that year that achieved a certain level of education, respectively) that we will try to analyse.

The first test we will use is the likelihood ratio test. We implement this test by computing the deviance for each cohort, which is $D_C = 2(l_1 - l_2)$, with l_1 being the log-likelihood of a saturated model and l_2 the log-likelihood of the model we want to test (see, for example, Macdonald et al. (2018)). For a binomial distribution we have:

$$D_C = 2\sum_i \left[c_c(i,c;e)\log\left(\frac{c_c(i,c;e)}{c_c(i,c)\hat{R}_C(i,e,c)}\right) \right.$$
$$\left. +(c_c(i,c)-c_c(i,c;e))\log\left(\frac{c_c(i,c)-c_c(i,c;e)}{c_c(i,c)-c_c(i,c)\hat{R}_C(i,e,c)}\right)\right], \quad (14)$$

where $c_c(i,c;e)$ is the number of successes (number of people that reported the education level e in the interview), $c_c(i,c)$ is the total number of people interviewed, and $\hat{R}_C(i,e,c)$ is the smoothed ratio. The C subscript means that all these quantities are defined for cohorts, and the sum over i that we are aggregating over the cohort dimension, similar to the sums in equations (6) and (8). Now we assume this test statistic D_C follows a χ^2 distribution with as many degrees of freedom as data points we have for the cohort minus one (since the initial ratio $\hat{R}_C(0)$ was fitted using the unsmoothed ratios given by $R_C^{CPS} = c_c(i,c;e)/c_c(i,c)$). This will give us a p-value for the probability that, if the smoothed ratios $\hat{R}_C$ were representative of the whole population, unsmoothed trajectories like the ones in the CPS, R_C^{CPS}, have a deviance like the one observed or greater. If this p-value is small, then it is not likely that our smoothed ratios are close to the true ratios for the whole population.

The second test we will use is the signs test. We compute the residuals $r_c(i) = \hat{R}_C(i) - R_C^{CPS}(i)$, the difference between the CPS ratios and the smoothed ratios for each cohort c at all times observed, i. We then check how many of these residuals are non-negative, $r_c(i) \geq 0$. If the model fits the data well, we expect the probability of finding non-negative residuals to be equal in practice to the one of finding negative ones (the probability of a residual being exactly zero being negligible). Therefore we can calculate the probability that, in a binomial distribution with success probability 0.5 and number of trials equal to the data points for the cohort, we have as many successes (non-negative residuals) as we observe or more, and check if we have too many or too few non-negative residuals compared with the expectations. In this case, very small or very large p-values (representing too many and very few non-negative residuals respectively) would mean that there is a bias in the smoothed ratios, which would be systematically above or below the unsmoothed curve.

Finally, we have the runs test. The runs test starting point is the sign of the residuals calculated for the signs test. We want to check how many times the sequence of $r_c(i)$ for a cohort changes signs along the trajectory (because we are dividing the residuals in negative and non-negative, any residuals equal to zero are considered to have positive sign; however, as mentioned earlier, the probability of this happening is negligible). This number will be one less than the number of runs of residuals with the same sign, u. If the residuals are truly random we don't expect them to show any structure. This structure would be represented by clusters of contiguous data being systematically above or below the smoothed curve. Therefore the test should check whether the number of runs observed is too small and not compatible with random ordering. The equations for the probabilities of observing a specific number of runs, $P[U = u]$, which can be

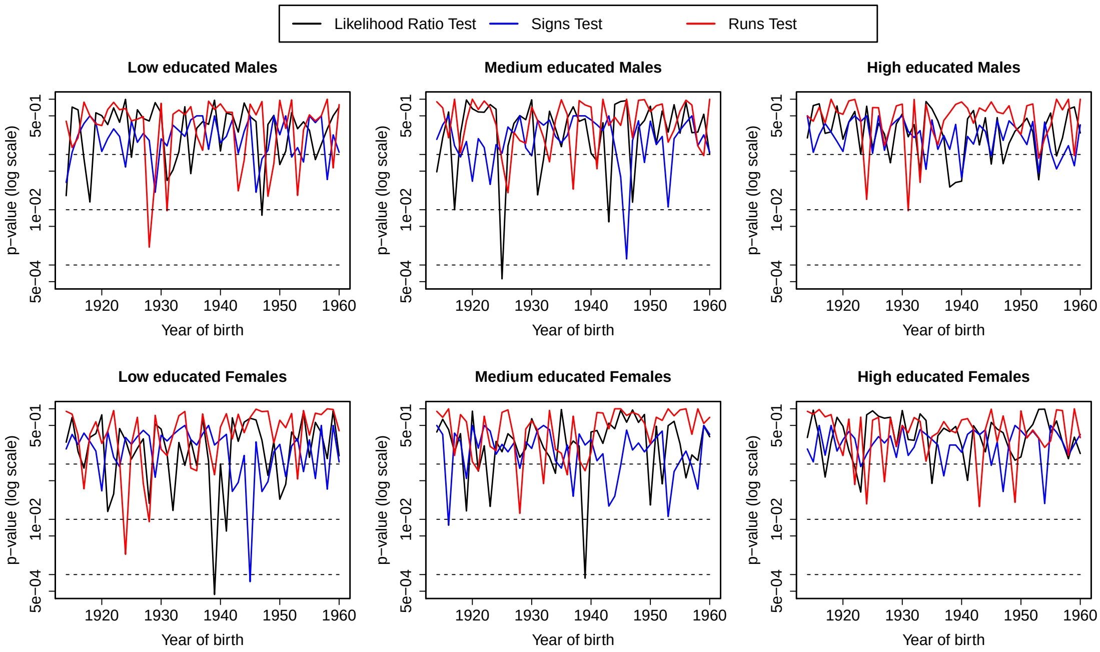


Figure 10: p-values obtained from applying the three tests to the ratios of educated people on a cohort by cohort basis, in logarithmic scale: likelihood ratio test (black lines); signs test (blue); runs test (red). The 90%, 99%, and 99.9% significance levels are marked by the dashed lines.

consulted in Macdonald et al. (2018), can be used to compute $P[U \leq u]$, the probability that the number of runs in a random sequence is less than or equal to our observation. This will give us a p-value representing how likely we were to find such ordered residuals given the number of positive and negative signs observed.

The p-values for all tests can be seen in figure 10. For each gender-education group we plot the p-value obtained using the three tests against calendar year, in logarithmic scale. The horizontal dashed lines mark the 90%, 99%, and 99.9% significance levels. Note that, due to the symmetric nature of the signs test (we would reject the null hypothesis if the p-value is either too big or too small), we do not plot p for this test, but $\min(p, 1-p)$. From the figure it is obvious that in most cases the tests cannot reject the null hypothesis that our unsmoothed ratios were produced by true population ratios given by our smoothed results with 99% confidence. We are analysing a total of 47 years of birth, 2 genders, and 3 education levels, which makes up a total of 282 cohorts. On these we are conducting 3 tests, so the number of p-values plotted in figure 10 is 846. It is perfectly understandable then that some of these p-values are smaller than 0.01 simply due to high number of tests conducted. However, it is encouraging that for not even a single cohort more than one p-value is smaller than 0.01 at a time, i.e., for cohorts with a relatively high difference between the unsmoothed and unsmoothed ratios (which therefore fail the log-likelihood ratio test) there is no systematic bias in the smoothed ratios (otherwise the p-value for the signs ratio would be small too); and in a few cohorts for which the p-value given by the signs or runs test is small (which could mean our smoothed ratios are biased) the p-value of the log-likelihood ratio test is relatively high (therefore, if this bias truly exists it must be really small or the deviances would be high). From this picture it looks reasonable to assume that our smoothed ratios are a good representation of the whole population.

5 Extrapolation to higher ages

Up to now we have restricted the maximum age in our analysis to 79. In fact, due to the smoothing procedure, and because the CPS only provided estimates of the population by single year of age up to age 79 in the 1989-2015 period, the oldest cohort included was the 1914 birth cohort. This choice ensured that for the oldest cohort analysed (whose members would be aged 75 in 1989) we had at least 5 CPS data points to which we could fit our smoothed ratios. The choice of the starting age was arbitrary, but age 40 was chosen so that we can expect the death rates to increase exponentially for the whole range of ages considered (otherwise the vanishing concavity condition would not be applicable). Fortunately, our recurrence relation for the ratios, equation (3), means we have a systematic way to extrapolate the ratios of educated people in a cohort up to any age for which we have death counts (separated by education level) and total exposures by single year of age. The CDC data and HMD respectively give us that information for ages beyond 100.

Extrapolating is obviously useful on itself as a way to extract information about mortality at high ages, but it will also prove to be important as a test of the quality of the results. A small over- or under-estimation of the initial ratios of educated people, or any issues with the quality of the data, could potentially be hidden at relatively young ages when exposures are big and mortality low. However, any such issue will have a very significant effect at very high ages, when the cohorts are almost extinct. As we will see the results of our extrapolation will force us to further modify our analysis due to concerns about data quality.

The extrapolation process is almost trivial to apply: once we have calculated our smoothed ratios of educated people for any range of ages we use equation (3) to estimate those ratios beyond the highest age for which we have CPS data. We can do this because the HMD gives us the $E_i(c)$ and the CDC data the $\Delta_i^{(e,c)}$ we need up to age 109.

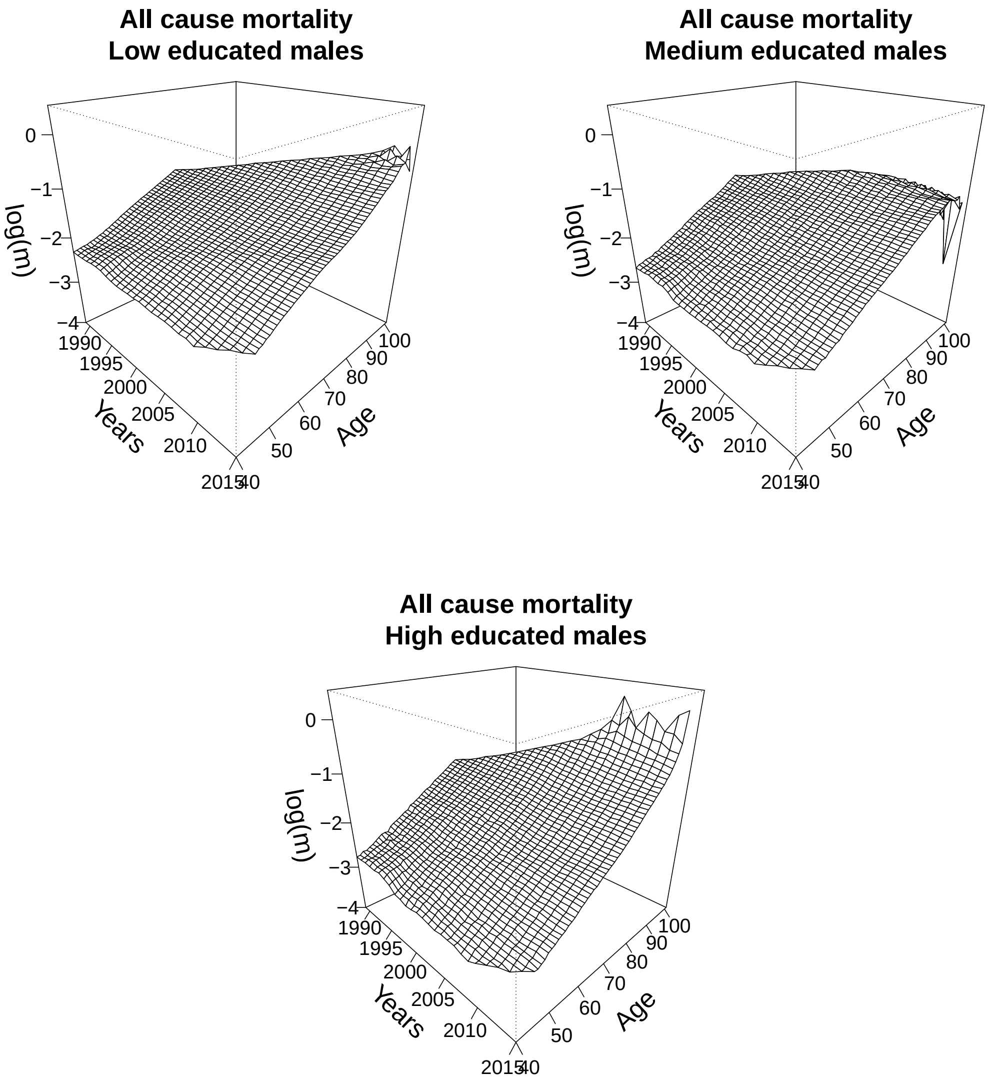
The resulting death rates for males are shown in figure 11. For each cohort we have extrapolated the death rates up to the highest age possible (i.e., the age of that cohort in the calendar year 2015). Simple visual inspection shows that there is some problem in the extrapolated area. For the low educated group we see a fast increase in mortality at very high ages and a noisy pattern, both of which are expected when we analyse ages close to 100. But the medium and high educated groups have a completely absurd behaviour. The mortality for group 2 reaches a *plateau* at very high ages and stagnates at a relatively low value, whereas the mortality for group 3 explodes at those same ages and even surpasses that of group 1.

This strange behaviour is extremely obvious if we plot the mortality of a cohort as it ages. A few examples are shown in figure 12. We see that for some cohorts the mortality of group 2 stops increasing, or even decreases, at very high ages; conversely, the mortality of group 3 grows very quickly as the cohorts age. For the first three cases the mortality of group 3 overtakes the mortality of group 2 around or before age 80, and in the last one both mortalities are quickly converging by that age (note the different scale in the x axis for each cohort). This can be caused by either an underestimation of the exposures or an overestimation of the number of deaths.

Something that can help us understand where the problem lies is comparing our smoothed ratios with the ACS data. If the problem is a systematic underestimation of the ratio of people in group 3 (with the corresponding overestimation of group 2) then this should show up clearly as a discrepancy between our smoothed ratios and the relatively noise-free ratios of the ACS. This comparison is shown in figure 13. Because of the restricted availability of ACS data we can only do the comparison for two of the cohorts shown in figure 12. Even with that the results seem to be clear: we are not underestimating the number of people in group 3. If there is no systematic bias in our smoothed ratios then the only reason why mortality would grow so quickly at high ages would be the misreporting of education in death certificates.

Misreporting of education in death certificates has been well studied, since comparability with other data sources was a key element to take into account when this new item was introduced. Rostron et al. (2010), building on the work of Sorlie & Johnson (1996), contains very valuable information regarding comparability between education reported in the CPS and in the death certificates. Thanks to the National Longitudinal Mortality Study (NLMS), death certificate information for CPS respondents from 1992 through 1998 who died during that period could be compared with their interview responses. With that data in hand the authors compare both sources of information on educational attainment of the individuals.

Of course, the first thing they note is that "[...] decedents without death certificate education information generally had lower educational attainment in the CPS than decedents with this information. For example, 49% of decedents without death certificate education information did not graduate from high school according to the CPS compared with 35% of decedents with death certificate education information". This means that our simple imputation, in which we assume CDC certificates are randomly missing education, is not completely accurate. People with lower education are more likely to have an "unknown" education in their death certificate than higher educated people. However, we have modified our imputation method to analyse two extreme cases (namely, deaths with unknown education are all ignored or all assigned to group 1) and neither of them changes the qualitative behaviour of the death rates for groups 2 and 3 seen in figure 12. This means that the problem with the excess of deaths in group 3 does not come from



All cause mortality
Low educated males

All cause mortality
Medium educated males

All cause mortality
High educated males

Figure 11: Death rates for males of all three education groups extrapolated to the highest ages possible.

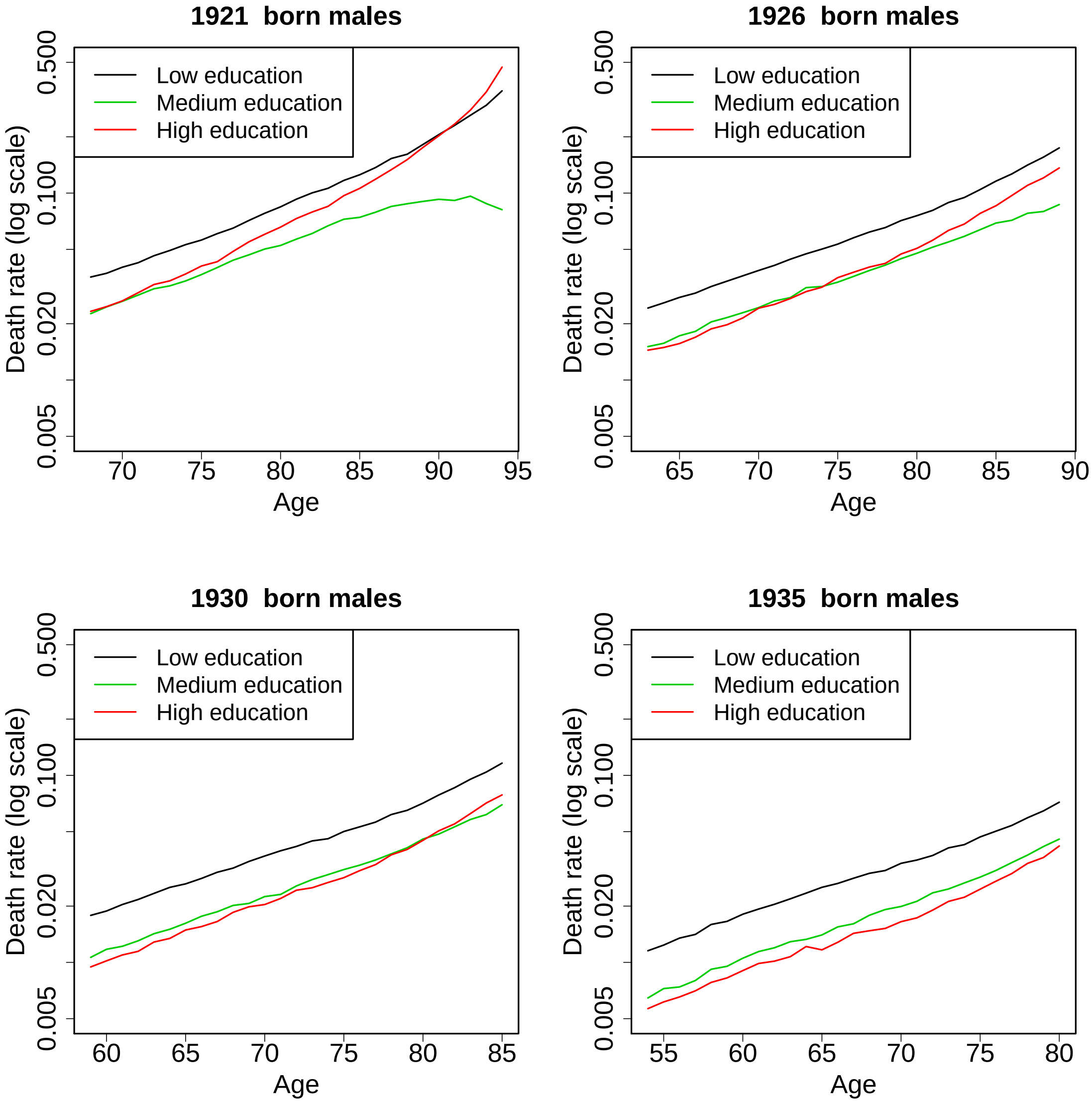


Figure 12: Death rates for males of different cohorts for all three education levels.

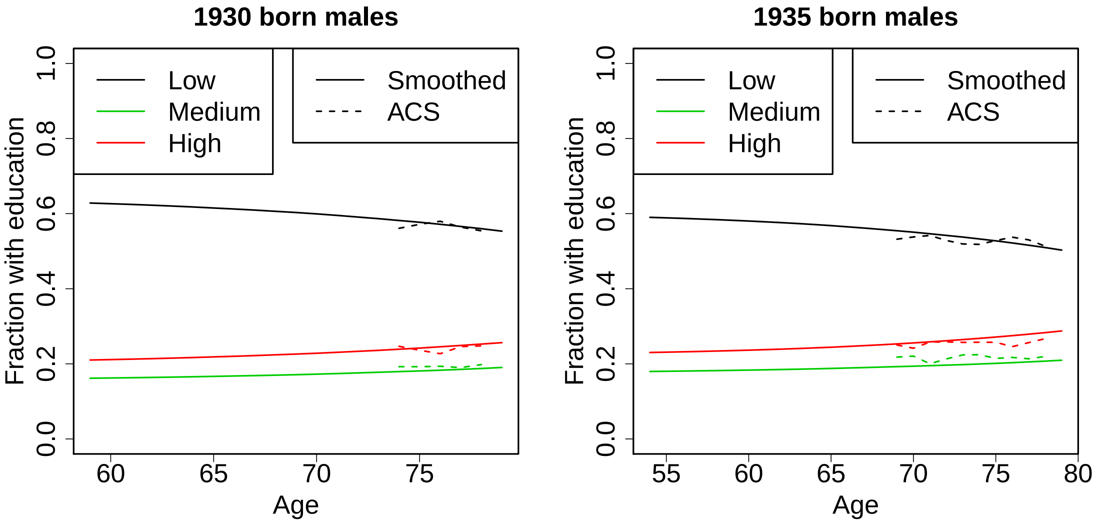


Figure 13: Comparison between raw ACS and smoothed CPS ratios.

the imputation procedure.

The authors also explore the problems of correspondence between the degree-based and years-of-schooling-based systems. Thanks to this they produce results that are very relevant for us: figures 1 and 2 of their paper show how their adjustment for misreporting corrects death rates as obtained from deaths reported using the 1989 and 2003 death certificate standards. For both males and females, regardless of how education was recorded, the corrected life expectancy at age 25 for high school graduates is bigger than before correction, whereas in turn the life expectancy for people without high school completed is smaller. This is caused by funeral directors writing a decedent's education down as "completed high school" when the person had never achieved their degree, as it is a very well known bias. However, we also see that the life expectancy for people with Bachelor's degrees or Graduate degrees (which together form our education group 3) as calculated using death certificates in the 1989 standard is increased when the correction is applied. However, life expectancy for people with some college is decreased by the correction. This is compatible with our observations: raw death counts of people with Bachelor's degrees or more are overestimated, whereas deaths of people with some college but no degree seem to be underestimated. For death certificates with the 2003 standard the authors chose to use three different education categories. In this case their "more than high school graduate" category, which corresponds to our groups 2 and 3 merged together, seems to be quite consistent.

There is another easy way to further diagnose our data. When every member of a cohort we are observing has died, the number of people who died in each education group divided by the total number of deaths should be equal to the initial number of people in each group divided by the initial number of people. This is, the quantity

$$R_C^D(i,e,c) = \frac{\sum_{j=1}^{i} D(t_0 + x_0 + j, x_0 + j, e, g)}{\sum_e \sum_{j=1}^{i} D(t_0 + x_0 + j, x_0 + j, e, g)}, \tag{15}$$

which measures the proportion that each education group has contributed to the total number of deaths in a cohort from the point they enter our analysis up to a time i, should converge to the initial value of the smoothed ratios $\hat{R}_C(0)$ as i increases ($D(t,x,e,g)$ is the number of deaths, as defined in section 3). Figure 14 contains plots of this quantity for certain cohorts. Solid lines represent how much each education group has contributed to the total number of deaths up to that age, dashed lines are the smoothed ratios of educated people, and the horizontal dot-dashed line marks the value of the initial ratio for each education group. Solid lines should approach the horizontal line as the cohort ages, but crossings are not expected and mean problems with the data.

The top two panels show what happens with the 1921 cohort. As expected, because lower educated people die younger, they make the bulk of the deaths at younger ages (the black solid line starts much higher than the dot-dashed line, whereas for the two other groups the opposite is true). As the cohort ages all three solid lines start approaching their theoretical limiting values, but around age 90 the solid line for group 3 has crossed over the dot-dashed line, whereas the solid line for group 2 seems to be getting stagnant still far from its supposed value at very high ages. This happens in both panels, which means the problem is not being introduced by the education imputation procedure. As we saw, our smoothed ratios are comparable to values extracted from other sources, which means the problem is likely to be people of group 2 being misreported as having a Bachelor's degree or more in their death certificates.

The two lower panels of figure 14 show what happens with two younger cohorts. Although the crossover has not happened yet for these, the tendency of the red line to approach its supposed asymptotic value much faster than the two other solid lines is clearly visible, and especially in the 1925 cohort (bottom left) it is clear that the crossing will eventually happen. This shows that the issue always appears at very high ages, not only for the older cohorts.

For all of this, we should conclude that it is the quality of the death counts data that is limiting the validity of our results. One possible solution, seeing how the results for the lower educated group are consistent in our analysis (the black solid lines do indeed approach its theoretical asymptote), and the conclusions of Rostron et al. (2010), is to combine the two higher educated groups and divide the whole population in just two categories, "never went to college" and "at least some college completed". For the cohorts shown in figure 12 we obtain the new death rates plotted in figure 15. In these there is still a clear hierarchy between the education groups which is maintained at all ages. Therefore mixing groups 2 and 3 in a single, broad higher educated group lets us obtain reliable results up to very high ages.

One last concern is how robust these results are. If our procedure is robust, then small changes to the raw data or the input parameters would not affect the results. However, if they produce a significative change in the final results then they should not be trusted.

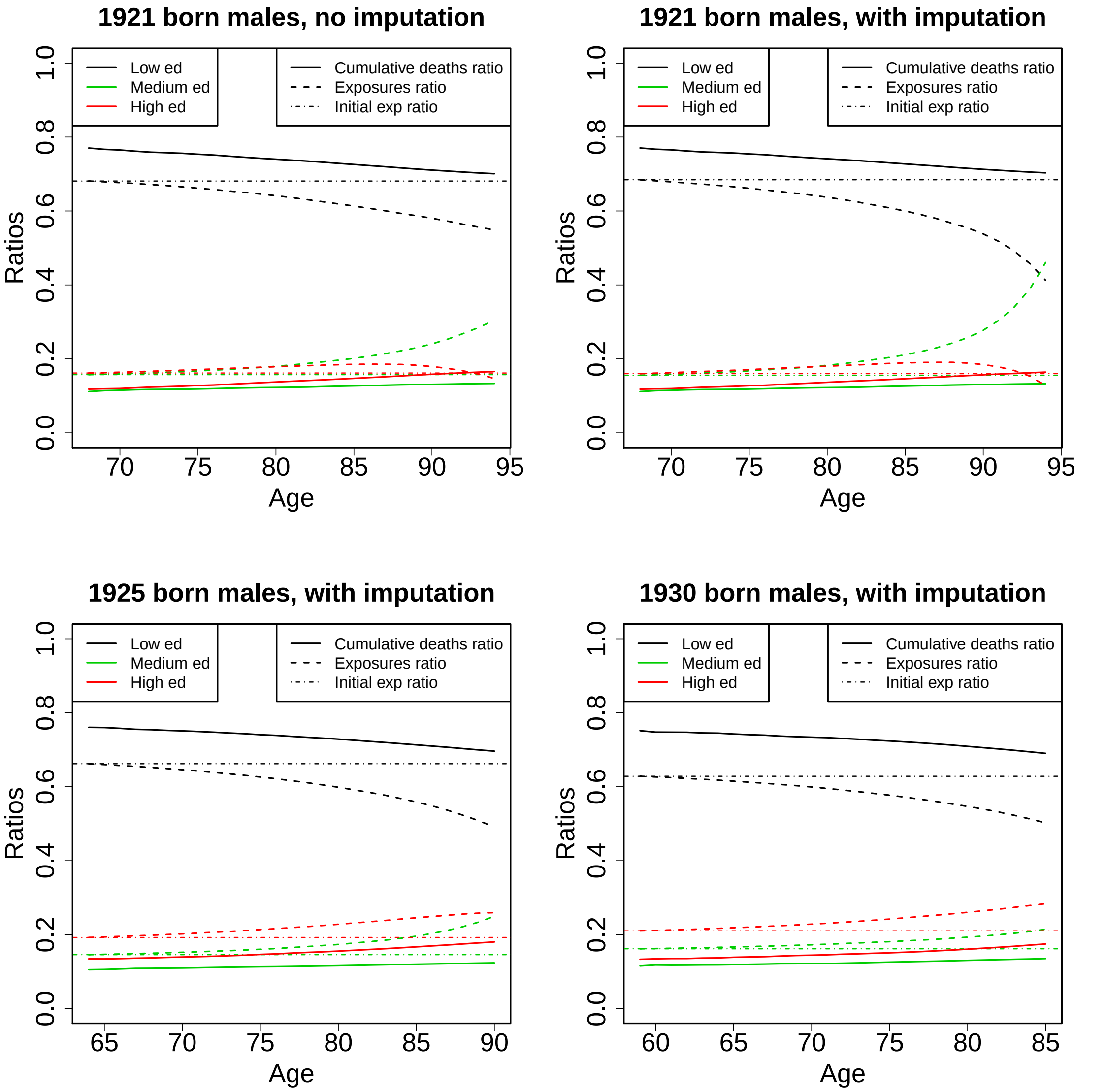


Figure 14: Figures comparing the ratios of cumulative deaths with the initial ratios of educated people.

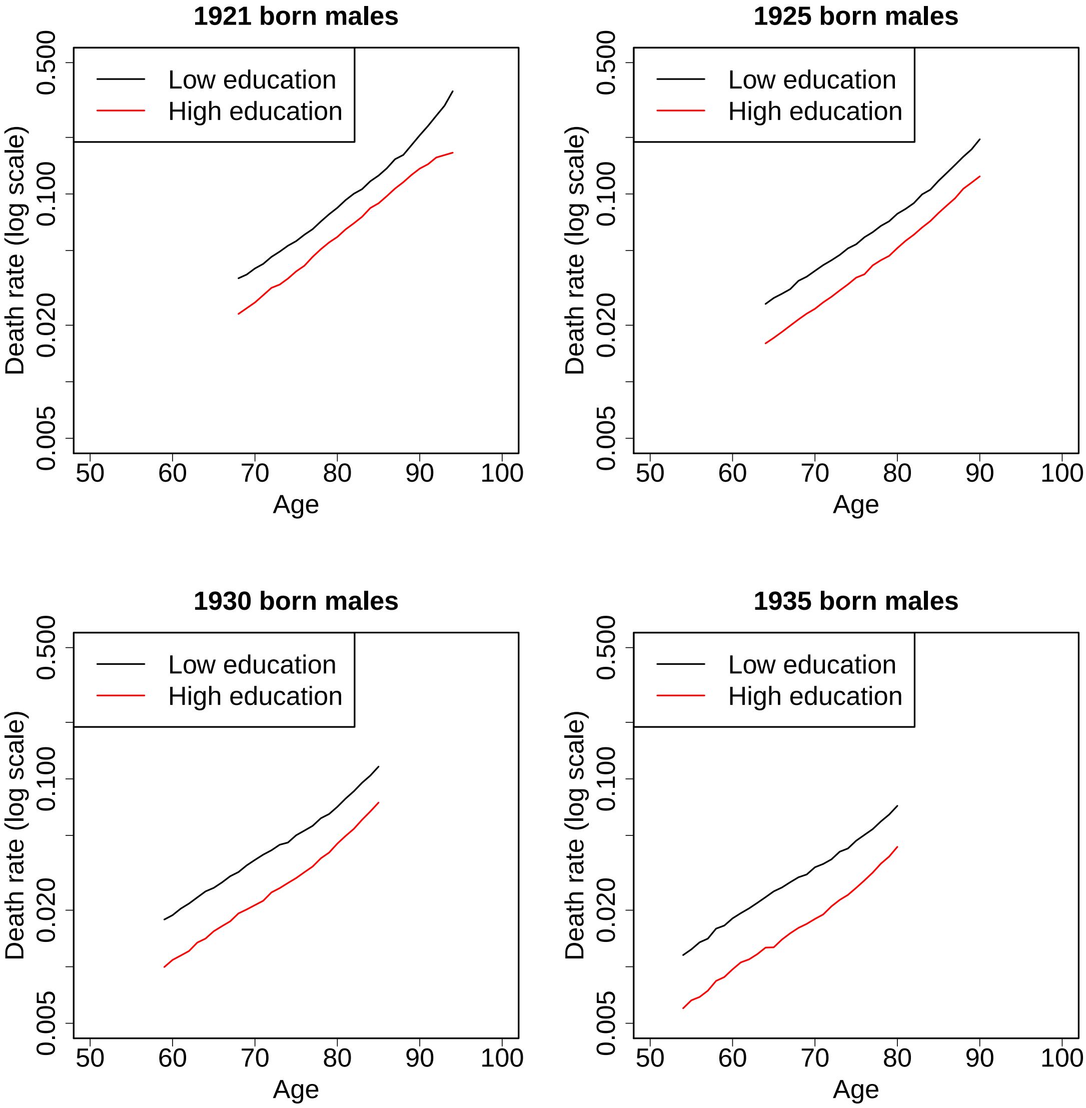


Figure 15: Death rates for males of different cohorts for the new two education levels.

If we have any robustness problems we would expect to find them specially in the oldest cohorts, the ones for which very few data points are available in the CPS, since the lack of data for long periods of time would make them more vulnerable to small changes in the data available and its treatment.

We introduce two changes in our analysis. Firstly, we move the starting year from 1989 to 1995. This will change the number of data points available for some of the oldest cohorts. We will also change the weights in our smoothing procedure, that we had (perhaps naively) overlooked so far. Instead of setting all of the ω_i in equation (9) to a constant we will use the estimated error in the unsmoothed ratios (given by equation (10)) to give a weight to each point. We will make the obvious choice $\omega_i = \sigma(\hat{R})^{-2}$ and see if this weighting affects the results obtained or not.

In figure 16 we see how the death rates for certain cohorts change as we change the method used to calculate them. The solid lines are the death rates when $\omega_i = \sigma(\hat{R})^{-2}$, and the dashed lines when ω_i =constant. For the cohorts for which it is possible (1921 and 1925) there are two sets of lines, one of which starts at a later age. These second death rates are obtained starting the analysis in 1995.

It is obvious that we have a problem with robustness in the 1917 and 1919 cohorts. It is also clear that the problem is only visible at high ages. As expected, small changes in the initial ratios of educated people create “trolls” or “phantoms” (as defined in Cairns et al. (2016)) that do not manifest themselves until the total cohort membership is very small. We also see how for younger cohorts the results seem to be robust for a wider range of ages: for the 1917 cohort we already have a noticeable discrepancy between the different death rates at age 90, whereas for the 1925 cohort there is barely any difference between them. There are two possible causes: one is that younger cohorts maintain higher memberships for longer (i.e., not as many people live for 90+ years if born in 1917, compared with people born in 1925); another difference is that, because the 1917 cohort enters our analysis at age 1989-1917=72, we only have 8 points to which we can fit our smoothed ratios, whereas for the 1925 cohort (which enters the analysis at age 1989-1925=64) we can use 16 unsmoothed ratios in the fitting procedure. More data points mean of course a better, more reliable fit.

6 Summary and Future Research

With the procedure described in this paper we have managed to obtain death counts, exposures, and, combining the two, death rates for the US population born between 1914 and 1970 for years 1989-2015, separated by educational attainment and cause of death. Even though some concerns remain about the reliability of our results at very high ages for the oldest cohorts, as discussed in the last paragraph of the previous section, some important insights have been gained thanks to the final data obtained.

Figure 17 shows the death rates for males and females at certain fixed calendar years. From these plots we can see that the gap in all cause mortality between different education groups exists at all ages analysed. It is also worth noting that the gap is wider at younger ages, meaning that education plays a bigger role in early mortality than it does

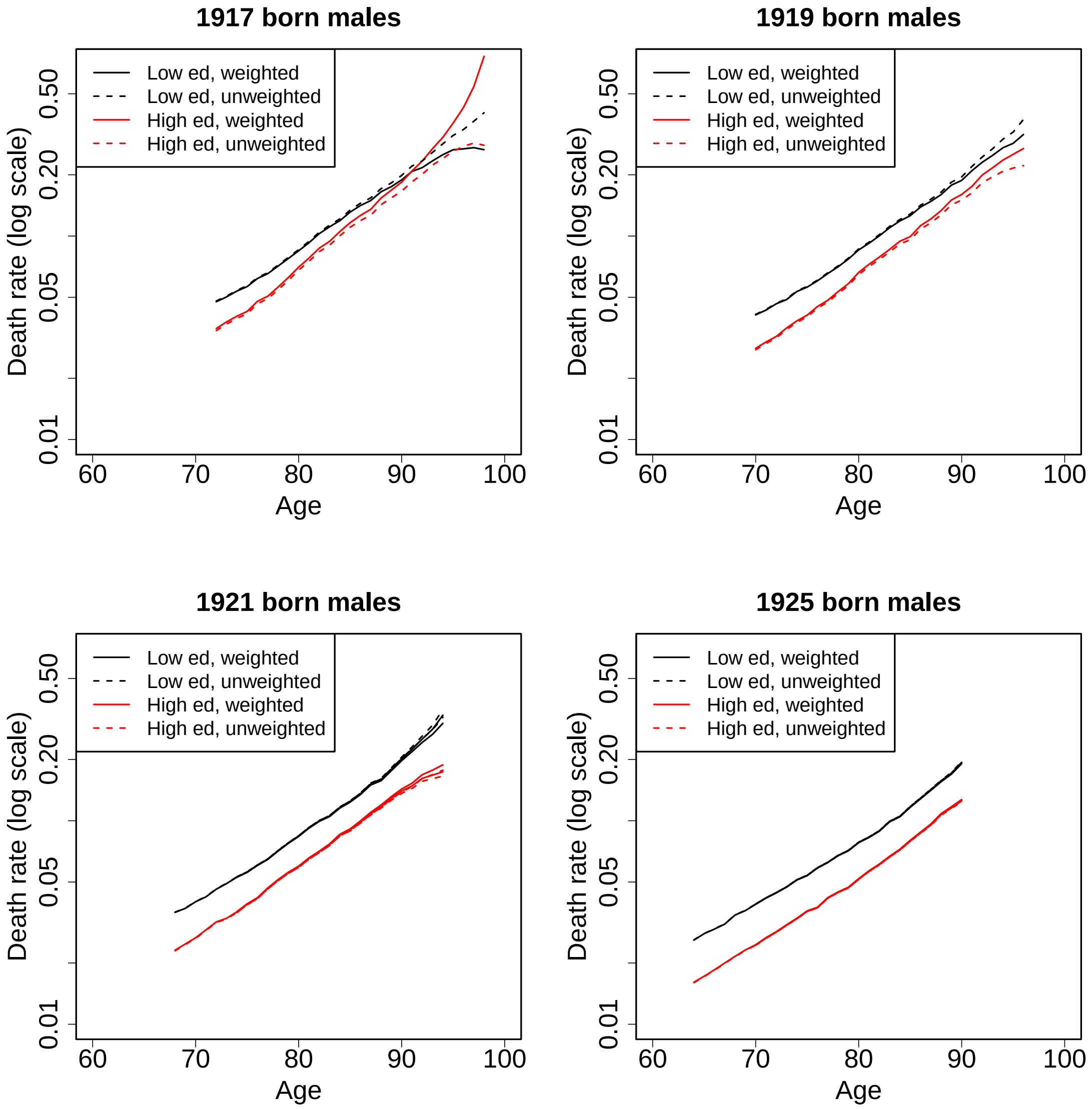


Figure 16: Death rates for males of different cohorts calculated in four different ways (1989-1995 as starting year, and weighted-unweighted fit of the smoothed ratios).

at higher ages.

Another interesting result can be seen in figure 18. In this figure we plot death rates for males and females at a fixed age against calendar year. For both genders we see that, in the period 1989-2015, there has been an increase in the mortality gap between the two education groups. This is mostly driven by a stagnation in the death rates of the lower educated group, which have been almost constant over the period of our anlaysis, while a mortality improvement is clearly seen in the higher educated population.

As possibilities for further research, a more thorough analysis of this data could highlight which causes of death are behind the trends observed, potentially giving some insight on which are the drivers of the diverging mortality scenario for different education groups. This data will also be useful in comparisons of total and cause of death mortality between different countries, since its granularity (death rates by single year of age and calendar year) offers greater flexibility than the age-standardised rates available elsewhere.

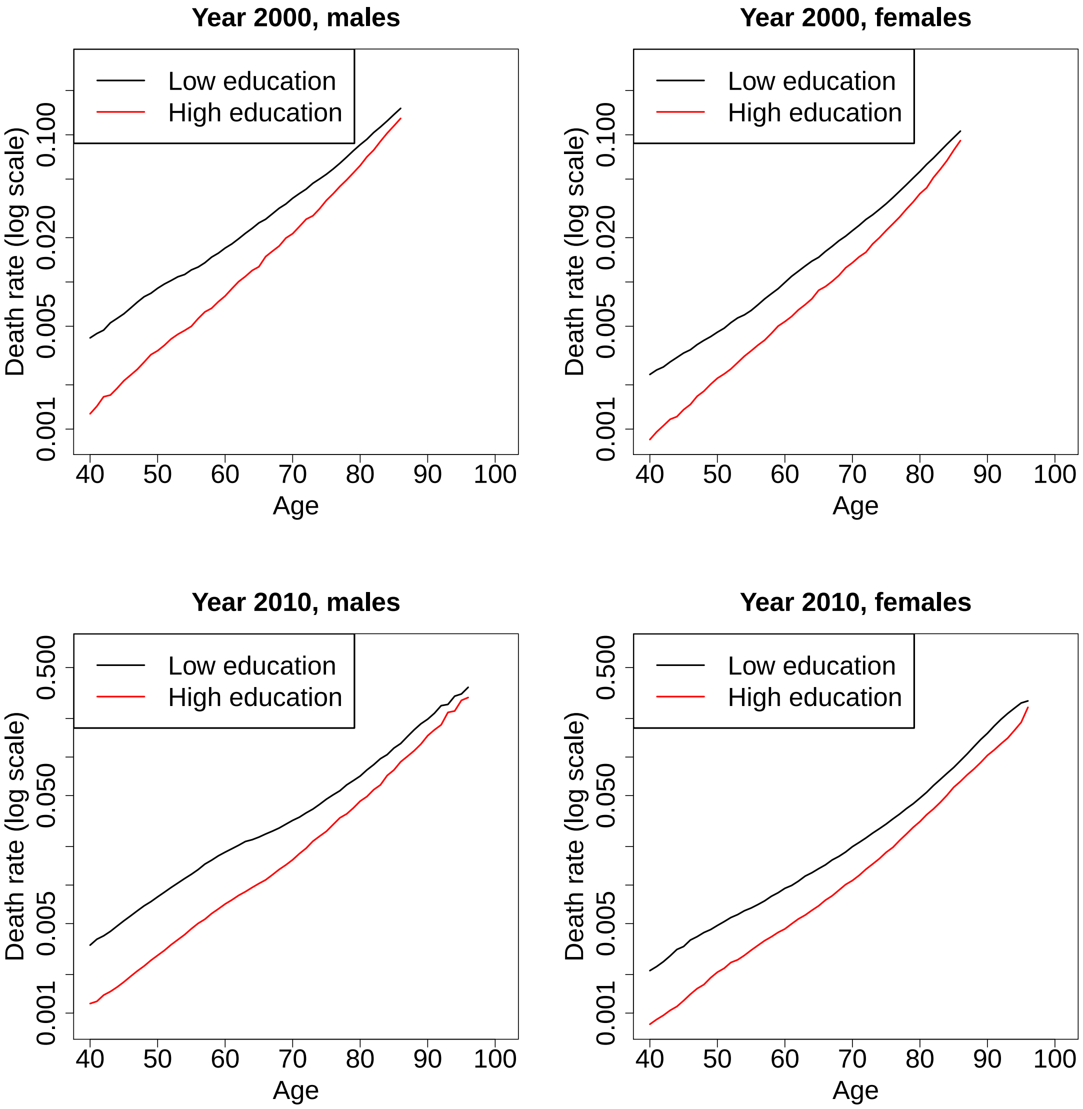


Figure 17: Death rates for males and females of both education groups against age at different calendar years.

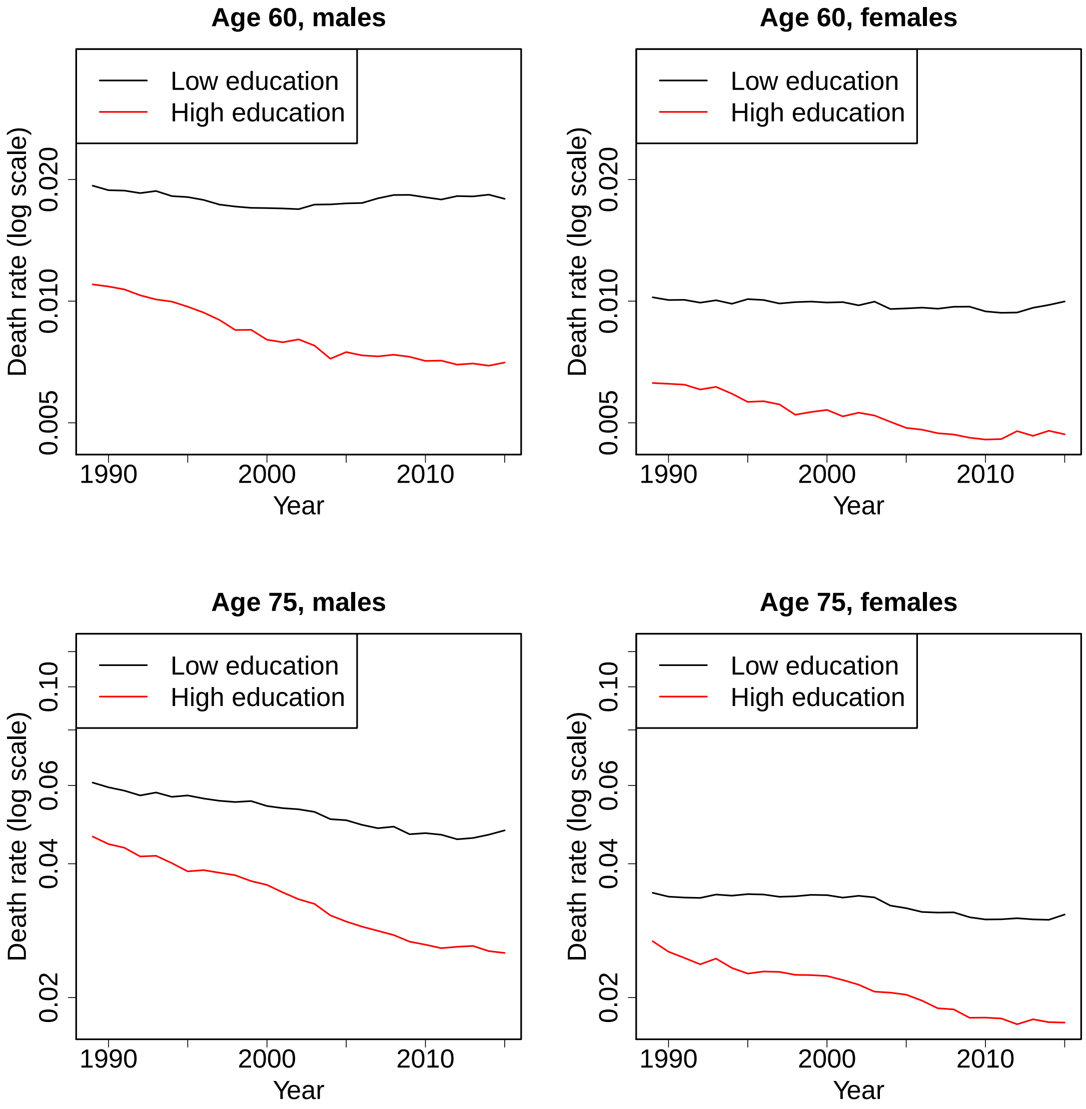


Figure 18: Death rates for males and females of both education groups against calendar year at different ages.

Bibliography

Anderson, R. N., Minino, A. M., Hoyert, D. L. & Rosenberg, H. M. (2001), 'Comparability of cause of death between icd-9 and icd-10: Preliminary estimates', *National Vital Statistics Reports* **49**(2).

Cairns, A. J. G., Blake, D., Dowd, K. & Kessler, A. (2016), 'Phantoms never die: Living with unreliable population data', *Journal of the Royal Statistical Society, Series A* **179**, 975–1005.

Case, A. & Deaton, A. (2015), 'Rising morbidity and mortality in midlife among white non-hispanic americans in the 21st century', *Proceedings of the National Academy of Sciences* **112**, 15078–15083.

Hoyert, D. L., Heron, M. P., Murphy, S. L. & Kung, H.-C. (2006), 'Deaths: Final data for 2003', *National Vital Statistics Reports* **54**(13), 109.

Jemal, A., Ward, E., Anderson, R. N., Murray, T. & Thun, M. J. (2008), 'Widening of socioeconomic inequalities in u.s. death rates, 1993–2001', *PLOS ONE* **3**(5), e2181.

Macdonald, A. S., Richards, S. J. & Currie, I. D. (2018), *Modelling Mortality with Actuarial Applications*, 1 edn, Cambridge University Press.

NCHS (1993), 'Public use data tape documentation: Multiple cause of death for icd-10 1990 data'.

NCHS (2016), 'National center for health statistics, vital statistics data (Accessed May 2017)'.
URL: *https://www.cdc.gov/nchs/data_access/vitalstatsonline.htm*

Olshansky, S. J., Antonucci, T., Berkman, L., Binstock, R. H., Boersch-Supan, A., Cacioppo, J. T., Carnes, B. A., Carstensen, L. L., Fried, L. P., Goldman, D. P., Jackson, J., Kohli, M., Rother, J., Zheng, Y. & Rowe, J. (2012), 'Differences in life expectancy due to race and educational differences are widening, and many may not catch up', *Health Affairs* **31**(8), 1803–1813.

Park, J. H. (1999), 'Estimation of sheepskin effects using the old and the new measures of educational attainment in the current population survey', *Economics Letters* **62**(2), 237–240.

Rostron, B. L., Boies, J. L. & Arias, E. (2010), 'Education reporting and classification on death certificates in the united states', *Vital and Health Statistics* **2**(151).

Sasson, I. (2016), 'Trends in life expectance and lifespan variation by educational attainment: United states, 1990-2010', *Demography* **53**, 269–293.

Sorlie, P. D. & Johnson, N. J. (1996), 'Validity of education information on the death certificate', *Epidemiology* **7**(4), 437–439.

U.S. Census Bureau (2000), ‘Current population survey design and methodology, technical paper 63’.

U.S. Census Bureau (2006), ‘Current population survey design and methodology, technical paper 66’.

US Department of Health and Human Services (1995), ‘Technical appendix from vital statistics of united states’, *National Vital Statistics Reports* p. 15.