The Common Cohort Effect Model for Cause of Death Data

Andrew J.G. Cairns*†
Maxwell Institute for Mathematical Sciences, and
Department of Actuarial Mathematics and Statistics, Heriot-Watt University

October 2, 2023

Abstract

This paper takes a highly granular US mortality dataset with 51 causes of death by age, year, sex and education level and uses a model-based approach to understand the drivers of mortality changes over the period 1989 to 2017. A new age-period-cohort model is proposed that models all causes of death jointly and which incorporates a small number of common cohort effects: effects that can be linked to underlying controllable risk factors such as smoking, giving insights into their impact on both specific causes and all-cause mortality.

Model outputs confirm that some causes of death have quite volatile mortality rates from year to year, but also allow us to quantify that volatility, while others are quite smooth. Comparison of estimated period effects also allows us insight into the relative rates of improvement in mortality due to medical advances for pairs of causes of death that have a common controllable risk factor.

Keywords: Cohort effect; Age-Period-Cohort model; CBDX model; Controllable risk factors

*Email: A.J.G.Cairns@hw.ac.uk

†This research forms part of the "Modelling Measurement and Management of Longevity and Morbidity Risk" research program funded by the Actuarial Research Centre of the Institute and Faculty of Actuaries, and the Society of Actuaries.

1 Introduction

Recent decades have seen significant changes in mortality rates in many countries. In general, this has been an improving trend, but, for some, the last 10 to 20 years have seen either a slowdown (UK) or even a reversal (US) of some of these improvements, further disrupted by the Covid pandemic. Recent years have also seen the emergence of important databases that facilitate analysis of these improvements, most notably the Human Mortality Database (2023) and the Human Cause-of-Death Database (2023) which cover a range of countries in a consistent and carefully-curated way (see, also, Barbieri, 2017). This now allows us the opportunity to investigate (a) which causes of death have had the greatest impact on changes in all-cause mortality (good and bad), and (b) which controllable risk factors have the most impact on specific causes of death.

Recent research on cause-of-death mortality divides into three interlinked strands. First, some researchers have projection as a main objective (e.g. Alai et al., 2018, Boumezoud et al., 2019 and Arnold and Glushka, 2021). In order to manage the number and reliability of time series parameters, these works tend to use a smaller number of cause-of-death groups and do not include cohort effects. A second theme, is the analysis of mortality inequalities. Much of the work in this direction has focused on all-cause mortality (e.g. Chetty et al., 2016, Li and Hyndman, 2021). But others have highlighted specific causes that exhibit high levels of inequality (e.g. Case and Deaton, 2015). It is only more recently that a more systematic approach has been taken to assess levels of mortality inequality across the full range of causes of death (The US Burden of Disease Collaborators, 2018; Cairns and Redondo Lourés, 2023; and GBD US Health Disparities Collaborators, 2023). A third line of work (including the present paper) aims to develop a detailed, model-based understanding of historical changes in mortality by cause of death using much more granular cause-of-death data (Villegas et al., 2023).

The objectives of this paper are as follows:

- To exploit the growing availability of mortality data by cause of death at a granular level to help understand the drivers of all-cause mortality.
- To develop a new model with a focus on understanding cohort effects that gives us insight into the dynamics of death rates by cause of death through time, by age and by subgroup.
- To understand better the nature of cohort effects in all-cause mortality.
- To understand better the role of controllable risk factors and their impact on individual causes of death and how these interact with cause-specific medical advances.
- To draw insights, through analysis of the modelled period effects, into the year-to-year volatility of death rates and the relative changes in mortality between different groups and by cause of death.

2 What is a cohort effect?

The modelling work in this paper has a particular focus on cohort effects.

In the mortality setting, if a birth cohort experiences higher or lower mortality than otherwise anticipated (taking account of age and the general time trend) then this is referred to as a cohort effect. In modelling terms, the simplest example is the age-period-cohort model

$$\log m(t,x) = \alpha(x) + \kappa(t) + \gamma(t-x) \tag{1}$$

(e.g. Osmond and Gardner, 1982) where $m(t,x)$ is the death rate in calendar year t at age x, $\alpha(x)$ is an age effect, $\kappa(t)$ is a period effect and $\gamma(t-x)$ is the cohort effect. Although this model has been used quite widely, it typically fits both all-cause and cause-specific mortality data very poorly across individual ages and years (see, for example, Cairns et al., 2009). A wide variety of alternatives have been proposed to remedy this problem. Renshaw and Haberman (2006) were the first to add a cohort effect to the age-period model of Lee and Carter (1992), with others proposed by Cairns et al. (2009) and Plat (2009). By including a richer combination of age, period and cohort effects, these models now model historical all-cause mortality rates much more accurately including accurate capture of cohort effects.

There are various reasons for the existence of a cohort effect in mortality data (e.g. Holford, 1991). Most prominently they can be caused by variation by cohort in the prevalence of *controllable* risk factors such as smoking or excess alcohol consumption. Typically there will be a smooth progression from one cohort to the next and a cohort that has, for example, high levels of smoking will experience higher rates of mortality from causes of death that have smoking as a significant risk factor compared to other cohorts with lower levels of smoking. This feeds through to a cohort effect in an age-period-cohort mortality model. Preston and Wang (2006) compare male and female mortality. They argue, first, that cohort effects are present in both male and female all-cause mortality data and, second, that the varying gap between males and females can be largely attributed to changes in smoking behaviour by cohort in the two sexes.

This behavioural explanation for cohort effects in mortality data contrasts with situations where an overly-simple model such as (1) is being used. In such circumstances, the cohort effect can sometimes act as a proxy for missing age and period effects in the model rather than represent a true effect linked to year of birth. (See Hunt and Blake (2014) for a discussion.)

3 Data

We use data for US males and females by education level (low education meaning up to and including high-school diploma; high education meaning some further education beyond high-school diploma). Deaths data have been taken from the US Centers for Disease Control and Prevention, National Center for Health Statistics (`https://www.cdc.gov/nchs/data_access/VitalStatsOnline.htm#Mortality_Multiple`). Exposures data by education have been developed using a combination of data from the Human Mortality Database (HMD, 2023; general exposures) and the Current Population Survey (education level). For further details, see Redondo Lourés and Cairns (2019, 2021).

Exposures and deaths are by single age (40 to 84) and single year (1989 to 2017). To ensure an accurate split between low and high education, Redondo Lourés and Cairns (2020) found that it was necessary to exclude cohorts with only a small number of observations.

Table 1: Cause of death groups.

1	Infectious diseases				
2	Cancer: mouth, gullet	3	Cancer: oesophageal		
4	Cancer: stomach	5	Cancer: colon	6	Cancer: rectum, anus
7	Cancer: liver	8	Cancer: pancreas	9	Cancer: other digestive system
10	Cancer: larynx	11	Cancer: lung, bronchus, trachea	12	Cancer: skin
13	Cancer: breast	14	Cancer: cervix	15	Cancer: uterus
16	Cancer: ovary	17	Cancer: other female genital	18	Cancer: prostate
19	Cancer: other male genital	20	Cancer: bladder	21	Cancer: urinary organs
22	Cancer: lymphatic etc.	23	Benign tumours	24	Cancer: other locations
25	Blood diseases	26	Diabetes		
27	Vascular dementia	28	Other mental illness	29	Parkinson's disease
30	Alzheimer's disease	31	Other diseases of nervous system		
32	Blood pressure + rheumatic fever	33	Ischaemic heart diseases	34	Non-rheumatic valve disorders
35	Other heart diseases	36	Cerebrovascular diseases	37	Circulatory diseases
38	Influenza	39	Pneumonia	40	Other acute respiratory infections
41	Chronic Obstructive Pulmonary Disease	42	Other respiratory diseases		
43	Alcoholic liver disease	44	Other liver diseases	45	Other digestive diseases
46	Diseases: skin, bone, tissue	47	Diseases: urine, kidney,...		
48	Suicide	49	Road/other accidents	50	Accidental Poisonings
51	Other causes				

Hence, only cohorts born between 1914 and 1970 inclusive are used in this study.

Deaths data are further subdivided into 51 cause-of-death groups listed in Table 1 (see Cairns and Redondo Lourés, 2023 for ICD-10 codes associated with each of the 51 causes). One reason for using a relatively high number of causes is that this allows us to pick out certain causes that have a single major risk factor, such as smoking as the major risk factor for both lung cancer and Chronic Obstructive Respiratory Disease (COPD). In turn, this will allow us to identify a smoking-specific cohort effect for mortality. If we use a smaller number of cause-of-death groups (e.g. all cancers merged into one group and all respiratory diseases into another) then other risk factors come into play, making it much more difficult to disentangle the impact of smoking from other risk factors.

In practice, each subpopulation (i.e. sex and education) will have fewer than 51 causes to be modelled. For example, some causes are sex specific (e.g. ovarian cancer), while others have too few deaths at certain ages or cohorts to allow the basic model (equation (2)) to be fitted.[1]

An empirical analysis of the main features of this dataset can be found in Cairns and Redondo Lourés (2023).

4 Modelling each cause of death individually: CBDX-I

We choose to model mortality for each cause of death in a consistent way by using the CBDX3 model (Dowd et al., 2020) in Equation 2. To start, we model each of the $N_{cod} = 51$ causes of death, c, individually:

$$\log m(c,t,x) \quad = \quad \alpha(c,x) + \sum_{k=1}^{3} \beta_k(x)\kappa_k(c,t) + \gamma(c,t-x) \tag{2}$$

[1]Specifically, high-educated females have zero deaths from Parkinson's disease in the 1969 birth cohort.

where $m(c,t,x)$ is the death rate for cause c in year t at age x, $(t-x)$ is the cohort year of birth, and $\beta_1(x)=1$, $\beta_2(x)=(x-\bar{x})$ and $\beta_3(x)=(x-\bar{x})^2-\sigma_X^2$ are pre-defined, parametric age effects with $\bar{x}$ and σ_X^2 being the mean and variance of the observed ages respectively.

For each cause of death, c, the model requires use of the following identifiability constraints (or equivalent):

$$\sum_y \gamma(c,y)=0,\ \sum_y (y-\bar{y})\gamma(c,y)=0,\ \sum_y (y-\bar{y})^2\gamma(c,y)=0\ \ \sum_y (y-\bar{y})^3\gamma(c,y)=0 \quad (3)$$

$$\text{and}\ \ \sum_t \kappa_1(c,t)=0,\ \sum_t \kappa_2(c,t)=0,\ \sum_t \kappa_3(c,t)=0 \quad (4)$$

with $\bar{y}$ being the mean year of birth for the observed cohorts.

The choice and complexity of the model (including a non-parametric age effect $\alpha(c,x)$ rather than not; three period effects, $\kappa_k(c,t)$, rather than 2 or 1; and a cohort effect, $\gamma(c,t-x)$) reflects two criteria. First, it helps ensure that a good fit (as discussed in Hunt and Blake, 2014) can be achieved across all causes of death including conditional independence of individual (c,t,x) cells. Second, even for causes where a simpler model is merited, the use of a common model helps us to compare results for different causes of death.

In Figure 1 we plot the fitted cohort effect for selected causes for low-educated females. The top left panel looks at causes that have smoking as the principal risk factor (lung cancer, COPD and cancer of the larynx). Although the fitted cohort effects are not identical, it is apparent that they have a very similar shape. The overall shape, and especially the trough to peak from 1950 to 1960 is similar to that for females' all-cause mortality in Villegas et al. (2023; figure 4 unsmoothed). For cancer of the larynx, the fitted cohort effect is very noisy, reflecting the fact that the number of deaths is about 1% of the numbers for both lung cancer and COPD: but the underlying shape appears to be similar. Similar observations can be made for males and females of different education levels. The top right panel shows three causes with excessive alcohol consumption as a *potential* principal risk factor. This is certainly the case for alcoholic liver disease. For liver cancer and other liver diseases the connection in the medical literature between alcohol consumption and the cause of death is less clear. But the close alignment of the three estimated cohort effects here suggests that excessive alcohol consumption is also a major risk factor for liver cancer and other liver diseases. The lower panel in Figure 1 shows the fitted cohort effect for three causes that are known to have different risk factors: lung cancer (smoking), alcoholic liver disease (alcohol) and diabetes (obesity, diet, exercise). Here, we see that the three fitted cohort effects are distinctly different.

The lower panel also includes the fitted cohort effect using all-cause deaths data (dotted line). This is different from the three cause-specific cohort effects, but we can ask the question is the all-cause cohort effect made up of some combination of cause-specific cohort effects?

These observations are consistent with the possibility that cohort effects in mortality data are linked to underlying risk factors such as smoking, excessive alcohol consumption, poor diet and exercise, etc.. Additionally, the upper panels of Figure 1 lead us to ask if we need 51 individual cohort effects? Alternatively, will a much smaller number of cohort effects be sufficient, with each one being associated with a different controllable risk factor?

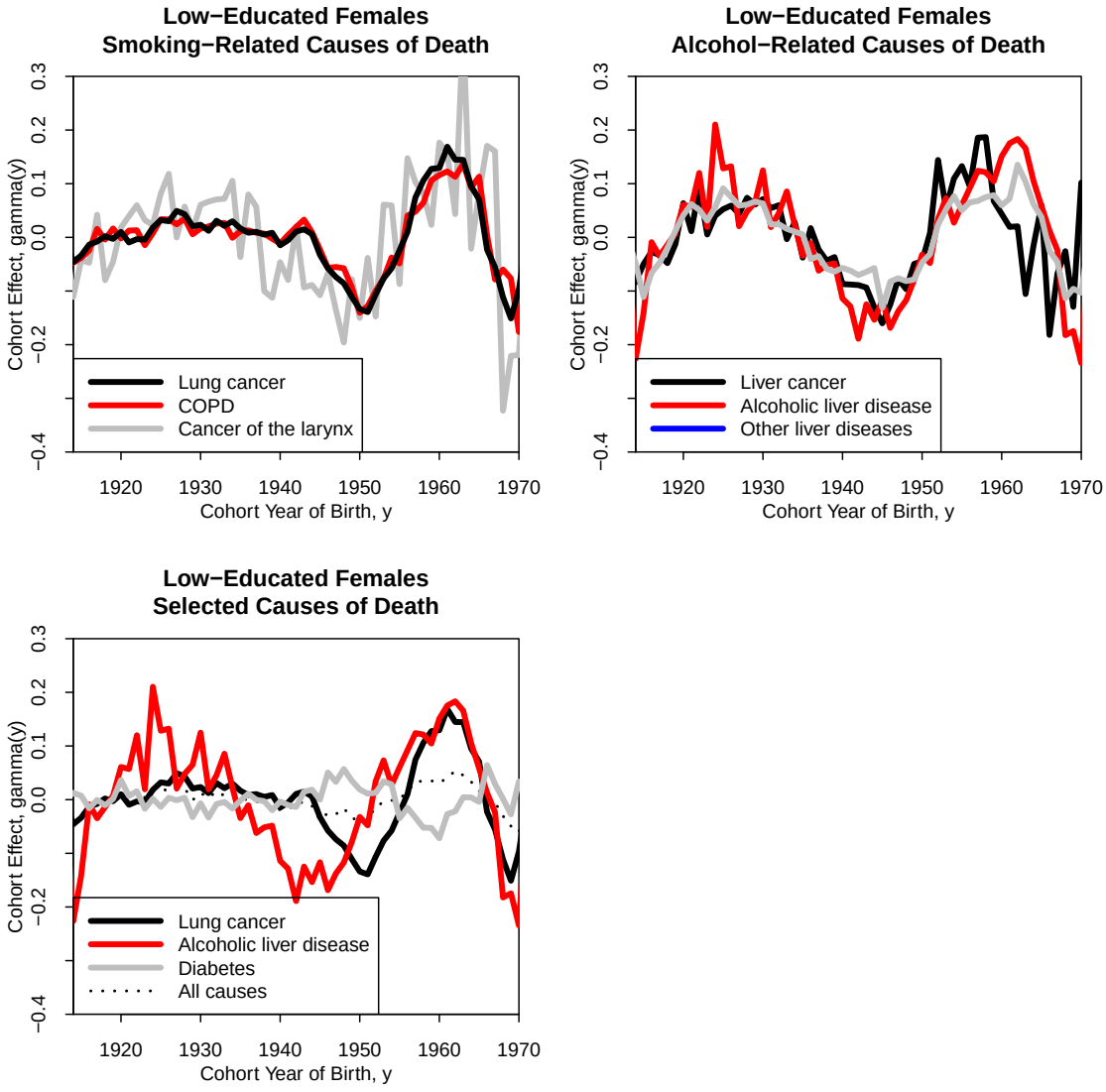

Figure 1: Fitted cohort effects (CBDX-I model) for selected causes of death for low-educated females. Top left: causes with smoking as a risk factor. Top right: causes with excessive alcohol consumption as a risk factor. Bottom left: causes with different risk factors plus all-cause mortality.

5 Modelling causes of death simultaneously: the common cohort effects model, CBDX-CCE

5.1 The common cohort effects model (CCE)

The preceding discussion suggests that it might be possible to achieve a good fit across all causes of death with many fewer than 51 distinct cohort effects. We, therefore, introduce the idea that there might be a small number n, of common cohort effects, $\chi_1(t-x), \ldots, \chi_n(t-x)$. Their use introduces the possibility that some or all of the common cohort effects are linked to specific controllable risk factors, such as smoking.

Here we model all N_{cod} causes jointly. Thus, for $c = 1, \ldots, N_{cod}$:

$$\log m(c,t,x) \;=\; \alpha(c,x) + \sum_{k=1}^{3} \beta_k(x)\kappa_k(c,t) + \underbrace{\sum_{j=1}^{n} \delta_j(c)\chi_j(t-x)}_{\gamma(c,t-x)} \tag{5}$$

where $\chi_1(y), \ldots, \chi_n(y)$ are the n common cohort effects that apply over all causes of death and the $\delta_j(c)$ control the contribution of the respective common cohort effects, $\chi_j(y)$, to the cause-specific cohort effect $\gamma(c,y)$ $(j = 1, \ldots, n)$. The proposed model, therefore, is a special case of that in equation (2). It is intended that the number, n, of common cohort effects will be relatively small: e.g. $n = 3$. For convenience, we will refer to the model with n common cohort effects as CCE-n (e.g. CCE-3).

5.2 Identifiability constraints

Comparing the models in equations (2) and (5), revised identifiability constraints are as follows:

- Period effects: constraints for the $\kappa_k(c,t)$ are the same as before (equation 4), so we have three constraints for each of the N_{cod} causes of death.
- Common cohort effects: constraints are the same as before, but the total number of constraints is much smaller: 4 constraints for each of the $j = 1, \ldots, n$ common cohort effects,

$$\sum_y \chi_j(y) = 0, \;\; \sum_y (y-\bar{y})\chi_j(y) = 0, \;\; \sum_y (y-\bar{y})^2\chi_j(y) = 0, \;\; \sum_y (y-\bar{y})^3\chi_j(y) = 0 \tag{6}$$

The constraints above can be regarded as being of a standard type in mortality modelling. In contrast, the structure of the model introduces some novel identifiability problems related to interchangeability of the $\chi_j(y)$. Define

$$\begin{aligned}
\gamma(c) &= (\gamma(c,1), ..., \gamma(c,N_y)) \;\; \text{(a row vector of length } N_y = \text{number of cohorts)} \\
\delta(c) &= (\delta_1(c), \ldots, \delta_n(c)) \;\; \text{(a row vector of length } n) \\
X &= \begin{pmatrix} \chi_1(1) & \ldots & \ldots & \chi_1(N_y) \\ \vdots & & & \vdots \\ \chi_n(1) & \ldots & \ldots & \chi_n(N_y) \end{pmatrix} \;\; \text{an } n \times N_y \text{ matrix.}
\end{aligned}$$

Then, for any non-singular $n \times n$ matrix A, we can note that

$$\gamma(c) = \delta(c)X = \delta(c)A^{-1}AX. \tag{7}$$

This means that we can change parameter estimates in equation (5) from $(\delta(c), X)$ (for all $c = 1, \ldots, N_{cod}$) to $(\delta(c)A^{-1}, AX)$ with no impact on any of the estimated cohort effects, $\gamma(c)$. To tackle this identifiability problem we considered two options. In both cases we first identify n causes of death $c_1, \ldots, c_n$ that have relatively high death rates and that have distinctively different cohort effects.

Option 1 Constraint 1: $\gamma(c_j, y) = \chi_j(y)$ for $j = 1, \ldots, n$, meaning that $\delta_j(c_i) = 1$ for $i = j$ and 0 for $i \neq j$. In this option, the only non-singular matrix, A, that satisfies (7) and the new constraint is the identity matrix itself.

Option 2 Constraint 2A: XX^T equals the identity matrix. Hence, the vectors $\chi_1, \ldots, \chi_n$ are orthogonal to each other, and each has magnitude 1. With this constraint, (7) still allows the matrix A to be an orthogonal matrix (that is, any matrix A satisfying $AA^T = I$), so we need additional constraints to restrict A to the identity matrix.

Constraints 2B.1: For $j = 1, \ldots, n-1$, $\delta_k(c_j) = 0$ for all $k > j$. For example, for $n = 3$, $\delta_2(c_1) = \delta_3(c_1) = 0$ and $\delta_3(c_2) = 0$.

Constraints 2B.2: For $j = 1, \ldots, n$, $\delta_j(c_j) > 0$ (without this, we could multiply all of the $\delta_j(c)$ and $\chi_j(y)$ by -1 without any impact on the fit)

The only matrix A that satisfies these additional constraints (2B.1 and 2B.2) is now the identity matrix.

Option 2 feels more complex than option 1, but, in numerical experiments, the numerical algorithm for option 2 converged significantly faster than option 1. The iterative scheme for option 2 is outlined in Appendix A.

On the other hand, option 1 gives us, potentially, a more immediate interpretation of the common cohort effects. For example, smoking is the only significant risk factor for lung cancer (cause #11) and excessive alcohol consumption is the main risk factor for alcoholic liver disease (cause #43). Thus, if we set $c_1 = 11$ and $c_2 = 43$ we have a natural interpretation for the first two common cohort effects: that is, $\chi_1(y)$ is the smoking cohort effect, and $\chi_2(y)$ represents an excessive-alcohol cohort effect. For option 2 with the same choices of $c_1 = 11$ and $c_2 = 43$ the interpretation of $\delta_1(c_2)$ and $\delta_2(c_2)$ is less clear. First, $\delta_1(c_2)\chi_1(y) + \delta_2(c_2)\chi_2(y)$ represents the alcohol-related cohort effect rather than $\chi_2(y)$ on its own. Second, there is a known association between smoking and alcohol consumption (see, for example, Beard et al., 2017), and so higher levels of smoking prevalence implicit in $\chi_1(y)$ might lead to higher levels of excessive alcohol consumption that would be reflected, first, in $\gamma(c_2, y)$, and then in $\chi_2(y)$. Additionally, the $\chi_j(y)$ are orthogonal under option 2, but there is no inherent reason why the cohort effects of interest, $\gamma(c_1, y)$ (smoking) and $\gamma(c_2, y)$ (alcohol) should be orthogonal.

For this study we chose $n = 3$ common cohort effects in combination with $c_1 = 11$ and $c_2 = 43$ (option 2 on its own does not require c_3 to be specified). In part, this reflects the fact that smoking and excessive alcohol consumption are the only two controllable risk factors with associated causes of death where the great majority of deaths are attributable to a single risk factor.

6 Numerical results

6.1 Model selection

Table 2: Maximum likelihood and BIC results for the CBDX-I and CCE-3. The two models are fitted to each of the four populations by sex and education level. The number of observations takes account of differing numbers of causes of death for males and females and individual (t, x) cells that are in the 1914 to 1970 cohorts only.

		Model					
		CBDX-I			CCE-3		
Population	# obs, N_{obs}	maximum log-lik, $\hat{l}$	effective # params, ν	BIC	maximum log-lik, $\hat{l}$	effective # params, ν	BIC
Males-Low	55440	-205252	8910	-253914	-206549	6003	-239335
Males-High	55440	-185460	8910	-234123	-186782	6003	-219568
Females-Low	59136	-209396	9504	-261609	-210775	6390	-245880
Females-High	59136	-185151	9504	-237365	-186435	6390	-221541

Table 2 focuses on the choice between the CBDX-I and CCE-3 models for the four populations. We use the Bayes Information Criterion $BIC = \hat{l} - \frac{1}{2}\nu \log N_{obs}$ where, for a given model, $\hat{l}$ is the maximum log-likelihood, ν is the effective number of parameters (that is, the total number of parameters to be estimated across all causes of death, minus the number of identifiability constraints), and N_{obs} is the number of obs. N_{obs} counts the number of (c, t, x) cells across all modelled causes of death, years and ages, but not counting young and old cohorts with two few observations.

Comparison of the BIC values shows that the CCE-3 model has a substantially higher BIC for each of the four populations confirming that, collectively across all causes of death, the CBDX-I model is significantly over-parameterised.

6.2 Fitted cohort effects

In Figure 2 we show fitted cohort effects under the CCE-3 model (lines) compared with earlier estimates using the CBDX-I model (dots) for low-educated females. In all cases (including those not plotted), we see that the fitted cohort effect under the CCE-3 model is much smoother than the original CBDX-I fit, especially for those causes of death with relatively small numbers of deaths such as cancer of the larynx. Additionally, the CCE-3 curves closely track the more noisy pattern in the CBDX-I fit. Smoothness of the cohort effect is a reasonable criterion for a mortality model given that cohort effects reflect mainly variation in underlying controllable risk factors in a given population, and these will typically vary slowly from one cohort to the next. The CCE-3 model achieves this smoothness for causes of death with low numbers of deaths (e.g. cancer of the larynx) by effectively borrowing information from other causes with large numbers of deaths (e.g. lung cancer). In the CBDX-I model we would need to incorporate a subjective smoothness criterion to achieve the similar results. So one conclusion is that the CCE-3 model mitigates the earlier

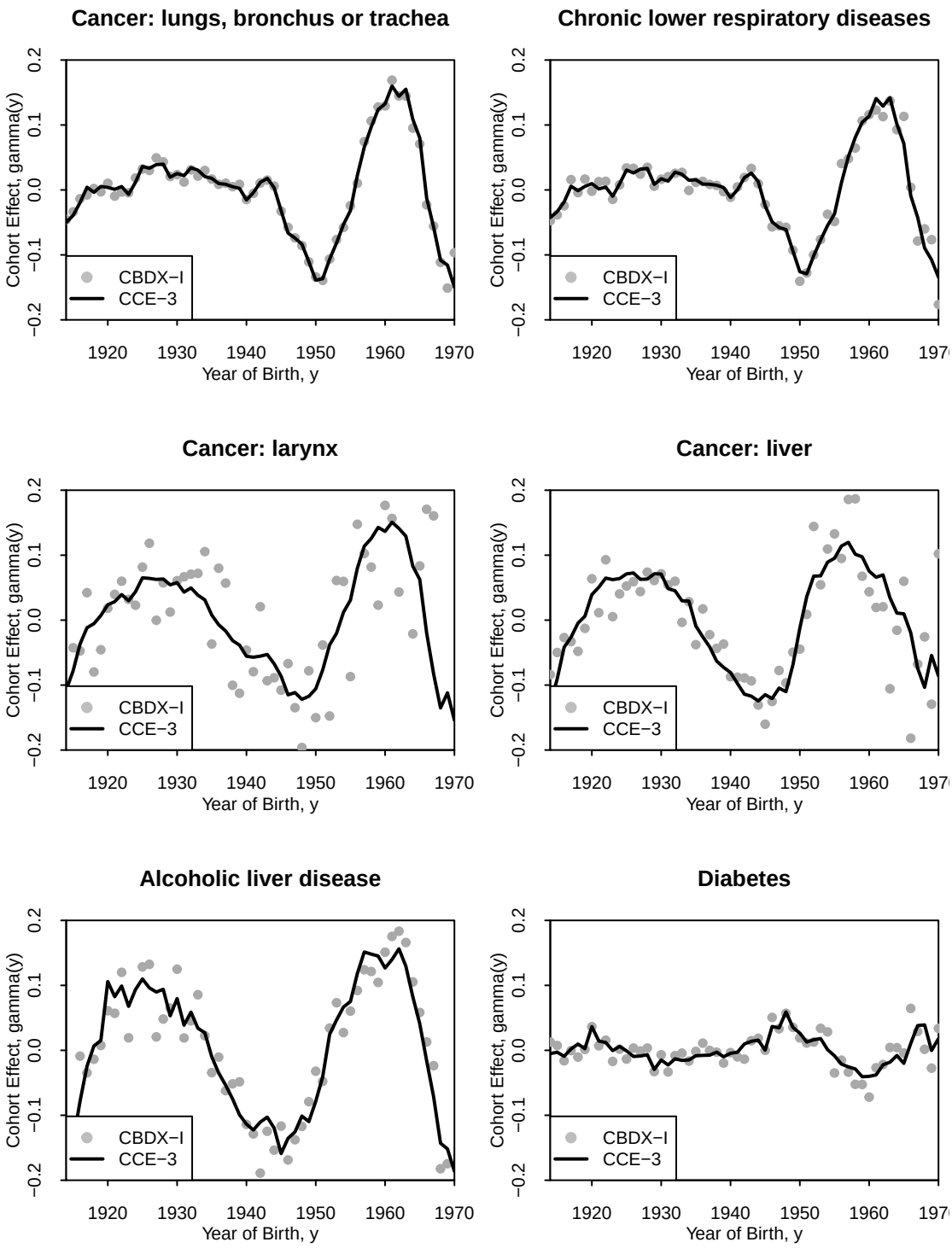

Figure 2: Fitted cohort effects for low-educated females for the CBDX-I model (dots) compared with the CCE-3 model (lines) for selected causes of death.

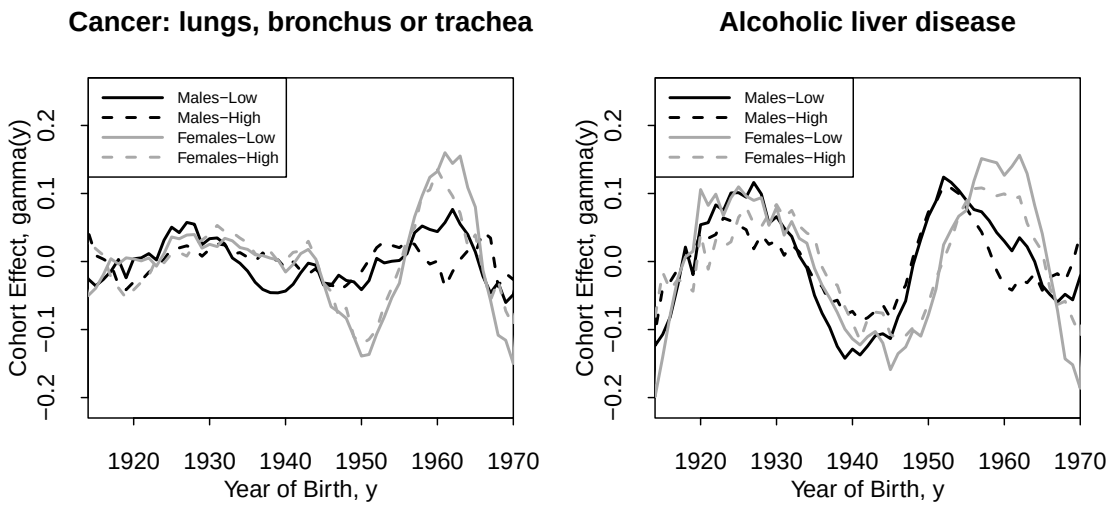

Figure 3: Fitted cohort effects for lung cancer (left) and alcoholic liver disease (right) for all four groups.

suggestion that the CBDX-I overfits the data to some extent especially for causes with relatively small numbers of deaths.

In Figure 3 we highlight what might be labelled as the smoking and excess-alcohol cohort effects for the four populations: i.e. lung cancer and alcoholic liver disease. Note that, while the smoking cohort effect (and similarly the alcoholic cohort effect) will reflect changes in the prevalence of smoking (excess alcohol consumption), it should not be interpreted as being directly equivalent to smoking prevalence for two reasons. First, identifiability constraints mean that some changes in smoking prevalence by cohort appear in the period effects (specifically linear, quadratic and cubic changes). Second, the impact of smoking on a cohort's lung cancer death rates is a combination of factors: smoking prevalence, the intensity and pattern of smoking over the lifetime of a cohort, the proportion of ex-smokers and time since cessation. So, our use of the expression 'smoking prevalence' is really a proxy for this more-complex combination.

The left-hand plot of Figure 3 shows some similarities in the smoking cohort effect. But a closer look reveals that low and high-educated females are more closely linked together than they are to males. The pattern for females suggest that after taking account of larger scale, potentially cubic changes in 'smoking prevalence', females born around 1950 had significantly lower prevalence than the general trend, while those born around 1960 had significantly higher prevelance. Everything else being equal, a swing in the cohort effect from -0.1 to $+0.1$ equates to a difference of about 20% in lung cancer death rates. For males the cohort effect (and therefore underlying smoking prevalence) does not vary by such large amounts, and we can see some deviation between the cohort effects between low and high-educated males.

Similar comments can be made about the excess-alcohol cohort effects (right-hand plot, Figure 3).

Comparison of the left and right-hand plots shows that the smoking and excess-alcohol cohort effects are quite different, with significant differences in the timing of peaks and

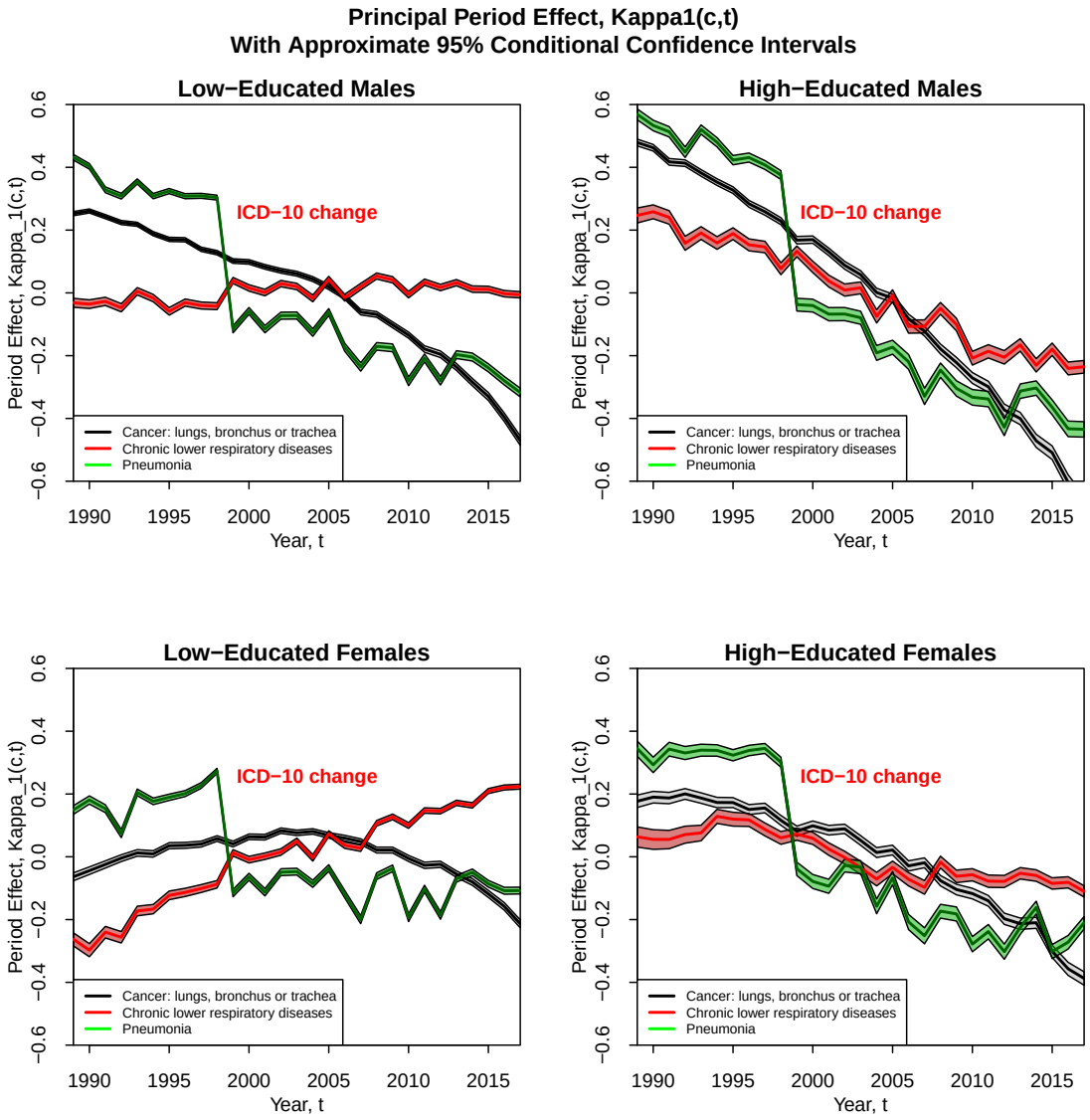

Figure 4: Fitted period effects $\kappa_1(c,t)$ for all four groups for selected causes of death. Changes in coding from ICD-9 to ICD-10 occurred between 1998 and 1999.

troughs, and the magnitude of these.

6.3 Fitted period effects

In the following discussion, we focus our attention on the primary period effect $\kappa_1(c,t)$ which affects the level of mortality for cause c across all ages and captures the headline changes in mortality from a specific cause through time.

6.3.1 Year to year volatility

In Figure 4, we compare time series of estimated $\kappa_1(c,t)$ for three causes: lung cancer, lower respiratory diseases (mainly COPD), and pneumonia. The change from ICD-9 to ICD-10 features in the results for pneumonia as a jump in $\kappa_1(c,t)$ reflecting the reallocation of some

cases of pneumonia to other causes. The plots also show approximate 95% conditional confidence intervals given all other parameter estimates. Confidence intervals are based on death counts from the given cause in a particular year, with fewer observed deaths resulting in a wider confidence interval. For example, there are many fewer lung cancer deaths in the high-educated than low-educated populations and so the confidence intervals are wider. For the three causes of death plotted, the confidence intervals are narrow enough that we can conclude that the year-to-year volatility in $\kappa_1(c,t)$ is genuine rather than just the result of sampling variation in the death counts.

For all four groups there is a striking, but not entirely surprising, difference between lung cancer, which is relatively smooth, and pneumonia, which is relatively volatile, with similar patterns of ups and downs for the four groups. COPD lies somewhere in between in terms of its volatility. Additionally, the pattern of year to year fluctuations in COPD approximately match those for pneumonia, even though it has a lower volatility. Lung cancer is a long-term illness and the relative smoothness of the plot suggests that short-term environmental fluctuations (e.g. weather patterns and seasonal flu) have relatively little impact on the final decline of an individual suffering from lung cancer. In contrast, pneumonia is a short-term and sometimes fatal illness (often triggered by influenza) which is highly sensitive to annual environmental fluctuations. COPD is also a long-term illness. But the plots suggests that the final decline of an individual can be accelerated by short-term environmental factors.

6.3.2 Relative changes in the prevalence of risk factors between groups

Figure 4 also allows us to make inferences about changes in smoking prevalence in one group *relative* to another. We focus here on COPD, but similar inferences can be made using lung cancer period effects. As a reference point, consider the period effect $\kappa_1(c,t)$ for low-educated males (top left). This is relatively flat. Now consider $\kappa_1(c,t)$ for high-educated males. This falls from around $+0.2$ to -0.2: a drop of 0.4 *relative* to low-educated males. If we assume that both groups have access to the same treatments and that most deaths are attributable to smoking, a potential explanation for this relative drop is that the prevalence of smoking in the high-educated group has dropped by about 40% relative to low educated males over the period 1989 to 2017. In turn this suggests that the high-educated group has been responding more strongly to the increasing public-health messages about the harmful effects of smoking.

Similar inferences can be made about low-educated females (prevalence rising relative to low-educated males) and high-educated females. Figure 6 allows us to make similar inferences about the prevalence of excess alcohol consumption in the four groups relative to each other.

6.3.3 Relative improvements in treatments

Figures 5 and 6 allow us to make inferences about improvements in the treatment of specific pairs of diseases *relative* to each other (rather than in absolute terms) where these have common controllable risk factors.

In Figure 5 we compare changes in lung cancer ($c = 11$) versus COPD ($c = 41$) mortality over time. At the first order, this is captured through changes in $\kappa_1(c,t)$. To aid discussion,

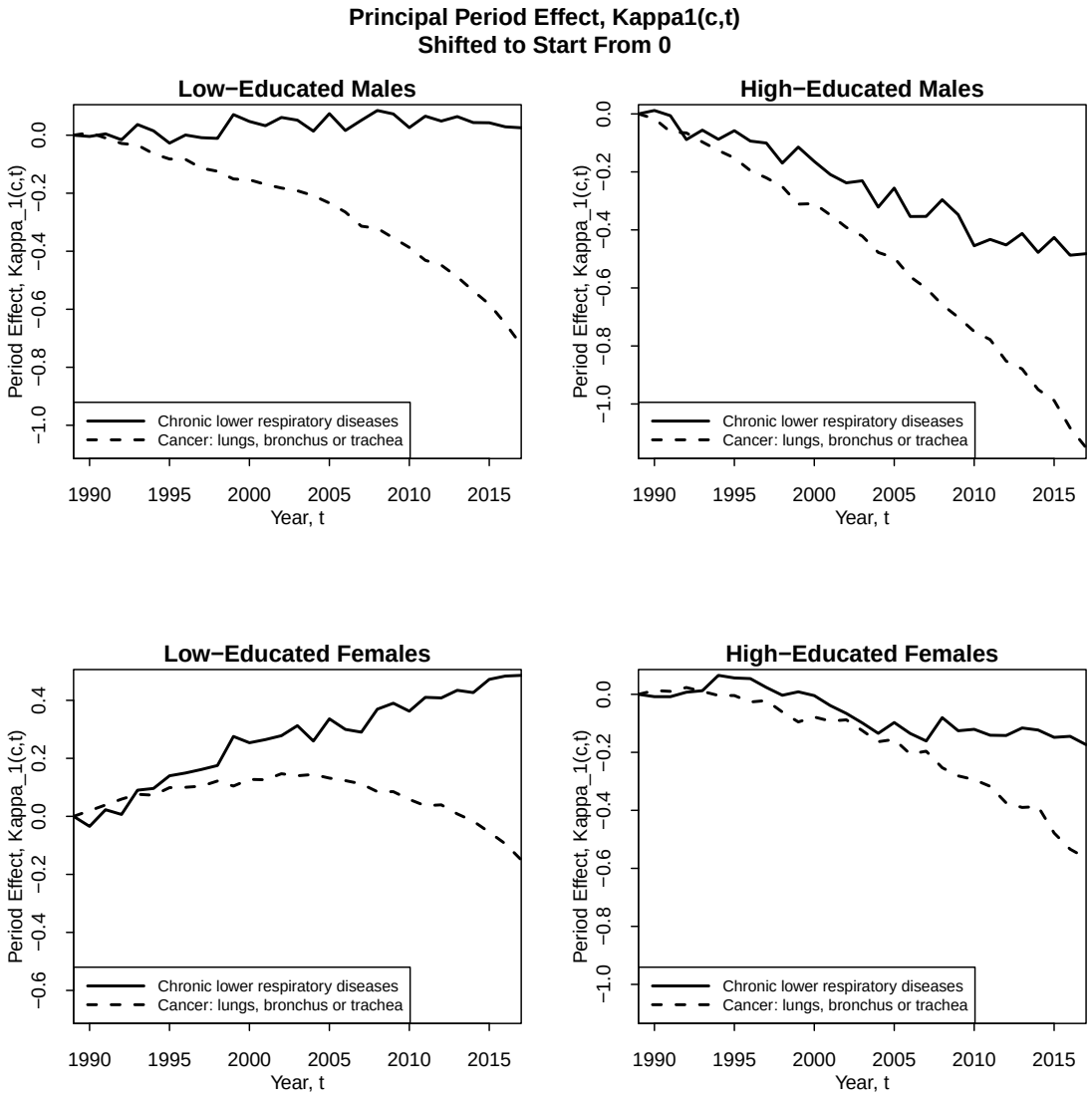

Figure 5: Fitted period effects the shifted $\tilde{\kappa}_1(c,t) = \kappa_1(c,t) - \kappa_1(c,1989)$ for all four groups for lung cancer and COPD.

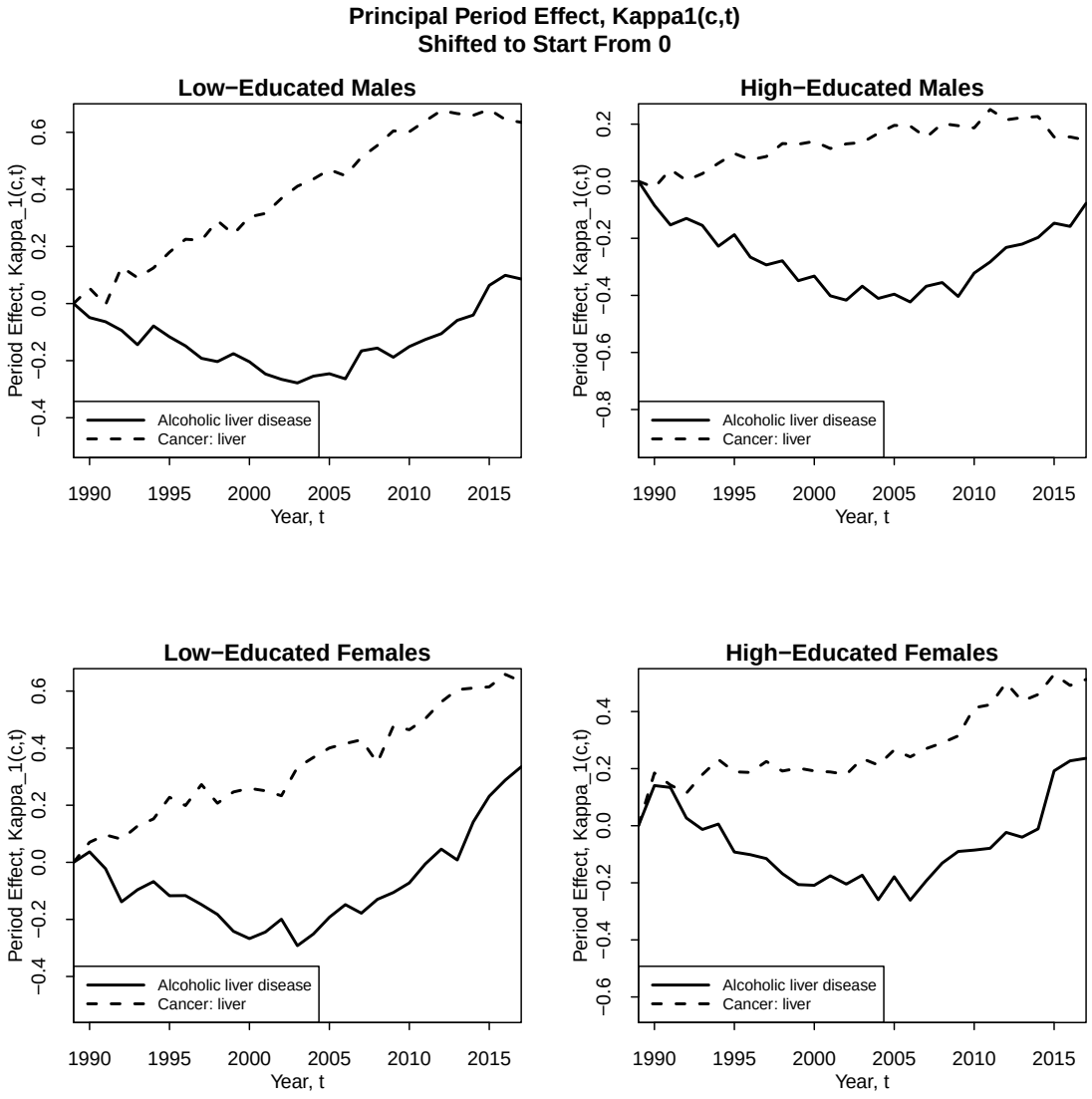

Figure 6: Fitted period effects the shifted $\tilde{\kappa}_1(c,t) = \kappa_1(c,t) - \kappa_1(c,1989)$ for all four groups for alcoholic liver disease and liver cancer.

period effects for the two causes have been shifted so that both start from 0: that is, we plot $\tilde{\kappa}_1(c,t) = \kappa_1(c,t) - \kappa_1(c,1989)$.

Data for these causes of death offers a valuable opportunity to consider relative improvements over time using $\tilde{\kappa}_1(c,t)$ again as a way to capture improvements over all ages. Smoking is, by far, the most significant risk factor for both causes of death. Consequently, if both causes of death had experienced the same level of improvements due to medical advances of various types, we would expect changes in $\tilde{\kappa}_1(11,t)$ and $\tilde{\kappa}_1(41,t)$ to approximately match each other. However, in Figure 5, we see a relatively-consistent divergence with $\tilde{\kappa}_1(11,t)$ falling at a faster rate than $\tilde{\kappa}_1(41,t)$. Our conclusion from this is that improvements in lung cancer mortality have happened at a faster rate than COPD, and the shape of the curves suggests a faster relative rate of improvement in the second half of the period. The graphical comparison does not allow us to quantify the absolute rate of improvement, as $\kappa_1(t)$ captures some elements of changes in smoking prevalence over time as well as medical advances. However, we can quantify the headline *relative rate of improvement* between the two causes for each population group. Males (low and high) and low-educated females all produce a similar result with a difference in the $\tilde{\kappa}_1(c,t)$'s around 0.65 to 0.75 by 2017: equivalent to an average relative rate of improvement of around 2.2% to 2.6% per annum. For high-educated females the gap is less wide equating to about 1.4% per annum, although the shape of the plot is similar.

Figure 6 shows the equivalent plots for liver cancer versus alcoholic liver disease. The upwards trajectories of $\kappa_1(c,t)$ for liver cancer point to significant increases in excess alcohol consumption in all four groups, especially the low educated, consistent with the more-general growth in deaths of despair highlighted by Case and Deaton (2015). Again the four plots exhibit similarities in terms of how the gap evolves over time, although the pattern match between the four groups is perhaps not as strong as Figure 5. Nevertheless, the similarities point to strong improvements in the treatment of alcoholic liver disease relative to liver cancer up to 2005. After 2005, relative improvements suggest a small reversal, with liver cancer catching up slightly. An alternative explanation is that deaths from the related cause of liver cirrhosis due to hepatitis C virus were, in the early years, being misclassified as alcoholic liver disease (Paula et al., 2010).

7 Summary

We have proposed a new age-period-cohort model (the Common Cohort Effect or CCE model) for modelling mortality by cause of death, sex and education level using highly-granular mortality data. Mortality records are subdivided into 51 causes of death, and this granularity allows us to focus on specific causes that have links to single controllable risk factors. A key feature of the model is the use of a small number of common cohort effects rather than 51 individual cohort effects. Use of the Bayes Information Criterion results in a strong preference for the CCE model with an conclusion that there is a significant degree of overfitting in the model with 51 individual cohort effects. Common cohort effects can be linked to specific controllable risk factors such as as smoking prevalence and this allows us to gain insights into the drivers of changes in mortality from individual causes of death as well as all-cause mortality.

The model outputs include period effects alongside the common cohort effects. Analysis

of the estimated period effects gives us further insights into mortality developments over time.

- Year-to-year volatility in period effects (and, therefore, death rates) varies considerably from one cause to another. The model allows us to quantify how much volatility there is around what might be a smooth underlying trend. Some causes of death proceeded by a long-term illness have low volatility. Others causes of death, such as pneumonia, are preceded by a short-term illness and are consequently more volatile. But others involving long-term illness such as COPD exhibit some volatility suggesting that short-term environmental factors can accelerate an individual's final decline.
- We can compare period-effect trajectories for different groups and make inferences about relative changes in the prevalence of smoking and excess alcohol consumption.
- We can compare period-effect trajectories for pairs of causes of death that have the same single risk factor and make inferences about relative improvements in the treatment of the two causes.

The focus of this paper has been on in-sample dynamics rather than out-of-sample projections. However, through this understanding, we will be better placed to develop scenarios for the future development of cause-specific and all-cause mortality. Additionally, it is intended that the work will be useful for developing strategies for improving mortality outcomes. We leave these outcomes for future work.

References

Alai, D.H., Arnold, S., Bajekal, M., and Villegas, A.M. (2018) Mind the Gap: A Study of Cause-Specific Mortality by Socioeconomic Circumstances. *North American Actuarial Journal*, 22: 161-181.

Arnold, S., and Glushko, V. (2021) Cause-specific mortality rates: common trends and differences. *Insurance: Mathematics and Economics*, 99: 294-308.

Barbieri, M. (2017) Expanding the Human Mortality Database to include cause-of-death information. Society of Actuaries. `https://www.soa.org/493831/globalassets/assets/files/research/projects/2017-hmd-cause-of-death-brief.pdf` (Accessed 29/9/2023)

Boumezoued, A., Coulomb, J.-B., Klein, A., Louvet, D., and Titon, E. (2019) Modeling and forecasting cause-of-death mortality. Society of Actuaries research report. `https://www.soa.org/4b140c/globalassets/assets/files/resources/research-report/2019/cod-mortality-f` `pdf` (Accessed 28/9/2023)

Cairns, A.J.G., Blake, D., Dowd, K., Coughlan, G.D., Epstein, D., Ong, A., and 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, 1–35.

Cairns, A.J.G., and Redondo Lourés, C. (2023) Higher-Age US Mortality By Education and Cause of Death: Trends, Inequality and Controllable Risk Factors. To appear in the *2023 Living to 100 Monograph*, Society of Actuaries. `https://www.soa.org/programs/living-to-100/monographs/`

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.

Chetty, R., Stepner, M., Abraham, S., Lin, S., Scuderi, B., et al. (2016) The Association Between Income and Life Expectancy in the United States, 2001–2014. *Journal of the American Medical Association*, 315: 1750-1766.
Dowd, K., Cairns, A.J.G., and Blake, D. (2020) CBDX: A workhorse mortality model from the Cairns-Blake-Dowd family. *Annals of Actuarial Science* 14: 445–460.
GBD US Health Disparities Collaborators (2023) Cause-specific mortality by county, race, and ethnicity in the USA, 2000–19: a systematic analysis of health disparities. *Lancet*, 402: 1065–82.
Holford, T.R. (1991) Understanding the effects of age, period, and cohort on incidence and mortality rates. *Annual Review of Public Health*, 12: 425–457.
Human Cause-of-Death Database (2023) French Institute for Demographic Studies (France) and Max Planck Institute for Demographic Research (Germany). Available at `www.causeofdeath.org`.
Human Mortality Database (2023) Max Planck Institute for Demographic Research (Germany), University of California, Berkeley (USA), and French Institute for Demographic Studies (France). Available at `www.mortality.org`.
Hunt, A., and Blake, D. (2014) A general procedure for constructing mortality models. *North American Actuarial Journal* 18: 116–138.
Lee, R. D., and Carter, L.R. (1992) Modeling and forecasting U.S. mortality. *Journal of the American Statistical Association*, 87: 659–71.
Li, H., and Hyndman, R.J. (2021) Assessing mortality inequality in the U.S.: What can be said about the future? *Insurance: Mathematics and Economics*, 99: 152-162.
Osmond, C., and Gardner, M.J. (1982) Age, period and cohort models applied to cancer mortality rates. *Statistics in Medicine*, 1: 245-259.
Paula, H., Asrani, S.K., Boetticher, N.C., Pedersen, R., Shah, V.H., Kim, W.R. (2010) Alcoholic liver disease-related mortality in the United States: 1980-2003. American Journal of Gastroenterology, 105: 1782-7.
Plat, R. (2009) On stochastic mortality modeling. *Insurance: Mathematics and Economics*, 45: 393-404.
Preston, S.H., and Wang, H. (2006) Sex mortality differences in the United States: the role of cohort smoking patterns. *Demography*, 43: 631-646.
Redondo Lourés, C., and Cairns, A.J.G. (2020) Mortality in the US by education level. *Annals of Actuarial Science* 14: 384–419.
Redondo Lourés, C., and Cairns, A.J.G. (2021) Cause of Death Specific Cohort Effects in U.S. Mortality. *Insurance: Mathematics and Economics* 99: 190-199.
Renshaw, A. E., and Haberman, S. (2006) A Cohort-Based Extension to the Lee-Carter Model for Mortality Reduction Factors. *Insurance: Mathematics and Economics*, 38: 556–70.
The US Burden of Disease Collaborators (2018) The State of US Health, 1990-2016 Burden of Diseases, Injuries, and Risk Factors Among US States. *Journal of the American Medical Association*, 319: 1444-1472
Villegas, A.M., Bajekal, M., Heberman, S., and Zhou, L. (2023) Key Drivers of Long-Term Rates of Mortality Improvements in the United States: Period, Cohort, and Cause of Death Analysis, 1959–2016. To appear in *North American Actuarial Journal*. DOI: 10.1080/10920277.2023.2167834

A Optimisation algorithm for the CBDX-CCE model

For a given set of death rates, $m(c,t,x)$, and exposures $E(t,x)$, deaths, $D(c,t,x)$ are assumed to have a Poisson distribution with mean $m(c,t,x)E(t,x)$. We maximise the log-likelihood function

$$l(\theta; D) = \sum_{c,t,x} w(c,t,x) \left\{ D(c,t,x) \log m(c,t,x) - m(c,t,x)E(t,x) \right\} + \text{constant}$$

where $w(c,t,x) = 0$ or 1 are weights that allow exclusion of those cohorts with very few observations, and θ is the full set of age, period and common cohort effects to be estimated (subject to identifiability constraints).
Example for $n = 3$ common cohort effects:

- For each cause of death $c = 1, \ldots, 51$:
  - Exact optimisation of $\alpha(c,x)$ given the current values of $\kappa_k(c,t)$, $\delta_j(c)$ and $\chi_j(t-x)$.
  - Update the $\kappa_k(c,t)$ (1 step) using Newton's method, given the $\alpha(c,x)$, $\delta_j(c)$ and $\chi_j(t-x)$.
  - Update the $\delta_j(c)$ using Newton's method, given the $\alpha(c,x)$ $\kappa_k(c,t)$ and $\chi_j(t-x)$. Exceptions: leave $\delta_2(c_1) = \delta_3(c_1) = \delta_3(c_2) = 0$ (constraint 2B).
- Update the $\chi_j(t-x)$ (exact optimisation), given the $\alpha(c,x)$ $\kappa_k(c,t)$ and $\delta_j(c)$.
- For each of the common cohort effects, j, apply the 4 constraints $\sum_y (y-\bar{y})^u \chi_j(y) = 0$.
- Apply the orthogonality constraint.

  Use of the Cholesky decomposition ensures that $\delta_2(c_1)$, $\delta_3(c_1)$ and $\delta_3(c_2)$ remain equal to zero.
- Repeat until the log-likelihood converges.