Cohort Effects in US Cause-of-Death Mortality Data

Andrew J.G. Cairns
joint work with Cristian Redondo Lourés

Heriot-Watt University, Edinburgh

Outline

- Introduction
- Data
- The CBDX model
- CBDX-I: Cohort effects for *individual* causes of death
- The Common Cohort Effects (CCE-n) model
- Results and discussion

Introduction

- Good quality data by cause of death is now available
  CoD $\times$ single or 5-year age group $\times$ single year
- Analysis:
  - Often use 5 to 7 CoD groups (e.g. cancers)
  - But a more granular approach can give more insight into what has happened in the past
- How do we make best use of this data?
- E.g. how can we exploit this data to get better insights into historical trends in all-cause mortality?

Data

- Details: Redondo Lourés and Cairns (2019, 2021)
- US males and females
- Sources: CDC (deaths); HMD, Current Population Survey (exposures)
- By sex
- By education level: low ($\leq$ high school graduation); high ($>$ high school)
- Single ages (40-84)
- Single years (1989-2017)
- 51 causes of death

- Excluded: the oldest and youngest cohorts (too few observations)
  Included: cohorts born in 1915-1970
- US analysis here $\Rightarrow$ what is potentially feasible for other countries e.g. using the HMD causes of death database (`causesofdeath.org`)

Cause of Death Groupings

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	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	Liver cirrhosis	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				

Detail $\Rightarrow$ able to separate causes with and without significant risk factors or inequality

E.g. cancers: some with strong single risk factors; some with multiple risk factors; some with no risk factors

Modelling reality: < 51 causes. E.g. males $\Rightarrow$ no ovarian cancer

Significant cohort effects and controllable risk factors

- Hypothesis:
  For a specific cause of death:
  - A significant cohort effect ⇒ one or more significant controllable risk factors
  - e.g. smoking, poor diet, exercise, alcohol etc.
  - bigger cohort effect ⇒ bigger relative risk associated with specific risk factors

  - Significant gap between high and low-educated also ⇒ one or more significant controllable or preventable risk factors
    even if there is no significant cohort effect
- 51 causes of death ⇒ greater insight into individual controllable risk factors
- Cause-of-death cohort effects + controllable risk factors
  ⇒ insight into all-cause mortality cohort effects

$$\log m(t,x) \quad = \quad \alpha(x) + \underbrace{\sum_{k=1}^{3} \beta_k(x)\kappa_k(t) + \gamma(t-x)}_{\text{CBD-M7 model}}$$

where

$$\beta_1(x) = 1, \quad \beta_2(x) = x - \bar{x}, \quad \beta_3(x) = (x-\bar{x})^2 - \sigma_x^2 \quad \text{(fixed M7 age effects)}$$

Seven identifiability constraints:

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

and

$$\sum_t \kappa_1(t) = 0, \quad \sum_t \kappa_2(t) = 0, \quad \sum_t \kappa_3(t) = 0$$

Model each of the N_{cod} causes of death, c, individually:

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

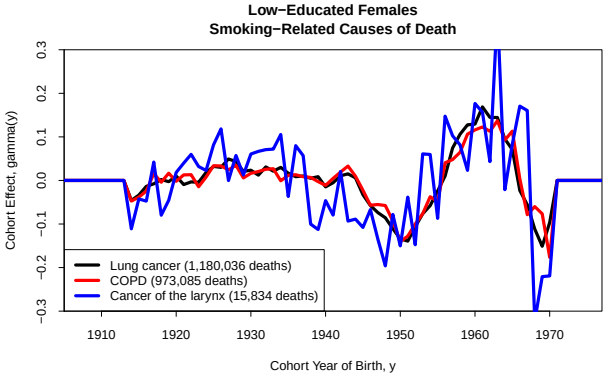
Identifiability constraints:

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

and

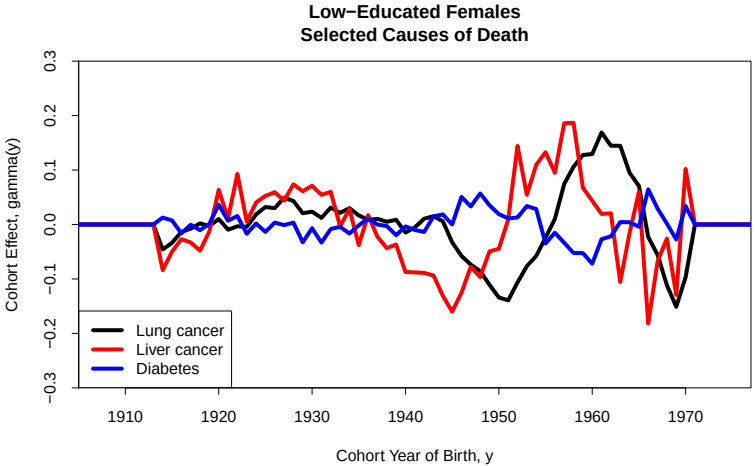
$$\sum_t \kappa_1(c,t) = 0, \quad \sum_t \kappa_2(c,t) = 0, \quad \sum_t \kappa_3(c,t) = 0$$

CBDX-I: Selected Cohort Effects, $\gamma(c, y)$



- Reported death counts: 1989-2017, ages 40-84
- Smoking is the main controllable risk factor for lung and laryngeal cancers, and COPD
- Very similar cohort effects
- But low death counts $\Rightarrow$ significant sampling variation in cancer of the larynx

CBDX-I: Selected Cohort Effects, $\gamma(c, y)$



- Other causes of death with different controllable risk factors have distinctly different cohort effects

Do we need 51 distinct cohort effects?

- Lung cancer, COPD and cancer of larynx $\Rightarrow$??? a common cohort effect "$\chi(\text{smoking}, y)$"
- $\chi(\text{smoking}, y)$ might have a single scaling factor for each cause of death
  scaling factor $\sim$ relative risk
- Other causes of death also have a very similar shape linked to other controllable risk factors
- So perhaps we just need a small number of cohort effects that reflect variation in a small number of controllable risk factors
  - $\chi(\text{smoking}, y)$
  - $\chi(\text{diet/exercise/obesity}, y)$
  - $\chi(\text{alcohol}, y)$
  - ....
- Be aware: cohort effects for risk factors might be correlated
  e.g. a tendency by cohort to lead a generally healthy or unhealthy lifestyle

Model the N_{cod} causes jointly with common cohort effects

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

where $\chi_1(y), \ldots, \chi_n(y)$ are n common cohort effects that apply to each cause of death

and the $\delta_j(c)$ control the contribution of each common cohort effect, $\chi_j(y)$, to the cause-specific cohort effect $\gamma(c,y)$

$$\log m(c,t,x) \quad = \quad \alpha(c,x) + \sum_{k=1}^{3} \beta_k(x)\kappa_k(c,t) + \sum_{j=1}^{n} \delta_j(c)\chi_j(t-x)$$

- Motivation: the n common cohort effects can be linked to n significant, underlying controllable risk factors
  e.g. smoking
- Each cause of death, c, has scaling factors $\delta_1(c), \ldots, \delta_n(c)$ attached to the common cohort effects
- Hypothesis
  Example: $\chi_1(y)$ links to cohort smoking prevalence & intensity
  Then, the size of $\delta_1(c)$ links to the *relative risk* of smoking for cause of death c.
  - If smoking is not a risk factor for CoD c then $\delta_1(c) = 0$
  - Bigger the relative risk $\Rightarrow$ bigger $\delta_1(c)$

- For each $j = 1, \ldots, n$:

$$\chi_j(y) \to \chi_j(y) + a_{j,0} + a_{j,1}(y - \bar{y}) + a_{j,2}(y - \bar{y})^2 + a_{j,3}(y - \bar{y})^3$$

with corresponding adjustments to the $\alpha(c, x)$ and the $\kappa_k(c, t)$ for all c which depend on $a_{j,0}$, $a_{j,1}$, $a_{j,2}$, $a_{j,3}$ and $\delta_j(c)$

$\Rightarrow$ similar identifiability constraints to the CBDX model

- $4n$ constraints on the n common cohort effects
- $3N_c$ constraints on the cause-of-death specific period effects

$$\begin{aligned}
\text{Define:} \quad \text{(row vector)} \ \gamma(c) &= (\gamma(c,1), ..., \gamma(c,N_y)) \\
\text{(row vector)} \ \delta(c) &= (\delta_1(c), \ldots, \delta_n(c)) \\
G &= \begin{pmatrix} \chi_1(1) & \ldots & \ldots & \chi_1(N_y) \\ \vdots & & & \vdots \\ \chi_n(1) & \ldots & \ldots & \chi_n(N_y) \end{pmatrix}
\end{aligned}$$

$$\begin{aligned}
\text{Then} \ \gamma(c) &= \delta(c)G \quad \text{in vector/matrix notation} \\
&= \delta(c)A^{-1}AG
\end{aligned}$$

where A is an arbitrary *non-singular* $n \times n$ matrix

- So we can make alternative linear combinations of the $\chi_j(y)$
- with corresponding adjustments to the $\delta_j(c)$
- with no impact on the cause-specific cohort effects, $\gamma(c,y)$

Possible solutions (n common cohort effects):
Focus (e.g.) on the first n *cause-specific* cohort effects

$$\text{Define } H = \begin{pmatrix} \gamma(1,1) & \dots & \gamma(1,N_y) \\ \vdots & & \vdots \\ \gamma(n,1) & \dots & \gamma(n,N_y) \end{pmatrix} \text{ and } \Delta_n = \begin{pmatrix} \delta_1(1) & \dots & \delta_n(1) \\ \vdots & & \vdots \\ \delta_1(n) & \dots & \delta_n(n) \end{pmatrix}$$

- Option 1: Constraining $\Delta_n = I_n$ ($n \times n$ identity matrix) avoids the $\delta(c)A^{-1}AG$ identification problem.
- Option 2: Alternatively constrain

$$\tilde{\Delta}_n = \begin{pmatrix} 1 & 0 & \dots & \dots & 0 \\ \delta_1(2) & 1 & 0 & \dots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \delta_1(n-1) & \dots & \ddots & 1 & 0 \\ \delta_1(n) & \dots & \dots & \delta_{n-1}(n) & 1 \end{pmatrix}$$

Identifiability problems: II (cont.)

Speeding up convergence:

- Instead of constraining causes $1, \ldots, n$
  choose to constrain causes of death $c_1, \ldots, c_n$ with distinctly different $\gamma(c, y)$
- The choice of $c_1, \ldots, c_n$ might depend on the underlying population
  e.g. depends on whether or not some controllable risk factors by cohort are highly correlated or not

- Option 2 seems to be faster than Option 1
  but still slow for some populations
- But Option 2 is an incomplete set of constraints
  so need to apply full constraints after convergence

- Option 3: Orthogonality constraint
  For $j, k = 1, \ldots, n$

$$\sum_y \chi_j(y)\chi_k(y) = \begin{cases} 0 & \text{if } j \neq k \\ 1 & \text{if } j = k \end{cases}$$

  Additionally, for a given $c_1, \ldots, c_n$:

$$\begin{aligned} \gamma(c_1, y) &= \delta_1(c_1)\chi_1(y) && (\gamma(c_1, y) \text{ constr. to depend on } \chi_1(y) \text{ only}) \\ \gamma(c_2, y) &= \delta_1(c_2)\chi_1(y) + \delta_2(c_2)\chi_2(y) \\ &\vdots \quad \vdots \quad \vdots \\ \gamma(c_n, y) &= \delta_1(c_2)\chi_1(y) + \ldots + \delta_n(c_n)\chi_n(y) \end{aligned}$$

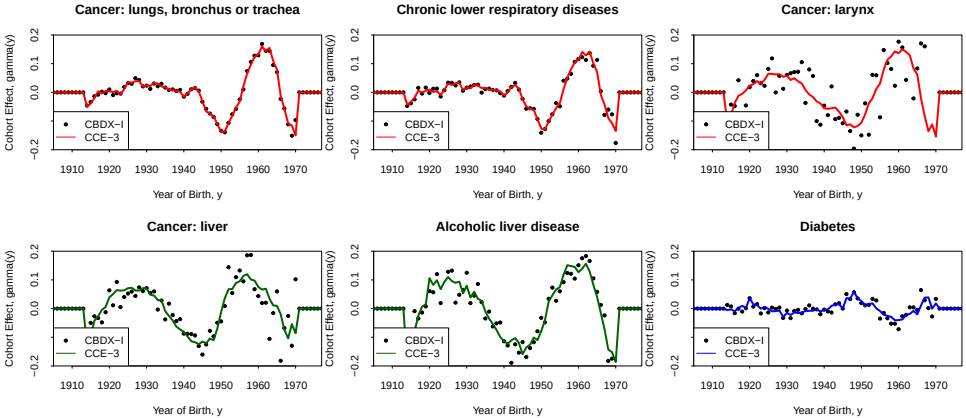
- Option 3 is similar to Option 2
- But orthogonality $\Rightarrow$ complete set of constraints
- Option 3 converges faster than Option 2

Results

	Model:	CBDX-I			CCE-3		
Group	# obs, N_{obs}	maximum log-lik, $\hat{l}$	effective # params, ν	BIC	maximum log-lik, $\hat{l}$	effective # params, ν	increase in BIC
Males-Low	**55440**	-205252	8910	-253914	-206549	6003	+14580
Males-High	55440	-185460	8910	-234123	-186782	6003	+14555
Females-Low	**59136**	-209396	9504	-261609	-210775	6390	+15729
Females-High	59136	-185151	9504	-237365	-186435	6390	+15824

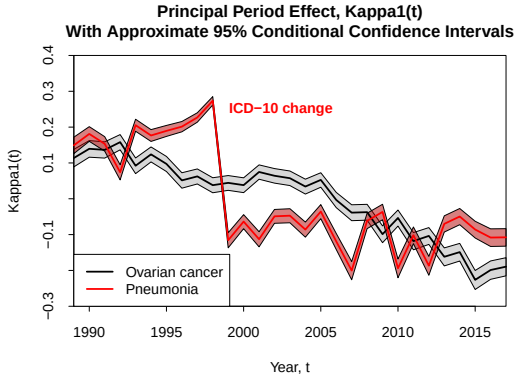
- $BIC = \log \hat{L} - 0.5\nu \log N_{obs}$
- All four populations: BIC $\Rightarrow$ CCE-3 is the best model
- Use of common cohort effects is also better at the level of individual causes of death

Examples: Low educated females; CDBX-I versus CCE-3



- Good correspondence between CCE-3 and CBDX-I
- But with reduced noise in the estimated cohort effects
- Distinctly different cohort effects are now clearer

Examples: Short-term versus long-term illness (low educated females)



- Pneumonia and ovarian cancer have similar number of deaths
- $\kappa_1(c,t) \Rightarrow$ headline variation in mortality rates at all ages
- But $\kappa_1(\text{pneumonia},t)$ is much more volatile than $\kappa_1(\text{ovarian},t)$
- Year-to-year variation for F-low highly correlated with F-high and M-low/high
- Hypothesis: higher volatility if short-term illness + susceptible to variable external factors

Discussion I

- Caution: e.g. smoking as a controllable risk factor
- $\gamma(c, y)$ is not the same as the underlying controllable risk factor
  - Impact depends on the prevalence and intensity of the controllable risk factor
    (e.g. 50% smoking prevalence: 40/10 heavy/light smokers different from 10/40)
  - Identifiability $\Rightarrow$ risk-factor$(y) \sim \gamma(y)+$ cubic(y)
  - Magnitude of $\gamma(c, y)$ depends on the relative risk linking the risk factor to CoD death rate
- If treatment of disease/illness is consistent between sub-populations
  then modelled period and cohort effects $\Rightarrow$ give insight into relative changes in the underlying behaviour

Discussion II

- Here: three common cohort effects.
  But is that the right number?
- Interpreting the common cohort effects, $\chi_l(y)$, is potentially challenging
  e.g. prevalence of controllable risk factors might be correlated
- Common cohort effects can help explain cohort effects estimated at the all-cause level
- Methodology can be adapted to HMD cause of death data
  - Need to handle 5-year age groups
    Robustness experiment: group US data into 5-year age bands and compare with single-age results
  - Smaller populations $\Rightarrow$ merge some smaller causes of death into coherent groups by controllable risk factors
    Robustness experiment: group US causes of death in the same way and compare results with 51-CoD results

Postscript: The impact of the Covid-19 pandemic

- 2020-21: very disruptive to the reporting of individual causes
  - So time series modelling just got a whole lot harder
  - How to make forecasts?
- Beyond 2021: possible scenarios
  - impact of "long covid" and related impairments
  - increased cancer deaths due to later diagnosis
  - more severe flu pandemic due to reduced immunity
  - fewer pneumonia deaths due to continued physical distancing and better hygiene

Summary

- Work in progress
- We propose the Common Cohort Effect model as a way to link cohort effects for different causes of death to underlying controllable risk factors
- US data: Three common cohort effects were found to be very effective
- Potential to add insight into the contribution of different causes to all-cause mortality improvements
- Potential to provide insight into the effect of specific controllable risk factors at the all-cause level

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

HERIOT WATT UNIVERSITY

Actuarial Research Centre
Institute and Faculty of Actuaries