Supplementary material to: Forecasting Causes of Death using Compositional Data Analysis: the Case of Cancer Deaths

A: The VECM model and Johansen trace test

The VECM-CoDA model extends existing mortality models by modelling multiple time trends for different causes of death. Because of dependence between causes it is important to model the interactions between cause specific time trends too. The Johansen trace test and cointegrated vector error correcting model (VECM) are standard methods for such modelling and these methods are described in this section together with some basic time series analysis methodology.

The VECM is often used in econometrics when forecasting macroeconomic variables. It is also used for mortality forecasting by Lazar and Denuit (2009) and Arnold-Gaille and Sherris (2013). Let $\mathbf{y_t}$ denote an n-dimensional random vector. Further, let $\mathbf{y_t}$ be integrated of order one meaning that the series is non-stationary but has stationary first differences. Note that the probability distribution of a stationary variable does not depend on time but only the length of the series. Cointegration is defined as the existence of a stationary relationship between two non-stationary variables, meaning that the two variables may drift apart in the short run but follow the same underlying pattern in the long run. This strong relationship can occur if the two variables are affected by the same underlying process. Cointegrating VECM determines whether cointegrating relationships can be found and provides estimates and forecasts of dependent time trends based on the number of cointegrating relationships. The cointegrating VECM is best understood from its equivalent, the vector autoregression model.

A vector autoregression model VAR(p) of order p estimates a system where $\mathbf{y_t}$ is described by p time lags of $\mathbf{y_t}$ in a system of linear relationships. Following the notation in Juselius (2006) and Lazar and Denuit (2009) the model can be written as,

$$\boldsymbol{y_t = \Pi_1 y_{t-1} + \Pi_2 y_{t-2} + \cdots + \Pi_p y_{t-p} + \epsilon_t}, \tag{1}$$

where $\mathbf{\Pi_i}$ is an $n \times n$ matrix that contains the correlations in the system. The VAR(p) can equivalently be written in its error correction form expressed in first differences, denoted by Δ such that $\Delta y_t = y_t - y_{t-1}$:

$$\boldsymbol{\Delta y_t = \Pi y_{t-1} + \sum_{i=1} \Gamma_j \Delta y_{t-i} + \epsilon_t}, \tag{2}$$

where $\mathbf{\Pi} = \sum_{i=1}^{p} \Pi_i - I_n$, I is an $n \times n$ identity matrix and equal to $\Gamma_j = -\sum_{j=i+1}^{p} \Pi_j$. Johansen (1988, 1991) shows that the number of cointegration relationships in the n-dimensional system is equal to the rank of $\mathbf{\Pi}$.

Three situations can occur for Π:

- Π has rank zero meaning there are no long run relationships among the series, but the series are non-stationary.

- Π has full rank which means that all of the series are stationary.
- Π has reduced rank, $r > 0$, thus there exist both stationary and non-stationary series and r stable long run relationships exist. (Juselius, 2006)

Deterministic elements are also incorporated in the VECM model when necessary to account for deterministic trends.

The methodology described and the Johansen test are used in a CoDA model to allow for several potentially dependent trends among the causes of death.

Results of the Johansen test

We test for co–integration in the first rank approximation, $k^1_{t,i}$, in the VECM-CoDA model and results for French females and Dutch males are shown in Table A1 and Table A2, respectively. The number of cointegrating relationships (r) is found by a sequence of tests. Following the procedure described in Juselius (2006), we start by testing $r = 0$, then $r <= 1$ and so on, until we fail to reject. This procedure gives the correct inference when determining r.

Table A1: Johansen trace test results for French females

Hypothesized no. of r	Test statistic	5%
r <= 6	7.42	12.25
r <= 5	15.83	25.32
r <= 4	28.10	42.44
r <= 3	57.11	62.99
r <= 2*	96.13	87.31
r <= 1*	147.20	114.90
r = 0*	220.66	146.76

Note: Asterisk indicates a non-rejected hypothesis at a 5% significance level.

Table A2: Johansen trace test results for Dutch males

Hypothesized no. of r	Test statistic	5%
r <= 9	6.86	12.25
r <= 8	17.91	25.32
r <= 7	34.18	42.44
r <= 6	54.21	62.99
r <= 5	81.54	87.31
r <= 4*	114.95	114.90
r <= 3*	157.02	146.76
r <= 2*	201.15	182.82
r <= 1*	257.07	222.21
r = 0*	344.46	263.42

Note: Asterisk indicates non-rejected hypothesis at a 5% significance level.

The results indicate that r equals 3 and 5, meaning that $n = 3$ and $n = 5$ stochastic trends are found, for French females and Dutch males respectively. The CT-CoDA model's assumption of one trend is thereby rejected as 3 and 5 trends were driving the system for the two populations. The number of co-integrating relationships is used when calculating the forecast from the VECM-CoDA model.

B: Selecting a subset of causes of death

In the analysis of selecting a subset of causes for the forecast of cancer deaths two statistical methods were used: elastic net and hierarchical clustering. We give a short summary of the two methods and further details can be found in the given references.

Elastic Net

We used the elastic net suggested by **?** to select a subset of causes when forecasting cancer. Elastic net is a shrinkage method used for variable selection in a linear regression. The elastic net method is best understood in relation to ordinary linear regression. Let us consider the response variable vector $\mathbf{y}$ of dimension $n \times 1$ and p predictors in a matrix $\mathbf{X}$ with dimensions $n \times p$ and the linear regression,

$$\mathbf{y} = \mathbf{X}\beta' + \epsilon, \tag{3}$$

where $\beta' = (\beta_{\mathbf{0}}, ..., \beta_{\mathbf{p}})$ is a vector of coefficients and ϵ an error term. Elastic net selects β using the following objective function:

$$\hat{\beta} = \arg\min ||\mathbf{y} - \mathbf{X}\beta'||^{\mathbf{2}} + \lambda_{\mathbf{1}}|\beta'| + \lambda_{\mathbf{2}}||\beta'||^{\mathbf{2}}, \tag{4}$$

where λ_1 and λ_2 are fixed non-negative penalty parameters. The objective function is the usual sum of squared residuals, typically used in OLS estimation for a linear regression, plus an L_1-norm penalty weighted by the constant λ_1 and a quadratic penalty weighted by the constant λ_2. By this, when a variable is not a good predictor for y, it will be assigned a weight of zero, i.e. $\beta_i = 0$.

The L_1-norm penalty corresponds to the penalty in the shrinkage technique called Lasso (**?**) and ensures variable selection of the most important predictors. The L_1-norm penalty has limitations as it does not allow for any grouping effect: that is, a group of correlated variables that are all important predictors of the response variable. The L_1-norm penalty will only select one of these predictors. To include this multicollinearity is potentially important when selecting a subset of causes as we want to reduce noise from causes that are not relevant for the prediction of cancer while keeping all the relevant ones. The quadratic penalty, ensures a grouping effect so that all variables in a correlated group are selected.

The elastic net method is used to select causes that predict cancer deaths for each age group, applying it to each age group separately.

Hierarchical Clustering

Cluster analysis seeks to determine related groups in a set of observations. The underlying idea is to find groups of observations, in a data matrix, which are characterized by a short distance to each other and large distance to other observations groups. We use a hierarchical clustering algorithm, based on a bottom-up approach where the most similar causes are linked first and thereafter merged with other pairs. More specifically, the algorithm is divided into three steps (cite matlab function/ documentation here).

- Step 1: The Euclidean distance between each cause is found and used to calculate a dissimilarity matrix.
- Step 2: The most similar pairs of causes are linked based on information from the dissimilarly matrix so that binary clusters are found. Next the binary clusters are linked based on the distance between each binary cluster. This process continues until all causes are linked and form a hierarchical cluster tree.
- Step 3: Finally, the specific cuts in the hierarchical cluster tree are determined i.e. how the specific binary cluster groups are connected. The binary clusters and links between them are divided into well–defined clusters based on where the causes are most densely packed. No specific number of clusters is selected, instead all links are specified creating natural divisions between the clusters based on the distance between them.

The hierarchical cluster algorithm is applied to each age group separately. Cluster diagrams are thereby constructed for each age showing which causes are most related in the data period. For further details about hierarchical clustering, see **?**.

Cause dependence and selection results

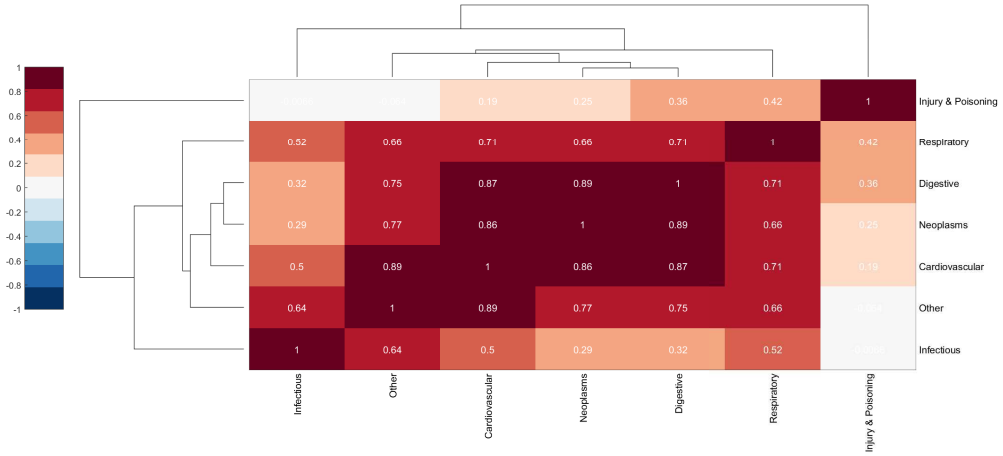
In this section we analyse whether all causes are important when forecasting cancer or whether a subset of them is more useful. In order to select a subset we first make a descriptive analysis of the dependence using hierarchical clustering analysis, and secondly by using elastic nets (**?**) we select which subsets of causes are potentially most important when forecasting cancer. The causes will be dropped one by one and 20 years out-of-sample forecast errors calculated rolling the origin, similar to the scheme described in Section 2.7. Dropping causes will automatically increase the proportion of cancer deaths in the same initial population. Thus, the RMSE cannot be used as an error measure for this analysis as it will be higher because of the relative increase in cancer. The Symmetric Mean Absolute Percentage Error (SMAPE), suggested by Tofallis (2015), is used instead but weighted with the average distribution of cancer deaths over age so that age groups with the most deaths are weighted the most.

We only consider sub-sampling of causes in the CoDA models; the LC model does not include dependence among the causes.

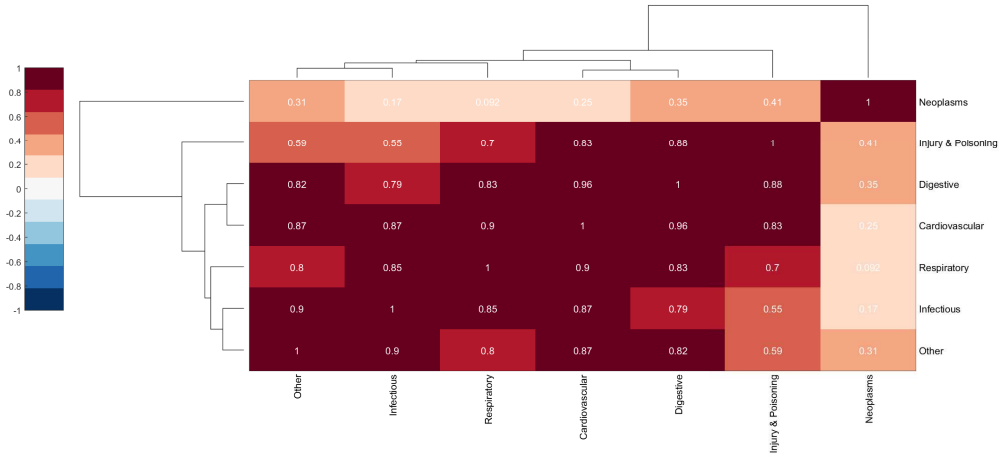
Hierarchical clustering results

The hierarchical clustering is applied using the centred age-specific deaths as this is the central input to the CoDA models. This section shows results for selected age groups as an illustration, but similar results are obtained for the other age groups.

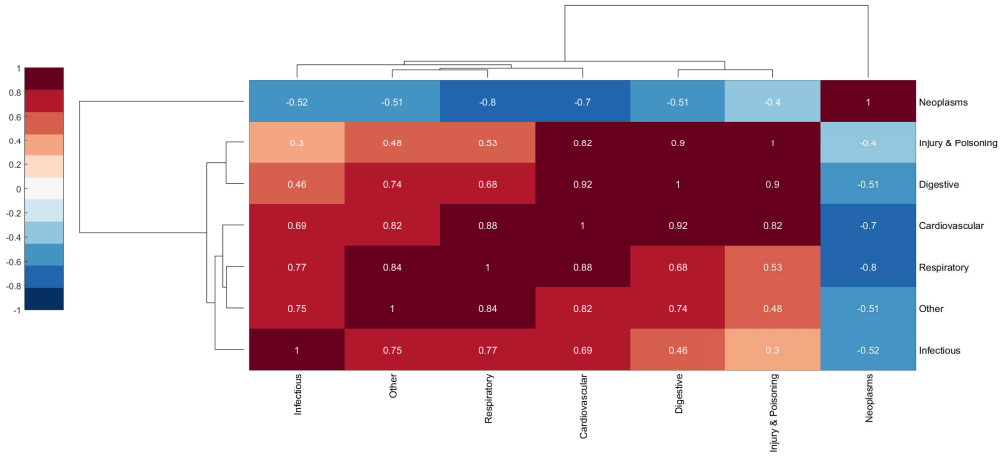
Figure B1: Hierarchical clustering of causes of death for French females for selected ages



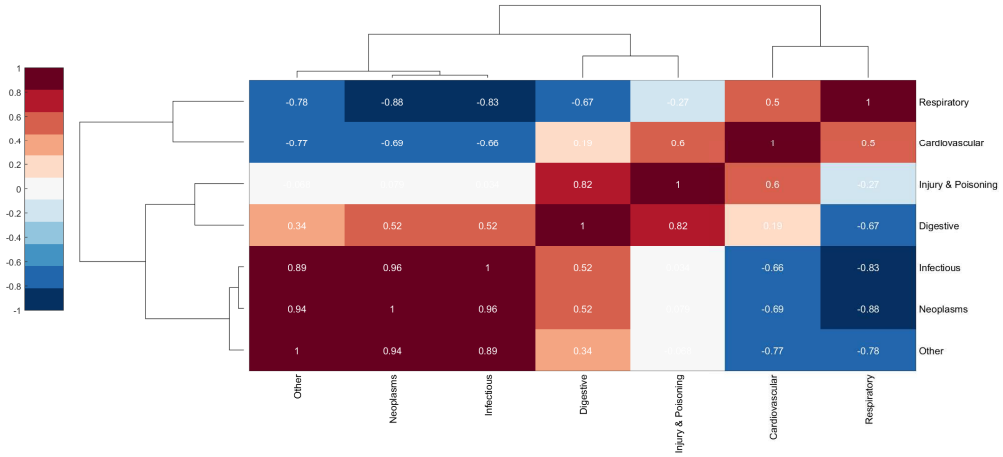
Age 25-29 years



Age 60-64 years



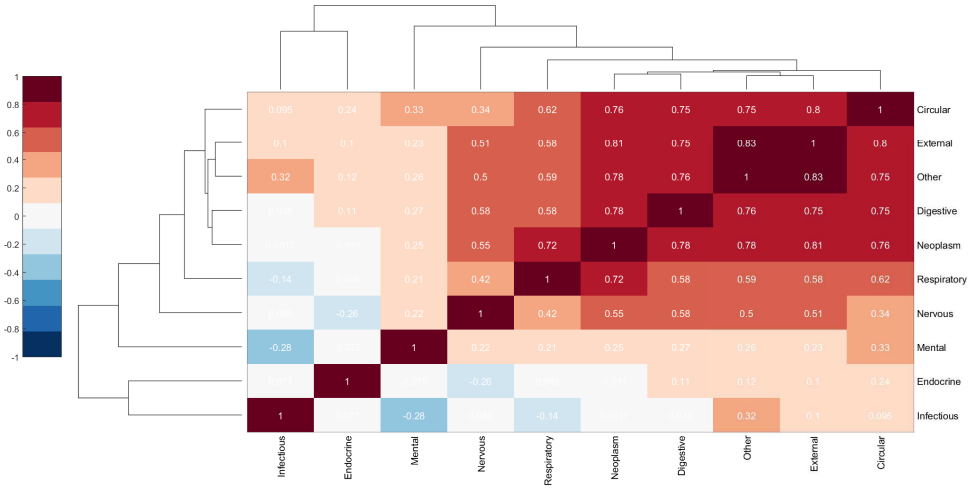
Age 70-74 years



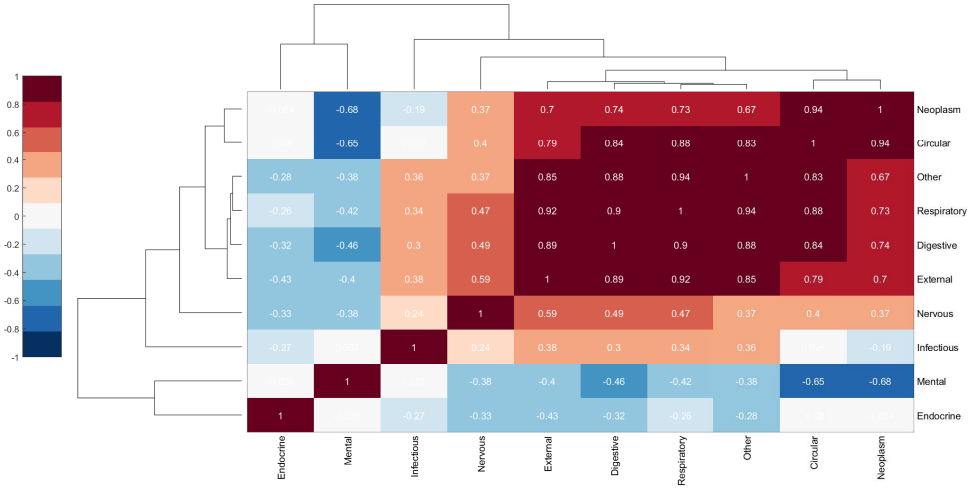
Age 80-84 years

Note: Colour from blue to red indicates correlations between the causes-of-deaths.

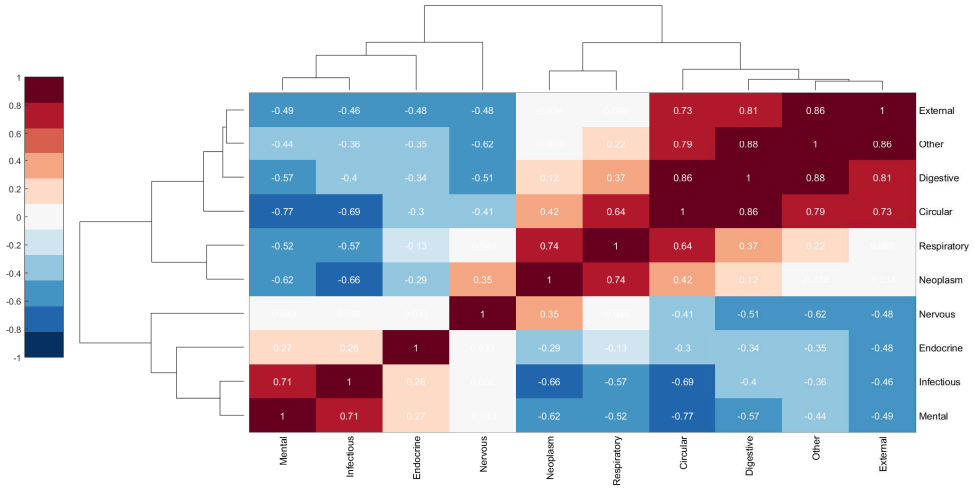
Figure B2: Hierarchical clustering of causes of death for Dutch males for selected ages



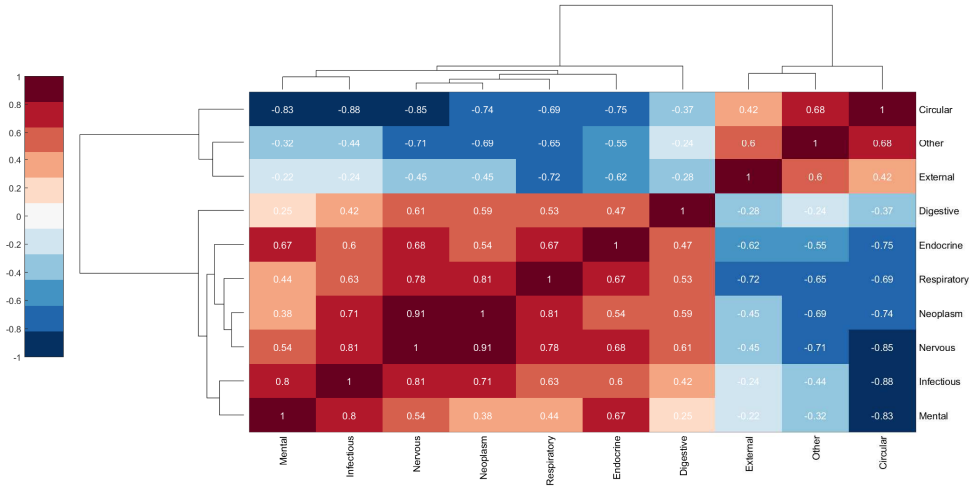
Age 25-29 years



Age 60-64 years



Age 70-74 years



Age 80-84 years

Note: Colour from blue to red indicates correlations between the causes-of-deaths.

Figures B1 and B2 show the results of the clustering analysis. The lines in the diagram indicate which binary or single groups of causes have the highest similarity. For example, for French females at age 25-29 cancer and digestive diseases are the most similar. The group consisting of cancer and digestive diseases is again most similar with cancer and so on. Injury and poisoning are least similar with any of the other causes, indicated with a line towards the rest of the causes.

Examining all the age specific cluster diagrams we find that no particular cause or set of causes is in a well defined cluster with cancer for all of the age groups. Instead, various causes cluster with cancer for different age groups: for example for French females digestive diseases cluster with cancer at ages 25-29 and with other

diseases and infectious diseases at ages 80-84. Thus, changes in the number of cancer deaths are not correlated with a particular cause but interact with the other causes differently across ages. That causes for some age groups are highly correlated emphasises the importance of modelling cancer deaths in relation to other causes.

Elastic net analysis

In order to identify possible subsets of causes for forecasting cancer, an elastic net analysis is carried out for each age group. Elastic net is a continuous shrinkage procedure for model selection (**?**) which determines a subset of the variables that is more useful for prediction of the dependent variable. Elastic net tackles multicollinearity among the causes so that if a group of highly correlated causes can predict cancer they are all selected. It is important to account for multicollinearity because all correlated causes might be important for the forecast of cancer. If multicollinearity is not accounted for only one of a group of correlated causes will be selected.

Table B1: Results from a Elastic net analysis for French females and Dutch males

	25-29	30-34	35-39	40-44	45-49	50-54	55-59	60-64	65-69	70-74	75-79	80-84	85-89	90-94	95-99	100+
French females																
Infectious diseases	0	0	0	0	0	-0.0365	-0.03672	0	0	0	0.0864	0.2673	0.2962	00951	0	0
Cardio-vascular	0.1362	0.1095	0.0930	0.0454	0.0060	0	0	-0.0228	0	-0.0010	-0.05144	0	0	0.4885	0.5206	0.6154
Respiratory diseases	0	0	0.0013	0	0.0005	-0.0109	-0.0371	-0.0614	-0.0570	-0.0999	-0.1309	-0.1055	0	0	0	0
Digestive diseases	0.1284	0.0840	0	0	0	0.02735	0.0083	0	-0.0215	0	0	0	0.0087	0	0	0.1373
Other diseases	0	0	0.0809	0.1135	0.1647	0.2042	0.1366	0.0802	0.0550	0.0454	0	0.2930	0.2747	0.2714	0.2905	0.0492
Injury and poisoning	0	0	0	0	0	0.1425	0.1660	0.1879	0.0920	0	0	0	0	0	0	0
R^2	0.80	0.77	0.59	0.72	0.73	0.65	0.61	0.40	0.23	0.70	0.81	0.93	0.95	0.95	0.95	0.91
Dutch males																
Infectious diseases	0	0	0	0	0	0	0	-0.0380	-0.1954	-0.1139	0	0	0	0.0457	0	-
Endocrine diseases	0	0	0	0	0	0	0	0	-0.0276	-0.0633	-0.0534	0	0	0.0442	0.0684	-
Mental diseases	0	0	0	0	0	0	0	-0.0108	-0.0552	-0.0419	-0.0345	0	0	0	0	-
Nervous diseases	0.0115	0.1120	0.0735	0	0	0	0	0	0	0.1333	0.2009	0.1855	0.1771	0.0459	0	-
Circulatory diseases	0.0508	0.1887	0.2542	0.3330	0.3519	0.3664	0.3340	0.2992	0.1834	0	0	0	0	0	0.1168	-
Respiratory diseases	0.1242	0.1138	0.0099	0	0	0	0	0	0.0092	0.3434	0.3137	0.1098	0.0204	0.2281	0.5073	-
Digestive diseases	0.1360	0.0443	0	0.0158	0.0082	0	0	0	0	0	0	0	0.2411	0.3140	0.0040	-
Other diseases	0.0903	0.2707	0.2276	0.1293	0.0836	0	0	0	0	-0.1089	-0.0429	0	0	0	0.3487	-
External	0.3343	0	0.1706	0.0888	0.0064	0	0	0	0	0	0	0	0	0	0	-
R^2	0.77	0.87	0.83	0.89	0.91	0.93	0.94	0.90	0.80	0.83	0.80	0.82	0.90	0.90	0.85	-

Table B1 shows the results from the elastic net analysis. Causes that are not correlated with cancer and hence not useful for the prediction of cancer in the elastic net analysis are given the weight zero. Highly correlated causes are given a high positive or negative weight. An R^2 measure is also reported for each age group.

Similar to the cluster analysis, the results vary by age and no single cause is unrelated to cancer for all ages. Thus, it is ambiguous which causes to drop when considering all ages. Instead we select causes to drop based on results for the age groups where most cancer deaths occur: that is the age groups 80-84 and 85-89 years for French females and for Dutch males the age groups 70-74 and 75-79 years. Hence the causes: injury & poisoning, cardio-vascular, digestive diseases and respiratory diseases are dropped one by one for French females, in the specified order. For Dutch males causes are dropped in the following order: external, digestive

diseases, circulatory diseases, mental diseases, endocrine diseases.

Table B2: Average Symmetric Mean Absolute Percentage Error (SMAPE) for observed vs. forecast cancer deaths 20-year out-of-sample rolling-window 's for French females and Dutch males, dropping selected causes

French females

Model	All included	Drop COD 7	Drop COD 3, 7	Drop COD 3, 5, 7	Drop COD 3, 4, 5, 7
CT-CoDa	0.06539	0.0484	0.0720	0.0999	0.0946
2S-CoDa	0.05658	0.0561	0.0771	0.0996	0.0827
VECM-CoDa	0.0749	0.0642	0.0813	0.0883	0.0817

COD1(Infectious diseases), COD2(Cancer), COD3(Cardio-vascular), COD4(Respiratory diseases), COD5(Digestive diseases), COD6(Other diseases), COD7(Injury and poisoning)

Dutch males

Model	All included	Drop COD 10	Drop COD 8, 10	Drop COD 6, 8, 10	Drop COD 4, 6, 8, 10	Drop COD 3, 4, 6, 8, 10
CT-CoDa	0.1351	0.1318	0.1293	0.2020	0.2443	0.2545
2S-CoDa	0.1068	0.0998	0.0970	0.1493	0.1629	0.1585
VECM-CoDa	0.1691	0.1557	0.1651	0.1819	0.1955	0.2215

COD1(Infectious diseases), COD2(Cancer), COD3(Endocrine diseases), COD4(Mental diseases), COD5(Nervous diseases), COD6(Circulatory diseases), COD7(Respiratory diseases), COD8(Digestive diseases), COD9(Other diseases), COD10(External)

Table B2 shows the SMAPE forecast errors for French females and Dutch males when dropping the specified causes one by one. The SMAPE falls slightly when dropping one cause but increases thereafter when dropping more causes. Hence, there is a trade-off between removing some non–relevant noise for the cancer forecast and introducing a forecast bias because of the dependency among the causes. As the SMAPE falls slightly when dropping one cause it shows that dropping causes can reduce the variance in the model leading to more accurate forecasts, but when too many causes are dropped the bias gets larger. It is outside the scope of this article to analyse this trade–off in further detail but the analysis presented clearly shows that the cause–dependency interacts with age, making it hard to reduce noise from non–relevant causes to cancer without introducing a significant forecast bias.

Conclusion of the selection of subsets of causes for forecasting cancer

By dropping causes that are less related to cancer at ages with most deaths we conclude that it is possible to improve the forecast accuracy by dropping a very limited number of causes. However dropping causes also introduces a forecast bias since causes are related differently across age groups. Thus, dropping more than a very small number of causes of death leads to a less accurate forecast of the number of cancer deaths.

C: Clark-West test

The significance of the out-of-sample analysis, described in section 3.3.2 in the main text, is tested using the Clark-West test. We test whether forecasts from the models 2S-CoDA, VECM-CoDA, and the LC models are significantly different from the CT-CoDA model using the Clark and West (2006) test. The Clark–West test defines the forecast error for two models denoted $m_i,\ i \in (1,2)$ by

$$e_{x,h,m_i} = \hat{d}_{h,x,m_i} - d_{h,x}$$

and the distance,

$$f = (e_{x,h,m_1})^2 - (e^2_{x,h,m_2} - (e_{x,h,m_1} - e_{x,h,m_2})^2). \tag{5}$$

Using the test statistic $CW = f/s.e.(f) \sim N(0,1)$, where s.e. is the standard error of f, it is possible to test whether two models produce the same forecast accuracy.

We test whether the models produce a significantly different forecast from the CT-CoDA model using the Clark and West (2006) test, averaging over age and the different out-of-sample forecast horizons also used in section 3.3.2 in the main text. p-values are reported in Table C1. Note that the RMSE in section 3.3.2 is measured in deaths and death rates, so it produces a weighted average whereas the Clark and West (2006, 2007) test is calculated as a simple average over the test statistics through ages. The RMSE thus weights age groups where most deaths occur in contrast to the flat weights in the test.

Table C1: P-values from the Clark and West test (Clark and West, 2007) of equal forecast accuracy between the CT-CoDa model and the 2S-CoDa, VECM-CoDa and LC models

	2S-CoDA	VECM-CoDa	LC
French females	0.0190	0.8407	0.0053
French males	0.0000	0.999	0.0015
Dutch females	0.4542	0.1563	0.0639
Dutch males	0.0000	0.0300	0.9904

The 2S-CoDA and LC models produce significantly different forecasts from the CT-CoDA model for the French population and only for the Dutch males for the 2S-CoDA. The VECM-CoDA model only produces different forecasts for Dutch males compared to the CT-CoDA model. Hence, the Clark-West test finds that the lower forecast error for the 2S-CoDA model compared to the CT-CoDA model is significant.

D: Figures

Figure D1: Weights used in the 2S-CoDa model, for French females

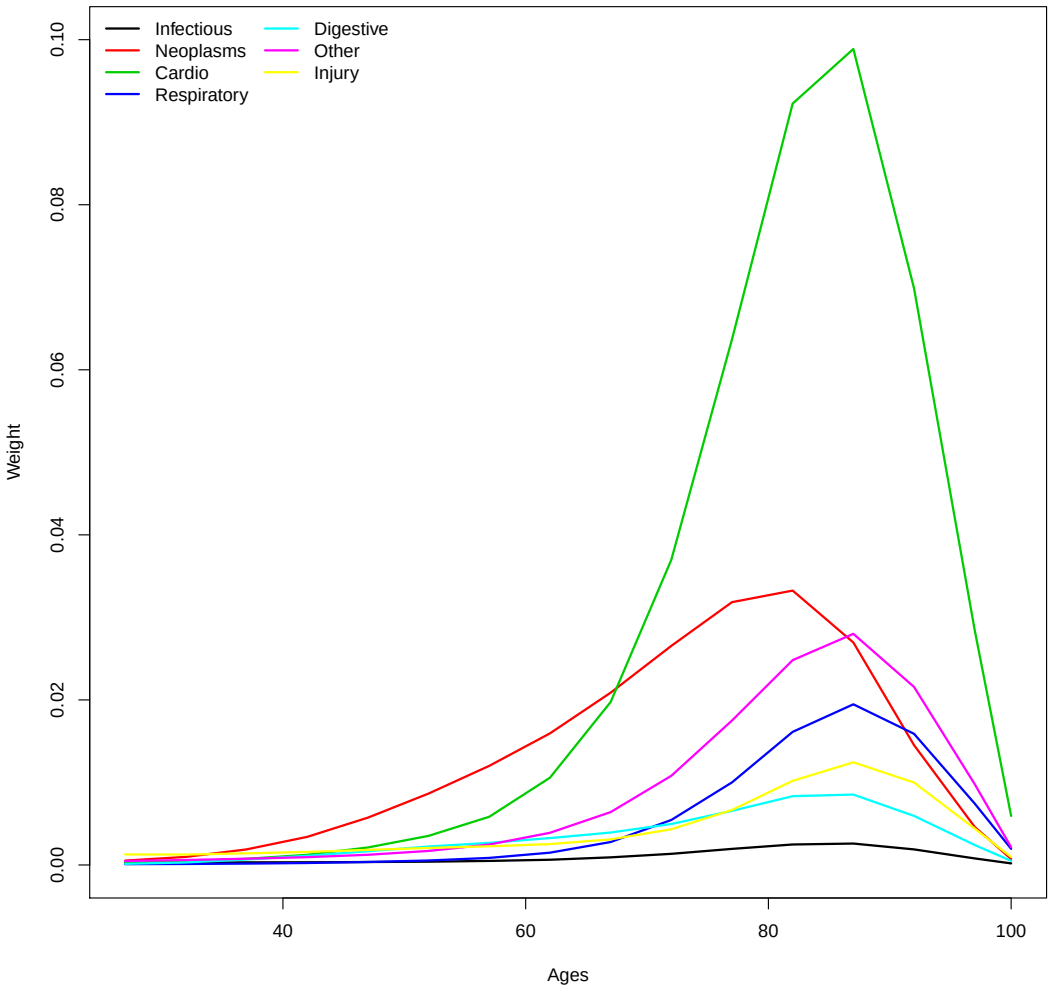


Figure D2: Geometric mean estimates, for Dutch males

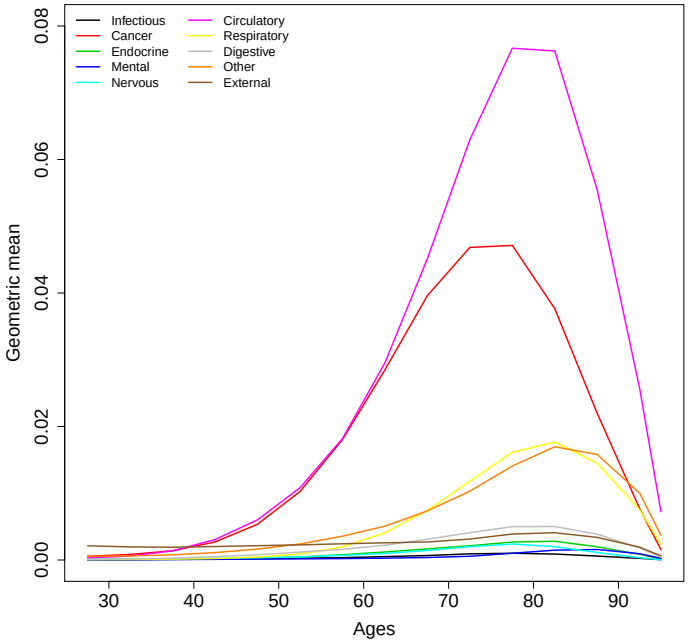
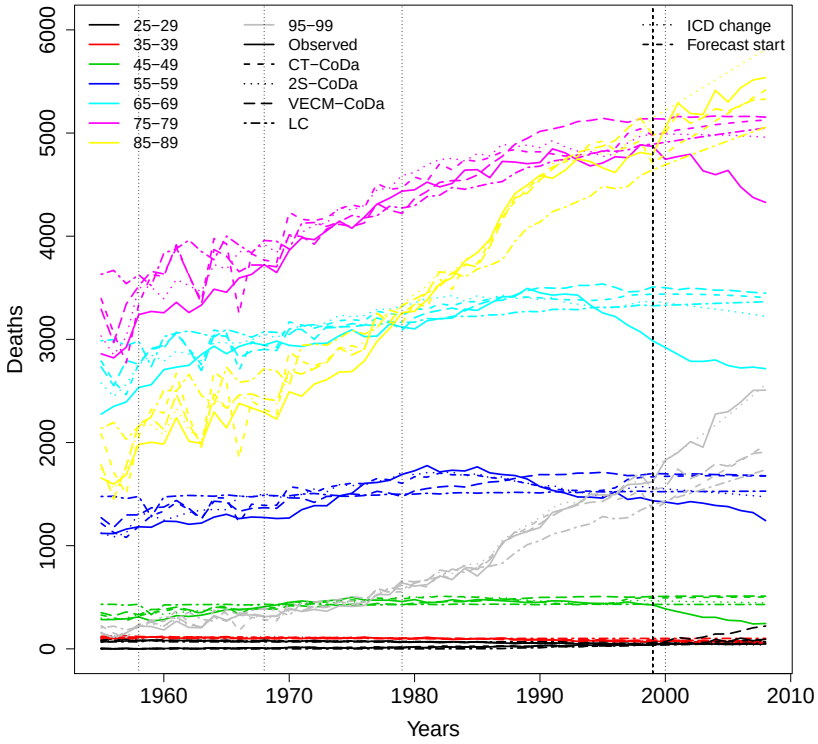


Figure D3: 10 year out-of-sample forecasts of cancer deaths across age groups for French males and Dutch females using the CT-CoDa, 2S-CoDa, VECM-CoDa, and LC models

French males



Dutch females

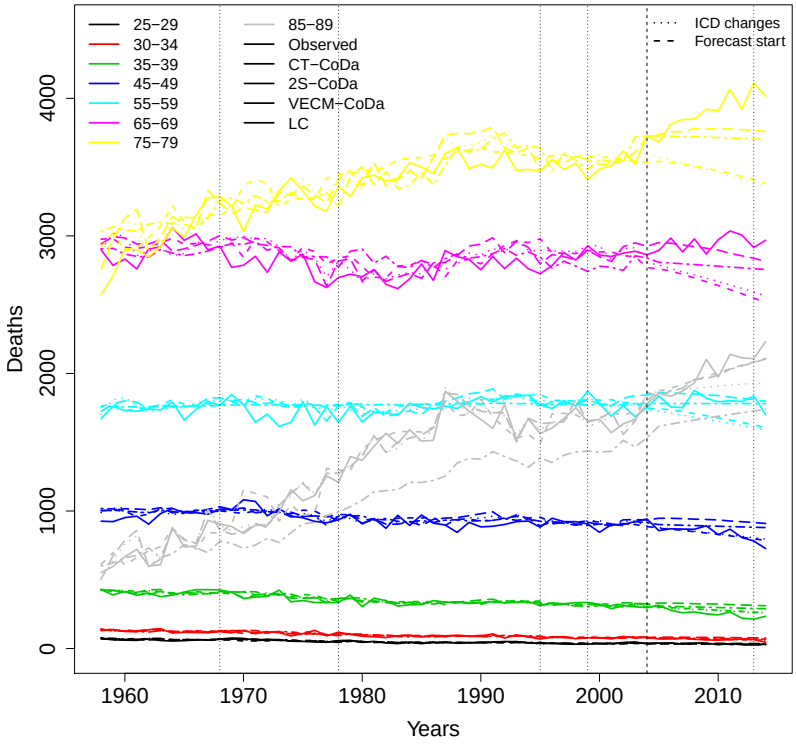


Figure D4: Distribution and centred distribution of cancer deaths for Dutch males in selected years

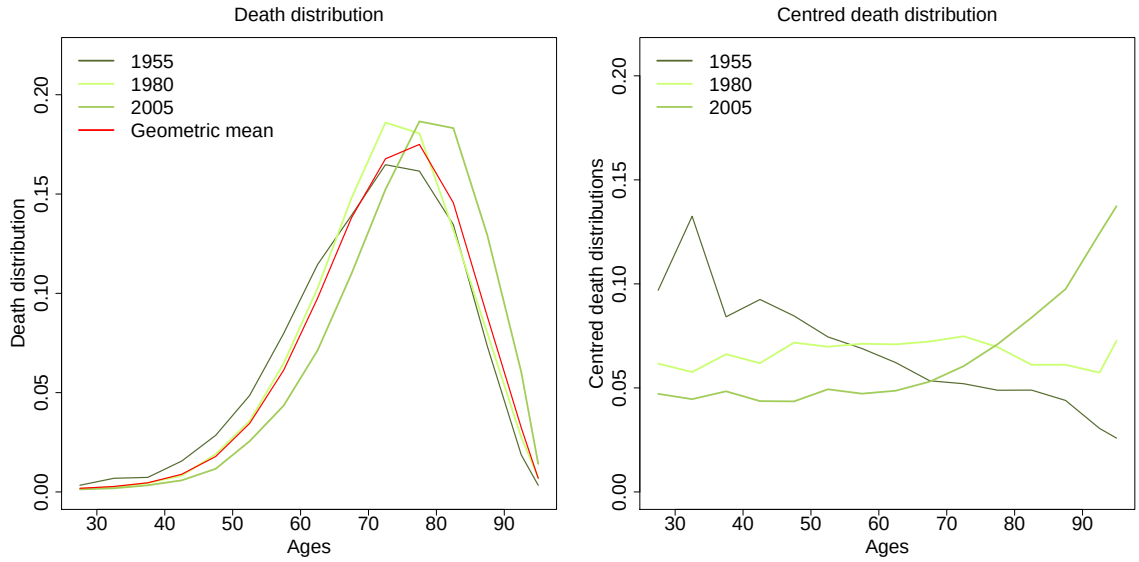


Figure D5: Rank 2 parameter estimates and forecasts for the CT-CoDa model for French females

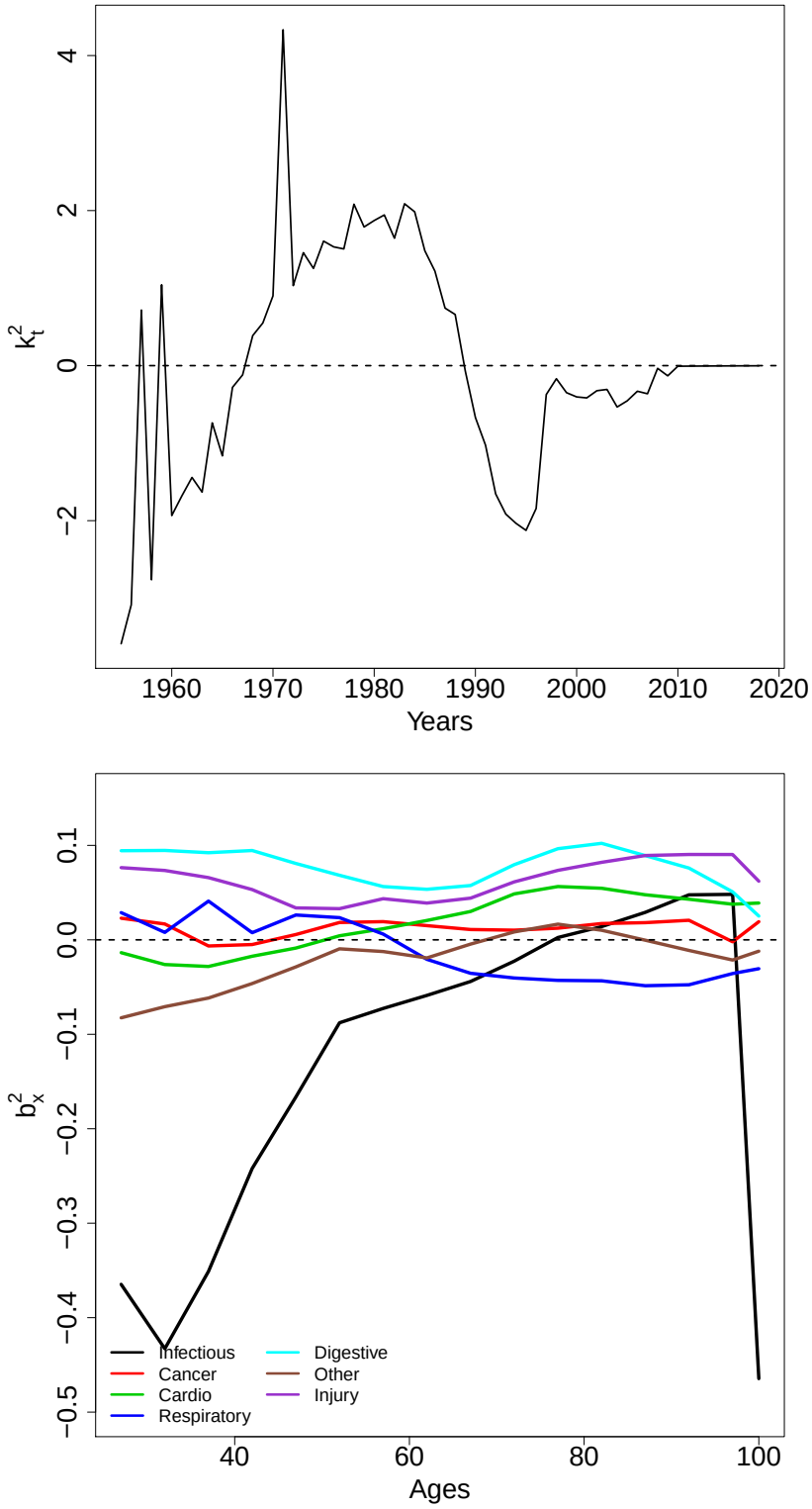


Figure D6: Rank 2 parameter estimates and forecasts for the VECM-CoDa model for French females

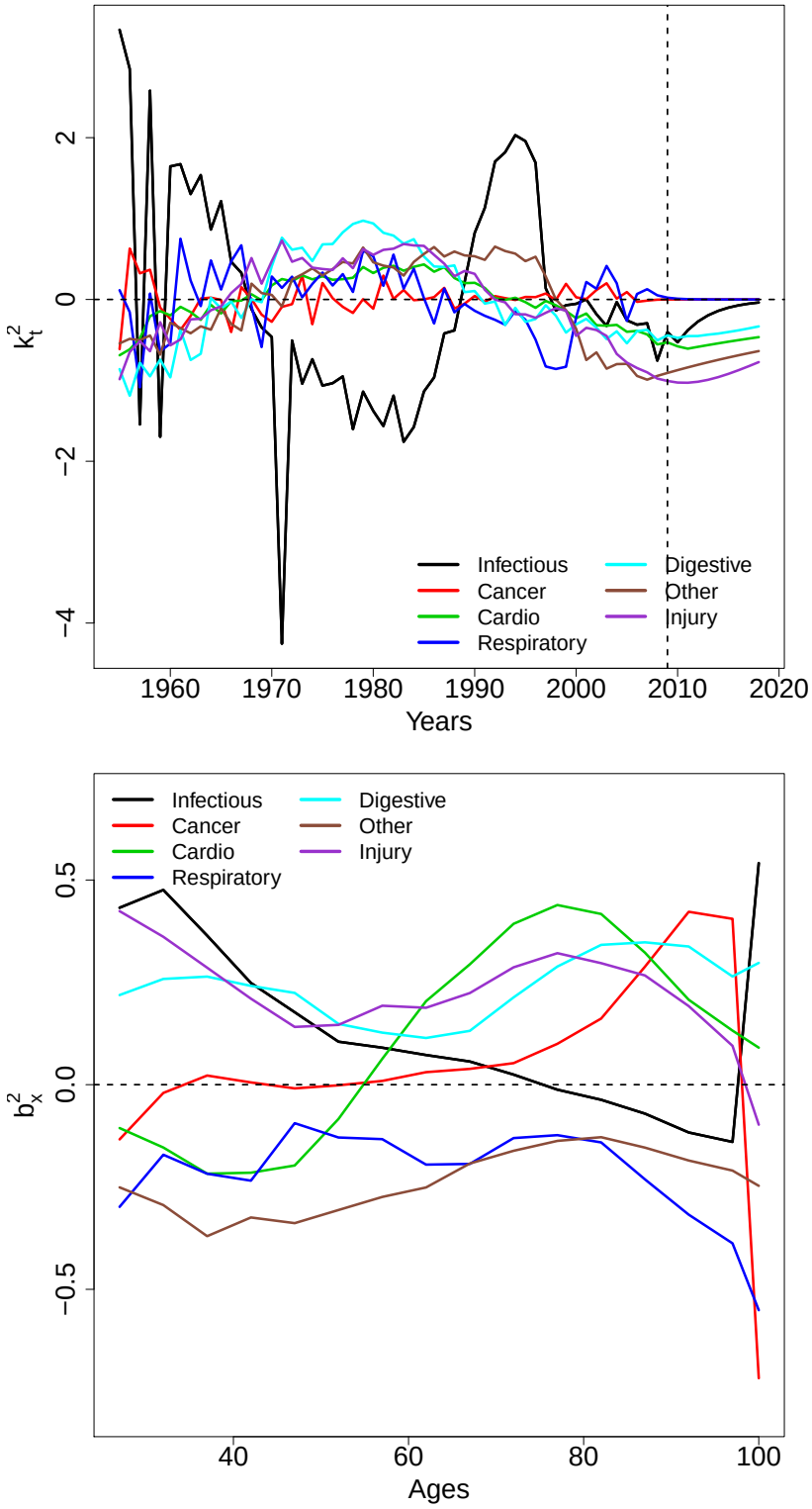


Figure D7: Rank 1 parameter estimates and forecasts for the individual decompositions in the 2S-CoDa model for French females

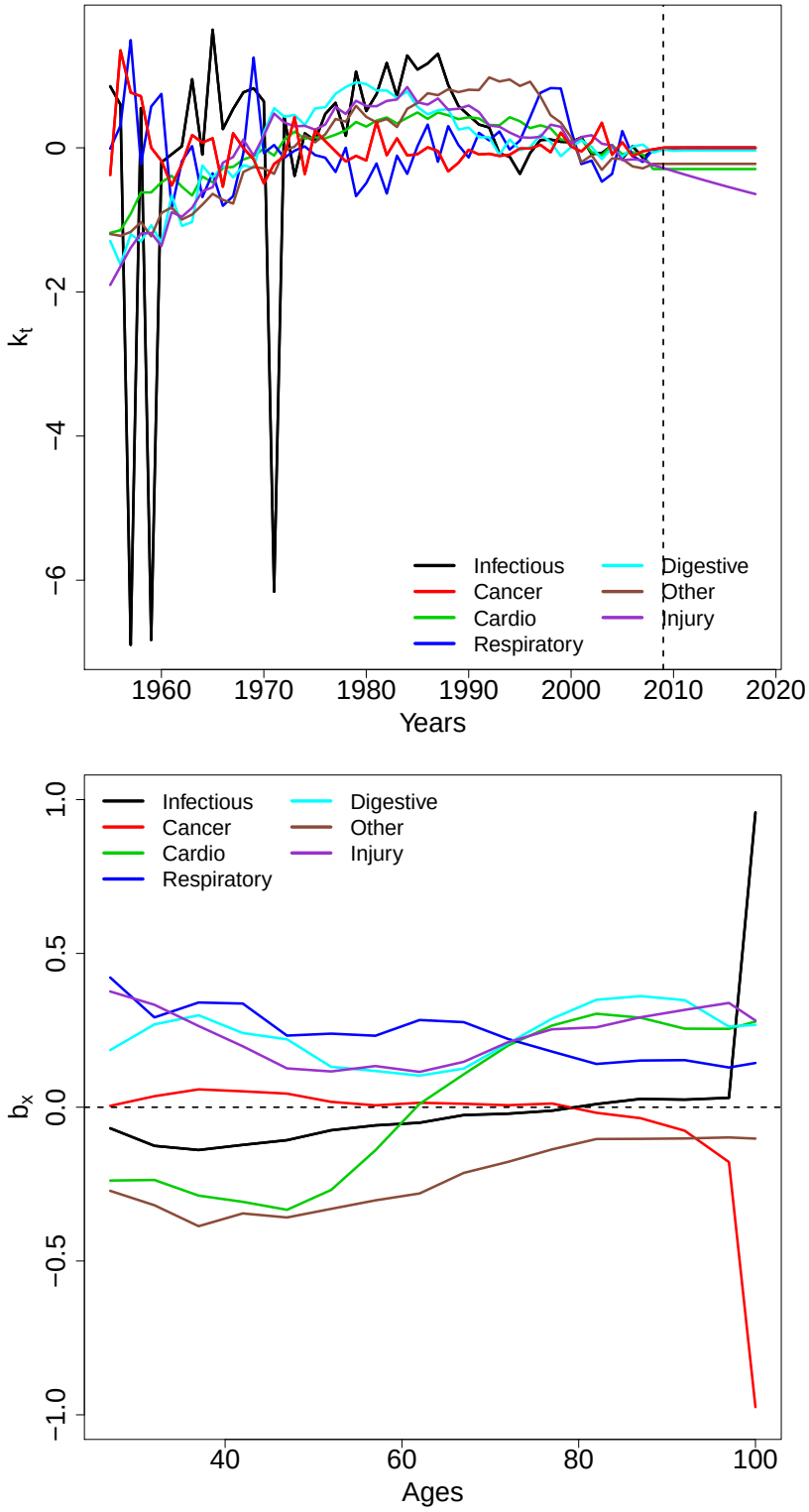


Figure D8: Residuals, for French females

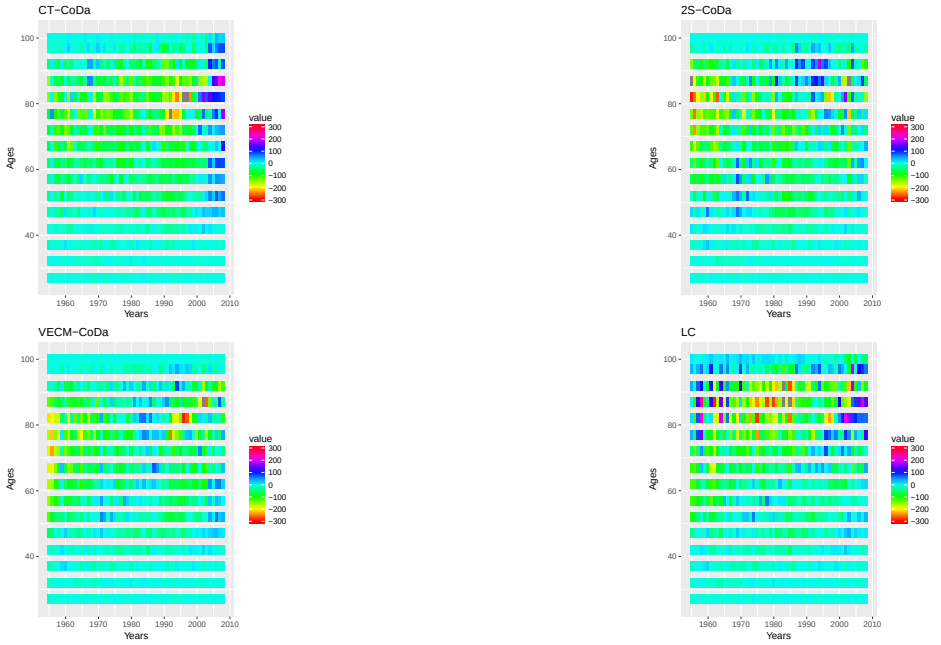


Figure D9: Residuals, for Dutch males

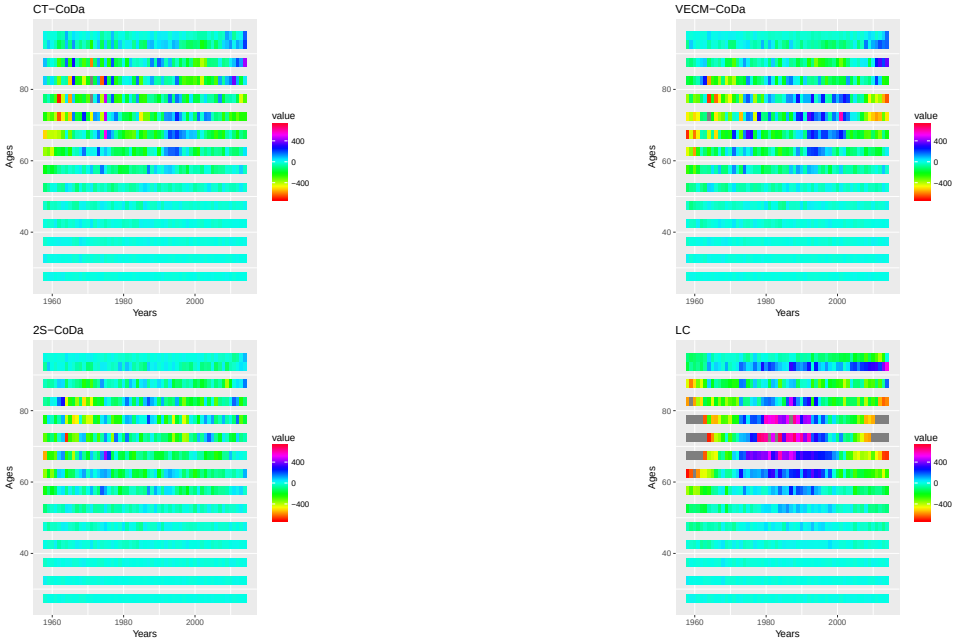
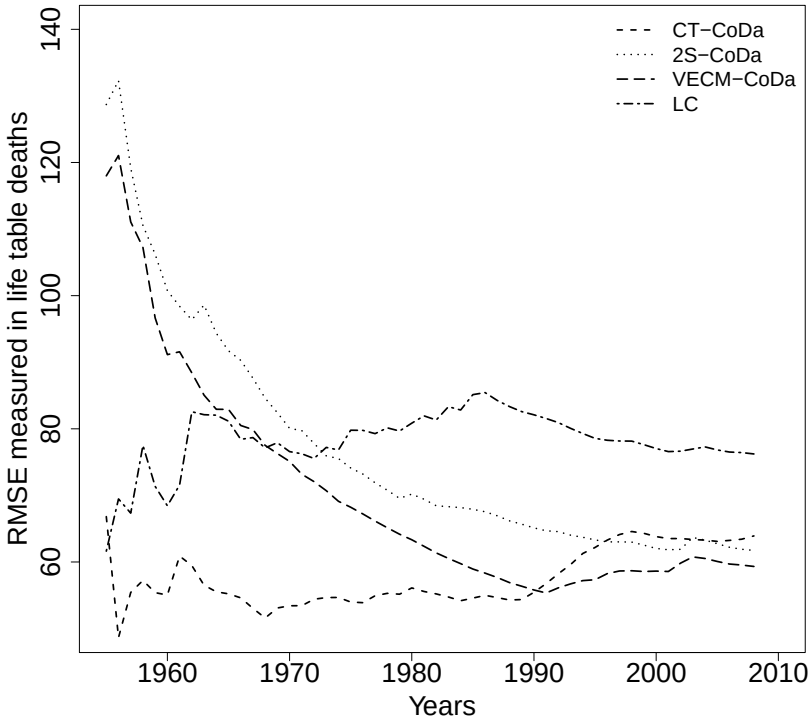


Figure D10: RMSE of observed vs. fitted cancer life table deaths for French females and Dutch males

French females



Dutch males

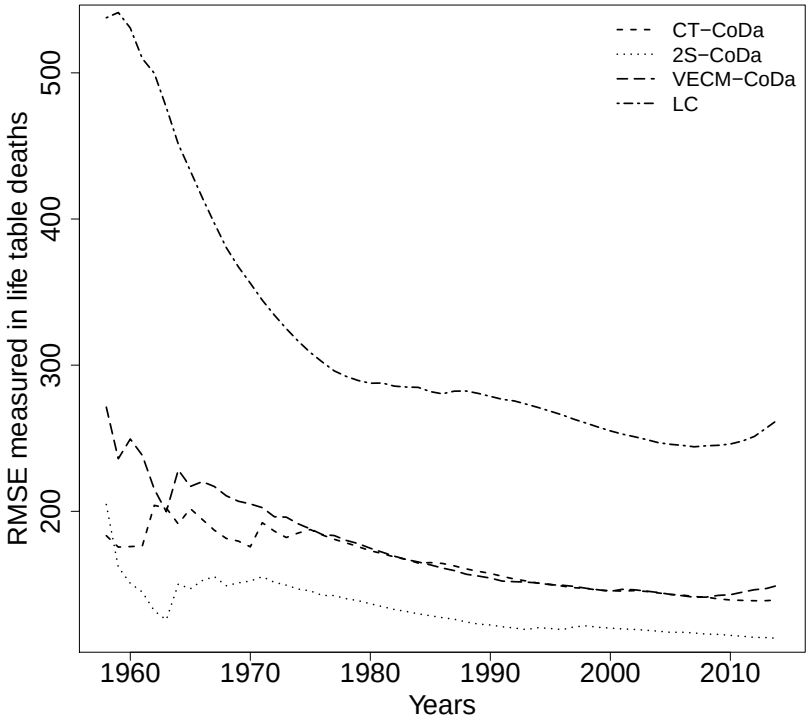
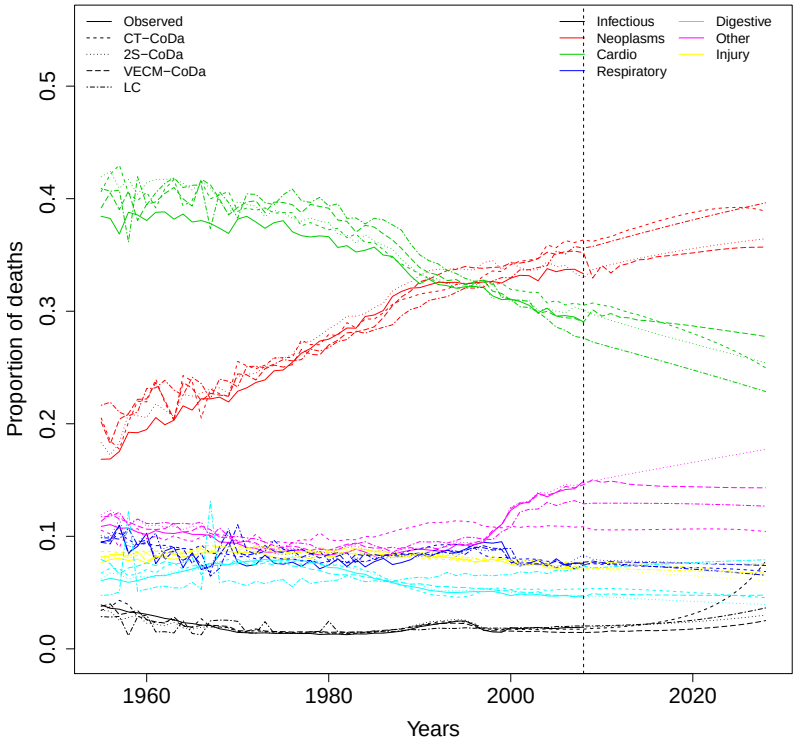
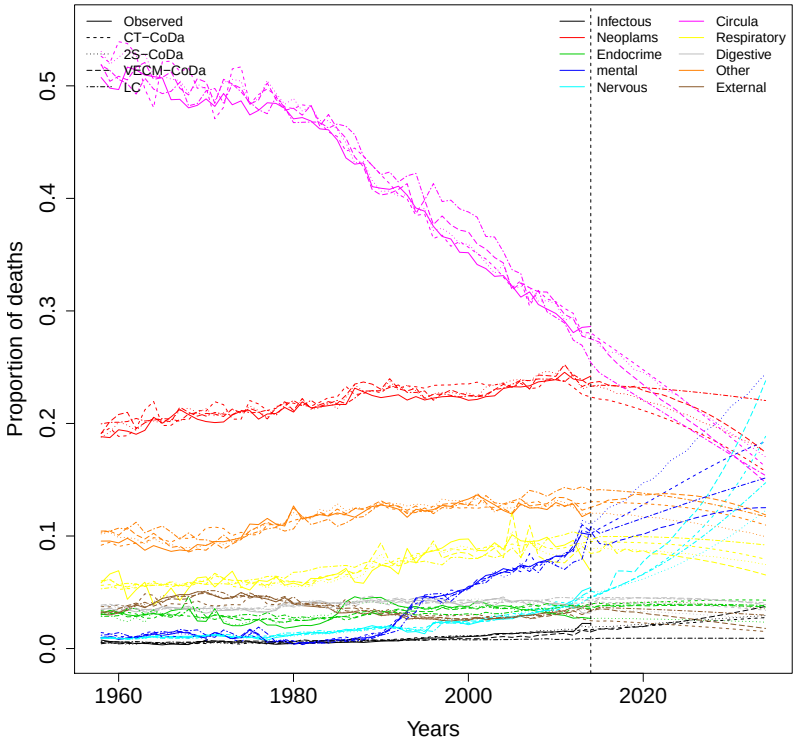


Figure D11: In-sample fits and 10-year out-of-sample forecasts of the proportion of deaths : French males and Dutch females

French males



Dutch females



E: Tables

Table E1: RMSE for observed vs. forecast proportions of cause specific deaths in the years 1955 to 2008 for French females

	Infectious	Cancer	Cardio	Respiratory	Digestive	Other	Injury & poisoning
CT-CoDa	0.000936	0.006872	0.020256	0.007401	0.003495	0.016259	0.003987
2S-CoDa	0.000919	0.006009	0.012634	0.004320	0.001557	0.003396	0.002058
VECM-CoDa	0.001886	0.005823	0.012728	0.005996	0.001795	0.005259	0.001932
LC	0.002632	0.006568	0.024737	0.013991	0.003219	0.013063	0.003919

Table E2: RMSE for observed vs. forecast proportions of cause specific deaths in the years 1958 to 2014 for Dutch males

	Infectious	Cancer	Endocrine	Mental	Nervous	Circulatory	Respiratory	Digestive	Other	External
CT-CoDa	0.001186	0.010853	0.002557	0.002122	0.002081	0.015318	0.008898	0.002026	0.007071	0.003400
2S-CoDa	0.001130	0.008910	0.001224	0.001947	0.001793	0.013829	0.008216	0.001743	0.003092	0.001290
VECM-CoDa	0.001914	0.010449	0.001389	0.000682	0.001901	0.014163	0.009116	0.001773	0.003333	0.001395
LC	0.003695	0.023872	0.002262	0.003900	0.002962	0.021685	0.012312	0.002138	0.006978	0.002565

References

Arnold-Gaille, S. and Sherris, M. 2013. Forecasting mortality trends allowing for cause-of-death mortality dependence. 17(4):273–282.

Clark, T. and West, K. D. 2007. Approximately normal tests for equal predictive accuracy in nested models. *Journal of Econometrics*, 138(1):291–311.

Clark, T. E. and West, K. D. 2006. Using out-of-sample mean squared prediction errors to test the martingale difference hypothesis. *Journal of Econometrics*, 135(1):155–186.

Johansen, S. 1988. Statistical analysis of cointegration vectors. *Journal of Economic Dynamics and Control*, 12(2):231 – 254.

Johansen, S. 1991. Estimation and Hypothesis Testing of Cointegration Vectors in Gaussian Vector Autoregressive Models. *Econometrica*, 59(6):1551–1580.

Juselius, K. 2006. *The Cointegrated VAR Model: Methodology and Applications*. Oxford University Press.

Lazar, D. and Denuit, M. M. 2009. A multivariate time series approach to projected life tables. *Applied Stochastic Models in Business and Industry*, 25(6):806–823.

Tofallis, C. 2015. A better measure of relative prediction accuracy for model selection and model estimation. *Journal of the Operational Research Society*, 66(8):1352–1362.