Cause of death specific cohort effects in U.S. mortality

March 24, 2021

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

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

We use a stochastic age-period-cohort mortality model to analyse US data for years 1989-2015 and ages, separated by gender, educational attainment, and cause of death. The paper focuses, in particular, on the fitted cohort effect for each sub-population and cause of death with two key findings. First, causes of death with a strong or distinctively-shaped cohort effect are also causes of death with significant, controllable risk factors, and that the fitted cohort effect gives us insight into the underlying prevalence of specific risk factors (such as smoking prevalence). Second, although each sub-population and cause of death has its own distinctive model fit, there are sufficient similarities between cohort effects to allow us to postulate that there is a relatively small number of underlying controllable risk factors that drive these cohort effects. The analysis then gives us insight into the modelled cohort effect for all-cause mortality.

Keywords

Stochastic mortality modelling, cause of death, US mortality, cohort effect, Bayesian methods, controllable risk factors.

Postprint license permissions

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

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 Introduction

Since the publication of Lee & Carter (1992), stochastic mortality models have been extremely useful tools for the modelling and forecasting of death rates. In the years since, many other models have been proposed, for example those including cohort effects by Renshaw & Haberman (2006), Cairns et al. (2009), or Plat (2009). In particular, Cairns et al. (2009) analysed the performance of several models using data from England and Wales and the United States.

Cairns et al. (2009) opened up discussion on the nature and importance of cohort effects in these models, a discussion that was developed further in Hunt & Blake (2014) who discuss in depth the model building process. Hunt & Blake (2014) emphasize that cohort effects should, typically, be modest in magnitude, picking up small but significant residual effects by cohort. In particular, for England and Wales it is now well known that people born around 1930 experienced faster mortality improvements than people born before 1925 or after 1945, see Richards et al. (2006) and Willets (2004).

As explained in Holford (1991), cohort effects typically appear due to changes in the exposure to risk factors for certain illnesses that are linked to when a person was born or through the development of good or bad habits during childhood or early adulthood that have an impact on later health. For example, people who smoke typically start doing so in early adulthood and keep this habit for most of their lives. However, the percentage of people who take up smoking varies considerably from cohort to cohort depending on the fashion of the day during early adulthood. Therefore we would expect to see a strong relationship between year of birth and deaths due to lung cancer, which will appear as a cohort effect when we fit our models. Fashions tend to change gradually rather than abruptly and so the resulting cohort effects will also typically be smooth functions.

1.1 Controllable risk factors

Almost all common individual diseases and causes of death have been extensively analysed in the medical literature to identify risk factors and the associated relative risk: that is, if an individual has a particular risk factor, how much higher is their death rate from a particular cause?

Risk factors can be classified, to some degree into three groups.

- **Controllable risk factors:** risk factors that are easily controllable by the individual, examples being smoking, diet, exercise and alcohol consumption.
- **Preventable risk factors:** risk factors for a particular cause of death that are much less controllable by the individual but which, nevertheless, can be reduced or prevented through other means such as vaccination. An example is Human Papilloma Virus (HPV) which is a very significant risk factor for cervical cancer: preventable through vaccination, but much less easily through changes in individual behaviour.

- **Non-preventable risk factors:** risk factors that cannot be controlled or prevented, such as genetic or racial factors, and personality traits such as conscientiousness.

In some contexts (for example, Cancer Research UK: www.cancerresearchuk.org), the first two groups are combined under the single heading "preventable", simply to highlight that significant intervention is possible.

Some risk factors fall between groups: for example, educational attainment and affluence. Arguably, something can be done to improve or change these characteristics. But, in reality, once most individual reach early adulthood their path has been set in terms of general education and, it is, arguably, the *ability to learn* from public health messages and consequently modify behaviour that makes a difference.

In this paper, it is the concept of controllable risk factors that is the most important as these are key drivers of cohort effects in mortality models.

1.2 Outline of the paper

In this paper we will fit a mortality model with a cohort effect term to the data described in Redondo Lourés & Cairns (2020). The data available consists of exposures and number of deaths subdivided by gender, education level, and cause of death, for the period 1989-2015, ages 50-75 and cohorts born between 1914 and 1965. We will use groupings of causes of death that have similar underlying risk factors, and fit the model separately for each of the four different groups (low/high educated males/females). The resulting cohort effects will give us insights on the lifestyle factors driving the widening mortality gap between different education groups, which has been widely reported in the literature, see for example Case & Deaton (2015), Olshansky et al. (2012), or Jemal et al. (2008).

2 Data modelling

The theory behind our modelling uses the standard assumption that the number of deaths in a single year and at a certain age is a Poisson random variable with expected value given by the death rate times the central exposed-to-risk:

$$D(x,t) \sim \text{Poisson}(m(x,t)E(x,t)), \tag{1}$$

where $D(x,t)$ is the number of deaths at age x in year t, $E(x,t)$ is the corresponding exposures and $m(x,t)$ is the underlying death rate. Additionally, we use the standard assumption that the $D(x,t)$ are conditionally independent of each other given the $m(x,t)$ and $E(x,t)$.

Now we need to specify the form of $m(x,t)$. Due to the wide range of causes of death we are going to analyse, we use a fairly flexible age-period-cohort model. This is a version of a model proposed by Plat (2009) which is suitable for modelling mortality over a limited range of higher ages (see, also, model CBD-X2 in Dowd et al. (2020)). In particular, we have:

$$\log m(x,t) = \alpha_x + \kappa_t^{(1)} + (\bar{x} - x)\kappa_t^{(2)} + \gamma_{t-x}, \tag{2}$$

where α_x is a non-parametric age effect, $\kappa_t^{(1)}$ and $\kappa_t^{(2)}$ are period effects, and γ_c is a cohort effect that is linked to year of birth $c = t - x$. (See, also, Cairns et al. (2019) (without the cohort effect) and Dowd et al. (2020) for further discussion of this and related models.) This model, in combination with the conditionally-independent Poisson assumption, allows us to write a likelihood function in terms of the parameters we want to estimate (the α_x, $\kappa_t^{(1)}$, $\kappa_t^{(2)}$, and γ_c) and the known exposures and number of deaths. As is the case with most stochastic mortality models, we need to impose some constraints on the parameters to ensure the uniqueness of the maximum likelihood estimate. The following transformations leave the death rates (and therefore the likelihood) unchanged:

$$\begin{pmatrix} \alpha_x \\ \kappa_t^{(1)} \\ \kappa_t^{(2)} \\ \gamma_c \end{pmatrix} \to \begin{pmatrix} \alpha_x + c_1 + c_2(\bar{x} - x) \\ \kappa_t^{(1)} - c_1 \\ \kappa_t^{(2)} - c_2 \\ \gamma_c \end{pmatrix} \tag{3}$$

$$\begin{pmatrix} \alpha_x \\ \kappa_t^{(1)} \\ \kappa_t^{(2)} \\ \gamma_c \end{pmatrix} \to \begin{pmatrix} \alpha_x + \phi_1 - \phi_2 x + \phi_3 x^2 \\ \kappa_t^{(1)} + \phi_2 t + \phi_3(t^2 - 2\bar{x}t) \\ \kappa_t^{(2)} + 2\phi_3 t \\ \gamma_c - \phi_1 - \phi_2(t-x) - \phi_3(t-x)^2 \end{pmatrix} \tag{4}$$

In particular, we see that constant shifts in the values of the $\kappa_t^{(1,2)}$ and γ_c can be absorbed in the α_x, and that linear and quadratic trends in the cohort effect can be absorbed by similar linear and quadratic transformations of the other parameters. This identifiability problem can be managed by imposing the following constraints:

$$\sum_t \kappa_t^{(1)} = 0 \ , \ \sum_t \kappa_t^{(2)} = 0, \tag{5}$$

$$\sum_c \gamma_c = 0 \ , \ \sum_c c\gamma_c = 0 \ , \ \sum_c c^2\gamma_c = 0. \tag{6}$$

Equations (5) mean that the fitted period effects will average to zero, and equations (6) that the fitted cohort effect will average to zero and also not have a linear or quadratic trend. These constraints will be imposed *a posteriori*, that is, we will first estimate the parameters using one of the methods described later in this section, and then we apply a transformation that ensures our estimates fulfil equations (5) and (6).

We use two approaches to the estimation problem. First, we use maximum likelihood estimation, implemented in R using StMoMo package (see Villegas et al. (2018) for details). The log-likelihood function is, up to a constant:

$$\begin{aligned} \ell_p &= \sum_{x,t} w(x,t) \Big[D(x,t)(\alpha_x + \kappa_t^{(1)} + (\bar{x} - x)\kappa_t^{(2)} + \gamma_c) \\ &\qquad - E(x,t)\exp(\alpha_x + \kappa_t^{(1)} + (\bar{x} - x)\kappa_t^{(2)} + \gamma_c) \Big] \end{aligned} \tag{7}$$

where the $w(x,t) = 1$ or 0 are weights that indicate if a particular observation is to be included or not (for example, $w(x,t) = 0$ for cohorts with four or fewer observations; see Cairns et al. (2009) and Dowd et al. (2020)). Confidence intervals are produced by bootstrapping using the function provided in the StMoMo library.

Typically, when modelling cause of death data, the maximum likelihood estimates of non-parametric functions are very noisy, and that noise makes it difficult to identify true patterns in the cohort effect. We, therefore, use a second approach in which Bayesian techniques are used in combination with time series assumptions about $\kappa_t^{(1)}$, $\kappa_t^{(2)}$, and γ_c. The purpose of this is to produce smoother, in-sample estimates of the cohort effects, in particular, by filtering out noise and allowing clearer identification of any significant patterns in the cohort effect. We will need to add new terms to the likelihood coming from the time series structure of $\kappa_t^{(1,2)}$ and γ_c, and priors for the unknown parameters in those time series, which will also need to be estimated. This will produce a log-posterior for all parameters that we can sample from to obtain the desired credible intervals. Due to the complexity of this log-posterior we need to use Markov Chain Monte Carlo (MCMC) methods to produce our sample. We reiterate that the only purpose of the time series in this model is to obtain a smoother in-sample fit than the one obtained through maximum likelihood estimation, and this procedure is not intended for death rate forecasting by cause of death.

Using the ideas of Cairns et al. (2011) we assume the vector of period effects $\boldsymbol{\kappa_t} = (\kappa_t^{(1)}, \kappa_t^{(2)})^T$ to follow a random walk with drift $\boldsymbol{\mu} = (\mu_1, \mu_2)^T$, and γ_c to be an AR(2) process of the form $\gamma_c = (\rho + \tau)\gamma_{c-1} - \rho\tau\gamma_{c-2} + \sigma_c\varepsilon_c$, where $\varepsilon_c \sim \mathcal{N}(0,1)$ for all c. To simplify the Markov chain algorithm we fix $\boldsymbol{\kappa_1} = 0$ and $\gamma_1 = 0$, and later convert our results to fulfil the constraints given in equations (5) and (6). In more detail, the full log-likelihood will be (ignoring constants):

$$\ell = \ell_p + \ell_{rw,\kappa} + \ell_{ar,\gamma} \tag{8}$$

where ℓ_p is given in equation (7), and the other two terms are the time series contribution to the likelihood:

$$\ell_{rw,\kappa} = -\frac{1}{2}\sum_{t=2}^{27}\left[(\Delta_{\boldsymbol{\mu}}\boldsymbol{\kappa_t})^T\Sigma^{-1}\Delta_{\boldsymbol{\mu}}\boldsymbol{\kappa_t}\right] - \frac{26}{2}\log\left(|\Sigma|\right), \tag{9}$$

$$\ell_{ar,\gamma} = -\frac{1}{2\sigma_c^2}\sum_{c=3}^{52}(\gamma_c - (\rho+\tau)\gamma_{c-1} + \rho\tau\gamma_{c-2})^2 - \frac{50}{2}\log\left(\sigma_c^2\right) - \frac{1}{2}\log(\sigma_{ac}) - \frac{1}{2\sigma_{ac}}\gamma_2^2, \tag{10}$$

where Σ is the covariance matrix associated with the random walk for the vector $\boldsymbol{\kappa_t}$; $\Delta_{\boldsymbol{\mu}}\boldsymbol{\kappa_t} = \boldsymbol{\kappa_t} - (\boldsymbol{\kappa_{t-1}} + \boldsymbol{\mu})$; ρ and τ are the two parameters of the $AR(2)$ model for γ_c, and σ_c^2 its variance; and σ_{ac}^2 is given by:

$$\sigma_{ac} = \frac{1+\rho\tau}{1-\rho\tau}\frac{\sigma_c^2}{(1+\rho\tau)^2 - (\rho+\tau)^2}. \tag{11}$$

All of these new parameters, namely Σ, $\boldsymbol{\mu}$, ρ, τ, and σ_c, are also to be estimated, and appropriate priors need to be provided for them. A standard choice for them would be:

$$\begin{align}
\Sigma &\sim \mathit{Inv\text{-}Wishart}(\nu, \mathbf{S}), \tag{12}\\
\mu_1 &\sim \mathcal{N}(0, \sigma^2_{\mu_1}), \tag{13}\\
\mu_2 &\sim \mathcal{N}(0, \sigma^2_{\mu_2}), \tag{14}\\
\sigma^2_c &\sim \mathit{Inv\text{-}Gamma}(a, b), \tag{15}\\
\mathrm{logit}(\rho) &\sim \mathcal{N}(0, \sigma^2_\rho), \tag{16}\\
\mathrm{logit}(\tau) &\sim \mathcal{N}(0, \sigma^2_\tau), \tag{17}
\end{align}$$

where ν, $\mathbf{S}$, $\sigma^2_{\mu_1}$, $\sigma^2_{\mu_2}$, a, b, σ^2_ρ, and σ^2_τ are hyperparameters that we need to fix in order to run our MCMC algorithm. Our results have been found to be fairly robust against changes in the values of these hyperparameters. However they need to be chosen carefully: for example, σ^2_c controls the level of smoothing we will achieve in the cohort effects. If its value is too large the result will be as noisy as the maximum likelihood estimate, but if it is too small we will see oversmoothing that destroys the real underlying trends we want to observe. Since the results we obtain for γ_c with this choice are very similar to the MLEs (as we will see in the next section) we do not find it necessary to use a different prior for these parameters.

Alternative ways of producing a smooth cohort effect, such as spline-fitting or two-step time series modelling have also been considered, but none of them were found to be better than the Bayesian approach. The former also have a somewhat stronger subjective component, since in the Bayesian method the role of the smoothing parameter is played by σ_c, which is not fixed to a specific value. Instead, it is estimated from the data with a certain level of flexibility given by its prior distribution.

These two models will be fitted separately to each of our four subpopulations (low-educated males, low-educated females, high-educated males, and high-educated females), and for each of the following groups of causes of death:

- Lung cancer.
- Lifestyle related cancers (mouth and gullet, stomach, gut or rectum, larynx, trachea, liver, and bladder cancers).
- Prostate and/or breast cancer.
- All other cancers (genitalia, pancreas, skin, urinary organs, lymphatic, benign or unspecified tumours, and all other locations of cancer).
- Chronic lower respiratory diseases (CLRD).
- Diabetes.
- All heart diseases.
- Cerebrovascular diseases.
- Other circulatory diseases.

- Alzheimer's, dementia and other mental illnesses.
- Accidental deaths (excluding accidental poisoning).
- Deaths of despair (chirrosis, accidental poisonings, and suicide).
- All other causes of death.
- All cause mortality.

There were two reasons to choose this grouping of causes of death. On the one hand, we tried to model groups that have a relatively large number of deaths and somehow homogeneous underlying risk factors. For example, lung cancer and CLRD are closely related to smoking, or heart disease to a combination of smoking and poor diet/sedentary lifestyle. On the other hand, we grouped causes of death so that the ICD standard change described in Redondo Lourés & Cairns (2020) would not result in a big discontinuity in the death rates between the years 1998 and 1999, which would complicate the modelling exercise.

3 Model suitability and goodness of fit

It is appropriate to ask if the APC model (equation 2) fits each dataset well and to have confidence that the estimated cohort effect is not being used as a substitute for a missing, third age-period effect. This has been approached throroughly through a combination of graphical diagnostics, analysis of mean squared standardised residuals, and testing for significance of the cohort effect. By way of example, consider the Chronic Lower Respiratory Diseases cause-of-death group (mainly this is Chronic Obstructive Pulmonary Disease, COPD) for lower-educated females. The results discussed below use the maximum likelihood estimates of the parameter values using StMoMo.

3.1 Graphical diagnostics

Selected graphical dignostics are plotted in Figure 1. The top-left panel of Figure 1 contains three elements by cohort year of birth $c = t - x$:

- Standardised residuals, $Z(x,t) = (D(x,t) - m(x,t)E(x,t))/\sqrt{m(x,t)E(x,t)}$ (black dots), for the reduced model with no cohort effect, $\log m(x,t) = \alpha_x + \kappa_t^{(1)} + (\bar{x} - x)\kappa_t^{(2)}$.
- The mean of the standardised residuals for each cohort, $\bar{Z}(c)$, for the reduced model (blue line).
- The fitted cohort effect, γ_c, for the full model (red line).

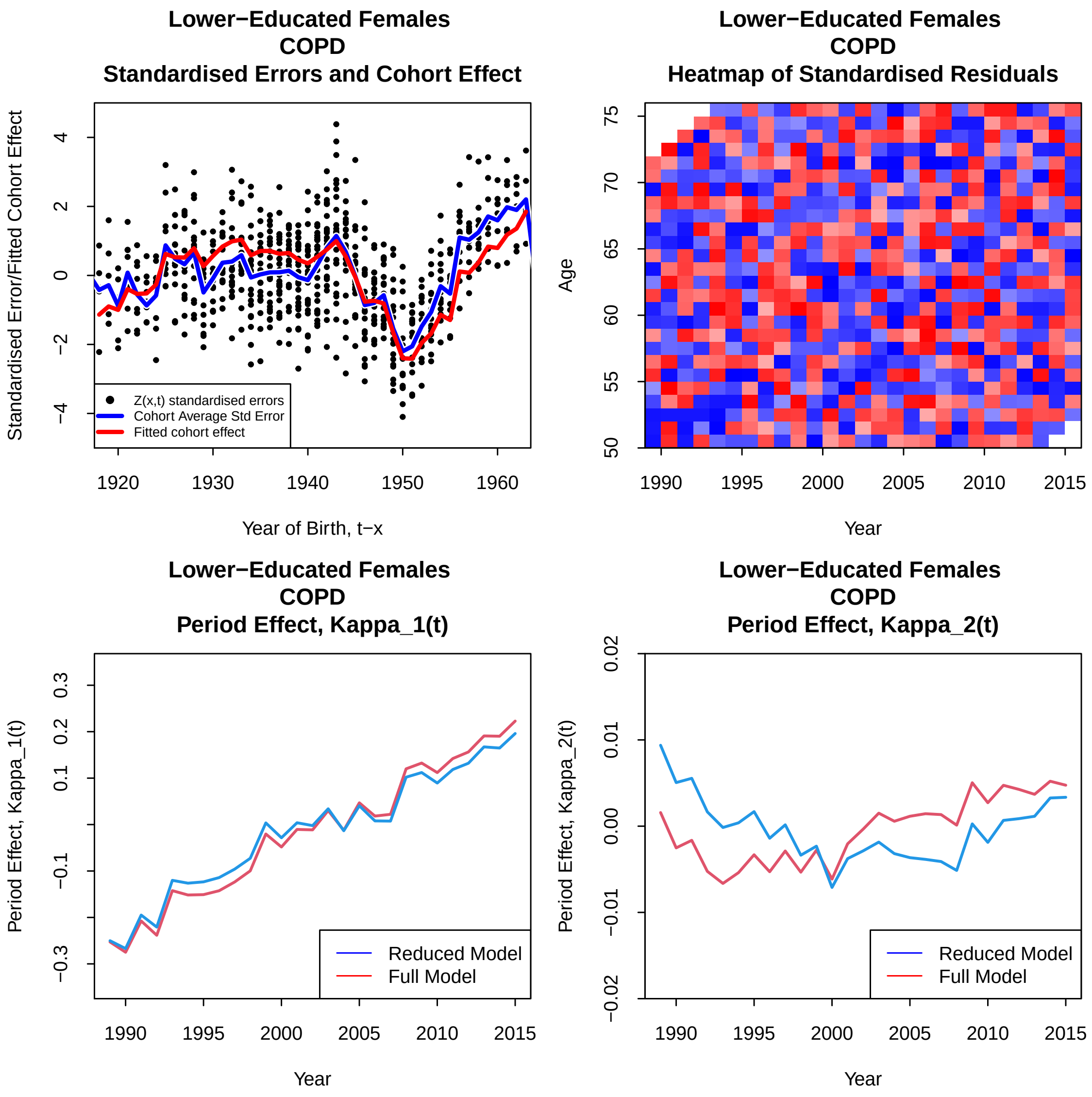


Figure 1: Fitting results for the full and reduced model for lower-educated females and CLRD as the cause of death. Top left: standardised residuals, $Z(x,t)$, plotted against year of birth $t-x$ for the reduced model (dots); mean of the $Z(x,t)$ for each cohort (blue line); fitted cohort effect, γ_c, for the full model (scaled to aid comparison). Top right: heat plot of the standardised residuals for the full model (red for positive values; blue for negative values; white for cohorts with zero weight). Bottom left: fitted values of $\kappa_t^{(1)}$ for the reduced (blue line) and full (red) models. Bottom right: fitted values of $\kappa_t^{(2)}$ for the reduced (blue line) and full (red) models.

In the spirit of Hunt & Blake (2014) and Dowd et al. (2020), for an age-period-cohort model to be adequately parameterised, the shape of the fitted cohort effect should approximately match the pattern of residuals against year of birth for the reduced (age-period) model with no cohort effect. (Conversely, if the two do not match, this is an indication that the fitted cohort effect is actually trying to capture age-period effects that the reduced model cannot adequately capture.) It can be seen in the top-left panel that the shape of the fitted cohort effect (which is scaled up to aid comparison) matches very closely the shape of the residuals for the reduced model.

The top right panel shows a heat plot of the standardised residuals for the full model. As can be seen, the residuals exhibit a very random pattern, and this is what we would expect if the model fits well and the conditional independence assumption between cells embedded in the Poisson assumption is correct.

The lower panels of Figure 1 compare the fitted $\kappa_t^{(1)}$ (left) and $\kappa_t^{(2)}$ (right) period effects for the full (red lines) and reduced (blue lines) models. When a cohort effect is added, we would expect the period effects to change slightly (as we see here). But if the differences are more marked (which we do not see here) then this would be a sign that the model is not suitable (in the sense of Hunt & Blake (2014)). The similarity of the period effects is consistent with the similarity observed in the top left panel for the cohort effects.

These graphical diagnostics have been repeated for all four education-gender pairings and for all 13 cause-of-death groups. And in all cases, the observations are similar to those above for low-educated females for CLRD.

3.2 Goodness of fit

Next, we consider the mean squared standardised residuals (MSSR):

$$MSSR = \frac{\sum_{x,t} w(x,t) Z(x,t)^2}{\sum_{x,t} w(x,t) - \nu}$$

where the $w(x,t)$ are the weights attached to each cell (0 or 1) and ν is the number of parameters estimated minus the number of identifiability constraints. If the model fits well then the MSSR will be close to 1.

For lower-educated females, Chronic Lower Respiratory Diseases, the MSSR is 1.05 indicating a good fit. We can also note that this compares favourably with some popular models at the all-cause level where higher MSSRs are typical (see, for example, models M2 and M7 in Sections 6.1.2 and 7.1.2 of Cairns et al. (2009)). Over all education-gender pairings and 13 cause-of-death groups, the MSSR ranges from 0.92 to 1.37 giving confidence that the model (equation 2) is suitable in all cases for the selected age and year ranges.

There are three conclusions from this section. First, the proposed model (equation 2) fits the data well or reasonably well for all education-gender pairings and causes of death. Second, following Hunt & Blake (2014), we are satisfied that, for the ranges of ages and years considered, the use of two age-period effects is sufficient to capture the majority of variation in the data. Third, the fitted cohort effect picks up the residual effects that exist when we fit the age-period-only model. This is robust, in the sense that the addition of the cohort effect does not distort significantly the fitted period effects.

4 Results: cohort effects for different causes of death

We will now present the fitted cohort effects for different population groups and causes of death. We will mainly focus on causes of death with well known controllable risk factors, which we will try to link to the observed patterns in γ_c.

Figure 2 shows the cohort effects for lung cancer and CLRD for all four population groups (low-educated males, high-educated males, low-educated females, and high-educated females). The fans show the 60%, 75%, and 90% confidence/credibility intervals. The orange fan in the background of each plot comes from maximum likelihood estimation and bootstrapping, whereas the grey fan in the foreground of each plot is the result obtained from the Bayesian approach. It can be seen that the Bayesian approach captures the main features of the MLE estimates, while avoiding most of the volatility resulting from the random sampling variation in the number of deaths.

It is interesting to note that, for each education-gender pairing, the shape of the cohort effect, γ_c, is very similar for the two causes of death. This is not surprising, since both lung cancer and CLRD are very strongly linked to cigarette smoking, (see, for example, US Department of Health and Human Services (2004), as well as specialist medical websites such as `www.cancerresearchuk.org` and `www.lung.org`). When we compare cohort effects for either lung cancer or CLRD for other education-gender pairings, we still see some similarities, but these are less strong than the lung-cancer/CLRD similarities noted above. This reflects the fact that cohorts within each education-gender pairing will have different smoking prevalences, and these differences show up in the fitted cohort effects. We also need to keep in mind that smoking *prevalence* is not the only driver of lung cancer and CLRD mortality. Mortality rates might also depend on the proportion of ex-smokers in a cohort (and when they ceased smoking) and what the mix is of heavy versus light smokers. So cohort effects will also reflect variations in these additional risk factors.

Finally, note that, due to the constraints introduced by equations (6), the cohort effect cannot have linear or quadratic trends, therefore these γ_c do not directly represent the proportion of smokers in each cohort (and the additional smoking risk factors). Any linear or quadratic trend in smoking prevalence will appear in the modelled period effects, while the fitted γ_c tells us about the local variations around a quadratic trend.

In Figures 3-6 we plot the cohort effects for a selection of causes of death for each of the gender-education pairings. The causes of death chosen have different links to different primary lifestyle factors (for example, smoking, poor diet, excessive alcohol consumption, lack of exercise) as well as intermediate drivers of mortality such

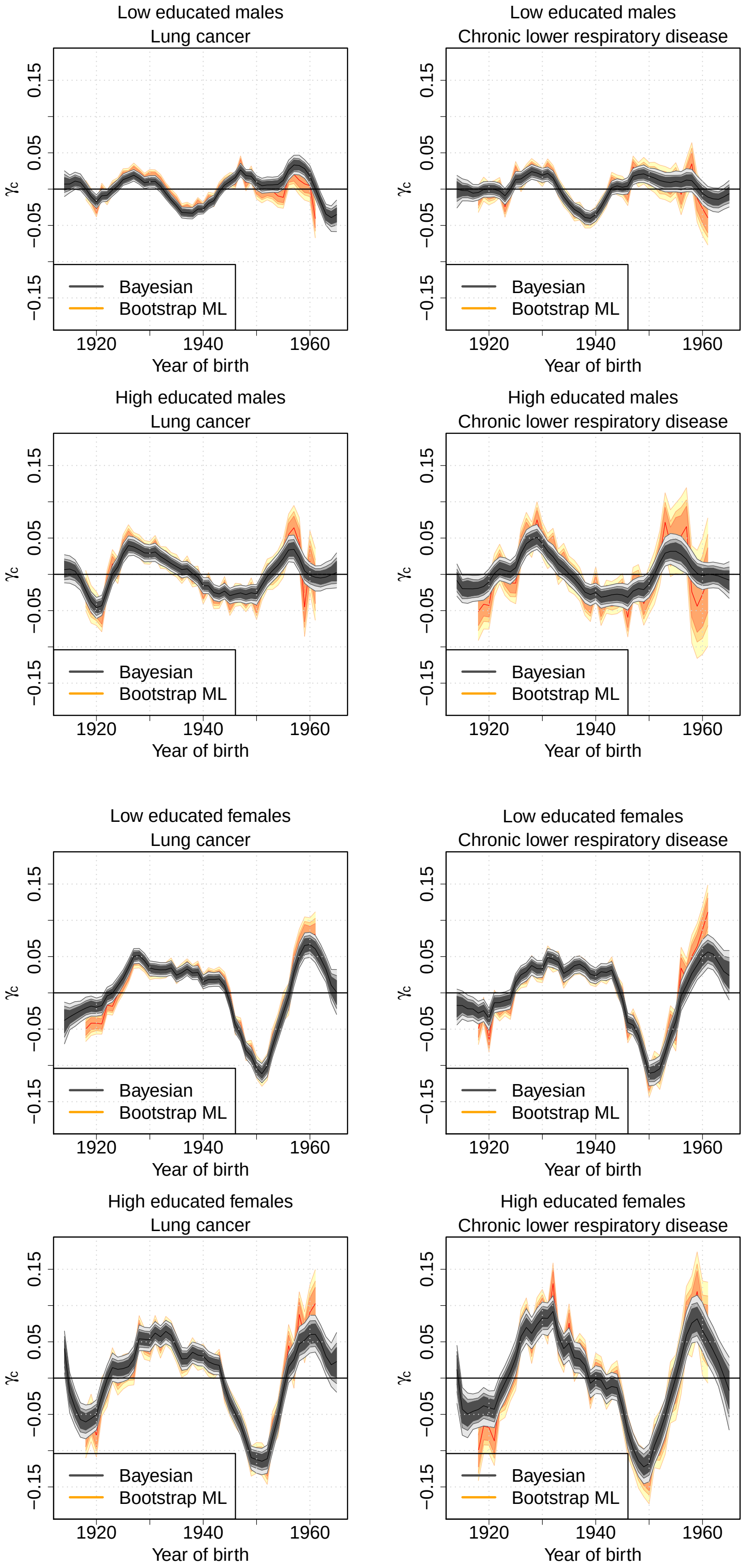


Figure 2: Cohort effects for lung cancer (left) and CLRD (right) for all four population groups, with 60%, 75%, and 90% confidence intervals. Orange and background is the result of the MLE-bootstrapping procedure. The grey fan is the result of the Bayesian method.

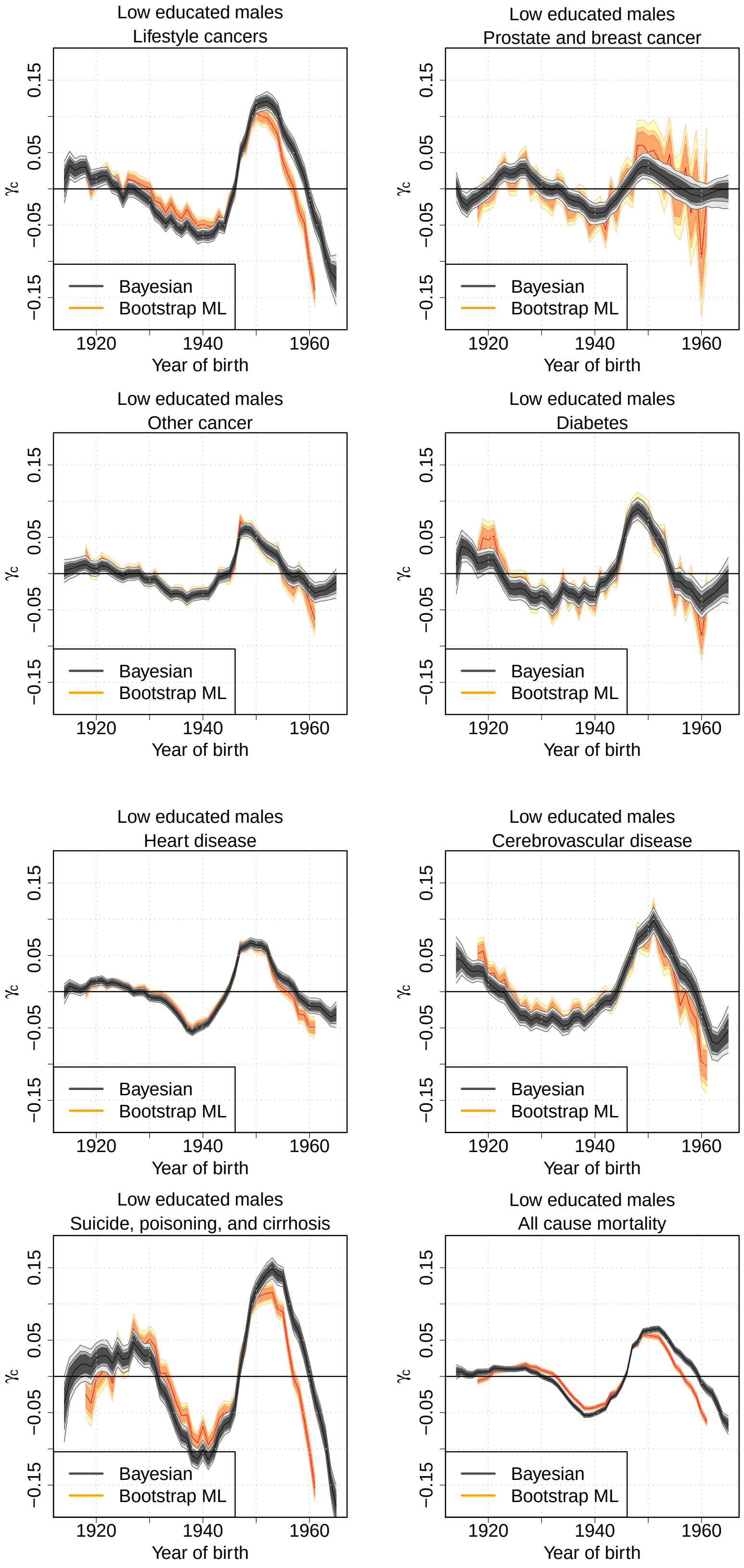
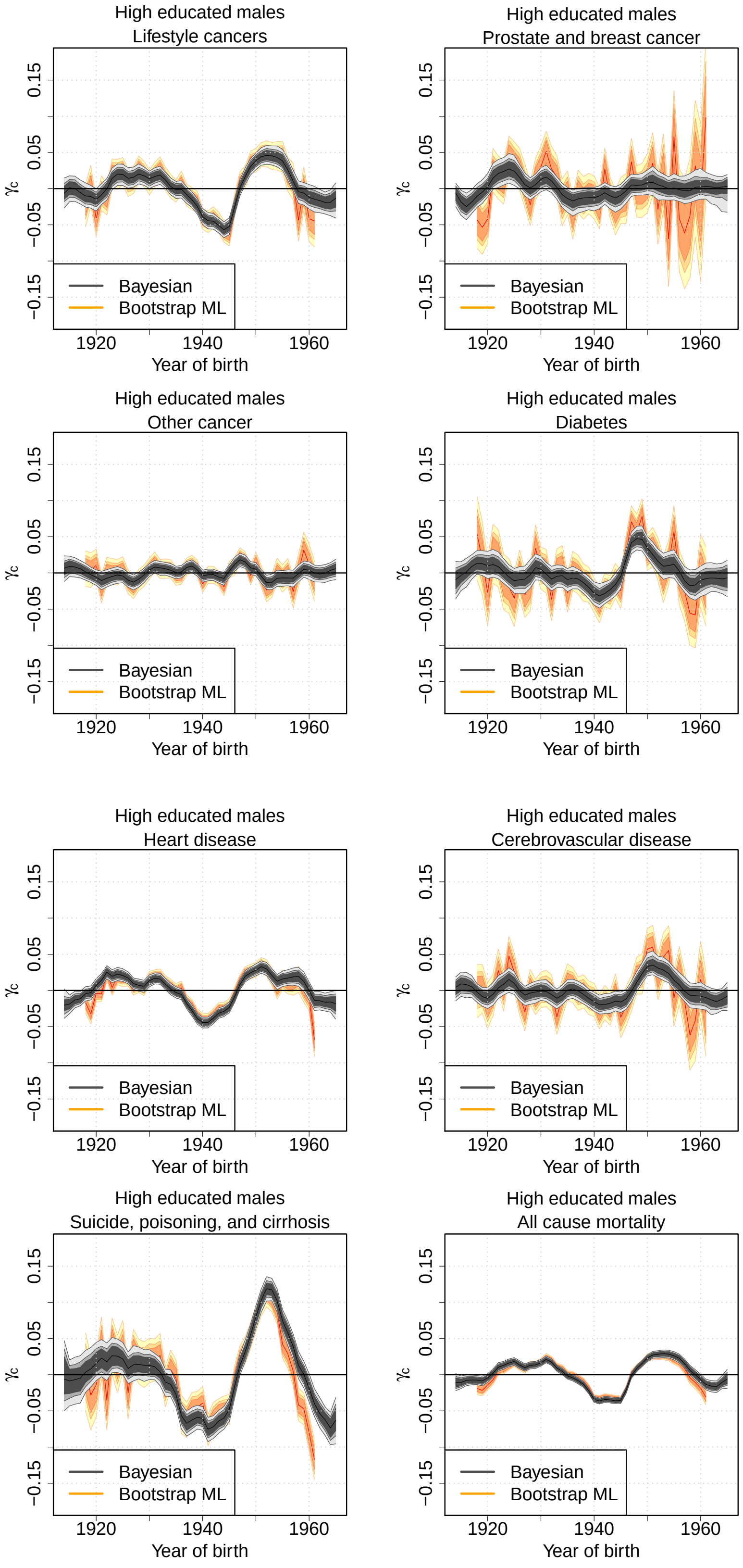


Figure 3: Cohort effects for several causes of death in low educated males, with 60%, 75%, and 90% confidence intervals. Orange and background is the result of the MLE-bootstrapping procedure. The grey fan is the result of the Bayesian method.



High educated males Lifestyle cancers

High educated males Prostate and breast cancer

High educated males Other cancer

High educated males Diabetes

High educated males Heart disease

High educated males Cerebrovascular disease

High educated males Suicide, poisoning, and cirrhosis

High educated males All cause mortality

Figure 4: Cohort effects for several causes of death in high educated males, with 60%, 75%, and 90% confidence intervals. Orange and background is the result of the MLE-bootstrapping procedure. The grey fan is the result of the Bayesian method.

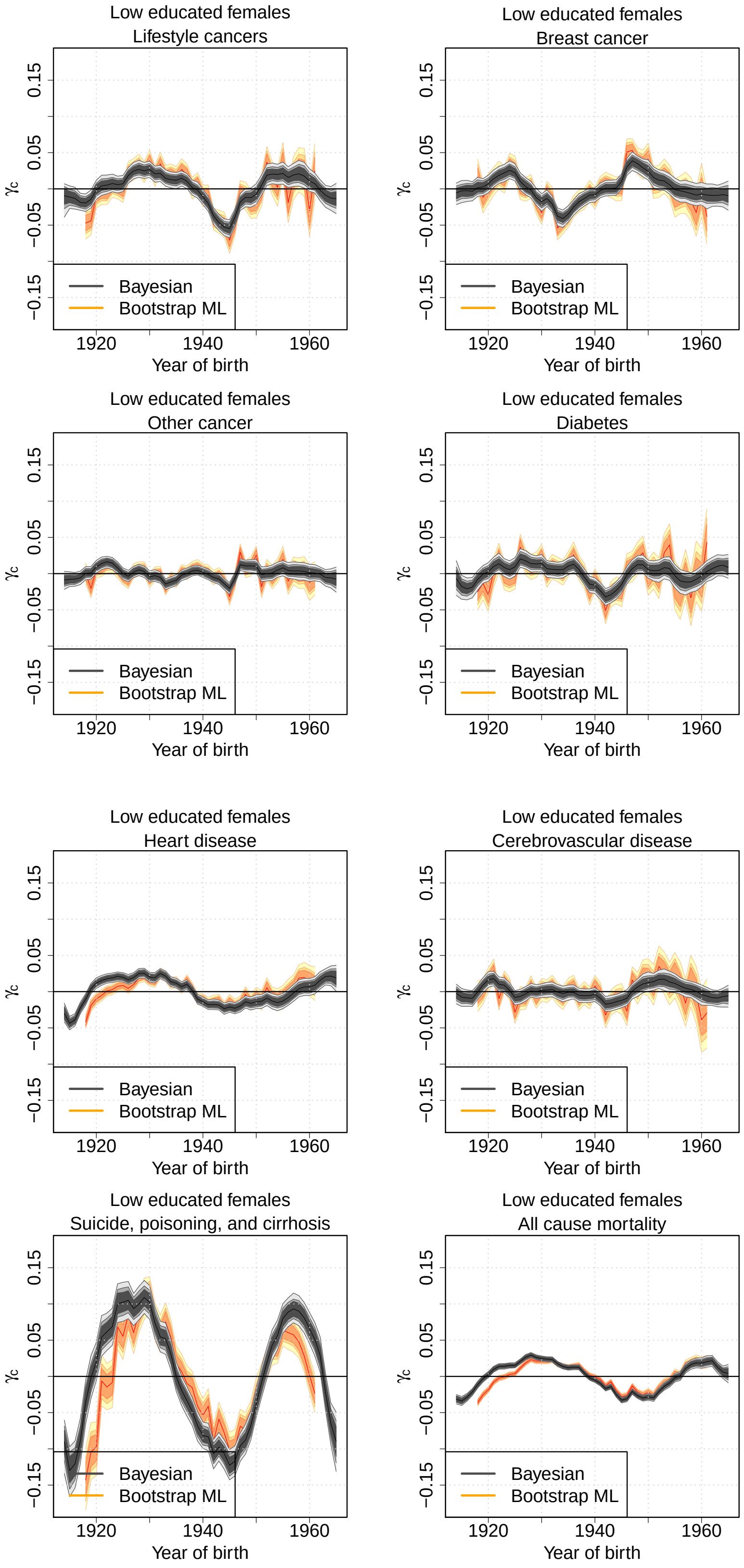


Figure 5: Cohort effects for several causes of death in low educated females, with 60%, 75%, and 90% confidence intervals. Orange and background is the result of the MLE-bootstrapping procedure. The grey fan is the result of the Bayesian method.

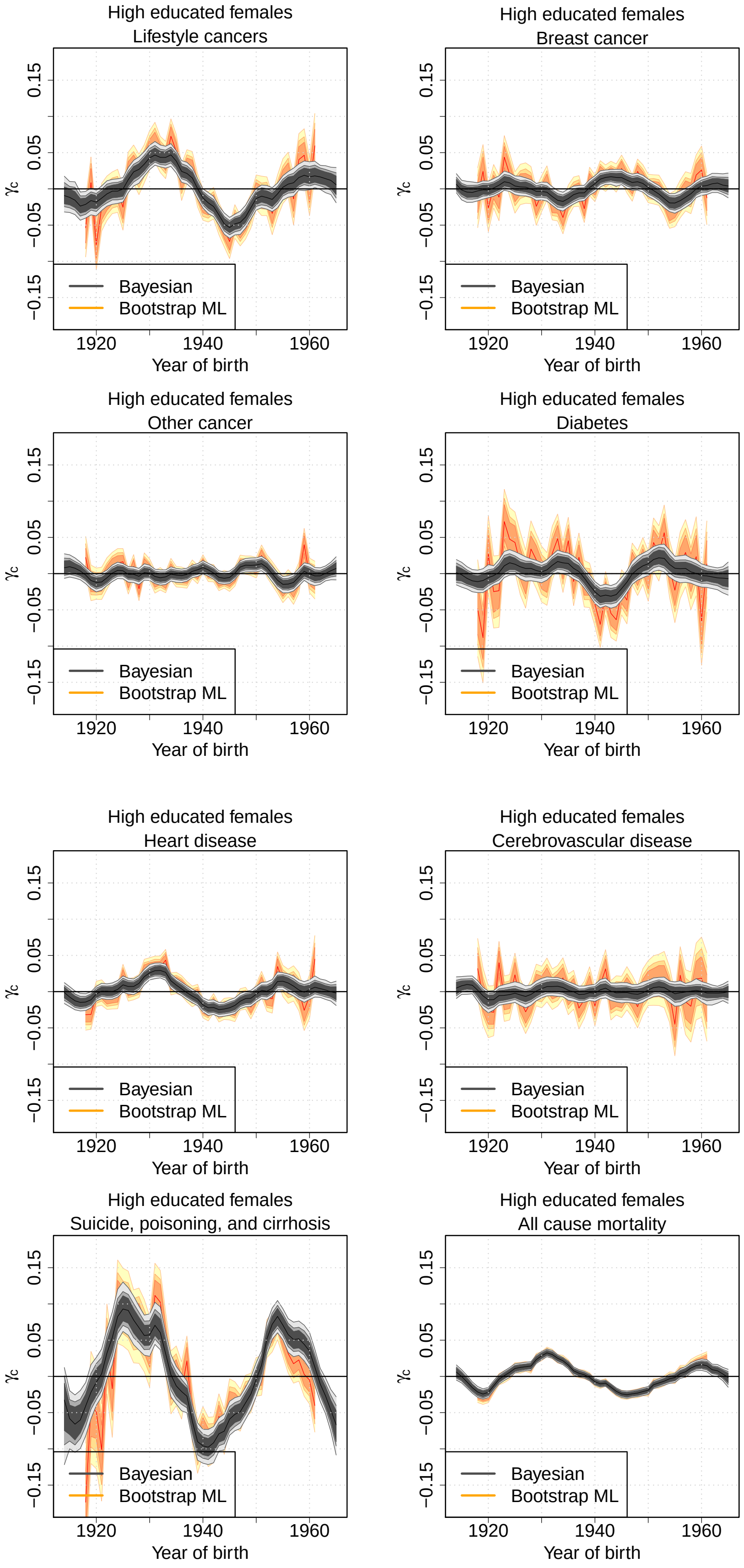


Figure 6: Cohort effects for several causes of death in high educated females, with 60%, 75%, and 90% confidence intervals. Orange and background is the result of the MLE-bootstrapping procedure. The grey fan is the result of the Bayesian method.

as obesity, high blood pressure, and other common co-morbidities. The different patterns that we see in the fitted cohort effects, therefore, give us indirect insight into the prevalences (and intensities) of these controllable risk factors by cohort.

The general picture is that, where we see a strong cohort effect (as examples: low/high-educated females, lung cancer; low/high-educated males, diabetes), the cause of death has significant controllable risk factors. Conversely, causes of death with no significant, or relatively weak, controllable risk factors will tend to result in only modest or weak cohort effects (for example, low/high-educated females and males, prostate and breast cancer). As such, we can postulate that the fitted cohort effects can be used as proxies for variations between cohorts in different health behaviours linked to controllable risk factors. Linked to this, it has been observed by Redondo Lourés & Cairns (2020) that whenever we observe a significant gap (e.g. lung cancer) between lower and higher-educated groups there is always a significant controllable risk factor, with a higher prevalence of the underlying behaviour in the lower-educated group. Conversely a narrow inequality gap (e.g. prostate cancer) is typically associated with causes of death with no significant controllable risk factors.

For lung cancer the link is straightforward: there is a single, significant controllable risk factor (smoking). For other causes of death there might be several significant controllable risk factors linked to multiple lifestyle factors. For example, apart from smoking, heart disease mortality has increased risk associated with several co-morbidities (obesity, diabetes, high blood pressure, high cholesterol and stress) that are, in turn, linked to poor diet and exercise as controllable risk factors. So the cohort effects for each cause of death will often be the result of a potentially complex combination of single-risk-factor cohort effects.

We also need to be mindful that a single-risk-factor cohort effect differs from the underlying lifestyle factor in a number of ways.

- A 10% increase, say, in the prevalence of the lifestyle factor might result in a 5% increase in mortality for one cause of death and a 20% increase for another.
- For a given cause of death the magnitude of this same increase might also depend on the average behaviour of those adopting the unhealthy lifestyle (the intensity). For example, on average, are smokers within a specific cohort heavy smokers or light smokers. In this sense, the fitted cohort effect reflects a combination of prevalence and average behaviour.
- With the given identifiability constraints in the model, a general linear or quadratic trend in the prevalence of a controllable risk factor by cohort will be transferred to the period effects, $\kappa_t^{(1)}$ and $\kappa_t^{(2)}$.

Now consider each education-gender pairing separately. For low educated males (Figure 3) we can see that γ_c sharply increases for cohorts born in the late 1940s for all of the causes of death shown. This might be related to the higher mortality experienced in later life by males who were drafted into the Vietnam war, see Schlenger et al. (2015). This is followed by a rapid decline for cohorts born in the 1960s. The strength of this effect, even for causes of death with weak or no known controllable risk factors, points out to a high prevalence of poor health behaviours in this group.

For high educated males (Figure 4) the picture is not so extreme. Although γ_c also increases for the Vietnam cohorts for most diseases highly correlated with controllable risk factors (diabetes, lifestyle cancers, heart disease, or sucide, poisoning, and cirrhosis), the increase is much smaller than for their lower educated peers. In contrast, there are no significant patterns in the cohort effects for prostate and breast cancer or other cancers where there are either no or at most weak controllable risk factors. For this group, the peaks and troughs for lung-cancer and CLRD (Figure 2) are somewhat different from those for other causes of death (Figure 4). This points to clear differences in prevalence of smoking as one controllable risk factor and other health behaviours that are risk factors for other causes of death.

For both low and high-educated females (Figures 5 and 6) cohort effects tend to be less pronounced than for males, and the patterns are a bit more varied. There is commonality in some cohort effects having a minimum around 1945 followed a few years later by a peak (lifestyle cancers, diabetes, cerebrovascular diseases and suicide, poisoning and cirrhosis for low-educated). Since this particular pattern is quite different from the lung cancer and CLRD cohort effects, this could point to variations in a different controllable risk factor and needs further detailed investigation. Interestingly, there are small but significant cohort effects for breast cancer which has no significant controllable risk factors. Again this merits further investigation, but possibilities include varying levels of access by cohort to screening and good quality healthcare following diagnosis.

5 Smoking: comparison with survey data

In general, it is hard to establish accurate information on the prevalence of specific controllable risk factors (rather than intermediate risk factors such as obesity, diabetes, high blood pressure etc.) that is measured or surveyed in a consistent way across the population and by cohort. An exception to this in the US is smoking prevalence data. But, even in this case, we will see that the relationship between smoking prevalence by cohort and the lung-cancer cohort effect is complex.

In the previous section, we singled out the results obtained for lung cancer and CLRD since a very high proportion of deaths due to these causes are linked to smoking. Because of this dependence on a single controllable risk factor, we argued that the cohort effects seen in Figure 2 could be linked to changes in smoking prevalence in the underlying populations, but that due to the identifiability problem the values of γ_c were not trivially related to the ratio of smokers.

We now consider data on the smoking habits of the US population form the National Health Interview Survey, run by the Centers for Disease Control and Prevention, CDC (2019). We use data for years 1990-2015, excluding 1996 (for which, along with 1989, data on smoking habits were not recorded.), and ages 30-79. Data are, otherwise, available by single age and calendar year and, to reduce the effect of sampling variation, we aggregate the data by cohort. For each cohort in the data we calculate the average ratio of "ever smokers" (i.e. current and ex-smokers), weighting each (age, year) cell by the number of respondents. Note that, as smokers tend to die earlier in life, the actual ratio of smokers *decreases* as a cohort ages. Since we are only observing early cohorts at very high ages we will be slightly underestimating the true ratio of ever smokers in these groups. However, since our mortality data covers similar ages and calendar years, we expect the resulting cohort effects to be similarly biased.

The same dataset has been analysed by Anderson et al. (2012) who use 5-year cohort groups rather than single years of birth. Comparison is by ethnic group rather than education level, but the broad shape of the proportion reporting ever-smoker status by cohort is similar to the ratios we report in Figure 7: for example, for females the dip in 1950 followed by a peak around 1960 and a sharp fall after 1960. While Anderson et al. (2012) focus on analysis of smoking behaviour by cohort, they also have the modelling of lung cancer mortality in mind as the ultimate goal.

Figure 7 shows, for each of the four subpopulations analysed, the average ratio of smokers (black line, left axis) and the median of the Bayesian estimate for the lung cancer cohort effect (red line, right axis). Even after averaging over all observations of a cohort the ratio of smokers is still very noisy (that is, the sample sizes for each cohort are relatively small). So, although we can easily make out the broad, long-term shape of the ratios, it is harder to make out short term fluctuations. As expected the cohort effect does not capture the general upwards or downwards trends due to the constraints imposed by (6), as explained earlier. After taking account of this, we can see that some features in the fitted cohort effect do have equivalent features in the smokers ratios. Most notably, for both groups of females, the minimum in the cohort effect in 1950 followed by a peak around 1960 has a matching dip and peak in the ratio of smokers. However, some other features in the cohort effect (e.g. the peak around 1928 for lower-educated females) do not have matching features in the ratio.

So although we might, tentatively conclude that smoking prevalence is a driver of the lung-cancer (and CLRD) cohort effects these plots suggest that the cohort effect must be a more complex combination of additional drivers. As remarked previously these might include the balance between current smokers and ex smokers, the *intensity* of smoking amongst current smokers (for example, Anderson et al. (2012) define intensity as average number of cigarettes smoked per day) and access to healthcare. A similar conclusion has been found by Jemal et al. (2018), who consider the link between current-smoking prevalence and lung cancer incidence. They find that smoking prevalence explains some of the variation in cancer incidence but not all. Complementing this, the analysis in Anderson et al. (2012) does indicate that, amongst current smokers, the average intensity of smoking does vary considerably from cohort to cohort (as well as the prevalence of smoking).

5.1 Discussion

Now we have a known link between lung-cancer and CLRD and smoking as the single dominant risk factor. But we have only found a loose link between the fitted

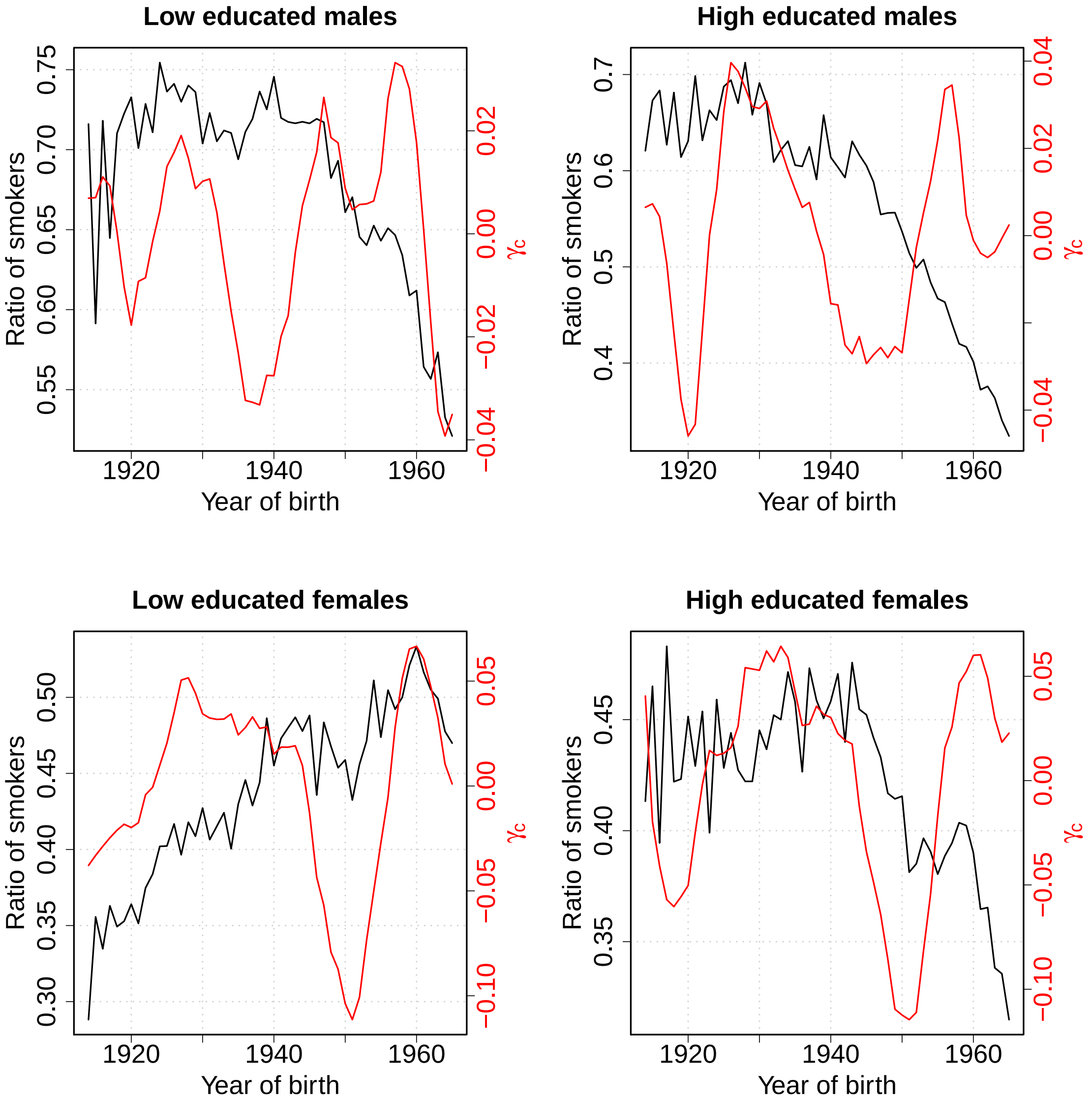


Figure 7: Average ratio of smokers (black line, left axis) and median of the bayesian estimate for the lung cancer cohort effect (red line, right axis) as a function of year of birth for the male and female lower and higher-educated subpopulations.

cohort effect and smoking prevalence. From this we infer that, even if it was possible to use surveys or other sources to measure the prevalence of other controllable risk factors, these might similarly be only loosely linked to associated cohort effects. In other cases, such as poor diet or excessive alcohol consumption, there is no clear threshold (as there is for smoking) that defines prevalence and, again, the "intensity" of poor diet or alcohol abuse above a threshold might vary by cohort as well as have an impact on mortality from specific causes. Consequently, survey data on the prevalence of particular risk factors might not serve as accurate predictors of cohort effects.

Instead, we propose that a mixture of cohort effects can be used as proxies for these drivers. In some cases, a cohort effect might be linked to a single risk factor such as smoking (and this might be a proxy for smoking behaviour rather than smoking prevalence). In other cases, a cohort effect might be linked to a combination of risk factors when these are quite highly correlated.

6 Conclusions

In this paper we have used statistical modelling in order to understand the lifestyle factors that drove the mortality inequalities described in Redondo Lourés & Cairns (2020). We have seen that cohort effects associated with specific causes of death are an important tool for the understanding of the underlying changes in health behaviours of the population being analysed. In particular, the cohort effects for causes of death with very similar risk factors show very similar trends, which can help us identify an underlying "lifestyle cohort effect".

We have also seen that cohort effects are much stronger for causes of death with lifestyle-related risk factors, further reinforcing the idea that cohort effects arise as a consequence of changes in the health behaviours of the underlying population. With this in mind, we can clearly see that the evolution of these health behaviours has been very different for males and females and, particularly in the case of males, for people of different socioeconomic status (for which education is used as a proxy in this work). This points to diverging lifestyles as the cause underlying the increase in the mortality gap in recent years. For example, the faster increase in γ_c for low educated males born in the late 1940s (who would be now in their early 70s) with respect to their higher educated peers, and for almost all preventable causes of death, points at lifestyle choices as the main driver of excess early mortality in low educated American males.

One downside of this method is that the existence of identifiability problems in the model somehow complicates the direct interpretation of the cohort effects, and comparison with survey data is not completely straightforward. How to transform the known trends of a lifestyle risk factor (for example, smoking prevalence) into a cohort effect that can be used in the model, or use the fitted cohort effect as a way to estimate the prevalence of these factors where survey data is not available, is an interesting problem that will be the subject of future research.

Bibliography

Anderson, C. M., Burns, D. M., Dodd, K. W. & Feuer, E. J. (2012), 'Birth-cohort-specific estimates of smoking behaviors for the u.s. population', *Risk Analysis* **32**, S14–S24.

Cairns, A. J. G., Blake, D., Dowd, K., Coughlan, G. D., Epstein, D., Ong, A. & Balevich, I. (2009), 'A Quantitative Comparison of Stochastic Mortality Models Using Data From England and Wales and the United States', *North American Actuarial Journal* **13**(1), 1–35.

Cairns, A. J. G., Blake, D., Dowd, K., Coughlan, G. D. & Khalaf-Allah, M. (2011), 'Bayesian Stochastic Mortality Modelling for Two Populations', *ASTIN Bulletin* **41**(1), 29–59.

Cairns, A. J. G., Kallestrup-Lamb, M., Rosenskjold, C. P. T., Blake, D. & Dowd, K. (2019), 'Modelling socio-economic differences in the mortality of danish males using a new affluence index', *ASTIN Bulletin* **49**, 555–590.

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.

CDC (2019), 'National health interview survey', https://www.cdc.gov/nchs/nhis/data-questionnaires-documentation.htm. Accessed: November 2019.

Dowd, K., Cairns, A. J. G. & Blake, D. (2020), 'Cbdx: A workhorse mortality model from the cairns-blake-dowd family', *Annals of Actuarial Science* **14**, 445–460.

Holford, T. R. (1991), 'Understanding the Effects of Age, Period, and Cohort on Incidence and Mortality Rates', *Annual Review of Public Health* **12**(1), 425–457.

Hunt, A. & Blake, D. (2014), 'A general procedure for constructing mortality models', *North American Actuarial Journal* **18**, 116–138.

Jemal, A., Miller, K. D., Ma, J., Siegel, R. L., Fedewa, S. A., Islami, F., Devesa, S. S. & Thun, M. J. (2018), 'Higher lung cancer incidence in young women than young men in the united states', *New England Journal of Medicine* **378**(21), 1999–2009. PMID: 29791813.
URL: *https://doi.org/10.1056/NEJMoa1715907*

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.

Lee, R. D. & Carter, L. R. (1992), 'Modeling and Forecasting U.S Mortality', *Journal of the American Statistical Association* **87**(419), 659–671.

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.

Plat, R. (2009), 'On stochastic mortality modeling', *Insurance: Mathematics and Economics* **45**(3), 393–404.

Redondo Lourés, C. & Cairns, A. J. G. (2020), 'Mortality In The US By Education Level', *Annals of Actuarial Science* **14**, 384–419.

Renshaw, A. & Haberman, S. (2006), 'A cohort-based extension to the Lee-Carter model for mortality reduction factors', *Insurance: Mathematics and Economics* **38**(3), 556–570.

Richards, S. J., Kirkby, J. G. & Currie, I. D. (2006), 'The Importance of Year of Birth in Two-Dimensional Mortality Data', *British Actuarial Journal* **12**, 5–38.

Schlenger, W. E., Corry, N. H., Williams, C. S., Kulka, R. A., Mulvaney-Day, N., DeBakey, S., Murphy, C. M. & Marmar, C. R. (2015), 'A Prospective Study of Mortality and Trauma-Related Risk Factors Among a Nationally Representative Sample of Vietnam Veterans', *American Journal of Epidemiology* **182**(12), 980–990.

US Department of Health and Human Services (2004), 'The Health Consequences of Smoking: A Report of the Surgeon General'.

Villegas, A. M., Kaishev, V. K. & Millossovich, P. (2018), 'StMoMo: An R Package for Stochastic Mortality Modeling', *Journal of Statistical Software* **84**(3), 1–38.

Willets, R. C. (2004), 'The Cohort Effect: Insights and Explanations', *British Actuarial Journal* **10**, 833–877.